The European Nutrigenomics Organisation

TOOLS AND METHODS FOR MASS SPECTROMETRY METABOLOMICS IN NUTRITION

December 12-14, 2007

Clermont-Ferrand, France

Organizing Committee

Augustin SCALBERT (INRA)
Jane HUBERT (INRA)
Claudine MANACH (INRA)
Jean-François MARTIN (INRA)
Estelle PUJOS (INRA)
Jean-Louis SEBEDIO (INRA)
Ivana BOBELDIJK-PASTOROVA (TNO)
Suzan WOPEREIS (TNO)

Sponsors

Groupe DANONE

Waters

Contact

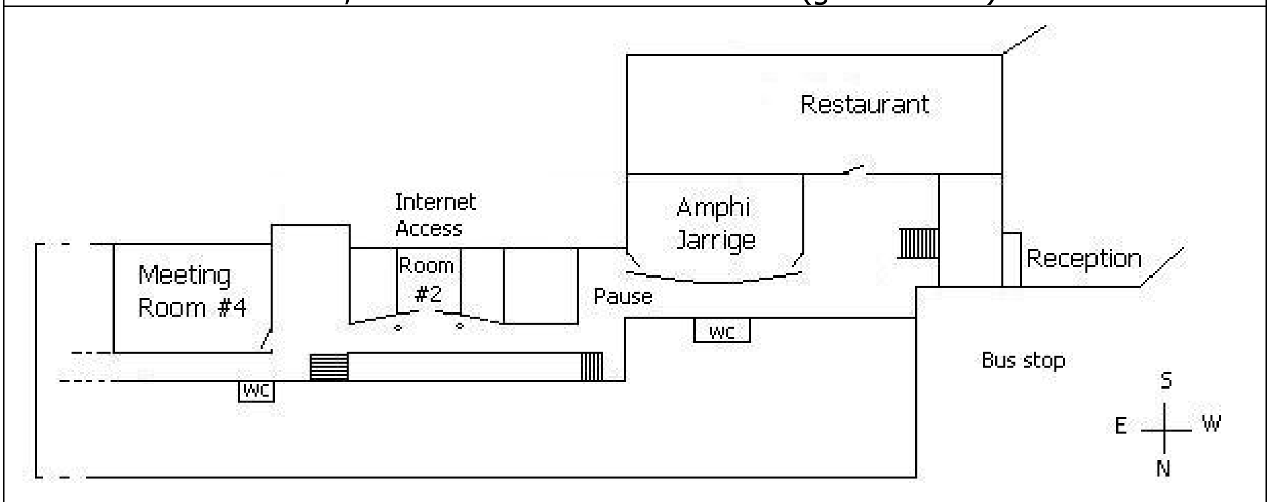
Augustin SCALBERT
INRA, Centre de Recherche de Clermont-Ferrand Theix
63122 Saint Genès Champanelle, France
Email: scalbert@clermont.inra.fr
+33 (0)4 73 62 47 87

Programme at a glance

	Wednesday Dec. 12th	Thursday Dec 13th		Friday Dec 14th	
9h		**Session 6** Room#4 Biological interpretation *Suzan Wopereis, Ben Van Ommen*	**Session 2** Amphi Jarrige Mass Spectrometry *Thomas Hankemeier*	**Session 4** Room#4 Data analysis *Bruce Kristal*	**Session 5** Amphi Jarrige Metabolite identification *David Wishart*
9h30					
10h					
10h30					
11h					
11h30		Pause		Pause	
12h		Reports sessions 2 & 6 Amphi Jarrige		Reports sessions 4 &5 Amphi Jarrige	
12h30					
13h	Registration	Lunch		Lunch	
13h30					
14h	**Lecture** Amphi Jarrige *Ben van Ommen*	**Session 3** Room#4 Data extraction *Elwin Verheij*	**Session 1** Amphi Jarrige Sampling strategy *Lorraine Brennan*	Conclusions *Oliver Fiehn*	
14h30					
15h	Pitfalls*				
15h30	Pause				
16h	Introduction of the 6 sessions by the moderators				
16h30		Pause			
17h		Reports sessions 1 & 3 Amphi Jarrige			
17h30					
18h					
20h… …	Dinner at Hôtel Mercure 82, boulevard François Mitterrand	Dinner at L'Odevie 1, rue Eugène Gilbert			

* With contributions of Manfred Beckmann, John Draper, Oliver Fiehn, Renger Jellema, Christophe Junot, Estelle Pujos.

INRA, Centre de Recherche de Theix (ground floor)



Session 1 : Sampling strategy

Moderator : Lorraine Brennan (UCD Dublin)

With the participation of:
Eleni Gika
Georgios Theodoridis
Doris Jacobs

Thursday, the 13th of December, 2007, 2 p.m.
Amphitheater Jarrige

Background:
In nutritional metabolomics one often has to deal with large variability in the samples compared to the changes of interest induced by the nutritional intervention. It is very important to limit variability induced by inappropriate sampling and storing of the samples.

Goal:
Share experiences with sampling in (nutritional) metabolomics studies. Establish advantages and disadvantages of the used methods.

Approach:
Discussion of at least the following topics:

Sampling
- Are 24 h urine samples appropriate for metabolomic studies?
- Bacterial contamination, addition and effects of preservatives on metabolic profiles.
- Collection of urine: the need to chill on collection and/or add a preservative.
- In the case of human subjects which is more appropriate: first void or second void samples.
- Is serum or plasma more appropriate to use?
- Which type of plasma is best for metabolomic studies?
- Extraction protocols for faecal water.

Storing and processing
- The effects of freeze thaw cycles on urinary metabolites.
- Effects of storage at room temperature, 4 °C, -20 °C and -80 °C
- Processing: dilution of samples, protein precipitation steps.

Result:
Proposal for best practice for nutritional metabolomics for sampling and storage of urine and plasma samples which is applicable in most studies.

Session 1 : Sampling strategy

Lorraine Brennan (UCD Dublin)

Urine: Classical nutritional studies use 24 h urine samples and we should discuss whether these are appropriate for metabolomic studies. Issues such as bacterial contamination, addition of preservatives and effects of preservatives on metabolic profiles should be discussed.
Collection of urine: the need to chill on collection and/or add a preservative.
In the case of human subjects which is more appropriate: first void or second void samples.

Blood: Is serum or plasma more appropriate to use? Which type of plasma is best for metabolomic studies?

Faecal samples: extraction protocols for faecal water

Storing and processing

The effects of freeze thaw cycles on urinary metabolites. Storage issues will be discussed: effects of storage at room temperature, 4 °C, -20 ° C and -80 °C will be addressed.
Processing: dilution of samples, protein precipitation steps.

Note: most of the above issues have been addressed using NMR based metabolomics.

In metabolomic studies, minimisation of unwanted sources of variation is important. The major sources in variation in metabolic studies can be broadly summarised as follows (Maher et al. 2007): (1) sample collection, storage and stability (2) sample pre-treatment prior to analysis (3) instrument variation/stability (4) intra-individual variations such as diurnal variation and stress (5) inter-individual variations due to genetic and environmental factors (6) inter-individual variations due to the presence/absence of disease. This section of the workshop will address some of the points above in the context of designing metabolomic studies for nutritional studies. Most metabolomic studies are designed to address a specific hypothesis and while a lot of attention is devoted to the design from the hypothesis view point it is easy to overlook sample collection, stability and storage issues. Recently, there have been a number of studies highlighting the importance of these points (Gika et al. 2007; Lauridsen et al. 2007; Maher et al. 2007; Teahan et al. 2006; Zhang et al. 2007). While a number of these studies have been carried out using NMR spectroscopy, the same issues and recommendations arising from these will also apply to LC-MS and GC-MS approaches.

When considering the design of a metabolomic study the first issue to consider is the time of the sampling. Diurnal variation has been documented in urine samples taken from healthy volunteers (Maher et al. 2007; Walsh et al. 2006). However, one can not simply decide to collect fasting morning samples as changes have been reported between samples collected with a 2 hour separation (Maher et al. 2007). Work from our laboratory has also shown differences between first void and second void urine samples. To minimise variation a very specific description of the collection of urine sample needs to be given such as "first void midstream urine samples".

The next issue to be addressed is the effects of preservatives on metabolic urinary profiles. Lauridsen et al investigated the effects of addition of NaN3 and NaF to human urine samples (Lauridsen et al. 2007). The authors concluded that for long term storage at -25 °C and -80 °C there was no requirement for the preservative. However, if preservative has to be added, for example due to extended storage at room temperature, the recommended preservative is NaN3 as it did not change the composition of the urinary profile.

With respect to blood sample collection recent studies have also highlighted the importance of how collection is performed. The generic advice of collection of fasting blood samples is not enough if systematic bias is to be avoided. In fact, details such as clotting time, clotting temperature and treatment conditions prior to centrifugation need to be considered. Teahan at al. demonstrated that clotting time had an impact on the metabolic profile and that clotting on ice delayed the observed changes (Teahan et al. 2006). When collecting serum samples for metabolomics studies one must record the clotting time and temperature at which the clotting occurred and standardise across all samples. Another issue which is often overlooked is that blood collection tubes can release materials into the samples and interfere with the Mass Spectrometry analysis (Drake et al. 2004). In addition, batch differences have been reported for certain vacutainers and again these concerns should be addressed in study design. With respect to plasma samples one must consider the anticoagulant to be used and possibility of unwanted peaks.

Sample storage and stability for human biofluids with respect to metabolomics are issues that have been addressed for NMR based and LC-MS based approaches. Lauridsen et al. showed that urine samples should be stored at or below -25 °C and recommended not to store at 4 °C for prolonged periods (Lauridsen et al. 2007). No beneficial effect of storing at lower temperature was seen for the duration of this study (up to 26 weeks) using NMR based metabolomics. Similar results were reported by Maher et al using NMR based metabolomics (Maher et al. 2007). A recent study showed that samples stored at -20 °C or -80 °C for up to one month were indistinguishable using LC-MS analysis (Gika et al. 2007). Short term storage of human urine samples at 4 °C was shown to be stable for up to at least 20 h using LC-MS and 24 h using NMR. The effects of freeze thaw cycles on the metabolic profiles have been investigated using NMR, LC-MS and GC-MS. In the case of the LC-MS study, freeze thaw cycles of up to 9 cycles did not impact on the clustering of the data in a PCA plot (Gika et al. 2007). Repeat freeze thawing for 7 days was shown to have little impact on the compounds measured in rat urine using GC-MS (Zhang et al. 2007). Studies in our laboratory showed that one freeze thaw cycle had minimal impact on the NMR spectra and the within bin correlation was above 0.97 for 90% of the regions studied. While freeze drying human urine samples has been common practice in the past, recent studies have shown that freeze drying results in considerable changes to the metabolic profiles (Lauridsen et al. 2007).

Other unwanted sources of variation that are important to consider in study design are recent consumption of certain foods and medication. Recently, it has been demonstrated that standardisation of the subjects diet reduced the variation in the urinary metabolic profiles (Walsh et al. 2006). In addition, there are many reports where diet, alcohol and medication have resulted in outlying samples. When setting out to design metabolomics studies it is essential that one considers food restrictions and that subjects record their dietary intake 24 h prior to biofluid collection and avoid the use of medication.

