

Annotation of LC/ESI-MS Mass Signals

Ralf Tautenhahn, Christoph Böttcher, Steffen Neumann
[rtautenh|cboettch|sneumann]@ipb-halle.de

Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany <http://www.ipb-halle.de>

Abstract. Mass spectrometry is the work-horse technology of the emerging field of metabolomics. The identification of mass signals remains the largest bottleneck for a non-targeted approach: due to the analytical method, each metabolite in a complex mixture will give rise to a number of mass signals. In contrast to GC/MS measurements, for soft ionisation methods such as ESI-MS there are no extensive libraries of reference spectra or established deconvolution methods. We present a set of annotation methods which aim to group together mass signals measured from a single metabolite, based on rules for mass differences and peak shape comparison.

Availability: The software and documentation is available as an R package on <http://msbi.ipb-halle.de/>

1 Introduction

Metabolomics and especially mass spectrometry have evolved into an important technology to solve challenges in functional genomics. The ambitious goal is the unbiased and comprehensive quantification of metabolite concentrations of organisms, tissues, or cells [Oliver98,Fiehn00].

The combination of chromatographic separation with subsequent mass spectrometric detection has emerged as key technology for multiparallel analysis of low molecular weight compounds in biological systems. Gas chromatography-mass spectrometry (GC/MS) based techniques are mature and well-established but restricted to volatile compounds, at least after derivatization. High-performance liquid chromatography-mass spectrometry (HPLC/MS) facilitates the analysis of compounds of higher polarity and lower volatility in a much wider mass range without derivatization. To complement GC/MS based profiling schemes towards a higher coverage of a systems' metabolome LC/MS based platforms have been developed recently [Roepenack-Lahaye04].

LC/MS techniques require an ionisation of the analytes out of the liquid phase under atmospheric pressure. This is in sharp contrast to GC/MS, where analytes are ionised and subsequently fragmented in the gas phase by electron impact (EI). Typical atmospheric pressure ionisation (API) techniques are electrospray ionisation (ESI) and atmospheric pressure chemical ionisation (APCI).

Positive-ion API spectra of low molecular weight compounds often comprise simple adduct ions with one or more cations (e.g. $[M + H]^+$, $[M + H + Na]^{2+}$) and cluster ions (e.g. $[2M + H]^+$, $[M + CH_3CN + H]^+$). For sensitive compounds fragment ions (e.g. $[M + H - C_6H_{10}O_5]^+$) can be observed due to a collision induced dissociation in the transfer region of the mass spectrometer. In negative-ion mode ionisation occurs by abstraction of positive ions (e.g. $[M - H]^-$) or adduct formation with anions (e.g. $[M + Cl]^-$).

For the identification of compounds it is a prerequisite to deconvolute a compounds' mass spectra from the GC(LC)/MS raw data. In case of EI huge spectral libraries exist where the extracted spectrum can be searched for. For soft ionisation techniques such as ESI to date such libraries do not exist, making the identification of known compounds and the structure elucidation of unknown compounds a serious bottleneck in LC/MS based profiling schemes. Thus, the automatic extraction of a components mass spectrum and elucidation of the chemical relations inbetween these spectra is a prerequisite for high throughput analyses and annotation of LC/MS datasets from metabolomics experiments.

This paper is structured as follows: first the workflow is explained, starting with the machine analysis, signal processing to the annotation with two complementary approaches. The last steps in the workflow are the disambiguation and conflict resolution of the results. In section 3 we evaluate the annotation on a real-world dataset measured from plant seed material. We finish with a conclusion and outlook.

2 Implementation

Our metabolomics pipeline consists of several consecutive processing steps. First the samples are run on our LC/MS platform, followed by a signal processing step which collects raw signals into centroid peak data. Since no identification is available for the mass signals, they have to be aligned based on their mass and retention time, such that the N peaks are recognisable across the M runs, producing a $N \times M$ matrix of intensities.

The annotation procedures operate on individual runs, but also take cross-sample correlation into account. The system is implemented in R and the Bioconductor framework, with some compute-intense tasks being placed in native C-code which is called through an R function.

