DISCUSSION PAPER

2017/22

PepsNMR for the ^1H -NMR metabolomic data
pre-processing

Martin, M., Legat, B., Leenders, J., Vanwinsberghe, J., Rousseau,
R., Boulanger, B., Eilers, P., De Tullio, P. and B. Govaerts

PepsNMR for the ¹H-NMR metabolomic data pre-processing

Manon Martin^{1*}, Benoît Legat², Justine Leenders³, Julien Vanwinsberghe⁴, Réjane Rousseau¹, Bruno Boulanger⁵, Paul H.C. Eilers⁶, Pascal De Tullio³, Bernadette Govaerts¹

Abstract

In the analysis of complex biological samples, control over experimental design and data acquisition procedures cannot ensure alone well-conditioned ¹H-NMR spectra with maximal information recovery for data analysis. A third major element affects the accuracy and robustness of the results: the data pre-processing/pre-treatment for which not enough attention is usually devoted, in particular in the metabolomic studies. The usual approach is to use proprietary software provided by the analytical instruments' manufacturers to conduct the whole pre-processing strategies. This widespread practice has a number of advantages such as a user-friendly interface with graphical facilities, yet it involve non-negligible drawbacks: a lack of methodological information and automation, a dependency of subjective human choices, only basic treatment possibilities and an absence of objective quality criteria to evaluate the pre-processing quality. This paper introduces PepsNMR to meet the needs, a R package dedicated to the whole processing chain prior to multivariate data analysis including, among other tools, solvent signal suppression, internal calibration, phase, baseline and misalignment corrections, bucketing and normalisation. Methodological aspects are discussed and the library is compared to the gold standard procedure with two metabolomic case studies. The use of PepsNMR on these data shows better information recovery and predictive power based on objective and quantitative quality criteria. Other key assets of the library are workflow processing speed, reproducibility, reporting and flexibility, graphical outputs and documented routines.

¹ *Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA/IMMAQ), Université catholique de Louvain (UCL), Belgium*

² *Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Université catholique de Louvain (UCL), Belgium*

³ *Laboratoire de Chimie Pharmaceutique, Université de Liège (ULg), Belgium*

⁴ *ClinBAY SPRL, Belgium*

⁵ *Arlenda, Belgium*

⁶ *Department of Biostatistics, Erasmus University Medical Centre, Rotterdam, The Netherlands*

*Corresponding author: manon.martin@uclouvain.be

Contents

1	Introduction and literature review	1
2	Methodology	2
2.1	PepsNMR Workflow presentation	2
2.2	Details on the pre-processing steps	3
	Raw signal • Phase errors and Group Delay suppression • Solvent signal suppression • Apodization • Fourier transform and chemical shift scale • Referencing with an internal standard • Zero order phase correction • Baseline correction • Negative values zeroing • Warping • Spectral window selection • Bucketing • Region removal • Region aggregation • Normalization	
2.3	Other PepsNMR features and functions	7
3	Quality criteria for PepsNMR evaluation	8
4	Results and discussion	8
4.1	Variance structure	9
4.2	Clustering and classification performances	9

5	Conclusions	10
6	Software availability	10
7	Datasets	10
7.1	Human Serum	10
7.2	Human Urine	10
7.3	Spectral pre-processing and outlier removal	10
	Acknowledgments	11

1. Introduction and literature review

Among the Omics sciences, the field of untargeted metabolomics holds a significant position for searching biomarkers in complex biological systems. This approach studies the response of an organism to an external stimuli, to determine the link between the experimental conditions and phenotypic changes of interest[19]. Its main advantage is to provide a metabolic snapshot of an organism, a tissue or cells, whereas its main

challenge is related to the treatment of large datasets that need powerful signal processing and statistical tools[1].

The more common analytical techniques used to generate metabolomic data are Gas Chromatography-Mass Spectrometry (GC-MS), Liquid Chromatography-MS and proton Nuclear Magnetic Resonance (^1H -NMR) Spectroscopy. They work complementarily, each having its own interest and limitations. Advantages of ^1H -NMR spectroscopy are that it requires low cost per sample and minimal preparation, it is generally untargeted thus unbiased, non-destructive, non-invasive, reproducible and fast [19, 28, 14].