Blood samples contain a range of low molecular weight compounds and proteins and efficient removal of the proteins is necessary prior to LC-MS or GC-MS analysis for metabolomic studies. Since human blood contains a range of low molecular compounds with a wide range of concentrations, stability and ability to bind to proteins the development of an extraction procedure is complex. Recently, Want et al. examined a range of deproteinization methods in combination with LC-MS and found that methanol precipitation was the most effective and reproducible approach resulting in the detection of over 2000 metabolite features and less than 2% protein (Want et al. 2006). A recent study investigated two protein precipitation methods for plasma samples and found the methanol method to be superior for use with UPLC-MS (Bruce et al. 2007). In addition, for use with GC-MS the methanol extraction procedure was found to be highly efficient and reproducible (A et al. 2005). Faecal samples are an example of another biofluid relevant to nutritional research which requires pre-treatment prior to metabolomic analysis. To date little has been published about faecal water preparation strategies and implications on metabolomic profiling.

All of the issues described above need to be considered and standard operating procedures (SOPs) drawn up for each metabolomic study. The importance of this can not be underestimated, especially in the case of multi-centre nutritional intervention studies, if meaningful results are to be obtained. In addition, the adaptation of a standard reporting system is essential for the description of the collection of biofluids relevant to nutritional research. To this end steps have been taken by the metabolomics standards initiative and a recent publication describes the reporting requirements for biological samples (Griffin et al 2007).

Session 2 : Mass spectrometry

Moderator : Thomas Hankemeier (Univ. of Leiden)

Thursday, the 13th of December, 2007, 9 a.m.

Amphitheater Jarrige

Background:
The number of applications of MS in nutritional metabolomics is growing and amongst these are examples of both target and comprehensive profiling. Furthermore, within these two categories, many different MS techniques can be used. All these possibilities can influence the type and amount of information present in the final dataset.

Goal:
Discuss the most commonly used MS methods, determine their application fields. Discuss the needs and possibilities of standardization of methods and results.

Approach:
Several questions will be addressed:

- What is most suitable for nutritional metabolomics, the analysis of a target list of metabolites, or a holistic profiling?
- If it is a target list, what will be the list?
- Do we want to have standardized methods, or only standardized results?

The answer on these questions will determine the analytics strategies. Several aspects are of importance:

- Choice of separation mechanism (LC, GC, capillary electrophoresis and various options to improve peak capacity)
- Quantification: how to deal with ion suppression? How to control the performance of the method? Dynamic range of MS
- Choice of ionisation mode (for LC-MS: EI, APCI, APPI, positive vs. negative mode, in-source fragmentation, miniaturization, for GC-MS: electron impact, chemical ionization)
- Choice of mass spectrometric detection method (Tof, Ion traps, FT-ICR)
- How to validate the method?
- How to assure comparability of data obtained?
- Are there any alternatives to mass spectrometry?

Result:
Establish recommended MS approach(es) (analysis, quantification, quality control etc.), for application in nutrition metabolomics studies.

Session 2 : Mass spectrometry

Thomas Hankemeier (Leiden University/Netherlands Metabolomics Centre)

Mass spectrometry (MS) has played an important role in the development of methods for profiling of metabolites. In short, broader screening approaches of metabolites combined with biostatistical tools for evaluation have been developed in the last century, and became a boost in the 70ies by profiling of bodyfluids with gas chromatography (GC) and GC-MS for, e.g., the investigation of in-born errors of metabolism. In the following years, soft ionization mass spectrometry, pyrolysis mass spectrometry, liquid chromatography (LC)-MS and NMR were used in combination with pattern recognition. In the 90ies, major improvements in new analytical technological were achieved and used for metabolite profiling, such as the improvement of LC-MS interfaces, the use of electrochemical detection (coulometric array technology) for LC, and capillary electrophoresis (CE)-MS. Combinations of different techniques to comprehensively measure the metabolome were developed for the profiling of bodyfluids, as it appeared not possible to analyse the wide range of metabolite classes with only one method. Parallel, in plant, major breakthroughs on the technology side have been realized. Recently, further improvements in separation methods have been realized and applied in metabolite profiling, e.g., LC-MS using smaller particles, longer monolithic columns, comprehensive GCxGC-MS, and the availability of improved high resolution MS detectors. Actually, MS is more sensitive than NMR, and only MS is further discussed in this workshop.

Different approaches in metabolite profiling (the term metabolite profiling is used in this document rather than metabolomics) are possible, a targeted approach, in which a preselected list of metabolites of interest are measured, and a non-targeted approach, in which all metabolites present are detected. The targeted approach has some advantages such as less complicated data pre-processing, and possibly better detection limits, however, one often has to choose carefully the list of relevant metabolites prior to a study. The question is what the relevant metabolites would be for the nutrition research area, for example, with regards to energy metabolism, inflammation and oxidative stress?

Next, the choice of the most appropriate analytical method comprises the choice of a separation method prior to MS, the interface between separation method and MS, and the type of MS detector.
Separation options are:

- Gas chromatography
- Liquid chromatography
- Capillary electrophoresis.
- Or, alternatively, no separation at all using rather a direct infusion approach.

Each approach has its advantages, and disadvantages. Important criteria for selection of a method are a.o. selectivity, dynamic range, detection limit, accuracy, precision, and price per sample.

For GC-MS, the advantages are high separation efficiency, reproducible retention times, possibility of universal electron ionization (EI) providing response for all metabolites and characteristic and reproducible mass spectral fingerprints, which can be identified when searching in databases (if they are present in the database, see workshop Metabolite Identification).
For GC-MS, the disadvantage is that a derivatization reaction is generally required to analyze medium polar and polar metabolites, and degradation of derivatized metabolites can occur during injection and separation. For example, less stable derivatives like some amino acids may be quantified reliable only if the whole system is in good shape.
To increase the peak capacity of the GC separation method, i.e. increasing the number of separated metabolites, prior to MS detection, comprehensive multi-dimensional GCxGC approaches are increasingly used.

For LC-MS, the advantage is that no derivatization is required for the analysis of also larger and polar metabolites, allowing fast analysis of also small sample volumes. A wide range of different detectors are available for LC-MS, partly developed during the proteomics wave in recent years.
For LC-MS, the disadvantage is that for more polar metabolites retention is less straightforward. Recently, hydrophilic interaction chromatography (HILIC) has been used for the analysis of polar metabolites. Another disadvantage is that the response factor of metabolites using electrospray ionization (ESI) and, to a lesser extent, atmospheric pressure chemical ionization (APCI), may depend on the presence of matrix compounds. This might be overcome to some extent by miniaturization of electrospray ionization to nanospray ionization, a better separation of metabolites, or the use of isotopically labelled metabolites in a targeted approach.
In LC-MS, different types of methods are described such as reversed-phase liquid chromatography (RP)-MS approach, a HILIC-MS approach and a lipid profiling approach, next to some more targeted approaches.
To increase the peak capacity in LC, options are smaller particles in LC columns (requiring often higher pressure of the LC), using longer columns such as monolithic columns. However, comprehensive multi-dimensional methods such as in GCxGC are less straightforward, as there are less efficient options for refocusing of peaks in the second dimension available compared to GCxGC.
In addition, identification requires a different strategy in LC-MS than in GC-MS, as mostly only the molecular ion is detected, and additional fragmentation experiments have to be conducted to obtain fingerprints of metabolites. However, these MS/MS spectra are not always as well comparable compared to the GC-EI-MS spectra.

For CE-MS, the advantage is that the separation efficiency is higher than in LC. However, the disadvantage is that migration times are often less reproducible for different type of samples, and that sensitivity due to the CE-MS interfaces mostly used is less compared to LC-MS.

The other options, direct infusion-nanospray ionization-high resolution MS has recently been reported for several applications. The advantage is high throughput, however, precautions have to be taken to compensate for possible ion suppression effects.

For GC-MS, there are generally less options to choose an MS detector. Most often, a quadrupole MS detector is used. For GCxGC, a fast time-of-flight (TOF) MS is used. Alternatives are triple quadrupole MS and ion trap MS detectors.

For LC-MS or CE-MS, a wider range of MS detectors are available, such as TOF, ion trap, triple quadrupole, ICR-FT, Orbitrap detectors, and hybrid combinations of those.

It should be mentioned that also matrix assisted laser desorption ionization (MALDI)-MS has been used for metabolite profiling, however, generally MALDI-MS of small molecules is less quantitative compared to ESI.

Another important aspect in metabolite profiling methods with MS is how to realize comparability of studies and data. Recently, a set of minimum requirements for reporting a metabolite profiling methods has been suggested (Sumner et al, Metabolomics 2007). However, the minimum requirements for the validation of metabolite profiling methods have not been described yet, and one often encounters publications where the validation of an analytical metabolite profiling method is not well described. Validation should include at least a description of:

- Calibration model (Linearity and range)
- Repeatability and intermediate precision
- Accuracy (use of matrix variation studies, etc)
- Lower limit of quantification

In addition, the measurement of uncertainty, or the control which samples, and which metabolites, are measured with an acceptable precision (and preferably accuracy) should be known. Several approaches are feasible, such as the use of routine quality control samples during a sample sequence.

It should be mentioned that there are some alternatives to mass spectrometry for at least special applications such as electrochemical detection and fluorescence after derivatization for LC detection for ultrasensitive analysis in low sample volumes. Obviously, another alternative is LC-NMR, but sensitivity is much less favourable compared to LC-MS,

Objectives of this workshop are:

- What is most suitable for nutritional metabolomics, the analysis of a target list of metabolites, or a holistic profiling?
- If it is a target list, what will be the list/how to come to such a list?

- Do we want to have standardized methods, or only standardized results?
- How to assure comparability of data obtained?
- How to validate LC-MS, CE-MS and GC-MS methods?
- Share experiences in LC-MS, GC-MS and CE-MS approaches:
    - Separation approaches used
    - Dynamic range challenge
    - Quantification: how to deal with ion suppression in ESI, and (less) in APCI?
    - Choice of mass spectrometric detection method (TOF, Ion traps, triple quad, FT-ICR, Orbitrap, hybrids, etc)
- Are there any alternatives to mass spectrometry?

Session 3 : Data extraction

Moderator : Elwin Verheij (TNO Zeist)

With the participation of:
Clemens Groepl

Thursday, the 13th of December, 2007, 2 p.m.

Room # 4

Background:
The application of LC-MS, GC-MS and NMR for metabolomics results in very rich and complex raw data often obtained for large numbers of samples. Data extraction, i.e. translation of raw data / signals into accurate and concise clean data, is a daunting task. The increasing popularity of high resolution mass spec, e.g. FT and Orbitrap systems and GCxGC-MS and high throughput systems (e.g. UPLC) results in the generation of even larger amounts of complex data.

Several data extraction strategies exist for the various analytical techniques, e.g. for NMR: binning, peak picking, deconvolution and for LC-MS/GC-MS: target processing, peak picking, deconvolution, etc. using software provided by instrument vendors, independent companies, or homemade tools (public domain or proprietary). All data extraction methods and softwares have their pros and cons with respect to critical issues such as throughput and data accuracy/quality.

Metabolomics collaboration is hindered by the application of a wide variety of data extraction strategies and tools, especially because N data extraction tools (and a multitude of user defined settings) applied to the same raw data results in at least N different clean data sets, and finally in at least N different statistical models.

Goal:
Improve possibilities for nutritional metabolomics collaboration by sharing experiences with different data extraction methods and proposing a standard or reference method(s).