2.1 Data acquisition and preparation

We analysed methanolic seed extracts of *Arabidopsis thaliana* by capillary high performance liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry (HPLC/ESI-QTOF-MS). In this setup the crude

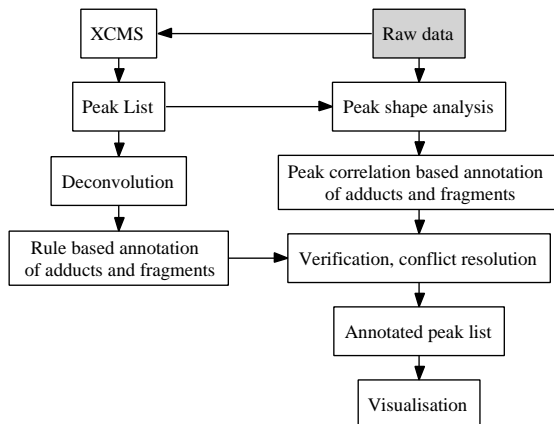


Fig. 1: Workflow of annotation procedure. The “Raw data” acquisition step (upper-right, grey) is carried out on a particular mass spectrometer, the remaining steps are vendor-independent.

seed extracts are first separated on a modified C18 phase applying an acetonitrile/formic acid-water/formic acid gradient at a flow rate of $5 \mu\text{L min}^{-1}$. Eluted compounds are detected with an API QSTAR Pulsar i equipped with an ion-spray source in positive ion mode between m/z 100 to 1000 at a mean resolution of 8000-9000 with an accumulation time of 2 seconds. 16 technical replicates were measured. Raw data files were exported with Analyst QS 1.0 and processed with the XCMS package [Smith06] using conservative parameters ($s/n \geq 3$, $\text{fwhm}=30$, $\text{bw}=10$).

2.2 Detection of known adducts and fragments

The observed signal s (measured in m/z) is the result of the molecule mass M modified by the chemical process during the ionisation:

$$s = \frac{nM + a}{z}$$

where n ($n = 1, 2, \dots$) is the number of molecules in the ion, a the mass of the adducts (negative for fragments) and z the charge.

For soft ionisation methods such as LC/ESI-MS, different adducts (e.g. $[M + K]^+$, $[M + Na]^+$) and fragments (e.g. $[M - C_3H_9N]^+$, $[M + H - H_2O]^+$) occur. Depending on the molecule having an intrinsic charge, $[M]^+$ may be observed as well. To scan for adducts and fragments with a known mass difference a predefined and user-extensible list (see Tab. 1) is employed.

Theoretically all isotope peaks, adducts and fragments belonging to a single components mass spectrum should have the same retention time. But since peak

Formula	Mass difference in amu
$[M + H]^+$	1.007276
$[M + Na]^+$	22.98977
$[M^+ + Na]^{2+}$	22.98977
$[M + K]^+$	38.963708
$[2M + Na]^+$	22.98977
$[M + H + Na]^{2+}$	23.9976
$[M + H - NH_3]^+$	-16.01872
$[M - C_3H_9N]^+$	-59.073499

Table 1: Examples of known adducts and fragments with their mass differences occurring in positive ion mode. The actual difference is calculated considering the charge and the number of molecules M in the observed ion.

detection works in the chromatographic domain and independently for all mass signals, they can differ. Especially in the case where chromatographic peaks are broad and noisy, their centroids may be detected in one of the neighbouring scans. To be robust against this effect, a “sliding retention time window” with a user specified width (e.g. $\Delta t = 4s.$, in our case equivalent to two scans) is used.

For each retention time window all possible combinations

$$M_{ij} = \frac{z_j s_i - a_j}{n_j}$$

of mass signals s_i within the time window and known adducts $A_j = (n_j, z_j, a_j)$, $j = 1 \dots J$ are calculated. Each group of similar masses M_{ij} (tolerance depends on the machine accuracy) yields one annotation hypothesis. All resulting reasonable adduct/fragment hypotheses are collected for subsequent verification, ambiguity removal and annotation.

Example: Given the mass signals $s_1, s_2, s_3 = (122.99, 138.97, 142.01)$ m/z which are observed in one retention time window. The differences $s_1 - A_{(Na)} = 122.99 - 22.98977 = 100.0002$ and $s_2 - A_{(K)} = 138.97 - 38.963708 = 100.0063$ support the hypothesis that s_1 and s_2 are adducts ($[M + Na]^+$, $[M + K]^+$) of a molecule M with an estimated mass of 100.0033 amu.

2.3 Verification of annotation hypotheses

The hypotheses described in the previous section can either be correct and reflect the actual chemical process during ionisation, or they are the result of co-eluting molecules having a known fragment/adduct mass difference by chance. Therefore a verification step is mandatory.

Intensity correlation across samples

Because of the theoretically fixed ratio between molecule and adduct intensities in all observations, a simple verification of the chemical relation for a given pair of mass signals is to calculate the correlation of the integrated peak intensities across all samples in which these peaks have been observed. The necessary alignment of two or more experiments is also part of the XCMS package.

