An efficient spectra processing method for metabolite identification from (1)H-NMR metabolomics data.

Daniel Jacob, Catherine Deborde, Annick Moing

An efficient spectra processing method for metabolite identification from ^{1}H-NMR metabolomics data

Daniel Jacob[1,2], Catherine Deborde[1,2] and Annick Moing[1,2]

[1]*INRA, UMR1332 Fruit Biology and Pathology, Centre INRA de Bordeaux, F-33140 Villenave d'Ornon, France*

[2]*Metabolome Facility of Bordeaux Functional Genomics Center, IBVM, Centre INRA de Bordeaux, F-33140 Villenave d'Ornon, France*

Corresponding author: Daniel Jacob, email daniel.jacob@bordeaux.inra.fr

tel +33 5 57 12 25 28, fax +33 5 57 12 25 41

Abstract

The spectra processing step is crucial in metabolomics approaches, especially for proton NMR metabolomic profiling. During this step, noise reduction, baseline correction, peak alignment and reduction of the 1D ^{1}H-NMR spectral data are required in order to allow biological information to be highlighted through further statistical analyses. Above all, data reduction (binning or bucketing) strongly impacts subsequent statistical data analysis and potential biomarker discovery. Here, we propose an efficient spectra processing method which also brings a helpful support for compound identification using a new data reduction algorithm that produces relevant variables, called buckets. These buckets are the result of the extraction of all relevant peaks contained in the complex mixture spectra, rid of any non-significant signal. Taking advantage of the concentration variability of each compound in a series of samples and based on significant correlations that link these buckets together into clusters, the method further proposes automatic assignment of metabolites by matching these clusters with the spectra of reference compounds from HMDB or a home-made database. This new method is applied to a set of simulated ^{1}H-NMR spectra to determine the effect of some processing parameters and, as a proof of concept, to a tomato ^{1}H-NMR dataset to test its ability to recover the fruit extract compositions. The implementation code for both clustering and matching steps is available upon request to the corresponding author.

Keywords: *^{1}H-NMR spectroscopy, spectra processing, metabolite identification, metabolomics*

Introduction

Nowadays, metabolomics which aims at studying the metabolite complement of organisms, tissues or biofluids is widely used in many research fields such as nutrition, toxicology, disease diagnosis, microbiology and plant sciences [1, 2, 3, 4]. Among the different analytical strategies, proton NMR spectroscopy (^{1}H-NMR) used in pioneering profiling studies [5, 6], remains largely used. This technology has been widely used as a high-throughput technique for non-targeted fingerprinting with little or no sample preparation. It has also been applied for targeted profiling and the absolute quantification of major metabolites, despite its relatively low sensitivity, taking advantage of its large dynamic range [7-10]. Today, many issues have been solved related to the spectra preprocessing steps as detailed in a recent review [11]. However, one of the major challenges for ^{1}H-NMR metabolic profiling remains the automatic assignment of metabolites from spectra. Due to huge data volumes, especially with a high-throughput strategy, it is essential to develop new assignment methods which greatly help the user, especially in plant kingdom where the knowledge about plant matrixes is lower than for mammalian biofluids. Although an expert user can successfully carry out this tedious task by manual assignment, this approach has several drawbacks. First, it is time consuming. Second, it strongly depends on the user's expertise level in several fields such as NMR spectroscopy, biochemistry and biology. The final list of identified metabolites may therefore depend on this expertise, thus introducing a bias.

To overcome this issue, two recent approaches have been described in recent publications [12, 13]. The first one is MetaboHunter [12], a user-friendly ^{1}H-NMR-based web server application. Concerning the automatic assignment of metabolites, the approach used by MetaboHunter consists in matching each compound from a reference compound library with the peak list of an NMR spectrum, exclusively relying on the peak positions within this spectrum. The strength of this method is to use little prior knowledge, apart from the biological source and some analytical parameters in order to choose the right reference compound library. Although it may produce good results as shown by the authors, this approach tends to provide too many candidates. While almost 100% of true-positives can be recovered in a mixture of a dozen compounds, the distribution of scores of true positives may span more than a hundred putative compounds [12]. This requires a tedious work for checking these numerous candidates**.** The second approach proposes an automatic method for identifying and quantifying metabolites in one-dimensional ^{1}H-

NMR spectra [13]. The method consists in constructing a model spectrum as the sum of carefully selected compounds and based on their deconvolution (*i.e.* Lorentzian shape). Contrary to MetaboHunter, this method takes into account the relative intensities of each peak for each compound spectrum versus sample spectrum. The strength of this approach is to bring an effective way to correctly estimate the quantification of each compound in the complex mixture spectra. However, as it requires a good prior knowledge about the compound list–together with the reference compound NMR spectra acquired in the same NMR conditions as the complex mixture, this approach does not allow metabolites to actually be identified, but rather allows the selected list to be validated afterwards.

To overcome this drawback, the present work proposes a new approach which attempts to gather the strengths of the two approaches described above, namely a low prior knowledge concerning the composition of complex mixture spectra, and a robust matching method which takes into account both the peak positions and their relative intensity. To fulfill this aim, we have developed a four-step workflow (Fig. 1) consisting in (i) spectra preprocessing, (ii) a new data reduction method called ERVA (Extraction of Relevant Variables for Analysis), followed by (iii) a clustering of latent variables approach (CLV) [14] to obtain new variables called buckets that correspond to latent compounds, and finally (iv) we try to match these latent compounds with each reference compound from a suitable library of spectra, the latter depending on the biological source and acquisition parameters (mainly solvent, pH, ionic strength and to a lesser extend pulse sequence). Our approach applies to metabolomics experiments based on ^{1}H-NMR profiling of a series of samples, *i.e.* with an experimental design. It therefore relies on a set of ^{1}H-NMR spectra, unlike the methods previously cited which treat only one spectrum at a time. At the final step of our approach, the proposed list of compounds thus produced for each cluster can also serve to build a specific reference compound library, for quantification using the method described in [13]. As a proof of concept, we have tested the efficiency of our approach by comparing the list of proposed compounds with known compounds in a set of simulated NMR spectra and with the result of a manual approach in a previously published study on tomato. The effect of several parameters has also been determined on the set of simulated spectra. The implementation code used to obtain the results presented in this article for the clustering and the matching steps is available upon request.

Methods

The following sections cover the four steps of the method, from raw NMR spectra to the proposed list of compounds, namely spectra preprocessing, data reduction, bucket clustering and matching (Fig. 1).

Spectra preprocessing

The purpose of the first step, spectra preprocessing, is cleaning artifacts in each spectrum, namely the noise and distortion of baseline shape. These two sub-steps are processed at the same time, in each spectrum separately and independently from other spectra. Noise reduction is performed with the help of a wavelet-based method [15] which relies on a multi-level decomposition of signal, Discrete Wavelet Transform. For baseline correction, an approach relying on a smoothed spectrum is used for baseline recognition and modelling. To complete the baseline model, an interpolation technique is employed over the signal area. Then the model is subtracted from the spectrum giving a flat baseline [16]. All other approaches for baseline correction and noise reduction can be used in conjunction with the subsequent steps of our approach, provided that some cautions are followed as discussed in the data quality section.

Spectra alignment and data reduction

Usually, ^{1}H-NMR spectra are exploited for profiling using their reduction into a set of variables corresponding to spectra regions also called “bins” or “buckets”. These buckets can be easily computed or more sophistically determined [11]. Ideally, a one-to-one correspondence for each of these buckets across all samples should be satisfied. However it is generally not the case because of uncontrolled changes in chemical shifts of NMR peaks due to slight differences in pH or ionic strength and other physicochemical interactions [17]. In order to avoid an alignment processing step, many approaches [18-20] attempt to slice the spectra so that each region common to all spectra contains the same peak but is not necessarily centered on this peak. Given that our assignment approach will rely on correlations between buckets issued from the data reduction step, herein we propose a new approach called ERVA for Extraction of Relevant Variables for Analysis (Fig. 2). First, peak alignment is performed before data reduction using a homemade script written in C language based on a method similar to ICO-Shift [21]. Then, to reduce the dimensionality of spectral data while attempting to retain core information, we have chosen a mathematical method

to extract relevant buckets. This method is based on a convolution product between a spectrum (S) and the second order derivative of the Lorentzian function (SDL).