Analytical and experimental variations that are unrelated to the biological perturbation under study deserve particular attention in metabolomic studies. They originate from sample collection, preparation and conservation, but also from data acquisition, and from intra-individual (*e.g.* nutrients) or inter-individual (*e.g.* genetic) variations[23], especially in human studies[20]. They can lead to inconclusive results or false biomarker discoveries in multivariate data analysis when incomplete or inappropriate methods in the foregoing steps are applied. Pre-processing, here used in its chemometric definition as all the data editing done before the data analysis[21], is used to enhance the Signal-to-Noise Ratio (SNR) but also to correct or minimize for instrumental artefacts and appropriately transform the data as interpretable spectral profiles and linked to the biological system under study.

Good data pre-processing is important. Commercial software such as TopspinTM, Chenomx or MestReNova remain the most used solutions for NMR data pre-processing but they do not provide advanced editing steps and require iterative, manual and arbitrary adjustments. These interventions will thus not necessarily improve the data quality and can impact the performances of multivariate data analysis built thereafter. On the other hand, free software that are mostly open-source exist for ^1H -NMR data processing (see Alonso *et al.*[1] for a review on the more important ones). Yet, their pre-processing tools remain basic and do not offer a global strategy dedicated to the binning-based chemometric approach for further data analysis[11]. Besides, the multiplicity of pre-processing scenarios and methods [3] that may arbitrarily affect the final conclusions should benefit from clear recommendations to guide the user's choice of when and how methods should be applied.

We introduce the R package PepsNMR (or Packaged Extensive Preprocessing Strategy for NMR data), for the (semi-automatic) pre-processing of 1D ^1H -NMR Free Induction Decay (FID) datasets. The objective of this software is to provide an exhaustive and flexible workflow to deal with typical features of raw ^1H -NMR data and cover the pre-processing (clean the raw instrumental data) and pre-treatment (transform the cleaned data for data processing) steps as defined by Goodacre *et al.*[16]. The advised pre-processing strategy transforms FID data through several steps into a binned spectral table that will enable the analysis of data with classic chemometric tools[25] to identify global trends in spectral

profiles.

Innovative and acknowledged methods are combined with well known algorithms in order to correct for instrumental artefacts and irrelevant biological variability, to increase the SNR, transform appropriately the data domain/scale and reduce its dimensionality.

Section 2 presents the pre-processing methodology: its motivation, the description of each step, and the package features. Section 3 briefly describes the quantitative criteria used to assess the quality of the pre-processed data. Then Section 4 illustrates and discusses the use of PepsNMR on the basis of two case studies. Thereafter, the conclusion summarises and brings forward key points of the methodology (Section 5) and finally, Sections 6 and 7 describe the software implementation and the datasets used in Section 4.

2. Methodology

2.1 PepsNMR Workflow presentation

PepsNMR originates from an initially unpublished Matlab program called “Bubble” and developed by Eli Lilly and one of the authors (Paul H.C. Eilers)[26, 30]. The library works as an open system with full flexibility: it provides a complete series of pre-processing functions to be run sequentially to transform the FIDs into the final spectra but each function could also be applied independently. Raw data in Bruker format can be imported directly, different methods and settings are available in most pre-processing steps and visualisation functions are available to monitor the process. The complete sequence of R functions is provided below in the advised order.

GroupDelayCorrection() The function corrects the frequency dependent — or first order — phase error due to the Group Delay induced by the Bruker's digital filter.

SolventSuppression() The variability of solvent residuals in the spectrum (especially for water) can mask informative signals from molecules of interest. This step removes the solvent signal (usually water) in the time domain using a Whittaker smoother.

Apodization() The apodization aims to increase the SNR of the spectrum (mainly) and/or its resolution. It multiplies the FID signal by a positive signal, often decaying. Several apodization functions are available in PepsNMR.

FourierTransform() The Fourier transform translates the complex FID signal in the time domain into a complex spectrum expressed in frequency (Hz). This function further calibrates the frequency scale and converts it into chemical shifts (ppm).

InternalReferencing() An internal compound (*e.g.* TMSP) is used for the chemical shift scale referencing.

ZeroOrderPhaseCorrection() In order to remove the frequency-independent phase shift, the spectrum is rotated to maximise the positiveness of its real part.

BaselineCorrection() To remove the spectral baseline, the

function uses asymmetric least squares with a roughness penalty.

NegativeValuesZeroing() Imperfect phase or baseline correction can lead to small residual negative values in the spectrum that cannot be interpreted. These negative values are thus set to zero.

Warping() The first global alignment solution provided in PepsNMR is a warping algorithm. It applies a warping function to spectra according to a reference spectrum. This warping function is a combination of polynomials and P-Splines which stretches/compresses the chemical shift axis to increase the similarity between the transformed spectrum and the reference spectrum.