Intensity correlation in the chromatographic domain

The adducts and fragments of the same molecule have the same intensity ratio also in each individual scan of the LC/MS measurement. Therefore their extracted ion chromatograms (EIC) are theoretically linearly dependent, and correlation or linear regression ($y = \alpha x + 0$, with subsequent evaluation of the factors α and the residuals) can be used to estimate the similarity between the chromatograms.

Peaks with a low intensity are more subjected to noise influence, their chromatograms can be very flat and/or jagged, resulting in a low correlation.

2.4 Exposing chemical relations by chromatogram correlation analysis

Even without any predefined mass differences, valuable chemical hypotheses can be obtained by analysing the chromatogram correlations across all samples. For this purpose EIC - correlations are calculated for *all* pairs of mass signals within each retention time window. This is done for each sample where the examined pair of mass signals was observed. As result of this computation, the distribution of peak shape correlations across the samples can be evaluated.

2.5 Ambiguity removal and conflict resolution

Since all chemical relation hypotheses are searched in a sliding retention time window, all duplicates, subset relations and conflicts between hypotheses have to be eliminated. In a naive approach, this would require to compare all members of all hypotheses against each other. To speed up this procedure, *signatures* are calculated for each chemical relation hypothesis. One hypothesis group H_k consists of several entries in the form (s_i, A_j) , which encode the annotation hypothesis. For each entry a signature s_{ij} is calculated as

$$\text{sig}_{ij} = p_1 i + p_2 j \quad (s_i, A_j) \in H_k,$$

with $p_1, p_2 > \max(i, j)$ being prime numbers. Furthermore, a hash value is created for each hypothesis group:

$$\text{hash}(H_k) = \sum_{(s_i, A_j) \in H_k} \text{sig}_{ij}.$$

Using these hash values, hypotheses groups containing the same annotations have the same hash value, and subset relations of hypotheses can be detected efficiently using the signatures.

Furthermore, some efforts are made to resolve chemical conflicts. For example $[M]^+$ and $[M + H]^+$ cannot appear both for the same molecule, because the molecule is supposed to have an intrinsic charge in the first case but not in the second. Therefore, relations of this kind can safely be removed.

3 Results

The evaluation is based on the arabidopsis seed dataset described above. The peak list exported from XCMS contains 1100 mass signals. The allowed mass tolerance for the annotation was $0.005 m/z$, the retention time window was set to 4 seconds. Larger windows (up to 10 seconds) had little effect.

3.1 Annotation of known mass differences

The rule set used contains 30 known mass differences covering typical adducts with protons, sodium and potassium cations and also 10 fragments with neutral losses. Using only this rule set about 200 signals are annotated as isotope peaks and there are more than 1000 competing annotations for adducts and fragments, especially for retention times with many coeluting substances.

For the verification step the threshold for the correlation of the intensities across the samples is specified as 0.6 and for the chromatogram correlation a minimum of 0.8 in 75 % of the samples is required.

After the verification step and ambiguity removal 10 % of the 1100 mass signals are annotated as isotope peaks and 20 % as adducts and fragments.

Fig. 2 shows an example of rule based annotation. Five signals belonging to the mass spectrum of feruloylcholin could be grouped in the peak list. Chromatogram correlations (Fig. 3) were used only for verification. Some combinations of low intensity signals show only a weak correlation, but the hypotheses is kept as long as the other correlations with these signals are above the threshold.

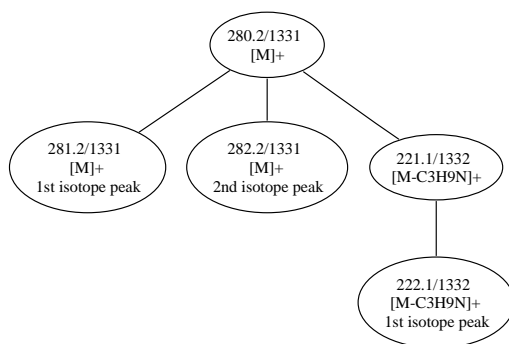


Fig. 2: Graph of mass signals. Nodes are labelled with mass/retention time and a short description, edges indicate an successful verification by correlation across samples and EICs.