The convolution product is defined as:

$$\mathrm{C}_\sigma[\mathrm{i}] = \sum_{k=-P}^{P} \mathrm{SDL}_\sigma\,[\mathrm{i}-\mathrm{k},\mathrm{i}].\,\mathrm{S}[\mathrm{i}] \qquad (1)$$

$$i \in [1, N]$$

Where S[i] is the value of the spectrum at point $i \in [1, N]$, N the size of the spectrum (number of data points), P defines the boundaries of the summation, and SDL_σ is the second derivative of the Lorentzian function, and defined as:

$$SDL_\sigma(x, x_0) = 16.\sigma.\frac{12.(x-x_0)^2 - \sigma^2}{\pi.[4.(x-x_0)^2 + \sigma^2]^3} \qquad (2)$$

Where σ is called the Lorentzian width which is the width at half maximum, x_0 is the center of the Lorentzian function. Due to the fast decrease on both sides of the symmetrical function SDL_σ, the P parameter can be limited to one thousand points, which greatly speeds up the computation time. The convolution C_σ produces a new signal (Fig. 2). If we overlay this signal over the spectra, its zero-crossings increased by the value of σ each side, give the bounds for the regions to integrate in order to obtain relevant buckets. As we need a one-to-one correspondence for each bucket across all samples and if we assume all spectra well aligned, thus the sum of spectra is also assumed to include all information from all spectra. Therefore, the sum of spectra can serve as a reference spectrum to determine the relevant buckets common to all spectra. As a result, all relevant information are extracted, free of any non-significant signal. Concerning noise, the minimum value of the signal resulting from the convolution C_σ in a ppm range considered as noise (typically from 10 to 11 ppm) can be taken as a threshold above which no bucket will be considered and used for denoising. Mathematically, applying such a convolution product on a spectrum is similar to partial wavelet decomposition. The second derivative of the Lorentzian function plays the role of a wavelet, and the σ parameter plays the role of the level of signal decomposition. In our method, the main difference is that only one decomposition level is taken into account. The second derivative of a Lorentzian was chosen because: i) a NMR spectrum is a sum of Lorentzian, plus noise and distortion; ii) the second derivative of a Lorentzian is symmetric, and its integral is zero.

The σ parameter is the resolution parameter of the algorithm. It sets the detail level of the decomposition with a typical value of 0.0005 ppm.

Clustering of buckets

Thanks to their exact matching with the resonance peaks, the buckets now have a strong chemical meaning, since the resonance peaks are the fingerprints of chemical compounds. However, to assign a chemical compound, several resonance peaks are generally required in 1D ^{1}H-NMR metabolic profiling and for the same peak, one or more chemical compounds may correspond (peak overlapping). To discover the latent compounds, *i.e.* which buckets (resonance peaks) are linked together to form a probable chemical compound signature, an approach similar to the CLV approach is used [14]. As the CLV method, it involves two steps, namely a hierarchical clustering analysis followed by a partitioning algorithm. Both steps have been implemented with the R software (http://www.R-project.org) and using the IGRAPH package (http://igraph.sourceforge.net/index.html) for the partitioning step.

To generate relevant clusters (*i.e.* clusters possibly matching to chemical compounds), an appropriate correlation threshold has to be applied on the correlation matrix before its cluster decomposition. Below this threshold, the correlation coefficients are reset to zero. The threshold value depends on the 1D ^{1}H-NMR data set quality and spectra processing efficiency and its choice will be discussed in Results.

Matching the bucket clusters with compounds

The aim of the last step is to attempt to match each cluster of buckets, assumed to be a latent compound signature, with each reference compound spectrum of an appropriate library of reference spectra. A reference compound library is defined as a set of ^{1}H-NMR spectra, with one file per authentic compound, each file containing a list of peak positions (ppm) with the corresponding relative intensities. The Human Metabolome Database - HMDB [22] is a good example of such a collection already publicly available. A home-made library of spectra that have been experimentally acquired with authentic standard compounds is another possibility.

Relying on a reference compound library, the matching process tries to match each cluster (Source) with reference compounds (Targets). As output, it provides for each cluster a list of several putative chemical compounds along with an overall matching score, the highest score

being placed in the first rank, and so on. The matching process, and how to calculate the overall matching score for each cluster are described below and shown in Figure 3.

1. The Source is divided into sub-clusters based on distances between peaks, with the assumption that distances within a sub-cluster are almost constant and are lower than those between sub-clusters (Fig. 3A).
2. Then, for each sub-cluster within the Source, a window on the ppm scale is defined based on its first and last peaks (Fig. 3A).
3. For the Target, the corresponding sub-cluster, if it exists, must have a peak number greater or equal to that of the Source. Then a window on the ppm scale is defined based on its first and last peaks, extended of a $\Delta\delta$ ppm on each side (Fig. 3B).
4. By sliding the Source window over the Target window of a peak position in each loop, a score is computed for all possible combinations having at most one gap among the Target peaks (Fig. 3B). The score is based on the concept of "valid cluster" introduced in Chenomx NMR suite 6.0 [23] and described in [13]. The sub-cluster is valid if the best score is lower than 0.33, and the algorithm loops for the next sub-cluster.
5. Finally, a total score (called $Score_{cc}$) for the matching with the compound signature is computed taking into account all best scores of valid sub-clusters, as:

$$Score_{cc} = \frac{\sum_i (1 - S_i) * N_i}{\sum_i N_i} \quad (3)$$

where **i** is comprised between 1 and the number of valid sub-clusters, $\mathbf{S_i}$ is the score for the valid sub-cluster **i** and $\mathbf{N_i}$ is the number of peaks within the valid sub-cluster **i**.

The limitation to a single gap is related to the criterion based on the distance between peaks; more than one gap would imply a distance greater than the average distance within the sub-cluster. In addition, the concept of “valid cluster” is defined as an indicator of the goodness of fit and trust for a given peak cluster. Herein, the score value is used to sort the well-matched compounds rather than to give a threshold for rejection. $\Delta\delta$ is the major parameter of the method and plays the role of a ppm tolerance. A typical value is 0.02 ppm.

The score based on the concept of valid cluster only takes account of the valid sub-clusters. The cluster size also has to be taken into account. For that, we use the scoring function introduced in MetaboHunter [12]:

$$Score_{mcc} = \frac{N_{mcc}}{(1 + N_{cl})} \quad (4)$$

Where N_{mcc} is the number of matched peaks and N_{cl} is the number of peaks within the whole cluster.

Then, we combine (3) and (4) to obtain the new score:

$$Score_{cluster/cmpd} = \sqrt{Score_{cc} * Score_{mcc}} \quad (5)$$

Given that a cluster, especially a small one, may have a good match with many compound signatures, a correction is needed that also takes into account the matching between the compound and the complex mixture spectra in order to improve compound ranking. For that purpose, we proceed as described above to compute a matching score $\mathrm{Score}_{\mathrm{cmpd/mixture}}$ between a compound and the complex mixture spectra. Finally, the two scores are combined with a weighed sum to obtain an overall matching score:

$$Overall\ Score_{cluster/cmpd} = \frac{w_1.Score_{cluster/cmpd} + w_2.Score_{cmpd/mixture}}{w_1 + w_2} \quad (6)$$

where w_1 and w_2 allow each type of contribution to be weighed. Thus calculated, the overall matching score is used to sort the set of putative chemical identifications proposed as candidate compounds. This overall matching score measures the similarity between the cluster and a subset of resonances belonging to a reference compound. Therefore several putative compounds may match with the cluster including false positives. The rank information gives the position in the ranking of candidates for this overall matching score. The highest matching score written at the first rank in the candidate compound list gives the most probable compound.