WindowSelection() The function selects the informative part of the spectrum to be kept for the data analysis.

Bucketing() The second solution to minimise peaks misalignment across spectra and proceed to data reduction is soft bucketing. The function uses trapezoidal or rectangular interpolation to deliver equally-spaced bins with a specified length or number.

RegionRemoval() The function allows to set to zero specific spectral regions of no interest in the study.

ZoneAggregation() It is useful to apply a targeted data reduction for molecules with an unstable chemical shift (e.g. citrate). The routine aggregates a pre-defined interval of descriptors into a single symmetrical peak.

Normalisation() The function normalises intensities in the spectrum to avoid unwanted inter-sample variation.

2.2 Details on the pre-processing steps

This section describes in a nutshell the methodological features of the pre-processing steps. For each R function, underlying mathematical and statistical concepts along with their algorithmic implementation are explained and discussed.

2.2.1 Raw signal

Each FID (Figure 2) recorded with the ¹H-NMR instrument is a complex decaying signal $S = S_x + iS_y$ of m data points (where S_x and S_y are respectively the real and imaginary parts of the signal). It represents the net sum of distinct signal components from different proton nuclei. The theoretical form of each single signal can be written as a function of its offset ν , amplitude s_0 and transverse relaxation time T with an exponential decay over time t : $s_0 \exp(i2\pi\nu t) \exp(-\frac{t}{T})$. The function `ReadFids()` imports raw Bruker files into two matrices: the complex FID signal matrix and the Bruker acquisition parameters needed for the pre-processing steps.

2.2.2 Phase errors and Group Delay suppression

For a perfectly phased signal, its real part starts at a maximum value and its imaginary part starts at 0. The corresponding spectrum has a positive real part in absorptive mode and an imaginary part in dispersive mode with both positive and negative peaks. Due to technical reasons (Bruker digital filter, incorrect magnetisation, etc.) the real and imaginary parts of the signal result in a mixture of both absorptive and dispersive

mode line shapes[18]. They need to be phase corrected. The total phase error Φ is usually written as a sum of frequency-independent (ϕ_0) and frequency-dependent ($\phi_1(\nu)$) phase errors: $\Phi = \phi_0 + \phi_1(\nu)$ that are corrected respectively with the zero and first order phase corrections. The initial non-phased signal is: $S = S_{\text{phased}} \exp(i\Phi)$.

This first pre-processing step aims to partly correct the first order phase error introduced by the Bruker digital filter. Indeed this filter induces a pre-acquisition delay: the real part of the FID does not have a maximum value at $t = 0$. This delay, usually of tens of data points, is illustrated in Figure 1.

In PepsNMR, the delay is removed from the FID based on properties of the Discrete Fourier Transform (DFT)[27]: a circular shift on the time domain signal by $\tau > 0$ samples, called the Group Delay, is identical to multiplying the corresponding spectrum F by the linear term $\exp(-i2\pi\nu\tau)$ i.e. apply a frequency-dependent linear phase shift in the frequency domain. The signal is first Fourier transformed to be multiplied by the phase shift, then back-transformed with the inverse Fourier transform. This double transformation is mandatory in the common case of a non-integer Group Delay. The function `GroupDelayCorrection()` removes the first order phase shift with τ recovered from the spectrometer acquisition parameters and returns the processed FIDs.

Figure 1. The Group Delay recorded before the signal acquisition has to be removed from the FID.

2.2.3 Solvent signal suppression

Water is the most common solvent for biofluid samples. It has a rather broad resonance and its chemical shift is varying according to solution conditions. Usual solvent suppression techniques aim to reduce the magnitude of the solvent resonance before the NMR signal reaches the receiver. Three different approaches are[4]: pre-saturation of the water resonance, production of zero net excitation of the water resonance and destruction of the water resonance with pulsed field gradients. Nevertheless, residual peaks of water are still measured and their intensity and instability can mask other interesting compounds. Moreover and based on our experience, solvent suppression improves the efficiency of subsequent correction steps such as the baseline correction.

PepsNMR solves this common issue with the non-parametric modelling and subtraction of the residual water signal directly from the FID. Under the assumption that water is the main compound of the analysed samples, a discrete penalised least

Figure 2. The solvent residuals signal is first estimated with a smoothing function and then subtracted from the FID.

squares with second-order differences presented by Eilers[7] and based on the Witthaker’s graduation method [32] is used to model separately the real and imaginary parts of the water signal.