3.2 Deconvolution based on chromatogram correlation analysis

Using only the chromatogram correlations, highly correlated pairs of mass signals were combined to chemical relation hypothesis groups. We used only those pairs

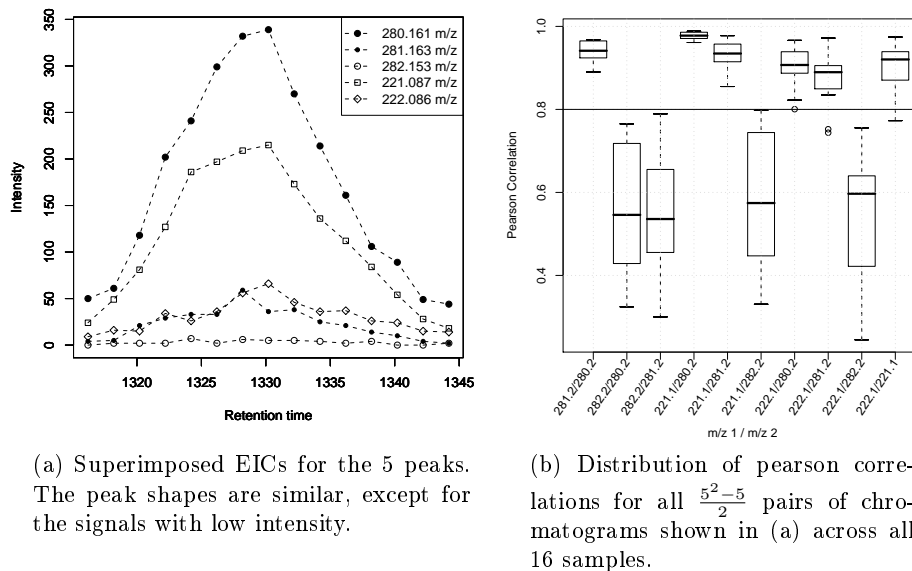


Fig. 3: Verification of the annotation shown in Fig. 2. Correlations with the low signal 282.2 m/z are below threshold.

with correlation higher than 0.8 in 75 % of the cases (12 out of 16 samples). Manual verification of one hypothesis group confirmed that all grouped signals belong to the mass spectrum of tryptophane (Fig. 4).

3.3 Run-time

The “Peak shape analysis” step in the workflow (see fig. 1) includes the creation of the extracted ion chromatogram (EIC) for each peak from each file using the generic XCMS functions. For each EIC this takes about 0.3 seconds on a standard 2 GHz PC. Depending on the number of peaks, files and the file size this preprocessing step can take up to several hours, which can be reduced by either spreading the computation across a compute cluster (we have good experience using the Sun Grid Engine 6.06) or by using a dedicated and optimized EIC extraction routine. Independent of the calculation method employed, we are currently creating a data warehouse for preprocessed LC/MS data, which includes the EIC for each peak and avoids recomputing EICs for the annotation step altogether.

Once the EICs are generated, the wall-clock run-time for the annotation (correlation, validation and graphics output) of the sample set of 16 files \times 1100 peaks described above is 120 seconds on a standard 2 GHz PC.

4 Conclusion

In metabolomics research the large gap between fingerprint data of unknown mass signals and profiling data for a limited number of metabolites needs to be narrowed down. The peaklists of LC/MS measurements can contain a few thousand peaks, and manual inspection of all of them is simply impossible. The downstream bioinformatics analysis such as hierarchical clustering or self-organising maps is difficult if a large number of observations is caused by artefacts of the analytical process.

The annotation is a valuable addition to the commonly used peaklists, based on both an extensible set of rules and peak shape analysis. Even if no chemical identification is possible, the truly interesting signals become more obvious.

Our annotation framework is also capable of including further sources that are aimed towards metabolite identification, such as the mass decomposition tool IMS [Böcker06] or exact masses from compound libraries such as the metabolic pathways database KEGG [Goto97,Kanehisa06] or KnapSACK [Shinbo06].

Acknowledgements

We thank Dierk Scheel, Edda von Roepenack-Lahaye and Jürgen Schmidt for their valuable discussions. The work is supported under BMBF grant 0312706G.

References

- [Böcker06] Böcker, S., Lipták, Z. and Pervukhin, A. Decomposing metabolomic isotope patterns. In *WABI 2006 Proc. of Wabi 2006 – 6th Workshop on Algorithms in Bioinformatics*. 2006.
- [Fiehn00] Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R. and Willmitzer, L. Metabolite profiling for plant functional genomics. *Nature Biotechnology*, vol. 18:115, 2000.
- [Goto97] Goto, S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K. and Kanehisa, M. Organizing and computing metabolic pathway data in terms of binary relations. *Pac Symp Biocomput*, pp. 175–186, 1997.
- [Kanehisa06] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, vol. 34(Database issue):354–357, Jan 2006.
- [Oliver98] Oliver, S., Winson, M., Kell, D. and Baganz, F. Systematic functional analysis of the yeast genome. *Trends Biotechnol*, vol. 16(9):373–378, Sep 1998.