For matching, w1 and w2 parameters have been introduced to adjust the contribution weight that takes into account the matching between the compound and the complex mixture spectrum in order to improve compound ranking. To fix the w1/w2 ratio, it is tempting to give more weight to the matching score between the compound and the complex mixture spectrum (higher w2) for size 2 clusters, and less weight for larger clusters. However, even if the size of a cluster equals 2, it is based on a strong correlation (>0.9), and such clusters often correspond to particular regions easily recognizable by an expert user. The doublet belonging to alanine (1.48 ppm) for instance is such a typical example. Moreover, in dense peak regions, peak overlapping affects peak intensities so that the patterns of each compound within such regions are intermixed. Although part of a compound has a good matching score within such a region, this does not guarantee that it is a true-positive

compound. Therefore we give the same w1/w2 ratio whatever the cluster size and the best value empirically found is 4.

Results and Discussion

The validation of our approach for an automatic assignment of metabolites described in Methods is mainly based and discussed on sets of simulated ^{1}H-NMR spectra. In this way, we could evaluate the influence of data quality (digital resolution, noise) and data processing (baseline correction, peak alignment) on the final results of the approach. Thereafter, using sets of simulated spectra, we compared our data reduction method to another widely used method, tested our matching process with two different compound libraries, and evaluated the choice of several parameters, namely the correlation threshold for the clustering step and the ppm tolerance parameter for the matching step. Finally, as a proof of concept, our approach was applied to a tomato ^{1}H-NMR dataset previously published [24].

Sets of simulated or real NMR spectra

In order to make it possible to vary different spectra parameters such as digital resolution, noise or peak misalignments within a set of simulated ^{1}H-NMR spectra, 17 reference compounds were chosen with their relative concentrations corresponding roughly to a metabolite profile of tomato fruit (Table 1). Then each spectrum was calculated as the sum of the 17 weighted normalized e spectra of reference compounds with a total intensity equal to one. The ^{1}H-NMR spectra of reference compounds were picked from our own compound library (called DBREF6) which includes 82 ^{1}H-NMR spectra that have been experimentally acquired with authentic standard compounds (pH 6, 27°C, TSP as chemical shift reference, deuterated phosphate buffer solution as solvent, NMR field 500 MHz) (see Electronic Supplementary Material Table S1). Two simulated spectra groups were formed (groups 1 and 2) simulating biological response under the effect of a factor. For each group, three repetitions were calculated by adding variation For each group, three replicates were calculated by adding a variation with a standard deviation of 20% for each intensity, thus taking into account biological and technological variability independent of the variation between-groups. Thereby, by varying noise, digital resolution or even peak alignment, a set of six NMR spectra of compound mixtures was generated, divided into two groups of three repetitions (see Electronic Supplementary Fig. S1).

To a lower extent, we also used a tomato dataset previously published [24]. The aim of this study was to characterize metabolic changes in tomato flesh and seeds in relation to crucial changes in fruit growth and development patterns. A global approach to quantify compositional changes in metabolic profiles during fruit development and ripening was developed, including untargeted metabolic profiling of polar extracts through ^{1}H-NMR. It should be noted that each dried extract was titrated with KOD to pH 6 in deuterated 400 mM phosphate buffer solution as solvent, before ^{1}H-NMR analysis on a Bruker Avance spectrometer (500 MHz NMR field strength). The tomato experiment dataset is stored into MeRy-B data repository for plant metabolomics [25], and can be accessed online [http://www.bordeaux.inra.fr/pmb/projects/t06002].

Data quality and data processing

Parameters such as noise, digital resolution, baseline correction or peak alignment mostly impact on the data reduction step. However, the following steps, bucket clustering and matching of clusters with compounds, depend on the quality of the data produced upstream. Since the main benefits of ERVA method are 1) reducing the dimensionality of spectral data while attempting to retain core information; and 2) producing buckets centered on a single resonance, the discussion will focus on the consequences of data quality and processing on these expected benefits.

Digital resolution may have an effect on resonance discrimination due to a lack of points. Indeed, if two peaks are very close, the convolution signal may not cut the x-axis between the two peaks so that they are merged into a single bucket. This may happen especially when the digital resolution is low (i.e. with a resolution 16K) and the zero crossing is impeded due to the small number of points (see Electronic Supplementary Material Fig. S2). Therefore a digital resolution of at least 32K is recommended for good convolution results.

Noise level mainly has an impact on the processing of low intensity peaks. Filtering techniques such as Savitzky-Golay filters [26] are very efficient to drastically reduce noise and thus increase the signal over noise ratio (SNR), but they also alter peak shapes and especially the inflexion points (i.e. points where the second derivative changes sign), possibly leading to a less peak separation. In contrast, with a high level of noise, the threshold applied on the convolution signal may be so high that some peaks with low intensity are lost. So, there is a trade-off to be found between the SNR and the preservation of peak shapes. This trade-off can be reached by using the wavelet denoising method [15] which relies on a multi-level decomposition of signal

using the Discrete Wavelet Transform. The noise present in the signal can be attenuated while preserving inflexion points (see Electronic Supplementary Material Fig. S3). The latter denoising method is therefore recommended for spectra preprocessing before bucketing.

Because a baseline can be approximated as a straight line across the width of a peak (the convolution of a straight line with the second derivative of a Lorentzian function being zero, due to the property of convolution: f*g" = f"*g), baseline correction has no effect on the signal resulting from the convolution and therefore on the position of buckets. In any case, a badly-corrected baseline may have effects on bucket integration and subsequently on the clustering step by altering correlations between buckets and therefore the matching of clusters with a library of reference spectra.

Concerning peak alignment, in order to have buckets common to all samples, we need a one-to-one correspondence for each bucket across all spectra. Although the use of the sum of all spectra implies that all spectra are well aligned, very small chemical shifts variation (<=0.0015 ppm) are nevertheless allowed without major impacts on the position of the buckets relative to a perfect alignment. However, imperfect alignment may impact on the clustering step, due to lower correlations between buckets belonging to the same compound. If the spectral peaks have significant local chemical shift variation, a spectral peak alignment has to be done to align the majority of the peaks in the misaligned region. To the extent that the peak alignment process does not truncate peaks themselves, but merely shifts the spectra with respect to each others with the cutting and joining points located between peaks, there is no impact on the downstream data reduction process. However, if alignments involving alteration of the lower part of the peaks become necessary, impacts will remain relatively minor using our data reduction method (see Electronic Supplementary Material Fig. S4). Indeed, buckets produced by the ERVA method are mainly based on the central part of peaks.

Comparison of two data reduction approaches in the preprocessing step

Figure 2 shows that the ERVA bucketing method does not integrate the entire information contained in the spectra, since the bucket areas do not cover whole spectra. However, the ERVA method allows all useful information to be integrated. In order to assess the improvement brought by the ERVA method on the quantity and quality of information integrated in the reduced data, we

have compared it with a very efficient bucketing method, the Adaptive, Intelligent Binning Algorithm (called AIBIN) [18] that is widely used. For this purpose, we implemented the AIBIN algorithm described in [18], in a house-made program written in C language. The test results presented in this section are based on the set of NMR simulated spectra described above with a 32K resolution and a 60 dB (0.1%) SNR. Regarding resolution parameters, we chose r=0.1 for AIBIN and σ=0.0005 ppm for ERVA. After reducing noise using the wavelet denoising method described above, we have applied both binning methods, providing two data matrices. The resulting number of buckets provided by each method was very similar: 220 for ERVA and 219 for AIBIN. Given that the six simulated spectra are separated into two groups corresponding to the two levels of a fictive factor, we first computed the proportion of variance explained by the fictive factor (i.e. the ratio between the sum of squares between groups and the total sum of squares). We found 56.3% for ERVA and 55.6% for AIBIN (see Electronic Supplementary Material Fig. S5). These values are very similar but the difference (0.7%) is nevertheless higher than the noise variance (which was of 0.03% after reducing noise). This small difference is due to the AIBIN method itself. As shown in Figure 4, AIBIN may generate asymmetric buckets in the presence of identical but slightly overlapping resonances such as doublets or triplets (see Electronic Supplementary Material Fig. S6). Next, the variance explained by the first component of PLS-DA applied on the two datasets, after unit-variance and an Orthogonal Signal Correction (OSC) was 99.5% for ERVA and 99.2% for AIBIN which is again very similar (see Electronic Supplementary Material Fig. S5). However, in terms of data quality, both methods are not equivalent. The ERVA method produces buckets centred on resonances, contrary to the AIBIN method. Therefore, using ERVA method, the ppm tolerance value in the matching step will be more stringent than using the method AIBIN. As shown below, the ppm tolerance parameter has a strong impact on the matching process between clusters of buckets and compounds.