This method estimates the smoothed series W that minimises the Q criterion:

$$Q = \sum_{j=1}^m (S_j - W_j)^2 + \lambda \sum_{j=3}^m (\Delta^2 W_j)^2 \quad (1)$$

This expression is a balanced combination of two conflicting objectives, *i.e.* fidelity to the data and smoothness. The first term is the sum of squares of differences measuring the lack of fit of W to the real or imaginary signal S on the m data points, the second term penalizes the series roughness with λ the penalty parameter multiplied by the sum of squared second order differences (defined as $\Delta^2 W_j = W_j - 2W_{j-1} + W_{j-2}$) quantifying the roughness. The analytical solution is provided by Eilers[7] and the use of sparse matrices greatly reduces the computational time for large series such as ^1H -NMR signals. The function `SolventSuppression(Fid_data, lambda.ss = 1e6, ...)` operates the estimation of the water signal for both the real and the imaginary parts of the FID and returns the processed FIDs and the real and imaginary water signals. `lambda.ss` accounts for the parameter λ and its default value (1e6) is a practical choice, based on empirical results.

2.2.4 Apodization

The apodization multiplies the FID by a signal that changes its shape and emphasises distinct spectral portions[4] to enhance the sensitivity and/or the resolution of the spectrum. Due to its nature, the NMR signal intensity decays over time whereas the noise has a random fluctuation with a constant amplitude. Hence, the SNR is higher at the beginning of the FID and declines over time. An increase in sensitivity (higher SNR) can thus be achieved by multiplying the FID by a decaying signal. However, a faster decay of the FID (already dependent on the decay time T) implies broader line widths of the spectral peaks, lowering their height and thus their resolution since the area under the peaks remains unchanged. Therefore, optimal trade-off parameters for the apodization signal are needed to

combine both these objectives. A classic apodization step applies an exponentially decaying signal principally to enhance the SNR but other apodization factors can be used to enhance both sensitivity and resolution with a concave multiplying signal.

Several types of apodization functions are provided by PepsNMR with the option `type.apod` in `Apodization(Fid_data, type.apod = "exp", ...)`: decreasing exponential or gaussian functions, Hanning window, etc.. By default PepsNMR multiplies the FID by a decaying exponential function: $h(t) = \exp(-t/a)$ with a the decay parameter.

2.2.5 Fourier transform and chemical shift scale

At this stage, the recorded signal in the time domain S is ready to be converted into a spectrum in the frequency domain F . The Fourier transform (FT) provides a general solution to transform the complex signal sequence into a complex spectral sequence. It extracts each signal and translates it into peaks in a spectrum with specific height, position and width that depends on the amplitude of the signal, the frequency of the corresponding compound and the relaxation time T [18].

PepsNMR achieves this transformation with the function `FourierTransform()` that applies the Discrete Fourier Transform (DFT) for uniformly spaced samples with the Fast Fourier Transform (FFT) algorithm.

The DFT can be written as:

$$F_j = \sum_{t=0}^{m-1} S_t \exp(-2i\pi jt/m) \quad (2)$$

for $j = 0, \dots, m-1$.

The scale of the F vector has still to be defined. PepsNMR first expresses it in Hertz on the basis of the spectrometer acquisition parameters with the following formula: $\nu_j = \nu_S + \Delta\nu \times (j - (m-1)/2)$ where ν_S is the transmitter frequency (called SFO1 in the Bruker acquisition file) located at the center of the spectral window and $\Delta\nu$ represents the difference of frequency between 2 data points and is calculated from the spectral window width (or sweep width). Note that ν_S is a sum of ν_B , the basic transmitter frequency (set automatically depending on the spectrometer and the observed ^1H nucleus) and ν_O , the transmitter frequency offset, tuned by the user to shift the center of the spectral window from ν_B to ν_S .

The scale in Hertz is then converted into a chemical shifts vector δ expressed in parts per million (ppm), a dimensionless scale that is unrelated to the magnetic field strength of the spectrometer[18]. A chemical shift represents the frequency ν_j relative to ν_B , imposing the 0 ppm to be located at the ν_B frequency. The conversion equation is: $\delta_j = 10^6 \times \left(\frac{\nu_j - \nu_B}{\nu_B} \right)$.

A spectral matrix is now created where each row corresponds to a spectral profile and each column represents a spectral intensity at a specific chemical shift. By convention, the chemical shift along the horizontal axis is set as decreasing towards the right.

