7

Software Techniques for Enabling High-Throughput Analysis of Metabolomic Datasets

Corey D. DeHaven, Anne M. Evans, Hongping Dai and Kay A. Lawton
Metabolon, Inc.
United States of America

1. Introduction

In recent years, the study of metabolomics and the use of metabolomics data to answer a variety of biological questions have been greatly increasing (Fan, Lane et al. 2004; Griffin 2006; Khoo and Al-Rubeai 2007; Lindon, Holmes et al. 2007; Lawton, Berger et al. 2008). While various techniques are available for analyzing this type of data (Bryan, Brennan et al. 2008; Scalbert, Brennan et al. 2009; Thielen, Heinen et al. 2009; Xia, Psychogios et al. 2009), the fundamental goal of the analysis is the same – to quickly and accurately identify detected molecules so that biological mechanisms and modes of action can be understood. Metabolomics analysis was long thought of as, and in many aspects still is, an instrumentation problem; the better and more accurate the instrumentation (LC/MS, GC/MS, NMR, CE, etc.) the better the resulting data which, in turn, facilitates data interpretation and, ultimately, the understanding of the biological relevance of the results.

While the quality of instrumentation does play a very important role, the rate-limiting step is often the processing of the data. Thus, software and computational tools play an important and direct role in the ability to process, analyze, and interpret metabolomics data. This situation is much like the early days of automated DNA sequencing where it was the evolution of the software components from highly manual to fully automated processes that brought about significant advances and a new era in the technology (Hood, Hunkapiller et al. 1987; Hunkapiller, Kaiser et al. 1991; Fields 1996). Currently, software tools exist for the automated initial processing of metabolomic data, especially chromatographic separation coupled to mass spectrometry data (Wilson, Nicholson et al. 2005; Nordstrom, O'Maille et al. 2006; Want, Nordstrom et al. 2007; Patterson, Li et al. 2008). Samples can be processed automatically; peak detection, integration and alignment, and various quality control (QC) steps on the data itself can be performed with little to no user interaction. However, the problem is that the generation of data, together with peak detection and integration, is the relatively simple part; without a properly engineered system for managing this part of the process the vast number of data files generated can quickly become overwhelming.

Two major processes in metabolomic data processing are the verification of the accuracy of the peak integration and the verification of the accuracy of the automated identification of the metabolites that those peaks represent. These two processes, while vitally important to

the accuracy of the results, are very time consuming and are the most significant bottlenecks in processing metabolomic data. In fact, the peak integration verification step is often omitted due to the extremely large number of peaks whose integration must be verified.

2. Background

At the outset, running a metabolomics study is actually simple and straightforward. Samples are prepared for running on a signal detection platform, signal data is collected on samples from the instrumentation, the signals are translated into peaks, the peaks are compared to reference libraries for the identification of metabolites and those identified metabolites are then statistically analyzed with whatever metadata may exist for the samples. Alternatively, the entirety of the detected peaks resulting from the instrument signal data are statistically analyzed without metabolite identification prior to the statistical analysis.

Once statistical analysis is completed and the significant signals have been stratified and metabolites identified, biochemical pathway analysis is performed to gain insight into the original biological questions the study asked. Too often, when the metabolomic experiments do not provide meaningful biological results, the realization may come that there's so much variability in the data, it can't be used to address the original objectives of the study. Despite the methods and software provided by the various instrument vendors, it turns out that running a global, non-targeted analysis of small molecules in a complex mixture that generates high-quality data and provides answers to biological questions is challenging. Doing so in a high-throughput environment is significantly more challenging.

However, a high-throughput metabolomics platform that produces reliable, precise, reproducible, and interpretable data is possible. It simply requires the right process coupled with the right software tools. As with any high throughput process it is important to have a logical, consistent workflow that is simple, reproducible, and expandable without negatively impacting the efficiency of the process. It is important to know when human interaction is required and when it is not. Well designed and integrated software can efficiently handle the majority of the mundane workload, allowing human interaction to be focused only where required.

3. Approach

Metabolite identification is essential for chemical and biological interpretation of metabolomics experiments. Two approaches to metabolomic data analysis have been used and will be described in detail below. The main difference between the two approaches is when the metabolite is identified, either before or after statistical analysis of the data.

To date, the most commonly used method of processing metabolomic data has been to statistically analyze all of the detected ion-features ('ion-centric'). Ion features, defined here as a chromatographic peak with a given retention time and m/z value, are analyzed using a statistical package such as SAS or S-plus to determine which features vary statistically significantly and are related to a test hypothesis (Tolstikov and Fiehn 2002; Katajamaa and Oresic 2007; Werner, Heilier et al. 2008) . The significant ion feature changes are then used to prioritize metabolite identification. One issue with this type of approach is the convoluted

nature of the data being analyzed. In many cases the "statistically significant ion-features" are various forms of the same chemical and are therefore redundant information. Most biochemicals detected in a traditional LC- or GC-MS analysis produce several different ions, which contributes to the massive size and complexity of metabolomics data. In addition, there are an even larger number of measurements for each experimental sample which impacts the false discovery rate (Benjamini and Hochberg 1995; Storey and Tibshirani 2003).

In the 'chemo-centric' approach to metabolomics data analysis discussed here, metabolites are identified on the front-end through the use of a reference library comprised of spectra of authentic chemical standards(Lawton, Berger et al. 2008; Evans, Dehaven et al. 2009). Then, instead of treating all detected peaks independently, as is done in the ion-centric approach, the chemo-centric method selects a single ion ('quant-ion') to represent that metabolite in all subsequent analyses. The other ions associated with the metabolite are essentially redundant information that only add to data complexity. Furthermore, the statistical analysis may be skewed since a single metabolite may be represented by multiple ion peaks, and the false discovery rate increased due to the large number of measurements relative to the number of samples in the experiment. Accordingly, by taking a chemo-centric approach any extraneous peaks can be identified and removed from the analysis based on the authentic standard library/database. Since the number of features analyzed statistically contributes to the probability of obtaining false positives, analyzing one representative ion for each metabolite reduces the number of false positives. Further, the chemo-centric data analysis method is powerful because a significant amount of computational processing time and power can be saved simply due to data reduction.

The majority of work and complexity with the chemo-centric approach are: first, the generation of the reference library of spectra from authentic chemical standards; second, the actual identification of the detected metabolites using the reference library; and third, the ability for quality control (QC) of the automated metabolite identification, peak detection and integration. Notably, the QC of the automated processes is often overlooked. However, the QC step is critical to ensure that false identifications and poor or inconsistent peak integrations do not make their way into the statistical analysis of the experimental results. The generation of a reference library entry made up of the spectral signature and chromatographic elution time of an authentic chemical standard is relatively straightforward, as is the generation of spectral-matching algorithms that use the reference library to identify the experimentally detected metabolites. In contrast, performing the QC step on the automated processes, including peak detection, integration and metabolite identification, is time and human resource intensive.

Not to be overlooked, an issue with using a reference library comprised of authentic standards is dealing with metabolites in the samples that are not contained within the reference library. The power of the technology would be significantly reduced if it was limited to identifying only compounds contained in the reference library. Through intelligent software algorithms, it is possible to analyze data of similar characteristics across multiple samples in a study to find those metabolites that are unknown by virtue of not matching a reference standard in the library, and, in the process, group all the ion-features related to that unknown together by examining ion correlations across the sample set (Dehaven, Evans et al. 2010). One such method capitalizes on the natural biological variability inherent in the experimental samples, using this variation the metabolites and

their respective ion-features can reveal themselves and be entered into the chemical reference library as a novel chemical entity (Dehaven, Evans et al. 2010). The unknown chemical can then be tracked in future metabolomics studies, and, if important, can be identified using standard analytical chemistry techniques.

Without going into detail, it is important to note that the sample preparation process is critical. High quality samples that have been properly and consistently prepared for analysis on sensitive scientific instrumentation are of extreme importance. Ensuring this high quality starts with the collection and preparation of the samples. No software system is going to be able to produce high-quality data unless ample effort is focused on consistently following standardized protocols for preparing high quality samples for analysis.

The following discussion, examples and workflow solutions make use of GCMS or LCMS (or both) platforms for metabolomic analysis of samples, although the concepts in general could apply to a variety of data collection techniques. Software tools are also presented to demonstrate the application of the concepts that are discussed but the tools themselves will not be discussed in great detail. It is also important to note that achieving the greatest operation efficiency of the process relies on treating all of the experimental samples in a study as a set and not as individual files. By using tools to analyze and perform quality control on the samples as a single group or set it becomes much easier to spot patterns that can be useful to determine what is going on in the overall process.

4. Processing data files, peak detection, alignment, and metabolite identification

4.1 File processing can become a major hurdle

There is no shortage of software available on the market to read spectral data and detect the start and the stop of peaks, and the baseline, and then calculate the area inside of those peaks. Each instrument vendor provides some flavor of detection and analysis software with their instrument and several open-source and commercial efforts to read spectral data and produce integrated peak data regardless of vendor format are available (Tolstikov and Fiehn 2002; Katajamaa and Oresic 2005; Katajamaa and Oresic 2007). In almost all cases, these packages do a complete job of finding and integrating peaks and do so in a reasonable amount of time. Thus, the peak detection and integration process is not the rate-limiting step when it comes to data quality and automated processing.