Effect of the correlation threshold in the clustering process

Compounds involved in the same biochemical pathway may present high correlations between their resonances, but not usually as high as for resonances corresponding to the same molecule. Thus, the optimal correlation threshold is the one that discriminates these two types of correlations. Unfortunately, it does not usually exist a single threshold for such discrimination for all compounds because of different SNR for each resonance. In the process of bucket clustering,

the higher is the correlation threshold, *i.e.* close to one, the lower are the numbers of clusters and their size, and the higher is their reliability because of stronger correlation. If the threshold is reduced, then the number and size of clusters increase, until the biggest clusters start to agglomerate and at the same time new smaller clusters to appear (Fig. 5). To find a correlation threshold value allowing a maximal discrimination of compounds, we propose to process as follows: for each value of the correlation threshold included in the range [0.900 - 0.999] in steps of 0.001, apply the threshold on the correlation matrix, get the corresponding clusters of buckets, then compute the ratio between a) the size of the biggest cluster and b) the total number of clusters. The higher limit in the optimal range for the correlation threshold is obtained for the minimal value of the ratio. This allows a maximal discrimination of compounds before their aggregation. By decreasing the threshold, new clusters less correlated may appear while those among the most correlated can agglomerate but the total number of clusters decreases. A reasonable limit for the correlation threshold is obtained when the size of the biggest cluster exceeds 40. It corresponds to the maximum size of buckets that we can expect to be clustered for a given compound having a complex pattern.

Effect of the ppm tolerance parameter ($\Delta\delta$) in the matching process

The ppm tolerance parameter acts as a tolerance regarding the position of the peaks in the matching process, and allows some issues to be solved. First, the peaks of the compound reference spectra may exhibit small differences in intensity and position with those of the compounds in the complex mixture on account of different acquisition conditions, including pH. Second, the alignment process required by the ERVA method can slightly shift the actual positions of the peaks. Therefore, the minimal advised value for the ppm tolerance parameter is around 0.005 ppm in case of well-aligned spectra. To highlight the effect of the ppm tolerance parameter, we performed several tests of matching. Based on a set of ^{1}H-NMR simulated spectra with a 32K resolution and 60 dB SNR (0.1% of total variance), and after reducing noise with the wavelet denoising method and data reduction based on the ERVA method, we chose two values for threshold on the correlation matrix to perform the bucket clustering process: the first value was the higher limit of the correlation threshold yielding a maximal discrimination of compounds, as discussed in the previous section; the second value was smaller so that some new clusters less correlated could emerged. Then, with different values of the ppm tolerance parameter, we used and

compared two different reference compound libraries for the potential annotation of the simulated ^{1}H-NMR spectra set. The first compound library was the complete metabolite library of HMDB which included 905 compounds with experimental reference ^{1}H-NMR spectra on September 2012. The second compound library was issued from our own compound library, DBREF6, this library having served to generate the simulated ^{1}H-NMR spectra set (Table 1). For each performed test (4 values of ppm tolerance parameter, 2 thresholds of correlation and 2 reference compound libraries, i.e. 16 tests in total), we reported the number of true-positives out of the 17 compounds included within the simulated set, found with the highest score for at least one cluster, i.e. at rank 1 (Table 2, and Electronic Supplementary Material Table S2). When values are lower than 17, it means that either one or more false-positives have obtained a higher score or the reference compound was not matched with a significant score. Since the peak position of compounds is tolerated within a ppm range ($\Delta\delta$), it increases the likelihood of matching with the right compound on one hand, while promoting false-positives on the other hand. Therefore, the higher the number of compounds constituting the library is, the greater is the likelihood of correspondence with some of them. But as shown in Table 2, the higher the ppm tolerance parameter is, the greater is the likelihood of matching with false-positives. It's typically the case using HMDB as the reference compound library, due to the fact that this library gathers different types of metabolites and reference spectra acquired in several conditions. Therefore, the reference compound library should be chosen in accordance with the biological source (*e.g.* mammalians, plants, microorganisms) and the NMR acquisition parameters (NMR magnetic field strength, sample pH, temperature), and above all must be available. Figure 6 depicts the overall approach, from spectra bucketing to the proposal of candidate compounds and gives an example of output produced by our matching process.

Experimental Dataset on Tomato

As a proof of concept, our approach for an automatic assignment of metabolites described in Methods was also applied to a tomato dataset previously published as described in Methods. The major metabolites of each extract were identified after manual peak assignment using ^{1}H-NMR spectra from pure authentic compounds associated with comparison of published data by an expert user. Thirty-one compounds were thus identified [24]. The data set comprised 54 spectra.

First, spectrum preprocessing was applied on the tomato data set as described in Methods. The water spectral region between 4.7 and 4.95 ppm was excluded and the range of 0.5 to 9.5 ppm

considered. Then we proceeded to the spectra alignment before applying a data reduction based on the ERVA approach. The clustering step was performed on this data matrix with a correlation threshold equal to 0.96. The choice of this parameter was based on the strategy explained in Figure 5. We obtained 43 bucket clusters. To assess the results generated by the matching process, we relied on the list of 31 corresponding metabolites of the tomato study [24], found by an expert user along with their shape and location of the multiplet which have been taken into account for quantification. Among these compounds, 4 of them have either a single peak (such as fumarate and formate) or a single detectable peak (such as acetylcholine and choline) from the comparison. Indeed, given that our matching approach is based on clusters of buckets, their minimal size must be greater or equal to 2, so they cannot be found by our approach. Because more than 18% (almost 1 out of every 5 metabolites) are concerned, it represents a loss of metabolic information. An appropriate scoring function is needed for such a pattern. A way to follow could be the Probability-based matching (PBM) [27] used in mass spectrometry to search for candidates in libraries.

The list of 27 remaining metabolites will serve thereafter as "reference list" and a putative compound will be classified as true positive when present in this list. Then, the search of compounds was launched within the tomato data set using the DBREF6 library mentioned above by applying the matching method, with the ppm tolerance parameter, varying between 0.01 and 0.04 ppm. The clustering and matching steps taken as a whole run in less than 10 s, on an AMD Quad-Core 2.4 GHz processor to process the 54 spectra relying on the DBREF6 library. All results are summarized in Table 3, with 0.03 ppm for the ppm tolerance parameter. We only report results for the true positives, *i.e.* the candidate compounds corresponding with chemical compounds validated by the expert user in the reference list, along with their overall matching score (computed from eq. 6). Moreover, we only report the true positive compounds if they ranked in the top five (i.e. the five candidate compounds with the highest scores for each cluster). We found 25 true-positives compounds (more than 80% of the 31 compounds identified by the expert user), including 21 compounds at rank 1 (nearly 70%). Even if an expert validation is required, our matching method allows the most relevant metabolites to be found and then easily validated by the user. In addition, our approach generates valuable information, such as clusters of highly correlated peaks often revealing a chemical substructure, providing clues for less straightforward assignment tasks (Fig. 6).