Approach:
In order to make this workshop a success we invite experts to contribute to this workshop and present their view on the data extraction as described above.
Topics discussed in the session on data extraction will include:

- What's around for the various analytical techniques, and what are the pros and cons?
- Is there a need for organizing round robins / benchmarking studies?, and
- if so, how do we arrange this? (who will do what, selection of well characterized raw datasets, etc)
- What's the ideal situation, how do we get there?
- Standardization of tools?
- Clear standardized documentation of clean data sets, how was it obtained, what does the data represent, etc.

Result:
Criteria to be defined on what is a well characterized raw dataset. Based on these criteria, data set(s) should be selected for the testing of different extraction methods of potential interest.
Decision on a round robin/benchmarking study

Session 3 : Data extraction

Elwin Verheij (TNO Zeist)

Extraction of information from metabolomics data has always been a problematic process. This was true in the long history of 'metabolomics avant la lettre' and still holds today. For example, in the study of medicinal plants and plant cell biotechnology, the metabolomics approach was daily routine. Back then, the low tech gold standard was thin layer chromatography (1D or 2D) in the 'pre' HPLC, GC, MS and NMR era, because it was cheap, and provided detailed information about sample composition (primary and secondary plant metabolites). Metabolites were visualized by coloring reaction with a multitude of different spraying reagents and/or fluorescence and the data was captured by photography and notes in logbooks. Data extraction was a very laborious, inefficient and ineffective. Progress in analytical technology brought HPLC GC MS and NMR to these labs, but also in the early days of these techniques data was captured on paper recorder rolls, photographic plates, etc. Only after the computer revolution in the 80's digital data systems became commonplace to capture analytical data. We finally got digital data, and the main purpose of the software that came along with analytical instruments was visualization of data, and computer aided extraction of data in a manner similar to what was done in the past with rulers, pair of scissors and other utensils.

Currently NMR and MS (LCMS and GCMS) are the metabolomics workhorses. In the past decade advances in analytical technology have provided us with reliable instrumentation capable of producing vast amounts of very rich data at increasing higher speeds simply because we need this (that's what we told the instrument companies). The data burden is commonly regarded as one of the major bottlenecks in analytical laboratories, and is not specific for metabolomics. The complexity and richness of metabolomics data (and also mass spec proteomics data) makes the 'omics' community the hardest hit. Metabolomics is shifting more and more from NMR towards mass spectrometry based techniques, mainly because mass spectrometry is more sensitive (but not the most sensitive technique) and it offers additional separation power to chromatographic and electrophoretic separation techniques to obtain selective/specific comprehensive composition data of complex mixtures at high throughput and with a large dynamic range (unfortunately not large enough). NMR results in relatively simple 2-dimensional data (shift and intensity) but already here there have been strong debates on the best method for data extraction, in particular between the two most popular approaches: binning and peak-picking plus alignment. Without going into details both approaches have their merits and disadvantages, which have a major impact on the composition of the clean data going into statistical analysis workflows. This is the fundamental problem of data extraction: the same data, extracted with n different methods, gives n different data sets and probably n different statistical models (if a single statistical procedure is applied). In applications where major changes to the metabolome occur (e.g. toxicity models) the differences in the final outcome are relatively small. Unfortunately, in all other applications where metabolome changes are more subtle data extraction errors have a dramatic impact on the outcome of a study. Recently, deconvolution of NMR spectra into compounds has been added to these two. This approach is superior from a fundamental, theoretical perspective because it targets compounds (metabolites) and not signals.

Hyphenated mass spectrometry techniques such as LC-MS and GC-MS, and in particular GCxGC-MS, result in far more complex data than NMR (3D or 4D) due to the increased dimensionality. On top of that mass spectrometry instruments have a large number of operation and acquisition modes, and other factors increasing the data diversity, e.g.:

- centroid or profile data acquisition
- fixed scan time (ToF, Q) versus variable scan time (IT)
- positive, negative and alternating/mixed polarity
- full scan MS1 versus SRM in MS^n
- nominal mass resolution to ultra high resolution
- the ability to mix of all of the above
- proprietary data formats and instrument specific acquisition modes/experiments

All instruments come with software for quantitative data processing as it is used in routine analysis (e.g. bioanalysis, residue analysis, environmental analysis etc.). Unfortunately this software was not made for processing very rich comprehensive metabolomics LC-MS and GC-MS data (e.g. 10000 features in a single file) of unknown composition. This software can be used, but this limits the number of compounds to a few hundred. It should be realized that multiple target processing with the standard software results in the best data quality because the process is supervised and transparent. Integration errors are relatively easy to detect and correct through well designed and functional user interfaces. Furthermore, the output does not contain contaminant peaks, isotopes, (auto)adducts, fragments, multiple charge states etc., in other words it is very clean data. However, the price of good data quality is time, in the order of 2 to 4 weeks (or more), depending on the number of target metabolites and number of samples. This data extraction approach has in addition to its time consuming nature another more relevant major drawback. It does not take full advantage of the richness of the data and the conceptual aspect of metabolomics, the data is incomplete, information and markers are being lost in the process. The only solution to this is the use of brute force comprehensive automated data extraction tools.

Instrument companies have only recently become active in producing automated data extraction tools for metabolomics and proteomics (e.g. Waters, AB, Thermo, Agilent, Bruker) for selling Plug & Play Metabolomics Systems. Major disadvantages of these proprietary tools are 1) they only work for specific types of data and data formats, and 2) they are black box systems (little is known about the underlying algorithms) At the same time more and more 3rd party software is becoming available for automated data extraction. These programs offer more flexibility with respect to the format of the input LC-MS data which is a clear advantage when using instruments from different vendors because data from the different instruments can be processed in exactly the same way.

The proteomics and metabolomics communities (and also others) have been very active the last 15 years in developing their own tools mainly because commercial software was (is) not available and/or the poor performance of some of the available software. Some of these DIY data extraction tools are freely available on the web (OpenMS, mzMiner, MetAlign, etc.) and appear to be cheap solutions to our problems. In reality they are not cheap because they all require training and good understanding of what it is doing. On top of that issues such as support and long term continuity are not favorable.

The various automated data extraction tools (commercial, public domain and DIY) use different algorithms. Popular approaches include:
1) Peak detection and integration, followed by alignment in the feature space
2) Warping of the m/z - retention time plane (alignment at raw data level), background and blank subtraction, followed by peak detection and integration
3) De-isotoping of data after 1 or 2, e.g. correlation analysis, accurate mass data
4) Metabolite extraction by using spectrum information: deconvolution, spectrum libraries, accurate mass, etc.
5) ….

Analogous to NMR (see above) the extraction of metabolite information (#4) is strongly preferred compared to the extraction of all the features in the data: the possibility of getting 15000 features from a single LC-MS file seems very impressive, but these represent a much smaller number of metabolites.

A quick search on the web resulted in approx. 20 different tools for automated data extraction (commercial and public domain). None of these tools is generic, i.e. they can only be used for either low or high resolution data, centroid or profile data, etc. The functionality and performance of all these tools is a white spot on the metabolomics map, and it is widely known that all automated data extraction tools have their specific problems which result in data with a variable amount of errors. The fact that high quality raw data is being corrupted due to these errors is very alarming. Typical problems include:
- missed peaks
- integrated noise peaks
- database mismatches
- misalignment
- integration errors

These errors become more and more dominant in the data at low signal to noise ratios. All algorithms eventually stumble on classification problems such as: is this noise or signal?, is this peak A or B or neither of the two?, are these spectra the same?, etc. Unfortunately, many metabolites of interest happen to be present at low concentrations and at low signal to noise. These extraction errors have a major detrimental effect on the outcome of metabolomics studies. The majority of these tools, unlike the standard quantitation tools provided by instrument companies, are applied without supervision in batch processing mode, and typically don't have interactive review and error correction functionalities.

It is obvious that there is big need for improvements in the area of data extraction. The ultimate goal is a (set of) perfect tool(s), but it is expected this will take quite some time. Development of new better algorithms and software (GUI , database links, etc.) are relatively slow processes, and a lag between availability of the appropriate data extraction tools and rapid technological/methodological innovation is inevitable. A serious implication of this lag is that there is no point to waiting for the perfect solution. It is equally important to focus on workflow optimization to detect, characterize, and reduce data extraction errors for the currently available metabolomics data extraction tools, and to use this as input for improving the data extraction software.

The objectives of this workshop are :

- share experiences in data extraction
    - software used,
    - approaches, algorithms, and workflows,
    - detection of data extraction errors,
    - parallel workflows and combining results ('democratic' data extraction),
- should we setup a dedicated special interest group for data extraction (within NUGO, team up with other groups, e.g. Mx Society, proteomics community, …),
- do we see any value in benchmarking of tools: is this valuable? If yes, agree on an action plan, and
- to generate excellent ideas, whatever they may be!

Session 4 : Data analysis

Moderator : Bruce Kristal (Brigham and Women's Hospital, Boston)

Friday, the 14th of December, 2007, 9 a.m.

Room #4

Background:
In nearly any analysis of large datasets, such as those being generated in nutritional metabolomics, it becomes desirable to extract as much meaning as possible. In addition to a suitable design of the original study, this also relates to learning how to design data analysis experiments optimally for the different types of data and datasets of different structures.

Goal:
Begin to lay down the beginnings of a "best practice" document to help people run and interpret their own experiments, and to serve as more discerning referees to enhance the overall quality of data analysis in the field.

Approach:
Presentation and discussion of the following issues:

- Commonly used algorithms.
- Choice of one or more algorithms that are well suited to solving the problems at hand.
- Commonly used data scaling methods.
- Common fundamental mistakes and misunderstandings and how to avoid these.
- Use of background information such as the past century of biochemical pathway analysis for data analysis.

Result:
First materials to create a "best practice" document to help people run and interpret their own metabolomics experiments.

Session 4 : Data analysis

Bruce Kristal (Brigham and Women's Hospital, Boston)

Informatics Practice in Metabolomics: Establishing a Foundation for the Future

Human beings innately analyze large datasets with a speed and accuracy that far exceeds the best that any current computer can match (consider an infant's face recognition vs a computer's, for example). Nonetheless, we have not evolved, nor does it appear likely that we can be trained, to directly handle large, high-density numeric arrays, such as those generated by metabolomics or other -omics experiments. We therefore turn to a suite of computer-based approaches to try to understand the data. Among the basic schemes include classification, dimensional reduction, visualization, pattern recognition, and modeling. Systems biology-related approaches, such as pathway modeling, were handled in a different workshop, and will not be further addressed here.

The primary focus for this workshop was to first establish a common starting point (ie, what is informatics), and then begin to build a consensus of which mathematical approaches are usually problematic, and which are generally valid. Included in this is an effort to understand where the counter-examples are. The goal being to build a strong foundation for moving forward in developing a general best practices approach

So what is informatics? At some level, it is any mathematical tool that can be used to simplify and understand a complex (ie large number of variables) dataset. The first distinction one must attempt to understand is the difference between informatics and frequentist statistics (eg, t-tests, ANOVA). Informatics is a field whose major goal is to understand what happened and/or predict if it will happen again. In contrast, frequentist statistics tells us whether what occurred was likely to have occurred by chance. The second key point is that informatics is a term used to cover a broad series of approaches, such as clustering for de novo classification, principal components for dimensional reduction, support vector machines for pattern recognition, etc… We cannot cover each of these in any detail; rather we focused on general concepts.