As it turns out, the file processing problem is primarily a file management problem that is the result of two issues – human and machine. The first problem stems from human interaction, in that a human being can introduce more error and inconsistency than is acceptable. Optimally, a human should play no role in the naming or processing of instruments data files. Naming of instrument data files should take place within the system used to track sample information, a LIMS for instance. The LIMS or other sample tracking system should generate a sample list and run order for the samples to be run on the instrument using a consistent naming convention that can be easily associated with the sample in question. The second problem stems from both machine and human. The software performing the peak detection and integration must have the capability of automatically processing a data file when presented with it, then archiving the file when completed. And,

in high-throughput mode, it is best not to have humans manage data files, either in storage locations, or, as noted above, in naming. For consistency, it is imperative that the machines control this step; running one experiment on one machine may be manageable manually but running experiments in tandem or on more than one instrument can easily result in misnaming, file version problems, location mishaps, etc. if file management is not automated.

4.2 Manual integration of peak data is inadequate for high-throughput processing

Processing metabolomics data in a high-throughput setting requires automated processing of data files. While an SDK (software development kit) is provided by many instrument vendors, and there are commercial and open source packages for creating this functionality available (Smith, Want et al. 2006), not all vendor software permits this functionality. One of the main reasons automated peak integration works well is because it allows data to be rapidly uploaded and processed. Manual integration, while perhaps more accurate, dramatically slows the peak analysis process. Further QC and refinement of the automated peak integration can be performed more optimally later in the process, where, in practice, the bar for peak detection can be slightly lower. The reasons that the bar for peak detection can be reduced will be discussed below.

4.3 Alignment based on peak similarity inadequate, retention index should be used

Many of the software packages provide capabilities to align the chromatograms to account for time drift in an instrument. In many instances internal standards and/or endogenous metabolites are used across the analyzed samples to align chromatography based on their retention times, such that there is confidence that the same peak at the same mass is consistent among the data files. This approach should be avoided because while it works fine for peak analysis and chromatographic alignment on a single, small study it will only be applicable within that one study where retention times are quite consistent. This type of alignment approach makes it much harder to do a comparison to a reference standard library where a retention profile is used as matching criteria. The better choice is to opt for retention index (RI) calculation, which can correctly align chromatograms even over long periods of time where conditions can be vastly different dependent on the condition in these systems. Using a retention index method, each RT marker is given a fixed RI value (Evans, Dehaven et al. 2009). The retention times for the retention markers can be set in the integrator method and the time at which those internal standards elute are used to calculate an adjustment RI ladder. All other detected peaks can then use their actual retention time and adjustment index to calculate a retention index. In this way, all detected peaks are aligned based on their elution relative to their flanking RT markers. An RI removes any systematic changes in retention time by assuming that the compound will always elute in the same relative position to those flanking markers. Because of this, a unique time location and window for a spectral library entry can be set in terms of RI, thereby ensuring that metabolites don't fall outside the allowed window over a much longer period of time. Retention indices have predominately been used for GC/MS methods however this approach can also have great success for LC/MS data alignment as well. LC/MS is certainly more complex as certain metabolites and classes of metabolites show more chromatographic shift in their RI

markers than others, in these cases increasing the expected RI window of the library entry in conjunction with mass and fragmentation spectrum data is sufficient for accurate identification. The advantages over many of the widely available chromatographic alignment tools, eg. XCMS (Smith, Want et al. 2006), as it can be used to match against a RI locked library over long periods of time and can align data from different biological matrices without potential distortion from structural isomers.

4.4 Identifying metabolites

Metabolite identification is essential to the biochemical and biological interpretation of the results of metabolomic studies. Lists of integrated peak data are of little use unless a library of spectra is available to compare peak data with to identify the metabolites represented by those peaks. Publicly created and maintained databases do exist (Wishart, Tzur et al. 2007; Wishart 2011). However, the utility of these databases to identify metabolites of interest from metabolomics studies is currently limited for a number of reasons. First, due to the significant number of different instrument types, methods, and runtimes it is a nearly impossible task to account for every possible representative of the spectra and retention time for a given metabolite under all of these diverse conditions. Second, metabolomics experiments utilize a global non-targeted approach where the method is optimized to measure as many metabolites as possible in a wide range of biological sample types (i.e., matrices). Certain metabolites behave differently in one matrix than in another, or differently in the same matrix under different conditions, for example in response to an experimental treatment versus when non-treated. Third, there may be areas of the chromatogram with a high-degree of co-eluting metabolites. Public databases of metabolite spectra can provide useful information in many cases, especially when no existing library exists. However, the public information is limited and certainly not as informative or reproducible as generating an in-house chemical reference library using the same equipment and protocols as used to analyze the experimental samples.

While requiring a significant resource commitment, the generation of an internal library of authentic chemical standards is a worthwhile task with significant advantages for high-throughput metabolomics. An in-house library of authentic standards provides a clear representation of the spectra resulting from a metabolite on the same instrument and method used to analyze the experimental sample. A retention index for the internal library can be calculated and set, resulting in library entries that are fixed in time. Consequently consistent, reliable, standard spectra that do not change over time are ensured which, in turn, facilitates automated, high confidence metabolite identification.

Software for performing spectral library matching, much like peak integration software discussed above, is readily accessible (Scheltema, Decuypere et al. 2009). From open-source applications to commercial packages there are numerous choices. Many software packages use some type of forward or reverse (or both) fitting algorithm that use mass and time components to match peaks to metabolites of similar mass and peak shape within a time window. Due to their global, non-targeted nature, metabolomic studies are not optimized for any metabolite in particular, so a positive metabolite identification in a metabolomics analysis is almost never a binary decision. It is highly unlikely to simply have a positive yes or no for a metabolite identification, instead it is more likely to have a probability score associated with the identification. Quality control of the scoring is essential and one of the

most important aspects of metabolomics analysis, especially for running studies in high throughput.

4.5 Unnamed metabolites

A chemo-centric approach, based on a reference library, to high-throughput metabolomics is a powerful method to identify metabolites within biological samples. If there is any weakness to using in-house generated reference libraries it would be in the realm of identifying the redundant ion peaks that originate from metabolites that do not exist in the library. Methods available to identify and group these redundant ion peaks are limited (Bowen and Northen; Dunn, Bailey et al. 2005; Wishart 2009).

The most common approach is to rely on the chromatographic elution similarity between these redundant ions as well as looking for user defined mass relationships between the ions that are consistent with known chemical modifications. The effectiveness of this approach is limited in highly complex samples where metabolite co-elution is common. In such situations, there can be multiple metabolites eluting simultaneously which confounds identifying their respective ions based on elution. Another shortcoming of this method is the inability to identify unique modifications or fragments that are not known to occur.

A method that has yielded very good results for analyzing spectrometry data and fits well within the framework of high-throughput metabolomics is the QUICS method (Dehaven, Evans et al. 2010) This method to identify and quantify individual components in a sample, (QUICS), enables the generation of chemical library entries from known chemical standards and, importantly, from unknown metabolites present in experimental samples but without a corresponding library entry. The fundamental concept of this method is that by looking at detected ion features across an entire set of related samples, it is possible to detect subtle spectral trends that are indicative of the presence of one or more obscure metabolites. In other words, because of the natural biological variability of the metabolite in the study samples, by performing an ion-correlation analysis across all samples within a given dataset it is possible to detect ion features that are both reproducible and related to one another. Using the cross sample correlation analysis it is then possible to add the spectral features for that metabolite to the reference library. Then the metabolite can be detected in the future using that library entry, even though the metabolite is unknown, i.e., without an exact chemical identification. Importantly, this method captures any unknown metabolite because it does not require chemical adducts and/or fragment products to be previously known or expected. Another advantage is that statistical analysis can be used to determine whether or not the metabolite is significant or of interest. In this way the important unnamed metabolites can be focused on for the work of performing an actual identification which enhances efficiency and reduces the work to identifying the most important metabolites.

5. Quality control

The ability to perform thorough quality control on identified metabolites in metabolomics studies is extremely important. The higher the quality of data entering statistical analysis, the higher the probability that the study will provide answers to the questions being asked. This section will focus on three aspects of quality control – quality control samples (i.e.,

blanks, technical replicates), software for assessing the quality of metabolite identification, and software for assessing the original peak detection and integration. This last point may seem out of order but for reasons to be described results in an invaluable check of the peak quality.

5.1 Blanks – Identify the artifacts of the process

A commonly overlooked issue in biological data collection is the presence of process artifacts. A process artifact is defined as any chemical whose presence can be attributed to sample handling and processing and not originating from the biological sample. In all analytical methods chemicals are inadvertently added to samples. Artifacts can include releasing agents and softeners present in plastic sample vials and tubing, solvent contaminants, etc. One of the easiest and most efficient means of identifying artifacts is to run a "water blank" sample interspersed throughout the entire process alongside the true experimental samples. In this way, the water blank will acquire all the same process-related chemicals as the experimental samples. Consequently, identification and *in silico* removal of artifacts can be accomplished by identifying those chemicals detected at significant levels in the water blank when compared to the signal intensity in the experimental samples. If not identified and removed, process artifacts can inadvertently arise as false discoveries.

5.2 Technical replicates – Find the total process variation

The intrinsic reproducibility of a method is critical since it has considerable impact on the significance and interpretation of the results. For example, if a 20% change was detected between treatment and control samples but the analytical method had a 20% coefficient of variation (CV) for that measurement, concerns regarding the accuracy of the measurement would call into question the biological relevance of that change in measurement. On the other hand, if the analytical method had a 2% CV for that same measurement it is much more likely that the same 20% change is of "real" biological significance. Clearly, smaller analytical variability of the method enables small, yet meaningful, biological changes to be detected accurately and consistently. It is therefore critical to determine the analytical reproducibility/variability of a method for every compound/measurement.