Conclusion

In this article, we propose an efficient approach for the automatic assignment of metabolites from ^{1}H-NMR metabolomics profiles of complex mixtures and test it on sets of simulated and real spectra. Although these automatic assignments need to be validated by an expert user, they bring valuable information and a helpful support for compound identification and allow time to be drastically saved. Even for an NMR expert, the matching step may highlight regions of interest further quantified. Our approach can effectively serve as an intermediate step before automatic quantification of metabolites that uses methods either based on the complete reconstruction of the complex mixture spectrum from a library of carefully selected reference compounds [13, 28, 29] or based on Region of Interests (ROI) [30]. Indeed, these methods of quantification require a prior list of carefully selected compound, which may be a drawback if applied to samples from various types and origins. Moreover, since the clustering and matching steps of our approach produce automatic assignments, similarly to BLASTs on genomic data, they can possibly be used on sets of unannotated NMR spectra stored in data repositories such as Metabolights [31]. The ongoing development of standard formats, such as nmrML within the COSMOS international consortium (http://www.cosmos-fp7.eu), will facilitate the interconnection of our tools with quantification tools within a pipeline in a near future.

Upon request, we will provide the implementation code used to obtain the results presented in this article for the clustering step, based on R, and the matching step, written in C.

Acknowledgements

We thank Dr Stéphane Bernillon for critical reading of the manuscript and Dr Marianne Defernez for helping us to improve the abstract.

References

1. Roessner U (2012) Metabolomics InTech, Rijeka, Croatia

2. Hall RD (2011) Biology of Plant Metabolomics. Annual Plant Reviews vol 43, Wiley-Blackwell, Chichester

3. Bundy JG, Davey MP, Viant MR (2009) Environmental metabolomics: a critical review and future perspectives. Metabolomics 5:3-21

4. Nicholson JK, Lindon JC, Holmes E (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica, 29:1181-9

5. Brown FF, Campbell I D, Kuchel PW, Rabenstein DC (1977) Human erythrocyte metabolism studies by H-1 spin-echo NMR. FEBS Lett., 82:12-16

6. Nicholson JK, Buckingham MJ, Sadler PJ (1983) High-resolution 1H-NMR studies of vertebrate blood and plasma. Biochem. J., 211:605-615

7. Lindon JC, Nicholson JK (2008) Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. Trends Anal Chem 27:194-204

8. Ward JL, Beale MH (2006) NMR spectroscopy In: Saito K, Dixon R, Willmitzer L. (eds) Plant metabolomics. Springer-Verlag, Berlin

9. Schripsema J (2010) Application of NMR in plant metabolomics: techniques, problems and prospects. Phytochem Anal 21:14–21

10. Wishart DS (2008) Quantitative metabolomics using NMR. Trends Anal Chem 27:228-237

11. Smolinska A, Blanchet L, Buydens L, Wijmenga SS (2012) NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. Anal Chim Acta, doi:10.1016/j.aca.2012.05.049 in press

12. Tulpan D, Léger S, Belliveau L, Culf A, Čuperlović-Culf M (2011) MetaboHunter: an automatic approach for identification of metabolites from ^{1}H-NMR spectra of complex mixtures. BMC Bioinformatics 12:400

13. Mercier P, Lewis M, Chang D, Baker D, Wishart DS (2011) Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra. J Biomol NMR 49:307-323

14. Vigneau E, Sahmer K, Qannariand EM, Bertrand D (2005) Clustering of variables to analyze spectral data. J Chemom 19:122-128

15. Donoho DL and Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage, J. Amer. Statist. Assoc, 90:1200 1224

16. Golotvina S, Williams A (2000) Improved baseline recognition and modeling of FT NMR spectra. J Magn Reson 146:122-125

17. Pauli GF, Gödecke T, Jaki BU, Lankin DC (2012) Quantitative 1H NMR. Development and potential of an analytical method: an update. J Nat Prod 75:834-851.

18. de Meyer T, Sinnaeve D, van Gasse B, Tsiporkova E, Rietzschel E, de Buyzere M, Gillebert T, Bekaert S, Martins J, van Criekinge W (2008) NMR-based characterization of metabolic

alterations in hypertension using an Adaptive, Intelligent Binning Algorithm. Anal Chem 80: 3783-3790

19. Davis R, Charlton A, Godward J, Jones S, Harrison M, Wilson J (2006) Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform. Chemom Intell Lab Syst 85:144-154

20. Anderson P, Mahle D, Doom T, Reo N, del Raso N and Raymer M (2010) Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data. Metabolomics 7:179-190

21. Savorani F, Tomasi G, Engelsen SB (2010) icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, J Magn Reson 202:190–202

22. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S *et al*. (2007) HMDB: the human metabolome database. Nucleic Acids Res 35, D521–D526

23. Chenomx NMR Suite (2010) Chenomx Inc. Edmonton, AB, Canada. http://www.chenomx.com

24. Mounet F, Lemaire-Chamley M, Maucourt M, Cabasson C, Giraudel JL, Deborde C, Lessire R, Gallusci P, Bertrand A, Gaudillère M, Rothan C, Rolin R, Moing A (2007) Quantitative metabolic profiles of tomato flesh and seeds during fruit development: complementary analysis with ANN and PCA. Metabolomics 3: 273-288

25. Dumazet HF, Gil L, Deborde C, Moing M, Bernillon S, Rolin R, Nikolski M, Daruvar A, Jacob D (2011) MeRy-B: a web knowledgebase for the storage, visualization, analysis and annotation of plant NMR metabolomic profiles. BMC Plant Biol 11:104

26. Savitzky A and Golay JME (1964) Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Analytical Chemistry 36 (8):1627-1639

27. McLafferty F, Hertel R and Villwock R (1974), Probability based matching of mass spectra. Rapid identification of specific compounds in mixtures. Org. Mass Spectrom., 9: 690–702.

28. Astle W, de Lorio M, Richardson S, Stephens D, Ebbels T (2012) A Bayesian Model of NMR spectra for the deconvolution and quantification of metabolites in complex biological mixtures. J Am Stat Asso *in press*

29. Hao J, Astle W , de Lorio M, Ebbels T (2012) BATMAN—an R package for the automated quantification of metabolites from NMR spectra using a Bayesian Model. Bioinformatics *in press*

30. Lewis I, Schommer S, Markley J (2009) rNMR: open source software for identifying and quantifying metabolites in NMR spectra, Magn. Reson. Chem. 47:123-126

31. Steinbeck C, Conesa P, Haug K, Mahendraker T, Williams M, Maguire E, Rocca-Serra P, Sansone S-A, Salek R, Griffin J-L, (2012) MetaboLights: towards a new COSMOS of metabolomics data management, Metabolomics, 8(5): 757-760.