- [Roepenack-Lahaye04] Roepenack-Lahaye, E. v., Degenkolb, T., Zerjeski, M., Franz, M., Roth, U. *et al.* Profiling of Arabidopsis Secondary Metabolites by Capillary Liquid Chromatography Coupled to Electrospray Ionization Quadrupole Time-of-Flight Mass Spectrometry. *Plant Physiology*, vol. 134:548–559, February 2004.
- [Shinbo06] Shinbo, Y., Nakamura, Y., Altaf-Ul-Amin, M., Asahi, H., Kurokawa, K. *et al.* *Plant Metabolomics*, chap. KNAPSAcK: A comprehensive species-metabolite relationship database., pp. 165–181. *Biotechnology in Agriculture and Forestry*. Springer, 2006.
- [Smith06] Smith, C., Want, E., O'Maille, G., Abagyan, R. and Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Analytical Chemistry*, vol. 78(3):779–787, 2006.

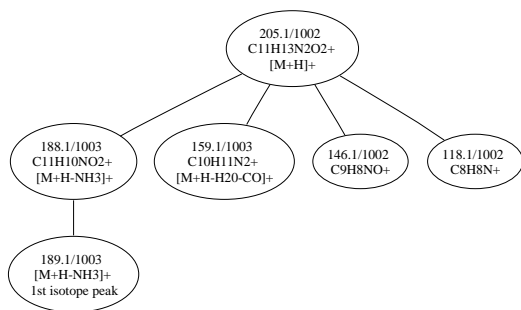


Fig. 4: Graph of extracted signals that belong to the mass spectrum of tryptophane. Nodes are labelled with mass/retention time and a short description, edges indicate a chromatogram correlation above threshold (see 2.4).

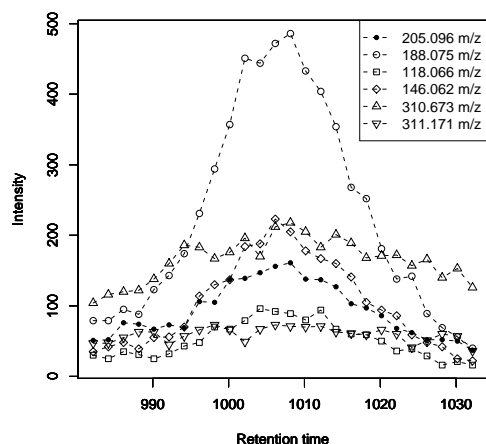


Fig. 5: Superimposed EICs for the 5 peaks shown in fig. 4 and also for the mass signals 310.7 and 311.2 m/z , which are co-eluting, but not chemically related to tryptophane.

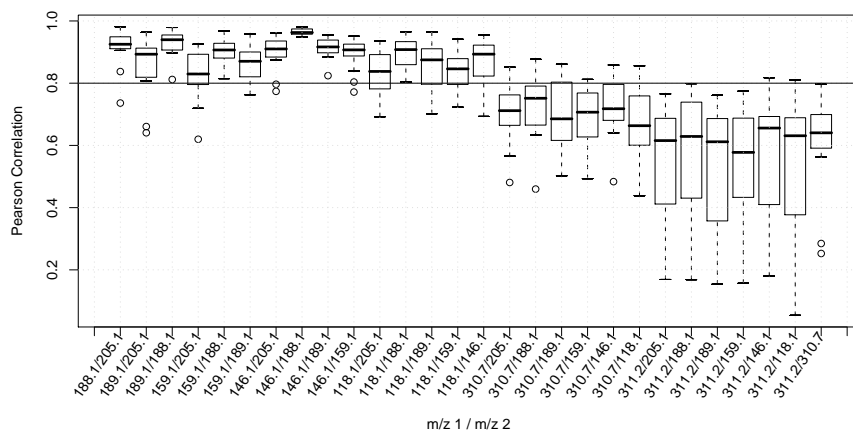


Fig. 6: Distribution of pearson correlations for all pairs of chromatograms shown in fig. 5 across all 16 samples. The mass signals 310.7 and 311.2 m/z are co-eluting, but not chemically related to tryptophane. It can be seen that the chromatogram correlations are significantly lower for *all* combinations with those signals.