By far the most common way to assess system stability and reproducibility is by use of internal standards. Internal standards can be measured throughout a study to monitor system reproducibility and stability. The drawbacks to this approach are that the number of standards is typically small and do not represent the myriad of chemical classes typically observed in a metabolomics analysis.

Another common approach to address method variability is by the use of technical replicates. With this approach the same biological sample is run multiple times, e.g., in triplicate, to determine method reproducibility. The advantage of this method over internal standards is the ability to determine the CV of the method for each compound detected within the matrix of the samples being analyzed. However, the disadvantage is that, while the replicate approach is extremely effective, it is also very time-consuming and of limited practicality in a high-throughput setting.

An extremely practical and efficient approach is to run a technical replicate of a sample composed of a small aliquot from all the samples in a study interspersed among individual experimental samples. An aliquot of each experimental sample is pooled, then an aliquot of the pooled sample mixture is run at regular intervals—every *n* number of experimental sample injections (*n* to be set by operator). An advantage of this pooled sample is that it provides CV information for all compounds detected in the study, in the matrix under study. Another advantage is that far less instrument analysis time is required which makes it far more practical in a high-throughput laboratory.

5.3 Quality control of automated metabolite identifications

Performing quality control (QC) for a given metabolite identification can be an exhaustive and time-consuming task. The work to perform QC on every metabolite identification in every sample within a metabolomics study can seem to be a nearly-impossible task. Considering a relatively small metabolomics study of 50 samples, with an average of 800 identified metabolites per sample, there would be 40,000 spectra to review for just that one study. Yet, as time-consuming as this process is, quality control of automated library calls is vital for ensuring accuracy and high confidence in the data which, in turn, enables meaningful biological interpretation of the results. A software package that can permit this process to proceed quickly and efficiently is critical in a high-throughput setting.

Visual inspection of all the samples in a study simultaneously enables rapid metabolite identification QC. By representing the sample data within a study as a single set in a visual manner and creating tools that quickly allow an analyst to investigate and manually accept or reject an automated metabolite identification, the task of performing quality control on even extremely large datasets can be accomplished rapidly and easily. An example of a visual data display is shown in Figure 1. In this example the panel across the top (Figure 1A) contains a list of all of the metabolites identified by the software in the experimental samples being analyzed. By highlighting one chemical, the structure for that compound is displayed in an adjacent window (Figure 1B). The default visualization for viewing a highlighted metabolite is broken down into a distinct method chart for each analytical platform method that was used to identify that metabolite. The display also shows the multiple analytical platforms where the metabolite was identified. In this example, the same metabolite identified on a GC/MS platform (Figure 1C), and LC/MS negative ion platform (Figure 1D) is shown. Within each chart, the individual sample injections, each with a unique identifier, make up the y-axis (Figure 1E). The x-axis represents the retention index (RI) time scale. Navigation of the interface involves scrolling down through the data table window (Figure 1A). From the interface it is also possible to review annotation regarding the highlighted metabolite (Figure 1F), view the analytical characteristics (e.g., Mass, RI) of the metabolite as well as toggle through RI windows containing ions characteristic of that metabolite (Figure 1G).

An example plot of data from the LC/MS negative platform is illustrated in Figure 2. In this example the samples are initially sorted by the sample type, namely process blank, technical replicate, or experimental sample. The dots within each method chart represent the detected ion peaks, and each point has associated peak area, mass to charge (m/z), chromatographic start and stop data which can be accessed by clicking on the individual dots, as shown in Figure 3.

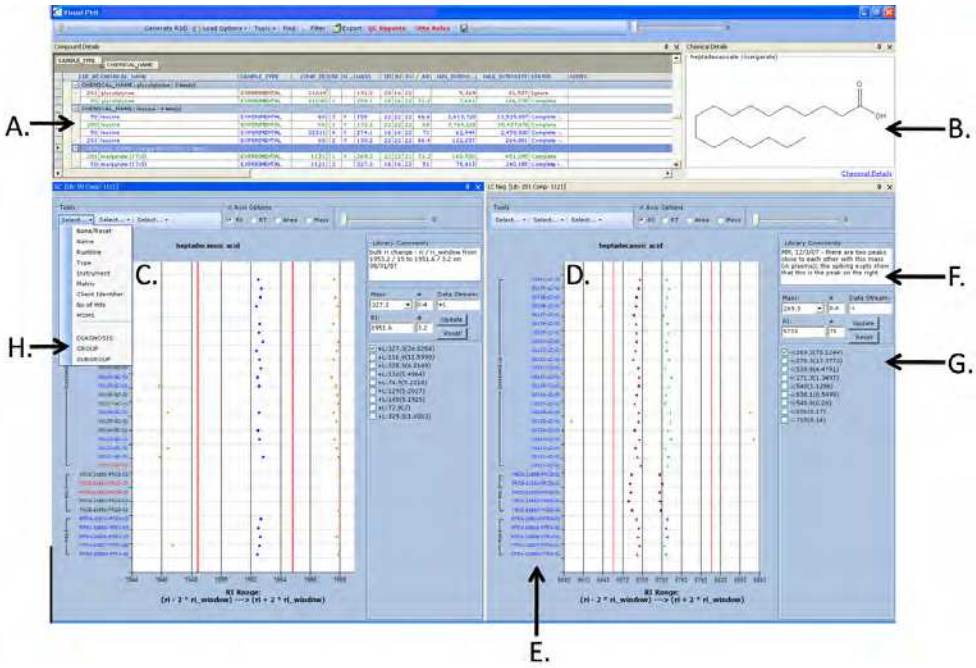


Fig. 1. Graphical user interface showing the view for the proposed identification of heptadecanoic acid. (A) Distinct list of identified metabolites for the loaded sample set. This list includes any metabolite identified at least once in any sample with the set. It also includes summary statistics such as averages for spectral scoring and chromatographic peak intensities, number of times detected, and status.(B) Chemical structure for displayed metabolite. (C) Data for the posed library identification heptadecanoic acid from the GC/MS method. (D) Data for the posed library identification heptadecanoic acid from the LC/MS negative ion method. (E) List of unique sample identifiers comprising the study. (F) Comment field for storing and displaying annotations that are relevant to the currently displayed metabolite. (G) List of other ion peaks that exist as part of the spectral library entry. (H) List of sample sorting options including associated sample metadata; diagnosis, group and subgroup.

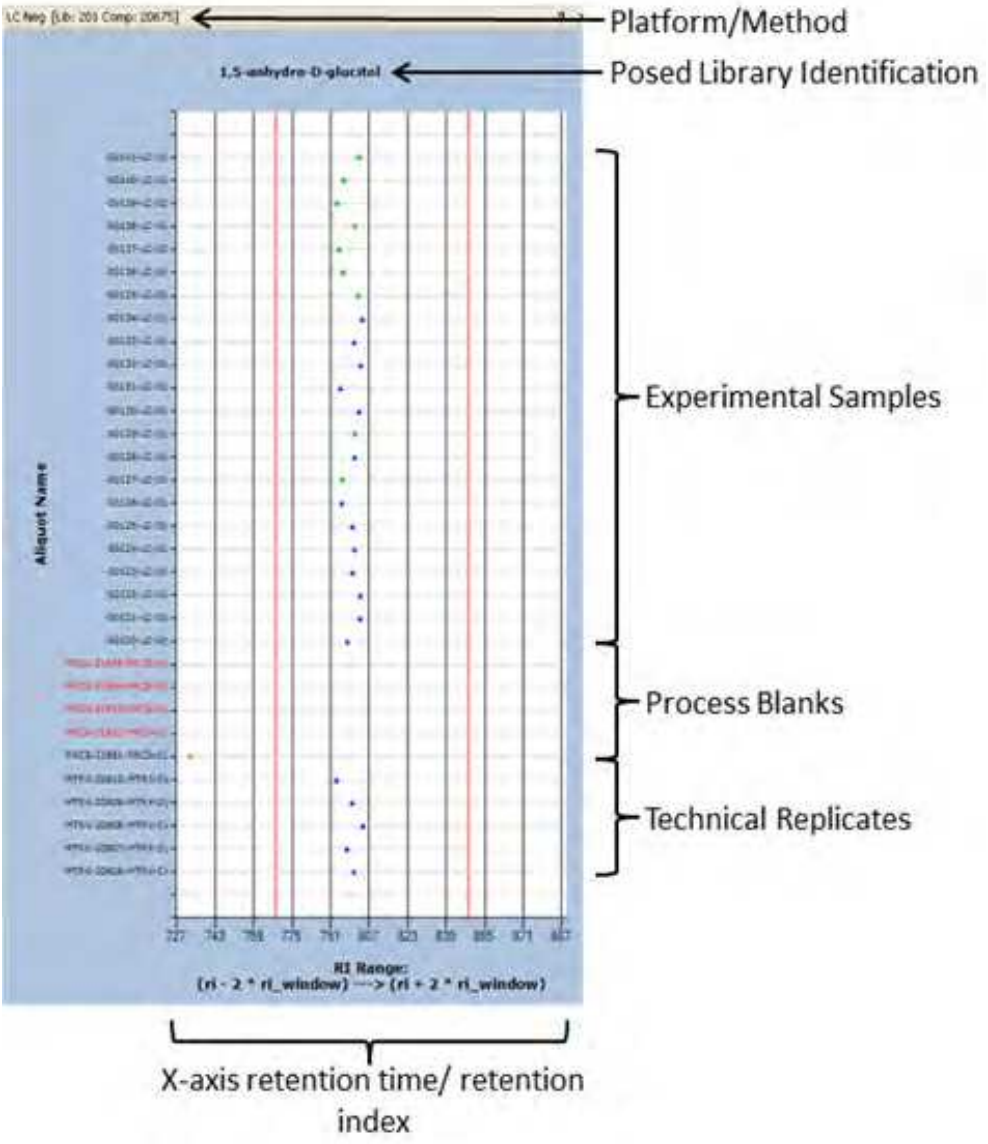


Fig. 2. Plot for LC/MS negative method. Individual samples in the sample set are displayed and sorted on the y-axis. Chromatographic retention time is presented on the x-axis.