One such concept lies in the need to very carefully define the question(s) of interest. For example, a general study of the metabolome present in people with and without Alzheimer's disease (AD) is inherently different than a study geared at building a profile that can distinguish AD from controls. This, in turn, differs from one that seeks to distinguish males with and without AD, or males between 75 and 90 with/without AD. Similarly, what constitutes a control group? Is someone with multi-infarct dementia included? Parkinson's associated dementia? Are they equally represented? Are numbers representative of the general population? This continues as one refines and defines progressively more specific/narrow criteria. Another choice is whether one wants to find the few most powerful single markers, or whether is most aimed at a general profile. Is the long-term goal clinical or scientific (this changes the acceptable options for model building). Defining group characteristics further differs from predicting group membership, and looking at inter-group subgroups opens other doors. Thus to reiterate – the central point is to know your question, and, if you are running the informatics for others, make sure you know the question that they are/were (NOT always the same…a recipe for disaster) trying to answer.

The choice of experimental question is – or, more precisely, needs to be, directly and inextricably linked with the experimental design. It is at best extremely difficult and, often, impossible to rescue a study that is poorly designed. By building the informatics in as part of the experimental design, it is possible to avoid studies that are doomed from the start and to make sure the study will be sufficiently powerful (statistically speaking) to make statements, and powerful enough at the over-fitting level

Another issue is choice of algorithm. Here, both scientific and two types of practical consideration are critical. Scientific decisions include making decisions between different algorithms and classes of algorithms. In some cases, the choice can be dictated by what was decided above (eg, a descriptive technique vs a discriminant technique) – in other cases the choice may "relate to the art of the user" Examples here include the choice between the different distance counting algorithms for clustering – or between hierarchal and partitional clustering. One line of practical issues centers on the available personnel resources. Some assays, such as the use of R or the running of genetic algorithms requires significant expertise. Thus, one of the practical issues is simply available expertise. The other practical issue relates to cost. The use of many of these algorithms can be through publicly available software, but other programs can cost 10,000 USD or more – a cost that can become prohibitive for a lab that wants to use multiple programs (ie, wants to use multiple approaches).

Another set of issues center on the consideration of an informatics workflow. Informatics workflows vary, often considerably, based on the questions being asked and the data one is working with, as well as, again, with "the art of the user". One potential workflow would look something like…

(i) Primary data manipulation: Data validation, data normalization with respect to experimental groups or standards or analytical issues, decisions about missing data handling, samples and variable level inclusion exclusion criteria.

(ii) Primary model building and analysis: Models are built for the sake of data quality decisions. Class specific analysis can be run to examine grouping within a class, eg, are there subgroups? Outlier decisions are also often made here, although such decisions may be made at multiple levels. Testing different scaling and transformation approaches is sometimes done here.

(iii) Unsupervised analysis. Here one simply hands the data to the computer and waits for brilliant results. Algorithms here include clustering, self-organizing maps, and principal components analysis. In general, the goal here is to understand the data at a global level, eg, outliers, overall structure, strength of differences, unexpected groupings (male female differences sometimes show up here, different trial sites often show up here.

(iv) Supervised analysis. Here one uses an initial set to build a model that one hopes will recognize the group in the future. These include tools such as kNN, SIMCA, PLS, PLS-DA, O-PLS-DA, Random Forest. The goal here is to build as powerful a classification model as possible. Note that these results are often not biologically interpretable.

(v) Machine learning. Here one essentially tells the computer…Here's a training set, with defined groups, and then lets the computer find one or more rules that can distinguish the groups. Often such studies are run iteratively, to attempt to reduce over-fitting. Three commonly observed approaches in other fields are artificial neural networks, genetic algorithms, and genetic programs. In some cases (eg, genetic algorithms) the results can often be interpreted. In others, such as many artificial neural nets, the data usually can't be.

(vi) Validation phase. Objective tests are used to define the quality of the model. Here one begins to find out just how many mistakes were made in the other phases. One sequence here might be internal validation, model optimization, External validation, optimization, 2^{o} validation.

It is notable, and deliberate, that the above section ended with the problems of overfitting. This is probably, today, the single greatest informatics problem that we observe. Basically, overfitting occurs when one uses an informatics approach to describe a dataset, for example using PLS-DA to find metabolites that distinguish two classes of interest. The problem lies in the ability of algorithms such as PLS-DA to find solutions in some spaces where no real solution exists. For example, PLS-DA can separate two groups comprised completely of random data. The gold standard here is replication in a blind, new dataset. There are at least four fundamentally different versions of over-fitting problems that we face as a field:

(i) People who truly have no idea what they are doing. The first is correctable only by broad training and exposure of these populations to correct practice(s). Note that many people often fall into this category in their early papers – the question is how many get beyond it.

(ii) People who are essentially cooking their data deliberately. This varies from people who are bordering on outright fraud to people who are deliberately taking advantage of every gray area in informatics.

(iii) People who are essentially over-replicating until it works. In this case, the individual build and tests a model, when it fails, they build another…This goes on iteratively until a model succeeds, which is then reported.

(iv) Datasets that are subtly non-independent/non-valid or non-independent, often in a way the informatician cannot know. An example of the former occurred to me when I was asked to analyze two datasets attempting to distinguish two groups. The datasets split beautifully, the only clue that something was amiss was a weird subgroup structure, turns out many of the controls were from a different study, thus the replication works because the second dataset was invalid. For the latter case, consider the situation where an informatician is analyzing a dataset blindly, and they are asked to analyze two sets of samples. They build a perfectly legitimate model on set one, and then blindly analyze the second set, and find their model holds. This is not necessarily useful, imagine, for example, the two groups being replicate samples.

One of the solutions, indeed arguably the best solution, to this suite of problems is full disclosure of what is done. The overall goal of the Metabolomic Standard Initiative (http://msi-workgroups.sourceforge.net/) is to generally enable different metabolomics groups to talk with each other . Within the informatics component of the standards initiative the goal is, in part, to request, and, ideally, eventually demand, sufficient information so as to be able to really understand what was done and catch key problems.

It is at this point, and only at this point, that we can legitimately get to another point – what is the most powerful way to attack metabolomics data. This may be broken into at least three pieces. Do we know, for any method, tools which always work? To my knowledge, the answer is no. (ii) Do we know for any type of project, tools which always outperform other tools (more power)? In general, in my experience, components based analysis (PCA, PLS, PLS-DA, etc always performs as well as or better than pure distance based algorithms, such as clustering and self-organizing maps, on metabolomics data. That said, we have looked at other data sets (non-metabolomics) where SOMs and clustering outperform projections. This suggests that we are not yet ready to answer this question. Do we know tools that yield more robust data? Again, our metabolomics data has generally done well with projections, but they can clearly overfit.

So where does this leave us? I believe that we can begin to define certain practices that are always invalid, or, alternatively, should always be used with extremely careful disclosure to describe the purpose of the analysis and the reason a normally invalid method was used (eg, to look for outliers or subgroups). In addition, we can use the current knowledge to begin to define a series of approaches which often work, and the optimal way of using each of these, thus establishing a solid platform on which to build this field.

Session 5 : Metabolite identification

Moderator : David Wishart (University of Alberta)

Friday, the 14th of December, 2007, 9 a.m.

Amphitheater Jarrige

Background:
Metabolite identification is an essential and often very complex and time consuming part of metabolomics studies. In addition to central databases containing compound information, many groups active in this area create their own databases, related to their own application. In general, the field could profit if the setup of the local databases is such that they can be used by others and shared between various groups.

Goal:
Exchange experience in storing and using identification information. In the end, increasing the number of identified and quantified metabolites in common matrices in nutritional metabolomics studies.

Approach:
Discussion and inventory of needs of the participants:

- GC-MS databases - what is needed to make them relevant to metabolomics and biomarker profiling.
- GC-GC-MS databases - how to prepare appropriate databases and search tools for metabolite identification and quantification.
- LC-MS and LC-MS/MS databases - what kinds of information need to be captured/displayed to make them widely useful.
- CE-MS, LC-IMS - can/should spectral databases for these technologies be established?
- Quantification of metabolite data by MS - how to do it?
- Automated compound identification by GC-MS and LC-MS/MS - how spectral databases can play a role.
- Defining biomarkers - is one compound enough? Are multiple compounds required? How to define a robust biomarker profile?
- Can/should biomarker profiles be databased?
- The role of private/public spectral databases and software in metabolomics and biomarker identification.

Result:
An overview of the approaches and available databases used by the participants. Create a list of items to be addressed in the paper of the best practice in identification.

Session 5 : Metabolite identification

David Wishart (University. of Alberta)

Metabolite identification is an essential part of any metabolomics experiment. Unfortunately, because metabolite ID is one of the most difficult and time consuming steps in metabolomics this crucial process is often deferred to the final stages of many studies and therefore left largely unfinished. In some cases, it is ignored altogether. Without formal compound identification, the discovery of any metabolically interesting patterns or clusters (via PCA or OPLS-DA) is largely meaningless.

The challenge of metabolite identification, especially for mass spectrometry, lies in the fact that there are potentially thousands of compounds that can match a given parent ion mass or a given atomic composition. The situation is made even worse given that most metabolomics experiments generate hundreds or even thousands of different masses. In the past, compound identification via mass spectrometry required the use of complementary analytical techniques, such as NMR, IR, or specialized chemical assays. However, this is now changing. In particular, with the introduction of improved separation technologies, higher resolution mass spectrometers, smarter "mass analysis" algorithms, more innovative chemical labelling schemes, more comprehensive MS databases and a better understanding of the "standard" metabolic composition of most organisms, it is now possible to confidently and rapidly identify many metabolites via MS [1-4].

MS-based metabolomics experiments can be done in any number of ways. The simplest approach is to study the biofluid or tissue extract by direct injection (thereby yielding a collection of parent ion masses or parent ion adducts). More commonly the mixture of interest is separated using gas chromatography (GC), liquid chromatography (LC) or capillary electrophoresis (CE) and then injected into a mass spectrometer. Prior to separation the mixture of interest may be trimethylsilated (for GC-MS) or it may be dimethylated or dansylated (for LC-MS) [2,4]. Separations can be done using one-dimensional, two dimensional or even multidimensional methods via gas phase, reversed-phase, ion exchange or normal phase methods. Furthermore, mass analyzers may employ parent-ion-only measurements, parent ion + electron impact (EI) fragmentation or parent ion + soft secondary fragmentation (MS/MS). Consequently a given MS-based metabolomics experiment can generate 3 general types of data or mass spectral tags (MSTs): 1) parent mass only; 2) parent mass + chromatographic retention time or 3) parent mass + fragment mass + chromatographic retention time. These properties, if properly documented, can allow identification of both previously known and hitherto unidentified compounds. The identification of unknowns is done by measuring the MSTs of postulated pure/authentic compounds that may match features of the unknown MST [2].

This document discusses the needs and requirements regarding data, databases and software for MS-based metabolite identification. In particular we have attempted to provide a framework or a series of recommendations regarding how metabolite identification can be facilitated in an MS-based metabolomics experiment. In doing so we have used the databases and software models developed for MS-based proteomics as a template for MS-based metabolomics.