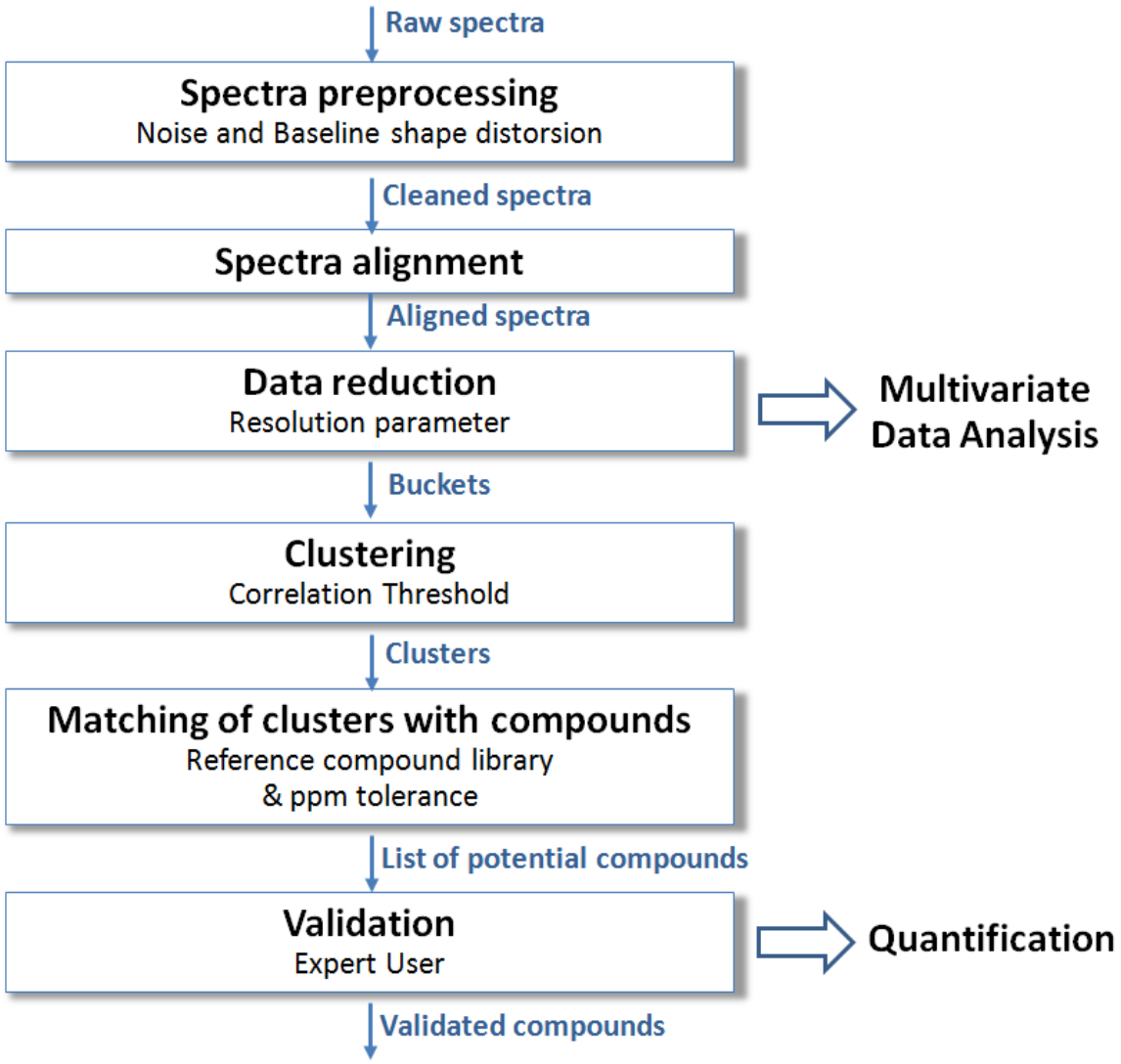

Figure 1 Overview of the different steps and outputs of the ^{1}H-NMR spectra processing method consisting in (i) spectra preprocessing and alignment, (ii) a new data reduction method called ERVA (Extraction of Relevant Variables for Analysis), followed by (iii) bucket clustering, and finally, (iv) cluster matching with compounds from a suitable library of reference spectra, in order to propose a list of compounds to be validated.

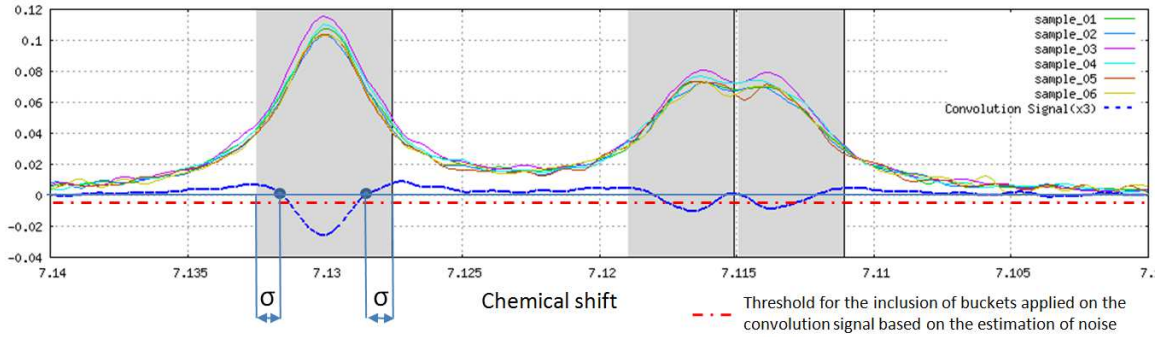

Figure 2 The data reduction algorithm is based on the sum of all NMR spectra included in the experiment set (6 spectra in the present figure). Then, a new signal (in blue) is computed which results from convolution between the spectra sum and a second derivative of a Lorentzian. The zero crossings of the resulting signal, extended each side by the value of σ (the full width at half maximum of Lorentzian function) give the bounds of the buckets. To take account of the presence of noise, the minimum value of the convolution signal (C_σ) in a ppm range considered as noise (typically from 10 to 11 ppm) can be taken as a threshold above which no bucket will be considered. Grey areas represent bucket regions. They do not cover the entire ppm scale because of elimination of non-significant regions.

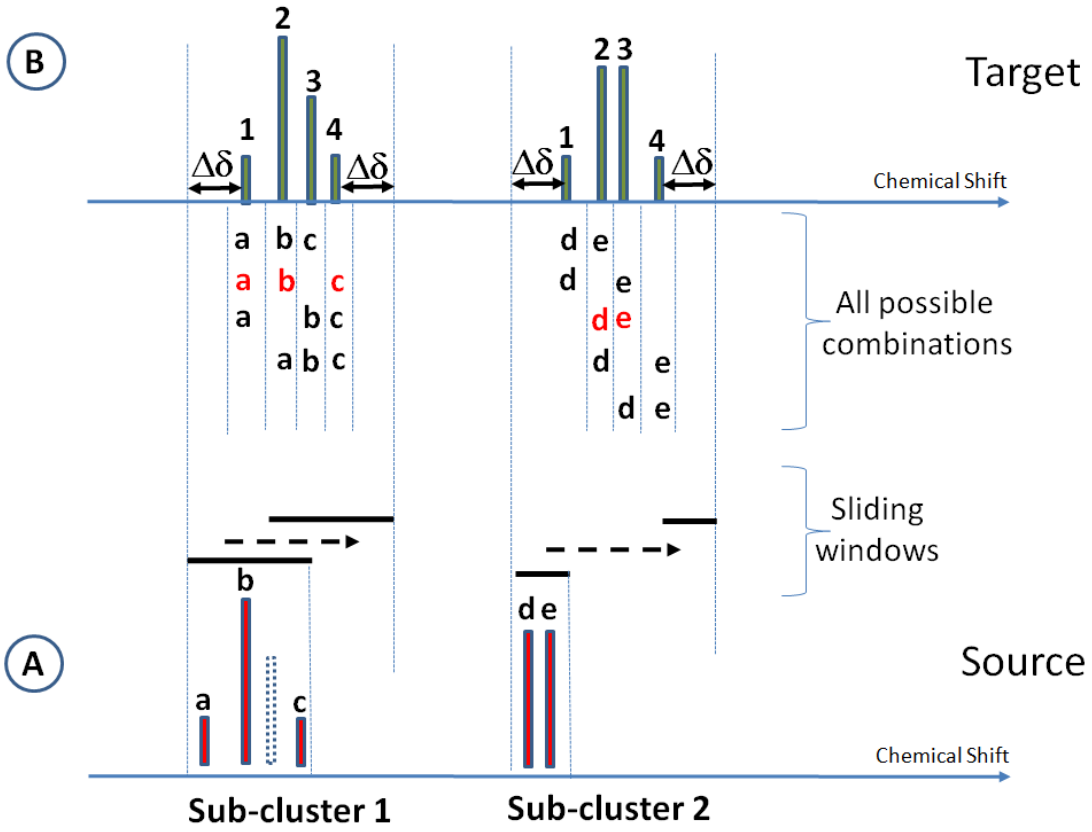

Figure 3 Getting the closest match between a cluster (Source) and a compound (Target) during the matching step. The source is divided into sub-clusters based on distances between peaks, with the assumption that distances within a sub-cluster are almost constant and lower than those between sub-clusters (A). Then, for each sub-cluster within the source, a window on the ppm scale is defined based on its first and last peaks (A). For the Target, the corresponding sub-cluster, if present, has a number of peaks greater or equal to that of the Source. Then, a window on the ppm scale is defined based on its first and last peaks, extended with Δδ ppm on each side (B). By sliding the Source window over the Target window of a peak position in each loop, (B) a score is computed for all possible combinations having at most one gap among the Target peaks (allowing a less correlated peak to miss in the Source as shown in sub-cluster 1), based on the concept of "valid cluster" [13]. Finally, the best score is retained (red) and a global score is computed.