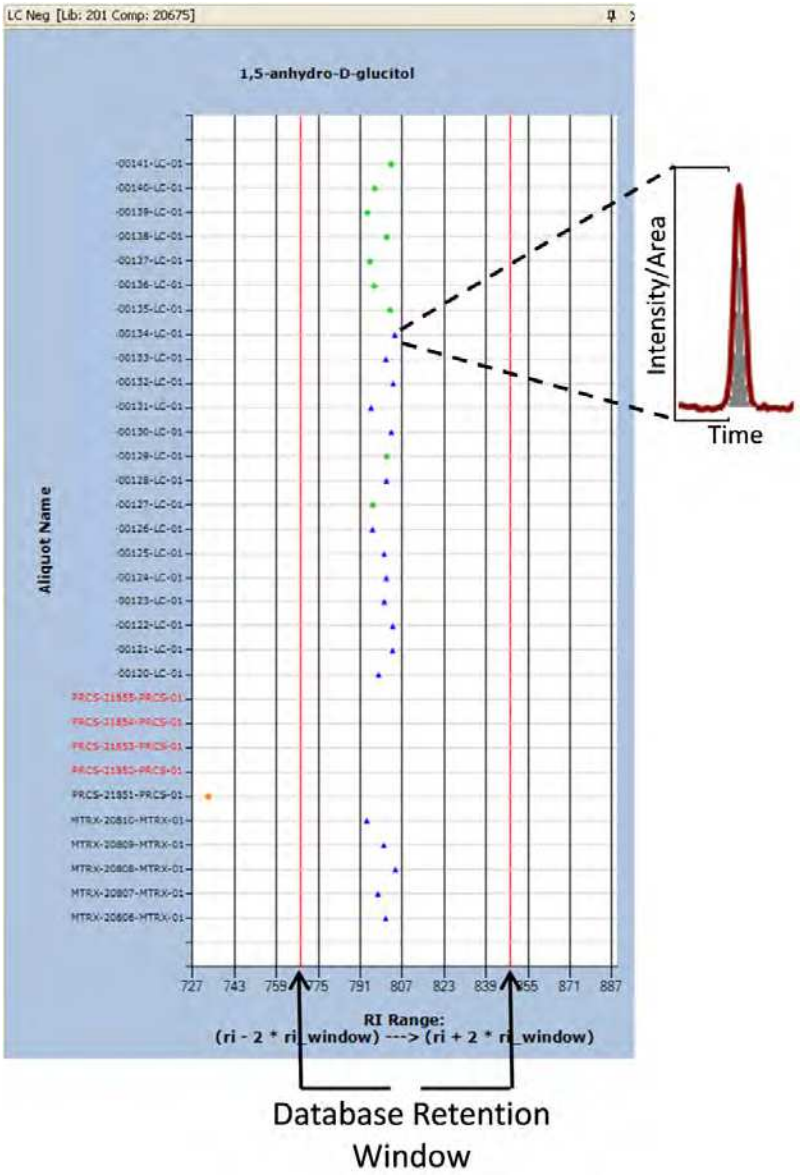


Fig. 3. Raw MS data can be accessed for each sample via graphical user interface links. Each "dot" represents the detected and integrated ion peak in the individual sample listed on the y-axis. Thus, each "dot" has an associated area, height intensity, chromatographic start, stop and apex retention time/retention index. Color and shape of "dot" are indicative of the quality of the match to the posed library identification (see chart 1) and can be used to launch underlying data such as raw MS data (insert). Colors of the samples listed on the y-axis also hold meaning (see chart 1).

This type of visualization permits the analyst to quickly verify the quality (i.e., QC) of the automated peak detection and integration software. Each dot that is representative of an ion peak can be individually removed/rejected from the proposed library identification. In this way, extraneously detected ion peaks in the window can be visualized and individually removed, as is the case shown in Figure 4. In this particular example there are two closely eluting ion peaks with the same mass. One of the peaks is 2-stearoylglycerophosphocholine (Figure 4, panel A) and the other is 1-stearoylglycerophosphocholine (Figure 4, panel B). In both panels the correct peak must be manually approved and the incorrect peak rejected (indicated by red dots). Stray detected ions can also be individually rejected from the identifications. In addition, the interface permits the interrogation of the integration quality of individual ion peaks since each dot is linked to the raw ion data as illustrated in Figure 4, panel C and Figure 5, panel B. In this fashion any potential inaccuracy in the automated detection and integration of individual peaks can be readily determined.

In addition to being able to curate each sample individually, the automated library identification for an entire sample set can be rejected. An example of this is shown in Figure 5. The presence of multiple dots for each sample in the RI window (Figure 5 A) coupled with the ability to view the underlying ion data (Figure 5B) makes it apparent that the automated metabolite identification was based on erroneous ion peaks that resulted from the integration of noise. As a result the automated call for the entire sample set was manually rejected by the analyst. Accordingly, with this visualization tool the analyst can rapidly determine the quality of the automated detection and integration and remove from the dataset any peaks which are of questionable quality.

In addition to being able to QC the automated peak detection and integration software, an interface such as this allows an analyst to visually inspect the quality of the library identification in each individual sample. In the graphical plot the "dot" representing the detected ion peak for the proposed metabolite identification is displayed in various color and shape combinations. Each combination of color and shape within each plot is an indicator of the quality of the automated metabolite identification, which greatly aids the analyst in making the quality assessment rapidly. Listed in Table 1 are possible color and shape combinations with the meaning for each. The quality assessment is based on spectral library matching logic (Evans, Dehaven et al. 2009). This graphical display allows the analyst to look at a proposed identification for a given metabolite made by the software and immediately determine its quality and confidence based on spectral match scores. In this way, the automated metabolite identifications for large datasets can be quickly evaluated by the analyst. An example of a proposed call for a group of ions in a sample set where the MS/MS spectral match was poor is shown in Figure 6. The low data quality is readily apparent by the preponderance of the red colored dots in the plot.

5.4 Quality control of automated integration

GC/MS and LC/MS/MS measurements of a bio-sample usually produce millions of ions, which are fragments and/or adducts/aggregates of the metabolites, artifacts from the system, and potentially false ions from background noise. Ideally, the false ions from the background noise are removed, using, for example, a Gaussian smoothing algorithm to filter them out of the dataset. The remaining ions are then integrated across time within a mass window to identify ion chromatographic peaks. Thousands of such ion chromatographic