MS-based Metabolite Databases

There are 3 types of MS-based Metabolite databases: 1) those that provide raw, unannotated MS (GC-MS or LC-MS) spectral data of biofluids or tissue extracts; 2) those that provide annotated MS (GC-MS or LC-MS) spectral data or MST's of biofluids or tissue extracts and 3) those that provide reference MS spectra or MST's (GC-MS or LC-MS) of pure compounds. The first two databases represent archival reference data of MS-based metabolomics experiments. Examples include Metlin [5], the Golm Metabolome database [6] and other databases [7]. They are very valuable resources and can be used to facilitate compound identification as well as general metabolomic analysis. However, the most useful databases for compound identification are those in category 3. These include the NIST GC-MS database and the HMDB LC-MS/MS library [8]. Minimally, these kinds of databases should include the name, chemical formula and monoisotopic mass (to 5 decimal places) of a significant number (>500) of metabolites or chemically modified (trimethylsilated) metabolites. Ideally these databases should also include chromatographic retention times or retention indices (for both GC and LC separations) as well as fragment ion mass data (EI or MS/MS fragmentation). As with similar kinds of proteomics MS databases, these metabolomics MS databases should be searchable by 1) elemental composition; 2) parent ion mass; 3) retention time/index; and/or 4) mass fragment pattern. A particularly useful addition to these kinds of databases would be the possibility of restricting the search to certain kinds of compounds (i.e. endogenous metabolites only, known mammalian metabolites only, toxins only, metabolites specific to a certain tissue or biofluid, etc.) or combinations of metabolite classes. Furthermore these databases should provide experimental details (images of spectra, collection conditions, MS spectrometer parameters, collision energies, instrument type, date of collection, lab/individual who collected the data) concerning the origin or source of their pure compound reference spectra. While it would be preferable if these MS-based metabolomics databases were freely accessible and freely downloadable (as many MS-based proteomics databases are), we appreciate the fact that commercial possibilities exist for the sale and distribution of these kinds of MS resources. Nevertheless, we would encourage the community (both academic and industrial) to do their best to make their database resources ***publicly*** accessible and adherent to these standards.

Because the number of compounds of interest to metabolomics researchers easily numbers in the thousands and because different organisms have profoundly different metabolomes, it is almost impossible for a single lab or a single investigator to have access or an interest in collecting referential MS spectra for ALL metabolites. Therefore, there is a need to establish a process by which individuals from many different metabolomic interests or backgrounds may contribute to a common respository of MS (GC or LC) reference compound spectra. While several compound reference databases already exist (NIST, the SDBS) these are general chemical databases and they are not limited to metabolites or compounds of biological interest. What is really needed is the equivalent of a GenBank or PDB for MS spectral deposition. Such a model allows users to deposit data to a common repository so that it can be searched, shared or used by other scientists. Already the NMR metabolomics community has started depositing their reference compound spectra into a common repository, called the BioMagResBank [9]. We would advocate for the establishment of a similar entity (lets call it the BioMSBank) or perhaps the expansion of the BioMagResBank to support submission (and searching) of referential MS (LC-MS and GC-MS) metabolite spectra. Clearly this would require long-term support by either a granting agency or perhaps the Metabolomics Society.

MS-based Metabolite Identification Software

Over the past decade, the trend for most "omics" software is towards web-based applications or web-server delivery. This has numerous advantages (accessibility, speed, improved maintenance, no need for installation support) and it is the way in which most MS-based proteomics software is now available – both commercially and academically. We believe that MS-based Metabolite Identification software should move to this model. We also believe that MS-based metabolite identification software should emulate many of the features used by MS-based proteomics software. This includes support for database or subdatabase selection, support for different mass list input formats, support for chemical modification corrections (in this case TMS or dimethylation), and inclusion of additional MST data (retention index, parent ion mass, fragment ion patterns, etc.). Obviously this kind of software needs to sit atop an appropriately maintained database of referential metabolite spectra (see above).

A separate kind of software tool that is unique to metabolomics (and small molecule MS work) is elemental or chemical formula prediction software. Given sufficiently high mass accuracy (~5 ppm) and resolution, it is possible to use the parent ion mass spectrum to determine the elemental composition and, in many cases, the identity or approximate identity of a compound. Recently Kind and Feihn have developed a series of 7 heuristic rules for chemical formula extraction and compound identification (or ranking) from high resolution MS spectra [1]. With the increasing availability of FT-ICR and Orbitrap MS instruments, there is a distinct possibility that this approach may offer a powerful adjunct, or even replacement, to MST-based metabolite identification. In particular, with the growing knowledge about what is found or findable in many metabolomes (~1000 metabolites in microbes, ~3000 metabolites in mammals), the use of chemical formula prediction, followed by rapid scanning for formula matches to known metabolite lists could prove to be a very fast, simple and robust method to metabolite identification – especially among organisms that have well characterized metabolomes. We would advocate that more effort be devoted to this particular kind of MS-based metabolite identification and that the software be made available through easily accessible web-server applications.

Metabolite Quantification and Standards

While MS-based approaches are widely recognized for their sensitivity and capacity to identify large (>100's) of metabolites, they are not generally recognized as being useful in metabolite quantification. This has often been the Achilles heel to many MS-based approaches. However, with the recent successes in using isotopic affinity tags (such as ICAT) to make MS-based proteomics reasonably quantitative, there is a distinct possibility that similar isotope tagging methods could make MS-based metabolomics equally quantitative. We believe that the issue of metabolite quantification should be made a priority in MS labs and that efforts that focus methodological improvements or software improvements to make rapid and accurate metabolite quantification possible should receive both encouragement and support from the metabolomics community. Indeed, identification without quantification is the bane of many analytical chemists – as well as many metabolomics researchers.

Recently the proteomics community has pushed for the establishment of standards to assess the performance of instruments, methods and labs in identifying and quantifying proteins from defined mixtures. We believe that similar initiatives should be undertaken in the metabolomics community. The use of defined metabolite mixtures (specific to plants, microbes and mammals) would provide a means to objectively assess the performance and reliability of algorithms, databases and protocols used in MS-based metabolite identification and quantification. Indeed, the use of standardized metabolite mixtures would provide much-needed validation of existing methods and a means of objective assessment of emerging methods.

References

1. Kind T, Fiehn O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics. 2007 Mar 27;8:105.

2. Kopka J. Current challenges and developments in GC-MS based metabolite profiling technology. J Biotechnol. 2006 Jun 25;124(1):312-22.

3. Tikunov Y, Lommen A, de Vos CH, Verhoeven HA, Bino RJ, Hall RD, Bovy AG. A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. Plant Physiol. 2005 Nov;139(3):1125-37.

4. Guo K, Ji C, Li L. Stable-Isotope Dimethylation Labeling Combined with LC-ESI MS for Quantification of Amine-Containing Metabolites in Biological Samples. Anal Chem. 2007 Nov 15;79(22):8631-8638.

5. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G. METLIN: a metabolite mass spectral database. Ther Drug Monit. 2005 Dec;27(6):747-51.

6. Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmüller E, Dörmann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D. GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics. 2005 Apr 15;21(8):1635-8.

7. Moco S, Bino RJ, Vorst O, Verhoeven HA, de Groot J, van Beek TA, Vervoort J, de Vos CH. A liquid chromatography-mass spectrometry-based metabolome database for tomato. Plant Physiol. 2006 Aug;141(4):1205-18.

8. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P,

Session 6 : Biological interpretation

Moderators : Suzan Wopereis, Ben van Ommen (TNO Zeist)

Thursday, the 13th of December, 2007, 9 a.m.

Room # 4

Background:
After having 'walked' through the metabolomics pipeline without failure and having found those effects caused by the nutritional intervention and identified all the relevant metabolites, it is essential to 'understand' what the data is telling you in order to make your study to a success. At the moment tools for the biological interpretation of metabolomics data are scarce. In this session we shall discuss how to improve the biological interpretation of metabolomics data.

Goal:
We will decide whether it would be useful for NuGO to create an intergrated workbench for metabolomics data interpretation, according to the Genepattern model (http://www.broad.mit.edu/cancer/software/genepattern/).

Approach:
Discussion of the following issues:

1. **Inventory of available tools**
   - Pathway analysis (Pathvisio, Metacore, other open source developed tools)
   - From pathways to processes (Metabo-go ontology)
   - Network biology (Metacore, …)
   - Multi-layer omics tools, data fusion
2. **Nutritional phenotype issues**
   - Plasma biochemistry – how can we relate the phenotype (commonly expressed as specific plasma biochemical measurements) to our metabolomics results?
   - Study design: challenges, plasma vs. urine, do we need a standard run in diet?
   - Biology driven analysis – Are we really measuring those compounds that can provide us the answers to our research questions or should we create a list with all relevant metabolites that (at least) should be measured?
   - Exploit "confounders" – Do we need an endocrine platform? Do we need to create subgroups based on geno- /phenotype? What factors influence the nutritional metabolome?
   - Fluxes – do we need information on the dynamics of metabolites for the biological interpretation?
   - How to separate endo- and exo–(food) metabolome
   - Validation studies – how can we validate our biological interpretation?
3. **Towards and integrated workbench?**
   Is it useful to create an integrated workbench for metabolomics data interpretation, according to the Genepattern model?

Result:
An overview of the approaches and tools for biological interpretation of metabolomics studies used by the participants. Decision on how biological interpretation in nutritional metabolomics studies can be improved and whether it would be useful if NuGO has an integrated workbench for metabolomics data interpretation.

Session 6 : Biological Interpretation

Suzan Wopereis, Ben van Ommen (TNO Zeist)

Background:

After having ‘walked’ through the metabolomics pipeline without failure and having found those effects caused by the nutritional intervention and identified all the relevant metabolites, it is essential to ‘understand’ what the data is telling you in order to make your study to a success. At the moment tools for the biological interpretation of metabolomics data are scarce. In this section we will discuss how to improve the biological interpretation of metabolomics data in nutrigenomics research.

Challenges:

<u>On the level of study design</u>

1. Nutrigenomics is the study of molecular relationships between nutrition and the complexity of molecular processes, commonly measured with “omics” techniques, with the aim to extrapolating how such subtle changes can affect human health (Chavez A, 2003; Müller & Kersten, 2003; van Ommen, 2004; Afman & Müller, 2006). Metabolomics as technology in nutrigenomics research could be of assistance in the unraveling of important disease mechanisms and also lead to the discovery of early biomarkers. One of the challenges for metabolomics in nutrigenomics research is to be able to seperate subtle metabolic changes from noise in your data. A good study design is crucial in this respect.

<u>On a single metabolite level</u>

2. *Do we understand the biological function of the metabolite in the matrix analysed?* This question can be addressed on two levels. The first is its “intrinsic” function. Glucose has a primary biological function as energy source. The second is its relevance related to the matrix. Increased glucose in urine indicates diabetes, while its increase in plasma is primarily correlated to the post-prandial state and insulin sensitivity. Many other metabolites have less known or more complex functions. Intra-organ concentrations have different relevance, and even intra-cellular compartmentalisation plays a role. Plasma carnitine is not straightforwardly related to mitochondrial carnitine concentration. Same story for glutathione concentrations.

   The HMDB database (http://www.hmdb.ca/; Wishart et al., 2007) provides a treasure of information on metabolite properties. NuGOwiki (http://www.nugowiki.org/) potentially adds to this knowledge base. Is this good enough? Is this redundant, as wikipedia steadily grows? Should we come to a concerted action in compiling a metabolome based biological function description? What can we learn from the clinical chemists?