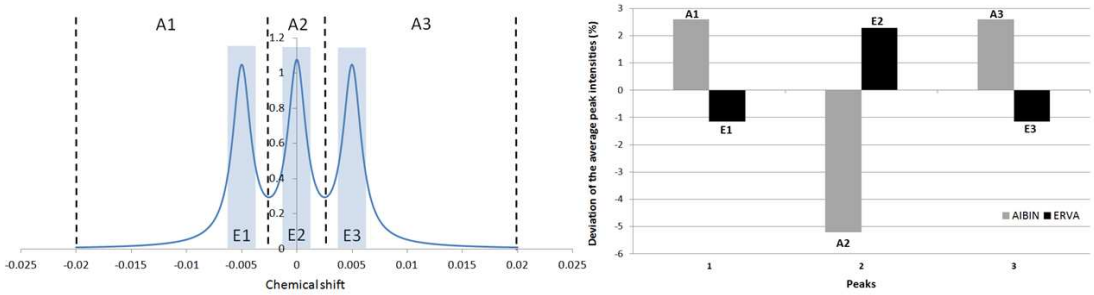

Figure 4 Comparison of buckets produced by ERVA and AIBIN binning methods based on the sum of three identical Lorentzians (FWHM=0.001 ppm) but shifted with a 0.005 ppm interval, so that they slightly overlap. On the left: the bins produce by the AIBIN binning method (A1,A2,A3) is delimited by the dotted lines, whereas those produce by ERVA binning method are shown by superposed shaded boxes (E1,E2,E3). First, integrations of ERVA's buckets provide values closer together than those obtained for AIBIN's buckets. Second, centres of buckets correspond to the centres of resonance peaks with the ERVA method unlike the AIBIN method. On the right: comparison of the percentage deviation from the average peak intensities for each buckets produced by both binning methods.

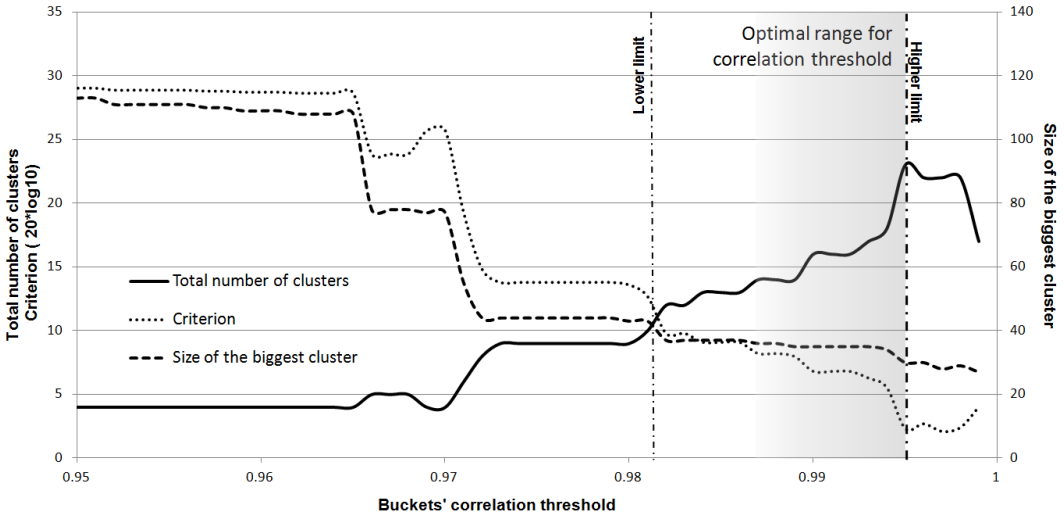

Figure 5 Effect of the buckets' correlation threshold on the size and number of bucket clusters obtained using the set of simulated NMR spectra (N=32K, SNR=70 dB, 220 buckets). The correlation threshold allowing a maximal discrimination of compounds (higher limit) is one that minimizes the ratio between a) the size of the biggest cluster and b) the total number of clusters (see Criterion). By decreasing the threshold, new clusters less correlated may appear while several among the most highly correlated will agglomerate but the total number of clusters will decrease. A reasonable lower limit for the correlation threshold is obtained when the size of the biggest cluster exceeds 40.

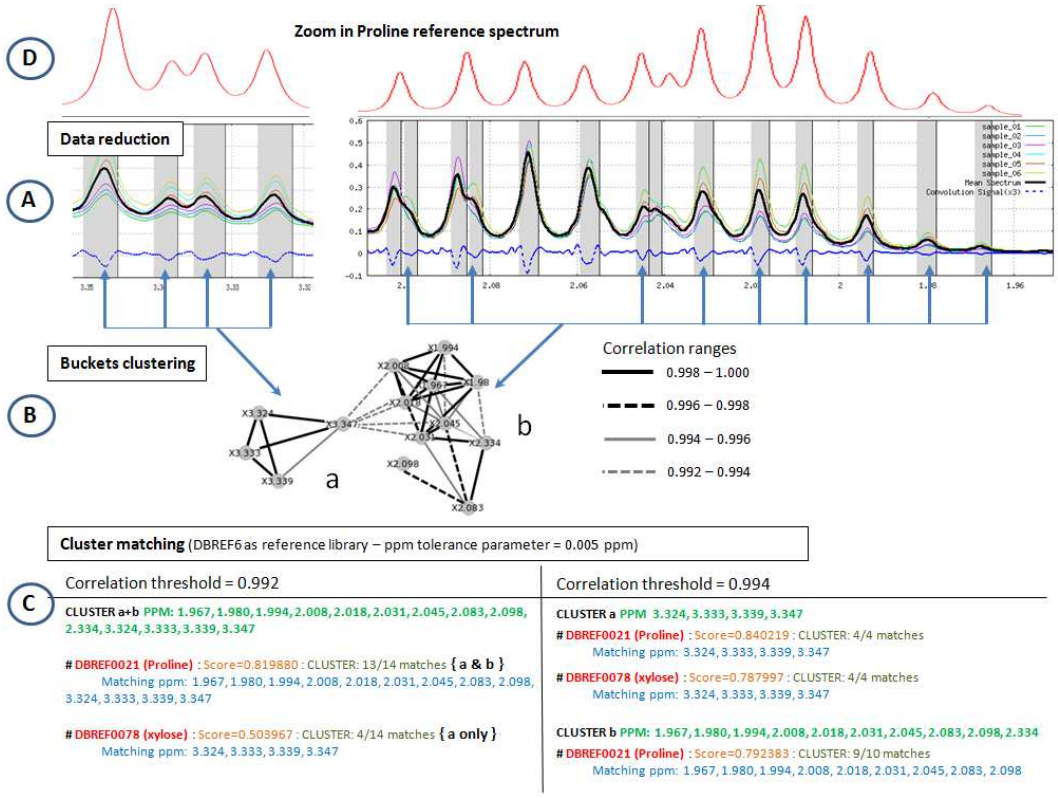

Figure 6 Illustration of the processing approach from spectra bucketing to the proposal of candidate compounds, using a set of six simulated NMR spectra (32K, SNR = 60 dB). First, the ERVA method of data reduction is applied to the spectra after noise processing, generating buckets as shown for two spectra regions in A. Second, the correlation matrix between bucket intensities is computed and a correlation threshold is applied for bucket clustering (B). The cluster containing 14 buckets shown in B includes 13 buckets shown in A. This cluster gathers two sub-clusters (a and b), each being intra-connected with higher correlations ($r>0.996$) than the interconnections ($r<0.994$). Third, matching of the cluster with compounds from DBREF6 database provides a list of candidate compounds (C) for two chosen correlation thresholds. Last, for validation, the reference spectrum of proline (D) is overlaid above the simulated spectra regions.