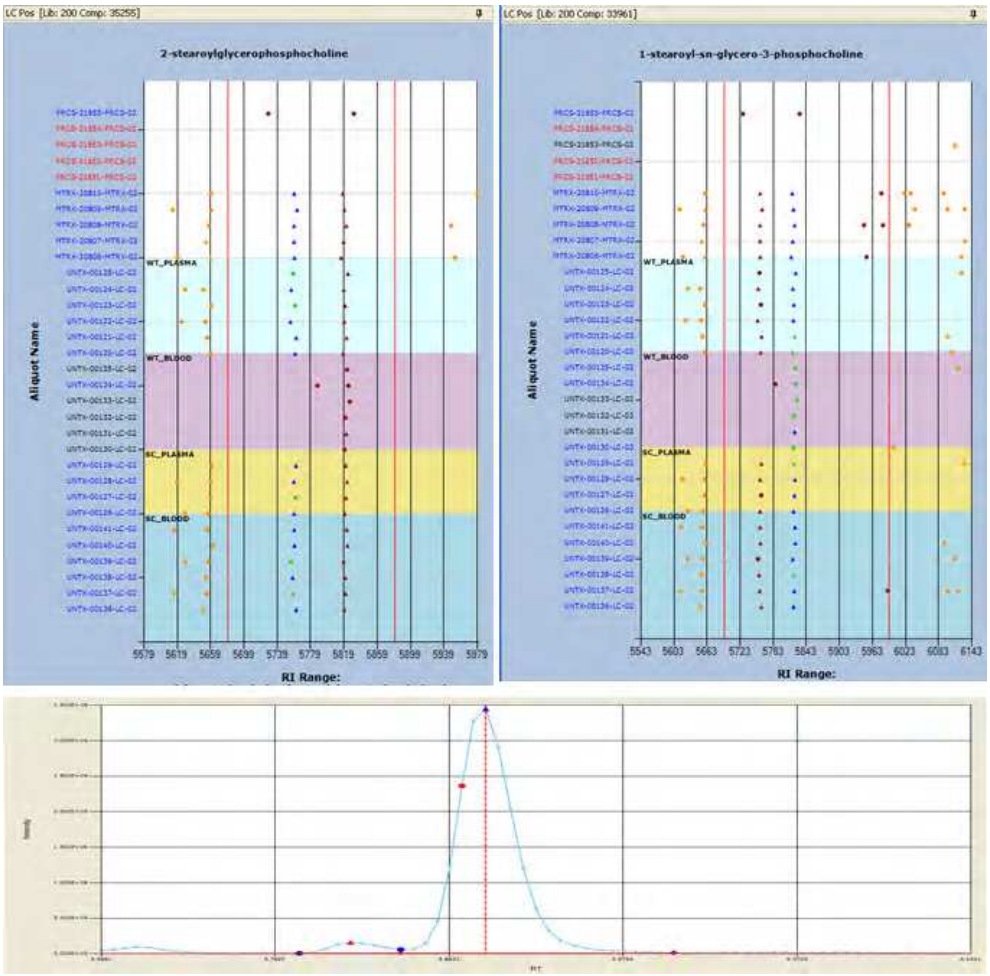


Fig. 4. Example of the visualization of closely related ion peaks. (A) Two possible ion peaks were detected in same retention window for 2-stearoylglycerophosphocholine. In this case the ion peak to the left is the correct peak (green and blue dots with arrow) and the peak on the right was rejected (red). (B) The peak on the right is actually 1-stearoylglycerophosphocholine. Therefore, the peak on the right is correct (green and blue with arrow) while the peak on the left is rejected (red). In addition ion peaks in dashed boxes are stray detected ion peaks not associated with the peak for 1-stearoylglycerophosphocholine ion peaks that were rejected (red). (C) The extracted ion chromatogram for one of the samples in the sample set for this ion shows the two peaks are well separated and accurately integrated.

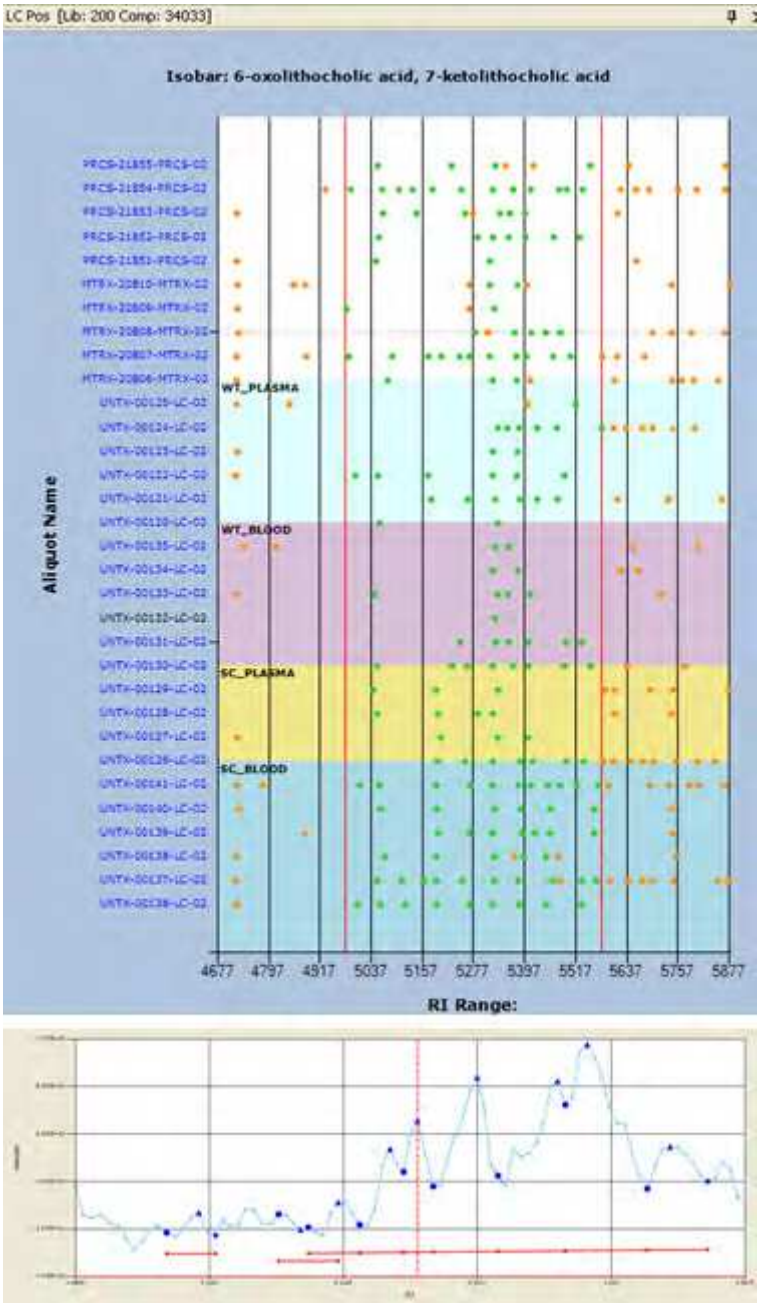


Fig. 5. Example of rejected identification for entire sample set. (A) Detected ion-peaks that result from (B) a noisy baseline as seen in one injection; the entire library identification for this LC/MS positive method is rejected.

Feature Color	Meaning
Blue	Identified metabolite with high confidence in the quality of the identification. These points represent the highest level of confidence for a metabolite identification.
Green	Identified metabolite with some confidence in the quality of the identification. These points require further research depending on the context.
Red	Identified metabolite with very low confidence in the quality of the identification.
Orange	Orange points represent a peak that was within and near the time window but was not called as the metabolite. There may be cases such as for very low-level peaks that the scoring was not adequate to make a call but taken in context with the rest of the dataset, the peak is indeed the metabolite in question.
Feature Shape	**Meaning**
Circle	Regular peak.
Triangle	For LC/MS a triangle represent that there exists underlying MS/MS data to confirm a peak's identity.
Other shape	Shapes other than a circle or triangle represent that a user has opted to also observe ion features that exist as part of the library entry other than the ion feature used for quantification.
Sample Label Color (y-axis label)	**Meaning**
Black	Black labels mean that there is one and only one ion feature representing the metabolite identification in the given time window.
Blue	Blue labels indicate that there is more than one ion-feature within the time window that may represent the metabolite identification.
Red	Red labels are shown when there are no possible hits for the given metabolite.

Table 1. Color/Shape combinations to demonstrate peak quality

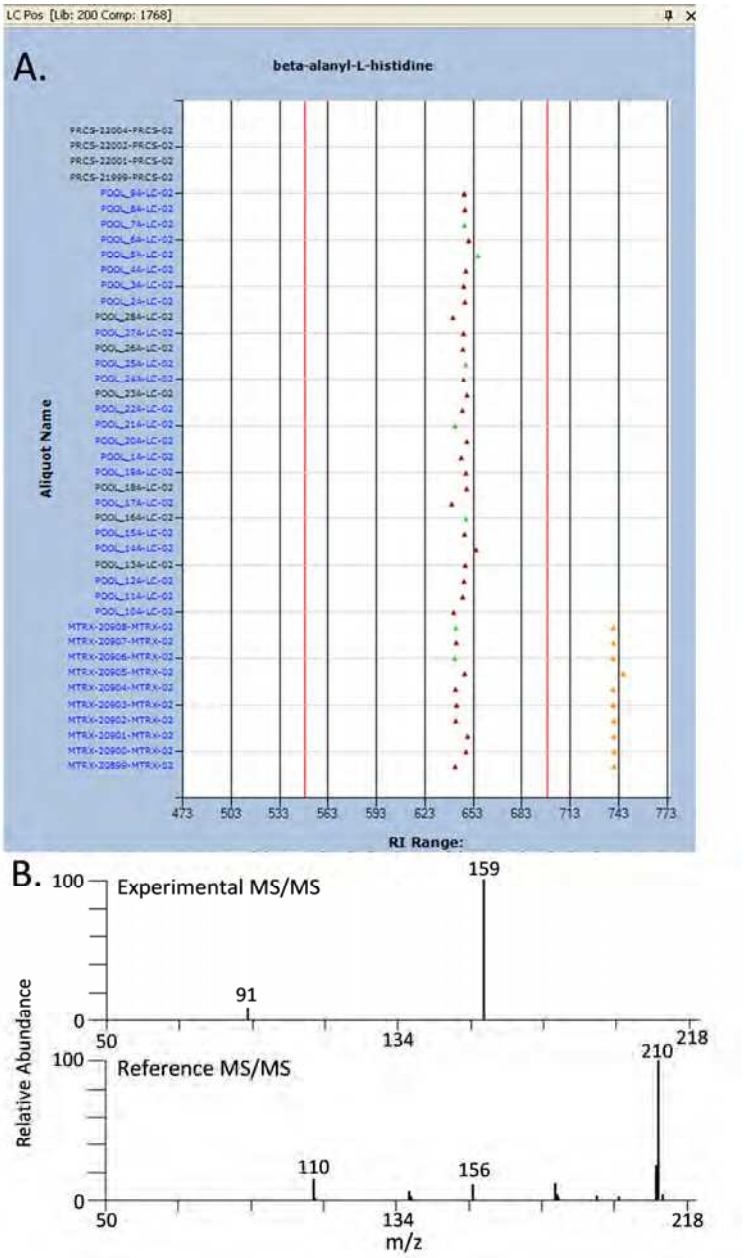


Fig. 6. An example of an entire metabolite call that was rejected because the MS/MS spectral match was poor. (A) Red color of dots indicates that the MS/MS spectral match was of low quality. (B) Experimental MS/MS spectrum from one injection compared to the reference library spectrum for beta-alanyl-L-histidine (carnosine).

peaks per sample are typically detected and integrated. Those peaks are organized into groups within a time window, adjusted/aligned by retention index from known internal standards to account for time drift and matched to library compounds (metabolites). For LC/MS/MS, secondary fragment ions of the primary quantification ion can also be used to match library compounds.