3. *Do we understand the relevance of the metabolite's changed body fluid concentration?*
   Can we translate organ processes to body fluid readouts? Biochemistry text books teach us about the glycolysis, but this does not occur in plasma. So, can we make a simple comparison between organ changes and plasma changes? Usually not. The body strives to maintain homeostasis, and organs need to work hard to keep plasma concentrations between acceptable boundaries, thereby showing much more fluctuation in (e.g.) liver. Inverse correlations may even be observed: oxidative stress produces a sharp drop of glutathione (GSH) in liver but due to liver damage, leaking of GSH may increase its concentration in plasma. Urine functions as an accumulating waste basket of exogenous compounds and their metabolites, while plasma levels of these compounds need to be as low as possible. So, translation is not straightforward. Statistical relationships between various compartments should be treated with care.

<u>On a complex level (profiles)</u>

4. *Do profiles really tell us more than the sum of the individual components?*
   Statistics will answer this question positively, but biology struggles with this truth. Biologists simply want to understand the relevance of each individual change. Do we therefore need to unravel the biochemical relationships between the components of a "profile"? Pathway tools like Pathvisio and biological network tools like Metacore (GeneGo Inc., St. Joseph, MI, USA) and IPA (Ingenuity Systems, Redwood City, CA, USA) provide a first attempt to connect these components but (also given the complex relationship between body fluids and organ biochemistry described above) are not (yet) up to the task. It looks like till now it is 90% brains and 10% tools.

5. *Can we separate the food metabolome from the health/disease effects hidden in the metabolome?*
   Macronutrients are metabolized and impact health. So, metabolomics reveals both and this creates problems, especially since many macronutrient metabolites act as bioactives (Gibney MJ, 2005). This complicates interpretation. Is it possible to divide the changes detected in the metabolome as a consequence of a diet intervention into changes deriving from the food and changes deriving from an effect on health?

6. *Matching metabolomics technology to biology*.
   In describing nutrition and health relationships, biologists often can produce a "wishlist" of metabolites to be identified and quantified. So far, metabolomics has primarily been a technology-driven science: metabolite panels were provided on physico-chemical properties, and usually these do not completely match with the biologists dream. Many times, they moaned at seeing the amino acids change again… Lipidomics provides a more biologically coherent metabolite panel and is a method of first choice for many papers. The inflammation-related oxylipids as separated and quantified by Newman and Pedersen is a jewel of a biologically relevant targeted lipidome (Newman JW, 2007). Such platforms make life easy for the metabolome-biologist.

Adding to this is that biological interpretation often has to be performed on a small number of metabolites. In discovery driven metabolomics platforms often limited numbers of peaks are identified. After multivariate data-analysis has been performed on the data, only the most relevant metabolites will go into the identification process. It is challeging to get a good biological interpretation based on only fragments of the overall picture.

7. *How can we deal with time-course metabolomics profiles?* Disease progression with early, intermediate and late stage development has been shown to have distinct time courses of metabolome profiles. Much of this cannot easily be explained.

8. *Linking the metabolome to other measurements*
   Although it may seem to increase complexity, a breakthrough may be obtained by integrating metabolomics with other phenotypic and possibly genotypic information. After all, we see metabolomics as a detailed phenotypic fingerprint, so let`s embed these observation in the science we understand better. Histology, function tests, gene expression, (targeted) proteomics etc. may help to further understand the metabolome biology.

Solutions:

Nutrition deals with metabolism, oxidation and inflammation as primary processes that maintain or affect the health – disease relationship. So, metabolites / metabolomics needs to focus on these processes, and (in case of human studies) in plasma. An integrated bioinformatics solution that helps to map and visualise profile results in this area, and links to knowledge bases might be useful. This means: translating metabolite changes to process changes, translating profiles to processes, translating plasma observations to organ (mechanism) processes. All of this sounds like science fiction but we might want to use this as an umbrella for toolkit development.

What should the nutritional metabolomics bioinformatics toolkit contain:

'Biological driven' metabolomics platforms: Biologist should create a wishlist of metabolites that are most relevant in the overarching processes metabolism, oxidation and inflammation. Metabolomics platforms should be developed that cover the most important compounds for the overarching processes.

Time-course metabolomics (+ fluxomics): The use of dynamic data instead of static data. This provides a dynamic view of changes in metabolic pathways and processes from the organs. In this way you are able to separate time effects – representing processes and pathology – from diet effects – representing concentration differences as cause from change in diets.

Use of challenge tests: The use of challenge tests – like for example the oral glucose tolerance test (OGTT) – in combination with a timecourse facilitates the detection of subtle effects in metabolomics data. When you measure at homeostatic level in subjects in studies in which only small treatment effects are expected you will probably only find differences in those components (proteins & genes) that are involved in the regulation of homeostasis – even in a fasted timecourse! The cause of this is the strong buffering capacity of plasma and many confounders. There are two reasons that a challenge test combined with a timecourse enables the detection of subtle alterations a) subtle alterations in the timecourse are repeatedly confirmed – without disturbing day to day variation and b) by perturbing a homeostatic metabolic situation, differences in response to a challenge become visible.

Metabo-ontology (MO): In the world of transcriptomics the Gene Ontology Consortium developed controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions. We should also have that for metabolites. This can than be used for metabolite enrichment statistics.

Interactomes: This represents the visualization of plasma metabolite changes in relation to organ specific processes and pathways into interactome maps. Organ specific affected processes have been translated into plasma specific metabolite profiles.

Integrated workbench: All the different data analysis tools needed for nutritional metabolomics are unified into an integrated workbench, following the example of Genepattern (Reich M, 2006). Genepattern is a software package developed for transcriptomics data analysis, which provides a comprehensive environment that can support (i) a broad community of users at all levels of computational experience and sophistication, (ii) access to a repository of analytic and visualization tools and easy creation of complex analytic methods from them and (iii) the rapid development and dissemination of new methods.

List of publications relevant to the workshop

Review

Dettmer, K., Aronov, P.A. and Hammock, B.D. (2007) Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews* **26**, 51-78.

1. Sampling, storing and processing of biological samples

Storage and stability of urine samples

Gika, H.G., Theodoridis, G.A., Wingate, J.E. and Wilson, I.D. (2007) Within-Day Reproducibility of an HPLC-MS-Based Method for Metabonomic Analysis: Application to Human Urine. *J. Proteome Res.* **6**, 3291-3303.

Zhang, Q., Wang, G., Du, Y., Zhu, L. and A, J. (2007) GC/MS analysis of the rat urine for metabonomic research. *Journal of Chromatography B* **854**, 20-25.

Lauridsen, M., Hansen, S.H., Jaroszewski, J.W. and Cornett, C. (2007) Human urine as test material in H-1 NMR-based metabonomics: Recommendations for sample preparation and storage. *Analytical Chemistry* **79**, 1181-1186.

Plasma/serum samples

Boernsen, K.O., Gatzek, S. and Imbert, G. (2005) Controlled protein precipitation in combination with chip-based nanospray infusion mass spectrometry. An approach for metabolomics profiling of plasma. *Analytical Chemistry* **77**, 7255-7264.

Want, E.J., O'Maille, G., Smith, C.A., Brandon, T.R., Uritboonthai, W., Qin, C., Trauger, S.A. and Siuzdak, G. (2006) Solvent-dependent metabolite distribution, clustering, and protein extraction for serum profiling with mass spectrometry. *Analytical Chemistry* **78**, 743-752.

Teahan, O., Gamble, S., Holmes, E., Waxman, J., Nicholson, J.K., Bevan, C. and Keun, H.C. (2006) Impact of analytical bias in metabonomic studies of human blood serum and plasma. *Analytical Chemistry* **78**, 4307-4318.

Jiye, A., Trygg, J., Gullberg, J., Johansson, A.I., Jonsson, P., Antti, H., Marklund, S.L. and Moritz, T. (2005) Extraction and GC/MS analysis of the human blood plasma metabolome. *Analytical Chemistry* **77**, 8086-8094.

Standardization

Griffin, J., Nicholls, A., Daykin, C., Heald, S., Keun, H., Schuppe-Koistinen, I., Griffiths, J., Cheng, L., Rocca-Serra, P., Rubtsov, D. and Robertson, D. (2007) Standard reporting requirements for biological samples in metabolomics experiments: mammalian/in vivo experiments. *Metabolomics* **3**, 179-188.

2. Mass spectrometry analyses

Reviews

Wilson, I.D., Plumb, R., Granger, J., Major, H., Williams, R. and Lenz, E.M. (2005) HPLC-MS-based methods for the study of metabonomics. *J Chromatogr B Analyt Technol Biomed Life Sci* **817**, 67-76.

Chambers, E., Wagrowski-Diehl, D.M., Lu, Z.L. and Mazzeo, J.R. (2007) Systematic and comprehensive strategy for reducing matrix effects in LC/MS/MS analyses. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **852**, 22-34.

LC-MS, quality control

Gika, H.G., Theodoridis, G.A., Wingate, J.E. and Wilson, I.D. (2007) Within-Day Reproducibility of an HPLC-MS-Based Method for Metabonomic Analysis: Application to Human Urine. *J. Proteome Res.* **6**, 3291-3303.

Wagner, S., Scholz, K., Sieber, M., Kellert, M. and Voelkel, W. (2007) Tools in metabonomics: An integrated validation approach for LC-MS metabolic profiling of mercapturic acids in human urine. *Analytical Chemistry* **79**, 2918-2926.

Waybright, T.J., Van, Q.N., Muschik, G.M., Conrads, T.P., Veenstra, T.D. and Issaq, H.J. (2006) LC-MS in metabonomics: Optimization of experimental conditions for the analysis of metabolites in human urine. *Journal of Liquid Chromatography & Related Technologies* **29**, 2475-2497.

Sangster, T., Major, H., Plumb, R., Wilson, A.J. and Wilson, I.D. (2006) A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabonomic analysis. *Analyst* **131**, 1075-1078.

Shen, Y.F., Zhang, R., Moore, R.J., Kim, J., Metz, T.O., Hixson, K.K., Zhao, R., Livesay, E.A., Udseth, H.R. and Smith, R.D. (2005) Automated 20 kpsi RPLC-MS and MS/MS with chromatographic peak capacities of 1000-1500 and capabilities in proteomics and metabolomics. *Analytical Chemistry* **77**, 3090-3100.

UPLC

Crockford, D.J., Lindon, J.C., Cloarec, O., Plumb, R.S., Bruce, S.J., Zirah, S., Rainville, P., Stumpf, C.L., Johnson, K., Holmes, E. and Nicholson, J.K. (2006) Statistical search space reduction and two-dimensional data display approaches for UPLC-MS in biomarker discovery and pathway analysis. *Analytical Chemistry* **78**, 4398-4408.

Nordström, A., O'Maille, G., Qin, C. and Siuzdak, G. (2006) Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: Quantitative analysis of endogenous and exogenous metabolites in human serum. *Analytical Chemistry* **78**, 3289-3295.

Rainville, P.D., Stumpf, C.L., Shockcor, J.P., Plumb, R.S. and Nicholson, J.K. (2007) Novel application of reversed-phase UPLC-oaTOF-MS for lipid analysis in complex biological mixtures: A new tool for lipidomics. *Journal of Proteome Research* **6**, 552-558.