# Compound	Abrev.	Name	Weighting coefficient for:		Fold Change between groups (%)	Peak number
			Group 1	Group 2		
1	Ala	alanine	0.8	0.3	-62.50	6
2	Asn	asparagine	6.7	7.4	10.45	12
3	Asp	aspartic acid	1.6	2.1	31.25	12
4	-	chlorogenic acid	2.1	1.6	-23.81	41
5	CitA	citric acid	10	12	20.00	5
6	Fru	fructose	1000	800	-20.00	39
7	GABA	GABA	1	1.2	20.00	11
8	Gluc	glucose	500	300	-40.00	44
9	Glu	glutamic acid	2.8	2.3	-17.86	29
10	Gln	glutamine	8.9	5.6	-37.08	17
11	Ile	isoleucine	0.1	0.15	50.00	28
12	MalA	malic acid	0.3	0.6	100.00	12
13	-	methylnicotinate	34	17	-50.00	8
14	Phe	phenylalanine	0.2	0.3	50.00	21
15	Pro	proline	3	6	100.00	38
16	Raf	raffinose	10	5	-50.00	35
17	Suc	sucrose	20	30	50.00	32
					Total of peak number	**390**

Table 1 Composition of the simulated ^{1}H-NMR spectra used for evaluating the robustness of different steps of our approach (data reduction, clustering and matching). 17 compounds were chosen with their relative concentrations corresponding to a metabolomics profile of tomato fruit. Two groups were formed (group 1 and 2) simulating biological response under the effect of a factor. A weighting coefficient was applied to the ^{1}H-NMR reference spectrum of the corresponding compound after normalization (total intensity equal to 1). The "Fold change" column gives the relative change between groups (Group 2 versus Group 1). The last column gives the number of resonance peaks constituting the pattern of the corresponding compound NMR spectrum. Thus, the total peak number is 319 per spectrum of mixture.

Reference compound library	Correlation Threshold	ppm tolerance parameter (Δδ)						Total of compounds
		0.005	0.01	0.02	0.03	0.04	0.05	
DBREF6	0.99 (21)	14	15	15	14	14	-	17 (0.01)
	0.994 (23)	15	16	14	14	13	-	
HMDB	0.99 (21)	-	4	9	11	11	9	14 (0.03)
	0.994 (23)	-	5	11	14	12	10	

Table 2 Effect of the ppm tolerance parameter (Δδ) in the matching process with compounds using two reference compound libraries (HMDB and DBREF6) for 2 different values of correlation thresholds (with the corresponding number of clusters in brackets) and applying the test on a set of simulated NMR spectra (32K, SNR = 70 dB). Values shown in the table correspond to the number of true-positives (out of the 17 reference compounds included within the simulated set) matched with the highest score for at least one cluster. A same number does not correspond necessarily to the same compounds. The last column shows the total number of reference compounds out of 17 found by combining the two threshold values of correlation, and for the best value of the ppm tolerance parameter (in parenthesis).

Metabolites	δppm (Multiplicity)	Clustering step		DBREF6			
		Corr. Threshold	BINS (ppm)	Rank (1)	Scores (2)	Ratio1 (3)	Ratio2 (4)
Alanine	1.48 (d)	0.96	1.479, 1.494	1	0.8295	2/2	6/6
Asparagine	2.92 (m)	0.96	2.893, 2.908, 2.941, 2.950	1	0.8557	4/4	12/12
Aspartate	2.81 (m)	**0.95**	2.804, 2.838	1	0.8312	2/2	12/12
Chlorogenic acid	7.68 (d)	0.96	5.304, 5.315, 5.325, 5.337, 5.348, 5.360, 6.399, 6.432, 6.960, 6.983, 7.132, 7.154, 7.656, 7.688	1	0.9031	14/15	37/41
Citrate	2.63 (dd)	0.96	2.555, 2.585	1	0.8315	2/2	5/5
Fructose	4.12 (m)	0.96	3.704, 3.724, 3.737, 3.893, 3.901, 3.915, 3.921, 4.007, 4.022, 4.048, 4.116, 4.124	1	0.8865	12/12	26/39
GABA	3.01 (t)	0.96	1.894, 1.909, 1.924, 2.296, 2.311, 2.325, 3.002, 3.018, 3.032	1	0.9330	9/9	11/11
Galactose	4.60 (d)	-		-	-	-	-
Glucose	5.25 (d) ; 4.66 (d)	0.96	3.394, 3.404, 3.413, 3.423, 3.431, 3.473, 3.486, 5.251, 5.259	1	0.7512	9/11	22/44
Glutamate	2.07 (m)	0.96	2.164, 2.172	3	0.7970	2/2	19/29
Glutamine	2.45 (m)	0.96	2.144, 2.151, 2.158, 2.442, 2.455, 2.470, 2.484	1	0.8958	7/7	17/17
Isoleucine	1.01 (d)	0.96	1.286, 1.298, 1.308, 1.322	1	0.8462	4/4	21/28
Lactic acid	1.33 (d)	0.96	1.352, 1.366	1	0.8141	2/2	6/6
Leucine	0.96 (t)	0.96	1.727, 1.741	3	0.7630	2/2	7/20
Malate	4.30 (dd)	0.96	4.293, 4.299, 4.312, 4.319	1	0.8376	4/5	12/12
Mannose	5.2 (d)	-		-	-	-	-
Phenylalanine	7.40 (m)	0.96	7.391, 7.416, 7.430	1	0.8603	2/2	21/21
Proline	1.98 (m)	0.96	1.975, 1.982, 1.990	1	0.8509	3/3	38/38
Pyroglutamic acid	4.18 (dd)	0.96	2.508, 2.517	1	0.8261	2/2	24/24
Raffinose	5.0 (d)	0.96	5.011, 5.019	1	0.7512	2/2	9/35
Sucrose	5.41 (d)	0.96	4.215, 4.232, 5.415, 5.424	1	0.8362	4/4	12/32
Threonine	1.33 (d)	0.96	4.253, 4.258, 4.266	3	0.8243	3/3	10/10
Trigonelline	9.13 (s)	0.96	8.083, 8.098, 8.112, 8.837, 8.852	1	0.7968	5/6	5/7
Tyrosine	6.91 (d)	0.96	3.072, 3.107	1	0.8261	2/2	16/16
UDP-glucose	5.61 (m)	0.96	5.974, 5.985, 5.991, 7.948, 7.964	1	0.8535	5/5	44/45
Valine	1.00 (d) ; 1.04 (d)	0.96	2.265, 2.274	2	0.7618	2/2	6/17
Xylose	4.59 (d)	0.96	3.323, 3.336, 3.345	1	0.8603	3/3	34/34

(1) Ranking based on the value of the overall matching score for the corresponding cluster.
(2) Overall matching scores (see eq. 6)
(3) Ratio of matched peaks: Cluster vs. Compound
(4) Ratio of matched peaks: Compound vs. mixed spectrum

Table 3 List of true-positive compounds generated by the matching step for the tomato ^{1}H-NMR data set, with 0.03 as ppm tolerance value, using the DBREF6 compound library. The rank information gives the position in the ranking of candidates for this overall matching score. The highest matching score written at the first rank in the candidate compound list gives the most probable compound. In contrast, when the rank value is greater than one it means that one or more false positives were matched with a higher score; when no score is given, it means that reference compound was not matched with a significant score.