Obviously, all the profiling and analysis of metabolites in biological samples are dependent on the accuracy and consistency of ion chromatographic peak detection and peak integration. However, GC/MS and LC/MS/MS measurements are complicated by a number of factors; for example, the co-elution of metabolites because of incomplete separation, the existence of artifacts from the system, the background noise, and the potential wide concentration ranges of metabolites in the sample. Such complexities can affect the detection and determination of the peak start, the peak end and the peak baseline. Incomplete separation can lead to shoulders on peaks on either the leading edge or the trailing edge of the main peak from metabolites present at higher concentrations. When compared to the baseline, the peak start and the peak end would be characterized by a baseline peak or a drop peak. Because of the complexity and variance inherent in biological samples, the same metabolite in different samples may have been automatically detected differently in regard to peak start, peak end and peak background. For example, in some cases, the ion chromatographic peaks for metabolites present in only trace amounts may not be well shaped, especially when a noisy background is present, so integration of such peaks might be quite variable from sample to sample. In other cases, the major ions for a metabolite may appear as a small shoulder on a larger ion peak and, as a result, may not even be detected during the automatic peak detection/integration process from sample to sample. In still other instances, a metabolite present in high concentrations may overload the column and distort the chromatographic profile, leading to peak splitting. For such high concentration metabolites the automatic library match may pick only one of the two peaks for quantification which will give an erroneously low value for the amount of the metabolite in the sample. Clearly, each of the above examples will lead to peak detection and integration inconsistency and inaccuracy across the samples in a sample set, which will potentially lead to wrong conclusions and wrong decisions in later analysis.

Global metabolomics has other challenges when it comes to peak detection. Unlike targeted metabolomics, global metabolomic profiling cannot be optimized for each metabolite that is present within a biological sample. Chromatography methods must be broad enough to detect as many of the metabolites in the sample as possible, regardless of chemical characteristics. Consequently, chiral compounds cannot be resolved and structural isomers are usually not well resolved in global metabolomics profiling. In downstream analysis for identified metabolites, structural isomers might better be combined to represent the metabolites, or, if one isomer is more crucial in elucidating the metabolism or biochemical pathway, consistently picking that one form across samples would ensure analytical consistency.

A software solution to detect and correct such inconsistencies in ion peak integration and library matching across samples in a sample set could be developed. After the quality control phase of the identification of the detected metabolites is completed, a deeper examination of the consistency of peak detection and integration could be performed to ensure consistency and accuracy. The quality control phase of automatically detected metabolites involves

providing a high-quality, filtered list of identified metabolites devoid of noise and artifactual metabolites to the end user (Figure 7). Sample set sizes range from a few samples to hundreds and even thousands of samples. Because hundreds of metabolites can be detected and measured in each sample, this type of quality control operates on the 'quant' ion peaks –those peaks detected in the samples that are used for quantification of those metabolites.

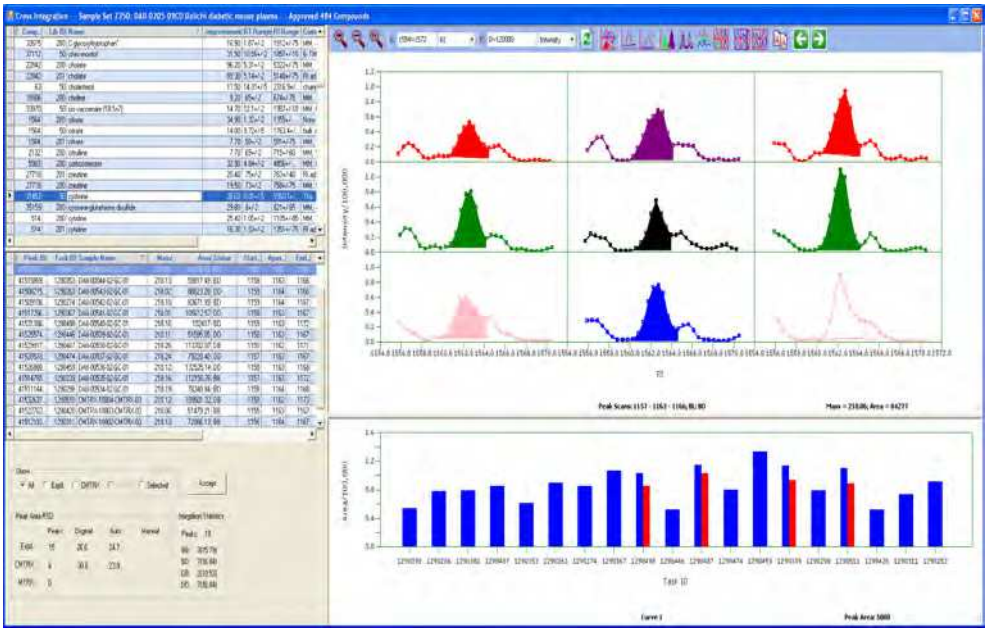


Fig. 7. Graphical View of Peak Integration for an Identified Metabolite: Identified metabolites (200~600) in the specified sample set (Upper Left); Quantitation peaks for selected metabolite in the samples in the sample set (Middle Left); Type of samples and Information about the sample peaks (Lower Left); Peak chromatograms (Upper Right); Sample peak area (blue for original integration and red for re-integrated (Lower Right).

The chromatograms of the ion peaks representing the quantitative mass from all of the samples in a set must be evaluated to determine if:

- the majority of the sample peaks are on the trailing edge of another peak,
- the majority of the sample peaks are on the leading edge of another peak,
- the majority are peaks that encompass two peaks in other samples, as a result of peak splitting.

Peak integration ranges are evaluated with alignment by retention index and the statistics of peak limits across the sample set. Accordingly, in addition to user specified manual correction, corrections in consistency and re-integration would be suggested and presented to the analyst for review and approval. Functionally, this type of software would give the end user a variety of methods to both investigate the automated integration and peak calls and to correct them as necessary. The software features must include:

- Automatic merging of approved peaks from the sample that match to the same library compound.
- Detection of shoulder peaks based on RI-aligned peak start or peak end distribution across the samples.
- Manual integration
- Manual peak splitting
- Show peak chromatograms in overlay mode or tabular mode for easy review/manual re-integration.
- Update peak integrations, peak recovery and library rematch

When an identified metabolite in a biological sample is at a sufficiently high concentration, it can overload the column and distort the chromatographic peak. Even though it may be out of the linear range, a consistent integration of the peak is still needed to characterize the group of samples. Distorted peaks tend to drive the integration software to identify a less than optimal peak to be used for quantification. In Figure 7, the peak for glucose was incorrectly split in a handful of samples by the automated peak integrator. By examining the consistency of the peak integration across the set of samples it is possible to easily identify and correct this situation. As shown in the example in Figure 8, this correction would improve the relative standard deviation from 20.1 to 7.4

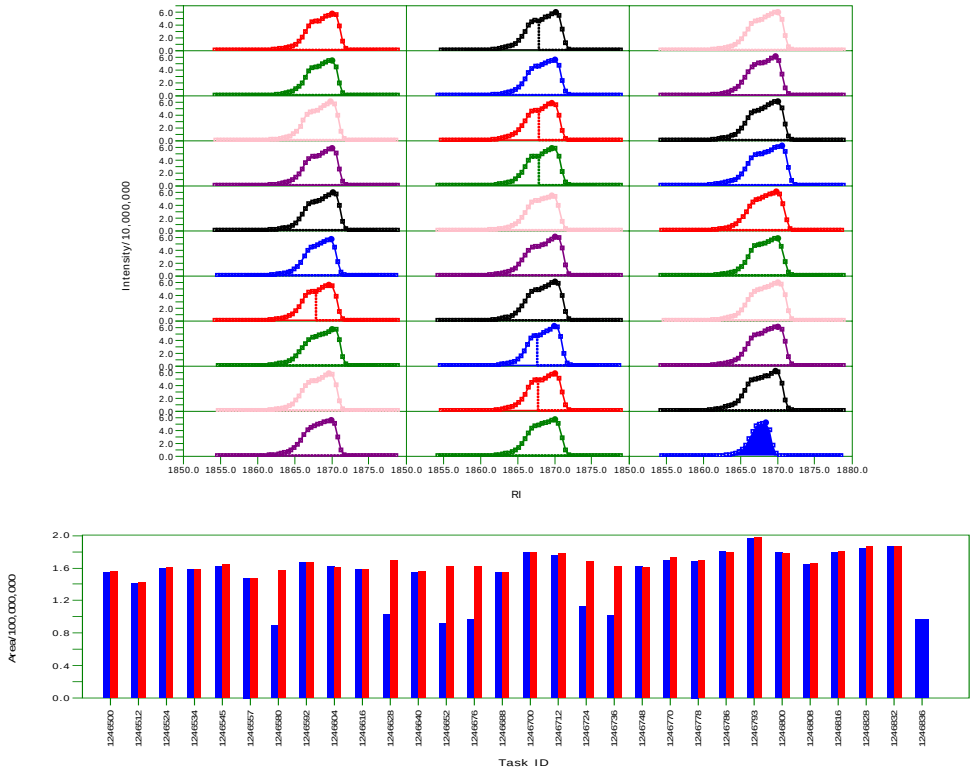


Fig. 8. Combining Peaks

As illustrated in Figures 9 and 10, small peaks on the leading or trailing side of a larger peak are often integrated inconsistently:

- Small shoulder peaks are detected
- Small shoulder peaks are not detected
- Small shoulder peaks are combined into the main peak

In Figure 8, the major peak on the left is identified as cysteine, whereas the shoulder on the right side is from threonate. In one sample, the small peak from threonate was inaccurately combined into the main peak for cysteine when it was automatically integrated, thus inadvertently increasing the response for cysteine in that sample. After re-integration the erroneous integration was corrected thereby restoring the correct integration for cysteine and permitting the independent detection of threonate in the sample as well.