Bruce, S.J., Jonsson, P., Antti, H., Cloarec, O., Trygg, J., Marklund, S.L. and Moritz, T. (2007) Evaluation of a protocol for metabolic profiling studies on human blood plasma by combined ultra-performance liquid chromatography/mass spectrometry: From extraction to data analysis. *Analytical Biochemistry* **In Press, Uncorrected Proof**.

FT-ICR

Kind, T. and Fiehn, O. (2006) Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics* **7**, 234.

HILIC (urine)

Idborg, H., Zamani, L., Edlund, P.O., Schuppe-Koistinen, I. and Jacobsson, S.P. (2005) Metabolic fingerprinting of rat urine by LC/MS Part 1. Analysis by hydrophilic interaction liquid chromatography-electrospray ionization mass spectrometry. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **828**, 9-13.

Idborg, H., Zamani, L., Edlund, P.O., Schuppe-Koistinen, I. and Jacobsson, S.P. (2005) Metabolic fingerprinting of rat urine by LC/MS Part 2. Data pretreatment methods for handling of complex data. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **828**, 14-20.

GC-MS

Zhang, Q., Wang, G., Du, Y., Zhu, L. and A, J. (2007) GC/MS analysis of the rat urine for metabonomic research. *Journal of Chromatography B* **854**, 20-25.

Jiye, A., Trygg, J., Gullberg, J., Johansson, A.I., Jonsson, P., Antti, H., Marklund, S.L. and Moritz, T. (2005) Extraction and GC/MS analysis of the human blood plasma metabolome. *Analytical Chemistry* **77**, 8086-8094.

Major, H.J., Williams, R., Wilson, A.J. and Wilson, I.D. (2006) A metabonomic analysis of plasma from Zucker rat strains using gas chromatography/mass spectrometry and pattern recognition. *Rapid Communications in Mass Spectrometry* **20**, 3295-3302.

GC-GC-Tof

Hagan, S.O., Dunn, W.B., Knowles, J.D., Broadhurst, D., Williams, R., Ashworth, J.J., Cameron, M. and Kell, D.B. (2007) Closed-loop, multiobjective optimization of two-dimensional gas chromatography/mass spectrometry for serum metabolomics. *Analytical Chemistry* **79**, 464-476.

Capillary electrophoresis

Lee, R., Ptolemy, A.S., Niewczas, L. and Britz-McKibbin, P. (2007) Integrative metabolomics for characterizing unknown low-abundance metabolites by capillary electrophoresis-mass spectrometry with computer simulations. *Analytical Chemistry* **79**, 403-415.

Soga, T., Baran, R., Suematsu, M., Ueno, Y., Ikeda, S., Sakurakawa, T., Kakazu, Y., Ishikawa, T., Robert, M., Nishioka, T. and Tomita, M. (2006) Differential Metabolomics Reveals Ophthalmic Acid as an Oxidative Stress Biomarker Indicating Hepatic Glutathione Consumption. *J. Biol. Chem.* **281**, 16768-16776.

MALDI-Tof

Edwards, J.L. and Kennedy, R.T. (2005) Metabolomic analysis of eukaryotic tissue and prokaryotes using negative mode MALDI time-of-flight mass spectrometry. *Anal Chem* **77**, 2201-2209.

Lipidomics

Shui, G.H., Bendt, A.K., Pethe, K., Dick, T. and Wenk, M.R. (2007) Sensitive profiling of chemically diverse bioactive lipids. *Journal of Lipid Research* **48**, 1976-1984.

Watson, A.D. (2006) Lipidomics: a global approach to lipid analysis in biological systems. *Journal of Lipid Research* **47**, 2101-2111.

Sommer, U., Herscovitz, H., Welty, F.K. and Costello, C.E. (2006) LC-MS-based method for the qualitative and quantitative analysis of complex lipid mixtures. *Journal of Lipid Research* **47**, 804-814.

Schwudke, D., Oegema, J., Burton, L., Entchev, E., Hannich, J.T., Ejsing, C.S., Kurzchalia, T. and Shevchenko, A. (2006) Lipid profiling by multiple precursor and neutral loss scanning driven by the data-dependent acquisition. *Analytical Chemistry* **78**, 585-595.

Complementary techniques

Williams, R., Lenz, E.M., Wilson, A.J., Granger, J., Wilson, I.D., Major, H., Stumpf, C. and Plumb, R. (2006) A multi-analytical platform approach to the metabonomic analysis of plasma from normal and zucker (fa/fa) obese rats. *Molecular Biosystems* **2**, 174-183.

Standardization (minimal reporting)

Sumner, L., Amberg, A., Barrett, D., Beale, M., Beger, R., Daykin, C., Fan, T., Fiehn, O., Goodacre, R., Griffin, J., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A., Lindon, J., Marriott, P., Nicholls, A., Reily, M., Thaden, J. and Viant, M. (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3**, 211-221.

3. Data extraction

Reviews

Brown, M., Dunn, W.B., Ellis, D.I., Goodacre, R., Handl, J., Knowles, J.D., O'Hagan, S., Spasic, I. and Kell, D.B. (2005) A metabolome pipeline: from concept to data to knowledge. *Metabolomics* **1**, 39-51.

Shurubor, Y.I., Paolucci, U., Krasnikov, B.F., Matson, W.R. and Kristal, B.S. (2005) Analytical precision, biological variation, and mathematical normalization in high data density metabolomics. *Metabolomics* **1**, 75-85.

Alignment

Vorst, O., Vos, C.H.R.d., Lommen, A., Staps, R.V., Visser, R.G.F., Bino, R.J. and Hall, R.D. (2005) A non-directed approach to the differential analysis of multiple LC-MS-derived metabolic profiles. *Metabolomics* **1**, 169-180.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. and Siuzdak, G. (2006) XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **78**, 779-787.

Katajamaa, M. and Oresic, M. (2005) Processing methods for differential analysis of LC/MS profile data. *Bmc Bioinformatics* **6**, Art. No. 179.

Kind, T., Tolstikov, V., Fiehn, O. and Weiss, R.H. (2007) A comprehensive urinary metabolomic approach for identifying kidney cancer. *Analytical Biochemistry* **363**, 185-195.

Normalization

Wagner, S., Scholz, K., Sieber, M., Kellert, M. and Voelkel, W. (2007) Tools in metabonomics: An integrated validation approach for LC-MS metabolic profiling of mercapturic acids in human urine. *Analytical Chemistry* **79**, 2918-2926.

Saude, E., Adamko, D., Rowe, B., Marrie, T. and Sykes, B. (2007) Variation of metabolites in normal human urine. *Metabolomics*.

Kemperman, R.F.J., Horvatovich, P.L., Hoekman, B., Reijmers, T.H., Muskiet, F.A.J. and Bischoff, R. (2007) Comparative Urine Analysis by Liquid Chromatography-Mass Spectrometry and Multivariate Statistics: Method Development, Evaluation, and Application to Proteinuria. *J. Proteome Res.* **6**, 194-206.

Craig, A., Cloarec, O., Holmes, E., Nicholson, J.K. and Lindon, J.C. (2006) Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry* **78**, 2262-2267.

Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H. (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in H-1 NMR metabonomics. *Analytical Chemistry* **78**, 4281-4290.

Bijlsma, S., Bobeldijk, I., Verheij, E.R., Ramaker, R., Kochhar, S., Macdonald, I.A., van Ommen, B. and Smilde, A.K. (2006) Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation. *Anal Chem* **78**, 567-574.

Baran, R., Kochi, H., Saito, N., Suematsu, M., Soga, T., Nishioka, T., Robert, M. and Tomita, M. (2006) MathDAMP: a package for differential analysis of metabolite profiles. *BMC Bioinformatics* **7**, 530.

van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K. and van der Werf, M.J. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *Bmc Genomics* **7**, Art. No. 142.

GC-GC-Tof

Shellie, R.A., Welthagen, W., Zrostlikova, J., Spranger, J., Ristow, M., Fiehn, O. and Zimmermann, R. (2005) Statistical methods for comparing comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry results: Metabolomic analysis of mouse tissue extracts. *Journal of Chromatography A* **1086**, 83-90.

4. Data analysis

Reviews

Brown, M., Dunn, W.B., Ellis, D.I., Goodacre, R., Handl, J., Knowles, J.D., O'Hagan, S., Spasic, I. and Kell, D.B. (2005) A metabolome pipeline: from concept to data to knowledge. *Metabolomics* **1**, 39-51.

Broadhurst, D. and Kell, D. (2006) Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2**, 171-196.

Kemsley, E.K., Le Gall, G., Dainty, J.R., Watson, A.D., Harvey, L.J., Tapp, H.S. and Colquhoun, I.J. (2007) Multivariate techniques and their application in nutrition: a metabolomics case study. *British Journal of Nutrition* **98**, 1-14.

Various methods

Smilde, A.K., Jansen, J.J., Hoefsloot, H.C.J., Lamers, R., van der Greef, J. and Timmerman, M.E. (2005) ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics* **21**, 3043-3048.

Cloarec, O., Dumas, M.E., Craig, A., Barton, R.H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J.C., Holmes, E. and Nicholson, J. (2005) Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic H-1 NMR data sets. *Analytical Chemistry* **77**, 1282-1289.

Ramadan, Z., Jacobs, D., Grigorov, M. and Kochhar, S. (2006) Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms. *Talanta* **68**, 1683-1691.

Fusion of data

Smilde, A.K., van der Werf, M.J., Bijlsma, S., van der Werff-van der Vat, B.J. and Jellema, R.H. (2005) Fusion of mass spectrometry-based metabolomics data. *Anal Chem* **77**, 6729-6736.

Crockford, D.J., Holmes, E., Lindon, J.C., Plumb, R.S., Zirah, S., Bruce, S.J., Rainville, P., Stumpf, C.L. and Nicholson, J.K. (2006) Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: Application in metabonomic toxicology studies. *Analytical Chemistry* **78**, 363-371.

Forshed, J., Idborg, H. and Jacobsson, S.P. (2007) Evaluation of different techniques for data fusion of LC/MS and H-1-NMR. *Chemometrics and Intelligent Laboratory Systems* **85**, 102-109.

Validation

Rubingh, C., Bijlsma, S., Derks, E., Bobeldijk, I., Verheij, E., Kochhar, S. and Smilde, A. (2006) Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics* **2**, 53-61.

Ioannidis, J.P.A. (2007) Is molecular profiling ready for use in clinical decision making? *Oncologist* **12**, 301-311.

Ioannidis, J.P.A. (2005) Why most published research findings are false. *Plos Medicine* **2**, 696-701.

Standardization

Goodacre, R., Broadhurst, D., Smilde, A., Kristal, B., Baker, J., Beger, R., Bessant, C., Connor, S., Capuani, G., Craig, A., Ebbels, T., Kell, D., Manetti, C., Newton, J., Paternostro, G., Somorjai, R., Sjöström, M., Trygg, J. and Wulfert, F. (2007) Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* **3**, 231-241.

5. Metabolite identification

BinBase

Fiehn, O., Wohlgemuth, G. and Scholz, M. (2005) Setup and Annotation of Metabolomic Experiments by Integrating Biological and Mass Spectrometric Metadata. *Data Integration in the Life Sciences, Proceedings*, 224 – 239.