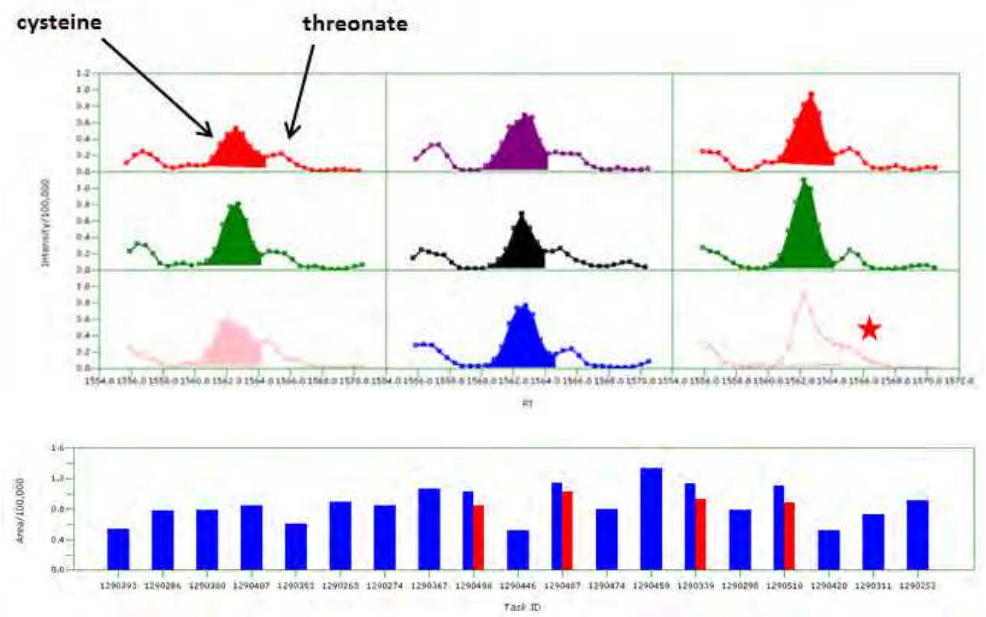
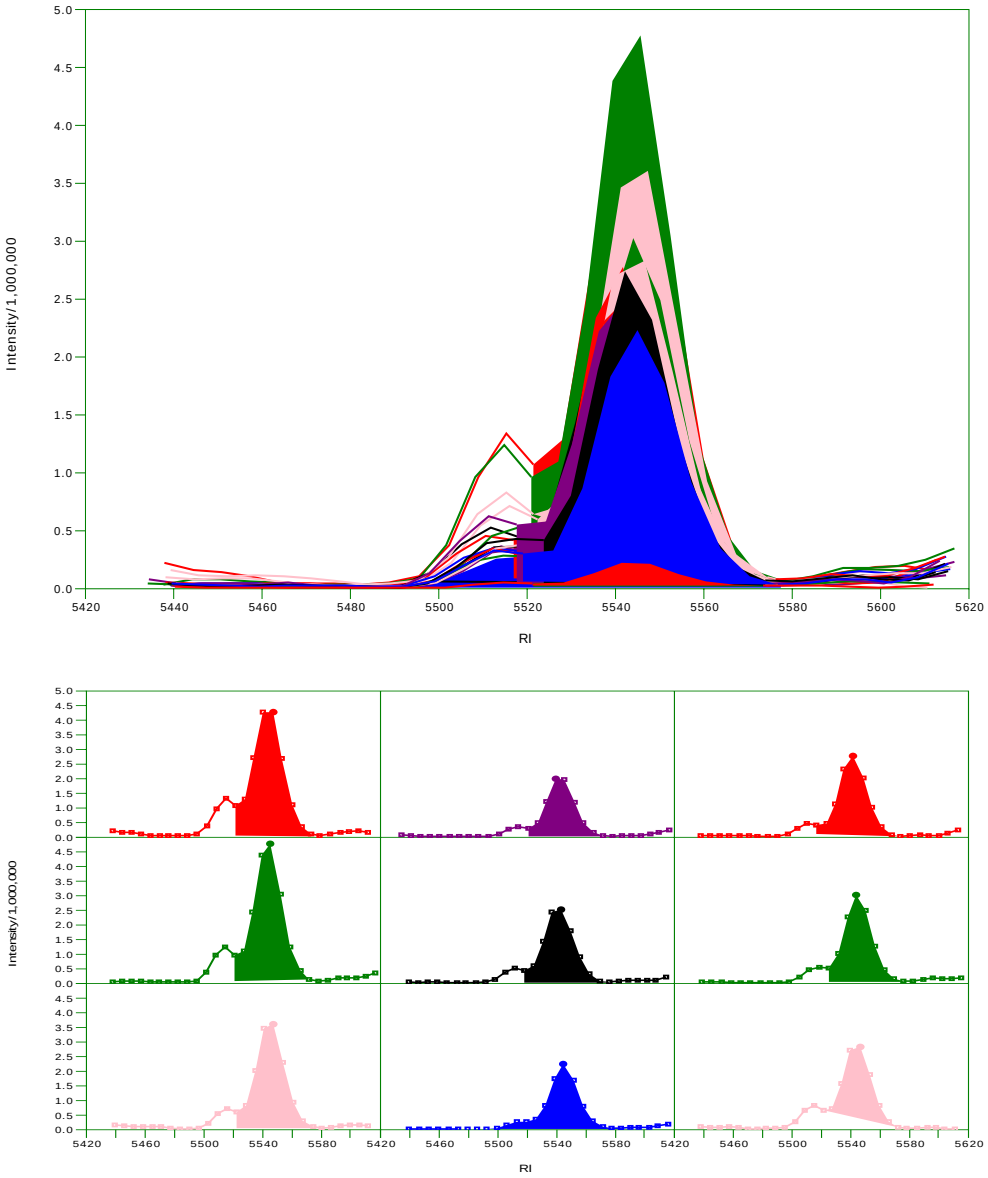


Fig. 9. Examples in inconsistent shoulder peaks. Splitting of shoulder (Upper panel); Area change after re-integration (Blue for automatic integration and red for re-evaluated integration (Lower panel).

In Figure 10, the major peak on the right is identified as 1-docosahexaenoylglycero-phosphocholine (1-DHGPC), whereas the shoulder on the left side is identified as 2-docosahexaenoylglycerophosphocholine (2-DHGPC). In one sample, the peak for 2-DHGPC was inaccurately combined into the peak for 1-DHGPC when it was automatically integrated. In another sample, the baseline was not calculated consistently. The curves at the lower right show the correction. After re-integration the erroneous integration was corrected and the small peak for 2-DHGPC was recovered.

Software that can detect inconsistencies in peak detection and integration across samples in a sample set can ultimately improve the accuracy in the integration of peaks that have been identified as metabolites; this in turn leads to lower CV's and more accurate statistical analysis which can contribute significantly to the elucidation of metabolism and metabolite pathway.



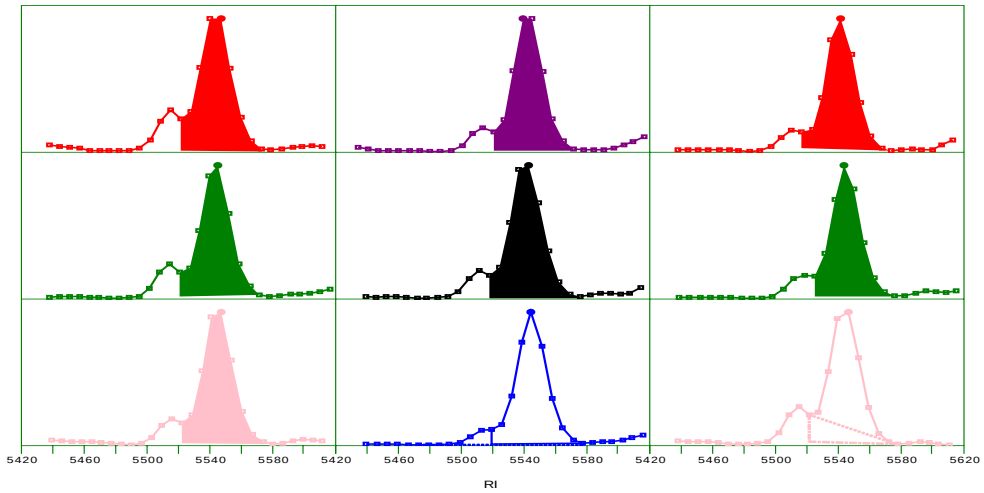


Fig. 10. Examples of inconsistent shoulder peaks

6. Conclusion

Metabolomics as a technology has demonstrated clear utility in a broad array of biological applications. The applications are not only in demonstrating simple metabolic comparisons between treated and control groups but in studies involving biomarker discovery, drug development/MOA/recovery, bio-processing, agricultural applications, consumer products, diagnostics, and so on (Sreekumar, Poisson et al. 2009; Berger, Kramer et al. 2007; Barnes, Teles et al. 2009; Boudonck, Mitchell et al. 2009; Ma, Ellet et al. 2009; Ohta, Masutomi et al. 2009; Watson, Roulston et al. 2009; Oliver, Guo et al.). The ability to run metabolomic studies in high-throughput has been a challenge thus far, not so much because of the complexity or size of the data, but because of the difficulty in generating reproducible data having low process variation that can be quantified, is devoid of artifactual components, and provides high confidence in the identification of metabolites. Without knowledge of the variability of the process on a metabolite by metabolite basis, it is not possible to determine the true biological variability and thus, cannot provide accurate answers to the questions that under investigation.

As demonstrated here, quality and the throughput of processing sample data for metabolomics studies do not need to be mutually exclusive. By taking an intelligent engineering approach to the data workflow, knowing when to automate a process and developing software solutions that are streamlined for this process, the processing of sample data for metabolomics studies can be done in significantly high volume and with high quality.

7. Acknowledgement

We gratefully acknowledge the work and contributions of all members of the Metabolon informatics, platform, project management, statistics and management teams for their dedicated work in building an enterprise metabolomics platform. CD, AE, HD, and KL are employees of Metabolon.

8. References

Barnes, V. M., R. Teles, et al. (2009). "Acceleration of purine degradation by periodontal diseases." *J Dent Res* 88(9): 851-855.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B* 57: 289-300.

Berger, F. G., D. L. Kramer, et al. (2007). "Polyamine metabolism and tumorigenesis in the Apc(Min/+) mouse." *Biochem Soc Trans* 35(Pt 2): 336-339.

Boudonck, K. J., M. Mitchell, et al. (2009). "Characterization of the biochemical variability of bovine milk using metabolomics " *Metabolomics* 5(4): 375-386.

Bowen, B. P. and T. R. Northen "Dealing with the unknown: metabolomics and metabolite atlases." *J Am Soc Mass Spectrom* 21(9): 1471-1476.

Bryan, K., L. Brennan, et al. (2008). "MetaFIND: a feature analysis tool for metabolomics data." *BMC Bioinformatics* 9: 470.

Dehaven, C. D., A. M. Evans, et al. (2010). "Organization of GC/MS and LC/MS metabolomics data into chemical libraries." *J Cheminform* 2(1): 9.

Dunn, W. B., N. J. Bailey, et al. (2005). "Measuring the metabolome: current analytical technologies." *Analyst* 130(5): 606-625.

Evans, A. M., C. D. Dehaven, et al. (2009). "Integrated, Nontargeted Ultrahigh Performance Liquid Chromatography/Electrospray Ionization Tandem Mass Spectrometry Platform for the Identification and Relative Quantification of the Small-Molecule Complement of Biological Systems." *Anal Chem* 81(16): 6656-6667.

Fan, T. W., A. N. Lane, et al. (2004). "The promise of metabolomics in cancer molecular therapeutics." *Curr Opin Mol Ther* 6(6): 584-592.

Fields, C. (1996). "Informatics for ubiquitous sequencing." *Trends Biotechnol* 14(8): 286-289.

Griffin, J. L. (2006). "Understanding mouse models of disease through metabolomics." *Curr Opin Chem Biol* 10(4): 309-315.

Hood, L. E., M. W. Hunkapiller, et al. (1987). "Automated DNA sequencing and analysis of the human genome." *Genomics* 1(3): 201-212.

Hunkapiller, T., R. J. Kaiser, et al. (1991). "Large-scale and automated DNA sequence determination." *Science* 254(5028): 59-67.

Katajamaa, M. and M. Oresic (2005). "Processing methods for differential analysis of LC/MS profile data." *BMC Bioinformatics* 6: 179.

Katajamaa, M. and M. Oresic (2007). "Data processing for mass spectrometry-based metabolomics." *J Chromatogr A* 1158(1-2): 318-328.

Khoo, S. H. and M. Al-Rubeai (2007). "Metabolomics as a complementary tool in cell culture." *Biotechnol Appl Biochem* 47(Pt 2): 71-84.

Lawton, K. A., A. Berger, et al. (2008). "Analysis of the adult human plasma metabolome." *Pharmacogenomics* 9(4): 383-397.

Lindon, J. C., E. Holmes, et al. (2007). "Metabonomics in pharmaceutical R&D." *Febs J* 274(5): 1140-1151.

Ma, N., J. Ellet, et al. (2009). "A single nutrient feed supports both chemically defined NS0 and CHO fed-batch processes: Improved productivity and lactate metabolism." *Biotechnol Prog* 25(5): 1353-1363.

Nordstrom, A., G. O'Maille, et al. (2006). "Nonlinear Data Alignment for UPLC-MS and HPLC-MS Based Metabolomics: Quantitative Analysis of Endogenous and Exogenous Metabolites in Human Serum." *Anal Chem* 78(10): 3289-3295.

Ohta, T., N. Masutomi, et al. (2009). "Untargeted metabolomic profiling as an evaluative tool of fenofibrate-induced toxicology in Fischer 344 male rats." *Toxicol Pathol* 37(4): 521-535.

Oliver, M. J., L. Guo, et al. "A sister group contrast using untargeted global metabolomic analysis delineates the biochemical regulation underlying desiccation tolerance in Sporobolus stapfianus." *Plant Cell* 23(4): 1231-1248.

Patterson, A. D., H. Li, et al. (2008). "UPLC-ESI-TOFMS-based metabolomics and gene expression dynamics inspector self-organizing metabolomic maps as tools for understanding the cellular response to ionizing radiation." *Anal Chem* 80(3): 665-674.

Scalbert, A., L. Brennan, et al. (2009). "Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research." *Metabolomics* 5(4): 435-458.

Scheltema, R., S. Decuypere, et al. (2009). "Simple data-reduction method for high-resolution LC-MS data in metabolomics." *Bioanalysis* 1(9): 1551-1557.

Smith, C. A., E. J. Want, et al. (2006). "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification." *Anal Chem* 78(3): 779-787.

Sreekumar, A., L. M. Poisson, et al. (2009). "Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression." *Nature* 457(7231): 910-914.

Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." *Proc Natl Acad Sci U S A* 100(16): 9440-9445.

Thielen, B., S. Heinen, et al. (2009). "mSpecs: a software tool for the administration and editing of mass spectral libraries in the field of metabolomics." *BMC Bioinformatics* 10: 229.

Tolstikov, V. V. and O. Fiehn (2002). "Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry." *Anal Biochem* 301(2): 298-307.

Want, E. J., A. Nordstrom, et al. (2007). "From exogenous to endogenous: the inevitable imprint of mass spectrometry in metabolomics." *J Proteome Res* 6(2): 459-468.

Watson, M., A. Roulston, et al. (2009). "The small molecule GMX1778 is a potent inhibitor of NAD+ biosynthesis: strategy for enhanced therapy in nicotinic acid phosphoribosyltransferase 1-deficient tumors." *Mol Cell Biol* 29(21): 5872-5888.

Werner, E., J. F. Heilier, et al. (2008). "Mass spectrometry for the identification of the discriminating signals from metabolomics: current status and future trends." *J Chromatogr B Analyt Technol Biomed Life Sci* 871(2): 143-163.

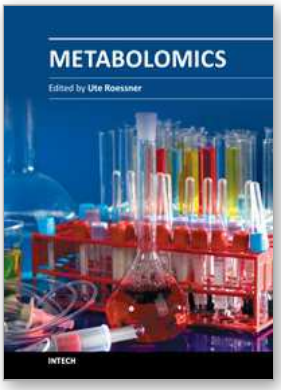
Wilson, I. D., J. K. Nicholson, et al. (2005). "High resolution "ultra performance" liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies." *J Proteome Res* 4(2): 591-598.

Wishart, D. S. (2009). "Computational strategies for metabolite identification in metabolomics." *Bioanalysis* 1(9): 1579-1596.

Wishart, D. S. (2011). "Advances in metabolite identification." *Bioanalysis* 3(15): 1769-1782.

Wishart, D. S., D. Tzur, et al. (2007). "HMDB: the Human Metabolome Database." *Nucleic Acids Res* 35(Database issue): D521-526.

Xia, J., N. Psychogios, et al. (2009). "MetaboAnalyst: a web server for metabolomic data analysis and interpretation." *Nucleic Acids Res* 37(Web Server issue): W652-660.



Metabolomics

Edited by Dr Ute Roessner

ISBN 978-953-51-0046-1
Hard cover, 364 pages
Publisher InTech
Published online 10, February, 2012
Published in print edition February, 2012

Metabolomics is a rapidly emerging field in life sciences, which aims to identify and quantify metabolites in a biological system. Analytical chemistry is combined with sophisticated informatics and statistics tools to determine and understand metabolic changes upon genetic or environmental perturbations. Together with other 'omics analyses, such as genomics and proteomics, metabolomics plays an important role in functional genomics and systems biology studies in any biological science. This book will provide the reader with summaries of the state-of-the-art of technologies and methodologies, especially in the data analysis and interpretation approaches, as well as give insights into exciting applications of metabolomics in human health studies, safety assessments, and plant and microbial research.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821