Identification through LC-MS

Kind, T. and Fiehn, O. (2006) Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics* **7**, 234.

Hegeman, A.D., Schulte, C.F., Cui, Q., Lewis, I.A., Huttlin, E.L., Eghbalnia, H., Harms, A.C., Ulrich, E.L., Markley, J.L. and Sussman, M.R. (2007) Stable Isotope Assisted Assignment of Elemental Compositions for Metabolomics. *Anal. Chem.* **79**, 6912-6921.

Identification through GC-MS spectra

Shellie, R.A., Welthagen, W., Zrostlikova, J., Spranger, J., Ristow, M., Fiehn, O. and Zimmermann, R. (2005) Statistical methods for comparing comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry results: Metabolomic analysis of mouse tissue extracts. *Journal of Chromatography A* **1086**, 83-90.

Tikunov, Y., Lommen, A., de Vos, C.H.R., Verhoeven, H.A., Bino, R.J., Hall, R.D. and Bovy, A.G. (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiology* **139**, 1125-1137.

Kopka, J. (2006) Current challenges and developments in GC-MS based metabolite profiling technology. *Journal of Biotechnology* **124**, 312-322.

Software

Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H. (2006) Metabolite projection analysis for fast identification of metabolites in metabonomics. Application in an amiodarone study. *Analytical Chemistry* **78**, 3551-3561.

Spasic, I., Dunn, W.B., Velarde, G., Tseng, A., Jenkins, H., Hardy, N., Oliver, S.G. and Kell, D.B. (2006) MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. *Bmc Bioinformatics* **7**.

Databases

Spasic, I., Dunn, W.B., Velarde, G., Tseng, A., Jenkins, H., Hardy, N., Oliver, S.G. and Kell, D.B. (2006) MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. *Bmc Bioinformatics* **7**.

Moco, S., Bino, R.J., Vorst, O., Verhoeven, H.A., de Groot, J., van Beek, T.A., Vervoort, J. and de Vos, C.H.R. (2006) A liquid chromatography-mass spectrometry-based metabolome database for tomato. *Plant Physiology* **141**, 1205-1218.

Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R. and Siuzdak, G. (2005) METLIN - A metabolite mass spectral database. *Therapeutic Drug Monitoring* **27**, 747-751.

Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmuller, E., Dormann, P., Weckwerth, W., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A.R. and Steinhauser, D. (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* **21**, 1635-1638.

Want, E.J., Nordstrom, A., Morita, H. and Siuzdak, G. (2007) From Exogenous to Endogenous: The Inevitable Imprint of Mass Spectrometry in Metabolomics. *J. Proteome Res.* **6**, 459-468.

Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D.D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G.E., MacInnis, G.D., Weljie, A.M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B.D., Vogel, H.J. and Querengesser, L. (2007) HMDB: the human metabolome database. *Nucleic Acids Research* **35**, D521-D526.

6. Biological interpretation

Metabolome varability

Fiehn, O., Wohlgemuth, G. and Scholz, M. (2005) Setup and Annotation of Metabolomic Experiments by Integrating Biological and Mass Spectrometric Metadata. *Data Integration in the Life Sciences, Proceedings*, 224 – 239.

Walsh, M.C., Brennan, L., Malthouse, J.P.G., Roche, H.M. and Gibney, M.J. (2006) Effect of acute dietary standardization on the urinary, plasma, and salivary metabolomic profiles of healthy humans. *American Journal of Clinical Nutrition* **84**, 531-539.

Stella, C., Beckwith-Hall, B., Cloarec, O., Holmes, E., Lindon, J.C., Powell, J., van der Ouderaa, F., Bingham, S., Cross, A.J. and Nicholson, J.K. (2006) Susceptibility of human metabolic phenotypes to dietary modulation. *Journal of Proteome Research* **5**, 2780-2788.

Inter-species comparison

Lamers, R.J.A.N., van Nesselrooij, J.H.J., Kraus, V.B., Jordan, J.M., Renner, J.B., Dragomir, A.D., Luta, G., van der Greef, J. and DeGroot, J. (2005) Identification of an urinary metabolite profile associated with osteoarthritis. *Osteoarthritis and Cartilage* **13**, 762-768.

Application to cohorts

Dumas, M.E., Maibaum, E.C., Teague, C., Ueshima, H., Zhou, B.F., Lindon, J.C., Nicholson, J.K., Stamler, J., Elliott, P., Chan, Q. and Holmes, E. (2006) Assessment of analytical reproducibility of H-1 NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP study. *Analytical Chemistry* **78**, 2199-2208.

Correlation analysis

Camacho, D., de la Fuente, A. and Mendes, P. (2005) The origin of correlations in metabolomics data. *Metabolomics* **1**, 53-63.

Pathway analysis

Curtis, R.K., Oresic, M. and Vidal-Puig, A. (2005) Pathways to the analysis of microarray data. *Trends in Biotechnology* **23**, 429-435.

Yetukuri, L., Katajamaa, M., Medina-Gomez, G., Seppanen-Laakso, T., Vidal-Puig, A. and Oresic, M. (2007) Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. *Bmc Systems Biology* **1**.

Pratical applications

Wikoff, W.R., Gangoiti, J.A., Barshop, B.A. and Siuzdak, G. (2007) Metabolomics Identifies Perturbations in Human Disorders of Propionate Metabolism. *Clin Chem*, clinchem.2007.089011.

Genepattern

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P. and Mesirov, J.P. (2006) GenePattern 2.0. *Nature Genetics* **38**, 500-501.

List of participants

Surname	First name	Institute / Organization	e-mail
Beckmann	Manfred	University of Wales, Aberystwyth, UK	meb@aber.ac.uk
Bernillon	Stéphane	INRA Bordeaux, France	stephane.bernillon@bordeaux.inra.fr
Bobeldijk	Ivana	TNO Quality of Life, Zeist, The Netherlands	ivana.bobeldijk@tno.nl
Brennan	Lorraine	UCD Dublin, Ireland	lorraine.brennan@ucd.ie
Casanova	Jean-Marie	Waters, Guyancourt, France	jean-marie_casanova@waters.com
Colquhoun	Ian	Institute of Food Research, Norwich, UK	ian.colquhoun@bbsrc.ac.uk
Comte	Blandine	INRA Clermont-Ferrand, France	bcomte@clermont.inra.fr
Defoort	Catherine	INSERM 476 INRA1260, Marseille, France	catherine.defoort@medecine.univ-mrs.fr
Dragsted	Lars Ove	Copenhagen University, Denmark	ldra@life.ku.dk
Draper	John	University of Wales, Aberystwyth, UK	jhd@aber.ac.uk
Dumazet	Hélène	CBiB Université Bordeaux 2, France	helene.dumazet@u-bordeaux2.fr
Ezan	Eric	CEA Sacley, France	eric.ezan@cea.fr
Fave	Gaelle	Newcastle University, UK	gaelle.fave@libertysurf.fr
Fiehn	Oliver	UC Davis, USA	ofiehn@ucdavis.edu
Gastaldi	Marguerite	INSERM Marseille, France	marguerite.gastaldi@medecine.u
George	Trevor	University of Reading, UK	t.w.george@reading.ac.uk
Gika	Eleni	AstraZeneca, Macclesfield, UK	elegik@yahoo.gr
Gröpl	Clemens	Free university Berlin	groepl@inf.fu-berlin.de
Haldar	Sumanto	Newcastle University, UK	Sumanto.Haldar@ncl.ac.uk
Hankemeier	Thomas	Leiden University, The Netherlands	hankemeier@chem.leidenuniv.nl
Hubert	Jane	INRA Clermont-Ferrand, France	jhubert@clermont.inra.fr
Jacobs	Doris	Unilever R&D, Vlaardingen, The Netherlands	Doris.Jacobs@unilever.com
Jellema	Renger	TNO Quality of Life, Zeist, The Netherlands	renger.jellema@tno.nl
Johansson	Anna	Lund University/Biomed Nutr, Sweden	anna.johansson@tbiokem.lth.se
Junot	Christophe	CEA Sacley, France	christophe.junot@cea.fr
Kristal	Bruce	Brigham and Women's Hospital, Boston, USA	bkristal@partners.org
Kristensen	Mette	National Food Institute/DTU, Søborg, Denmark	mehxy@food.dtu.dk
Llorach	Rafael	INRA Clermont-Ferrand, France	rafallorach@ub.edu
Lodge	John	University of Surrey, UK	j.lodge@surrey.ac.uk

Surname	First name	Institute / Organization	e-mail
Manach	Claudine	INRA Clermont-Ferrand, France	manach@clermont.inra.fr
Marsden-Edwards	Emma	Waters, Manchester, UK	emma_marsden-edwards@waters.co
Martin	Jean-François	INRA Clermont-Ferrand, France	martin@clermont.inra.fr
Martin	Jean-Charles	INSERM 476 INRA1260, Marseille France	jean-charles.martin@medecine.univ-mrs.fr
Marur	Vasant	Brigham and Women's Hospital, Boston, USA	vmarur@rics.bwh.harvard.edu
Mathers	John	Newcastle University, UK	john.mathers@ncl.ac.u
Newman	John W.	WHNRC / USDA, Davis, USA	john.newman@ars.usda.gov
Nugent	Anne	UCD Dublin, Ireland	anne.nugent@ucd.ie
Ommen (van)	Ben	TNO Quality of Life, Zeist, The Netherlands	ben.vanommen@tno.nl
Pedersen	Theresa	WHNRC / USDA, Davis, UK	theresa.pedersen@ars.usda.gov
Peissner	Wolfgang	LMU University of Munich, Germany	wolfgang.peissner@med.uni-muenchen.de
Pereira	Hélène	INRA Clermont-Ferrand, France	hpereira@clermont.inra.fr
Pujos	Estelle	INRA Clermont-Ferrand, France	epujos@clermont.inra.fr
Rasmussen	Lone Graasbøl	Copenhagen University, Denmark	logr@life.ku.dk
Rist	Manuela	Technical University of Munich, Germany	m.rist@wzw.tum.de
Scalbert	Augustin	INRA Clermont-Ferrand, France	scalbert@clermont.inra.fr
Seal	Christopher	Newcastle University, UK	Chris.Seal@ncl.ac.uk
Sebedio	Jean-Louis	INRA Clermont-Ferrand, France	jls@clermont.inra.fr
Shaobo	Zhou	University of Wales, Aberystwyth, UK	szz@aber.ac.uk
Theodoridis	Georgios	Aristotle University, Thessaloniki, Greece	gtheodor@chem.auth.gr
Ulmius	Matilda	Lund University/Biomed Nutr, Sweden	matilda.ulmius@tbiokem.lth.se
van Erk	Marjan	TNO Quality of Life, Zeist, The Netherlands	marjan.vanerk@tno.nl
van Iersel	Martijn	BiGCaT, Maastricht University, The Netherlands	mvaniersel@gmail.com
Vauzour	David	University of Reading, UK	d.vauzour@reading.ac.uk
Verheij	Elwin	TNO Quality of Life, Zeist, The Netherlands	elwin.verheij@tno.nl
Visser (de)	AJC (Ries)	IsoLife B.V., Wageningen, The Netherlands	ries.devisser@isolife.nl
Wishart	David	University of Alberta, Canada	david.wishart@ualberta.ca
Wopereis	Suzan	TNO Quality of Life, Zeist, The Netherlands	suzan.wopereis@tno.nl

Notes