2.2.6 Referencing with an internal standard

In order to have a more reliable scale, it can be referenced to a known standard, *i.e.* an internal reference compound whose chemical shifts are minimally influenced by external factors (*e.g.* temperature or concentration) and ideally located outside the spectral region to be clearly identifiable. The reference compound used in NMR is usually a silane derivative such as the Trimethylsilyl-propionate (TMSP) or the 4,4-Dimethyl-4-silapentane-1-sulfonic acid (DSS). The chemical shift of these standards peaks is referenced to 0 ppm by convention but other reference compounds may have a non-null ppm value.

In PepsNMR, the `InternalReferencing(RawSpect_data, shiftHandling = "cut", method = "max", range = "near0", ...)` function proposes two ways to locate the reference compound peak in each spectrum on a range of intensities (`range`) with the `method` argument: it selects either the maximum intensity or the first peak in the search range higher than a predefined threshold.

2.2.7 Zero order phase correction

For technical reasons explained above, the spectrum can still exhibit a zero order phase shift error of a certain angle ϕ_0 that can be expressed as $F = F_{\text{phased}} \exp(i\phi_0)$ where F_{phased} is the perfectly phased spectrum. This phase shift is constant for all signal vectors and then independent of the spectral frequencies.

Based on the principle that the real part of the perfect spectrum F_{phased} has a pure absorptive mode, an optimal angle ϕ_0 can be found automatically in PepsNMR by maximising an adequate criterion of positiveness of the real part of the spectrum.

PepsNMR proposes two positiveness criteria measured on a range R of descriptors: it can be quantified by the maximisation of the highest positive peak or by the *rms* criterion that represents the ratio between the sum of squares of positive intensities and the sum of squares of all intensities of the spectrum:

$$\phi_0 = \arg \max_{\phi \in [0, 2\pi]} \frac{\sum_{j \in R} (F_j^\phi)^2 \times \text{ind}_j}{\sum_{j \in R} (F_j^\phi)^2} \quad (3)$$

where $F^\phi = F \exp(-i\phi)$ is the real part of the rotated spectrum with a correction angle $-\phi$ and $\text{ind}_j = 1$ if $F_j^\phi > 0$ and 0 otherwise. The user can decide on which range R to maximise the positiveness criterion since particular areas that can be difficult to phase (especially the spectral water region) can be excluded. The function `ZeroOrderPhaseCorrection(RawSpect_data, type.zopc = "rms", ...)` rotates each spectrum of the spectral matrix and returns the processed spectra with their rotation angles. In addition a vector of angles can be directly provided by the user with the argument `type.zopc = "manual"`.

Note that residuals of first order phase shift can still be present. They are currently not removed by PepsNMR but are negligible after the Group Delay removal.

2.2.8 Baseline correction

Ideal spectra are expected to have a flat baseline but, even with the phase corrections, baseline artefacts can still appear. They are caused by multiple sources such as the presence of macromolecules, a not entirely linear electronic detection process or calibration errors for the 180° pulse [12]. It is critical to remove these artefacts by an appropriate methodology to enhance the warping efficiency and perform an accurate data analysis on meaningful and unaltered peak shapes. Several algorithms have been suggested to estimate the baseline to be removed from the spectra. They mainly involve: iterative penalized least squares methods, robust baseline estimation, iterative polynomial fitting, local means or wavelets transforms [22, 24].

The AsLS smoothing algorithm[9] in PepsNMR allows a flexible baseline estimation with fast and reproducible computations. It is a nonparametric penalised method that combines asymmetric weighting of deviations from the trend and penalty on non-smooth behaviour of the estimated series.

This generalisation of the Whittaker smoother provides a robust baseline estimate with no artefact at the end of the spectrum compared to a polynomial fit[21].

The baseline estimator Z of a signal F , is found by minimising the objective function (4) similar to Equation (1) where the first term represents the lack of fit to the data now measured by asymmetric least squares and the second term is the roughness penalty. The asymmetric weighting will favour positive corrected intensities as positive deviations are less weighted than negative ones.

$$Q = \sum_{j=1}^m \psi_j (F_j - Z_j)^2 + \lambda \sum_{j=3}^m (\Delta^2 Z_j)^2 \quad (4)$$

In Equation (4), ψ_j is a specific weight for the frequency j : if $F_j > Z_j$, $\psi_j = p$, else $\psi_j = (1 - p)$ where $p \in [0, 1]$ is close to 0. The function `BaselineCorrection(RawSpect_data, lambda.bc=1e7, p.bc=0.05, ...)` estimates iteratively the weights ψ_j and the smoother Z with a gradient algorithm[8] and subtracts the estimated baseline function from the spectrum: $F^* = F - Z$. An illustration of the process is given in Figure 3.

Since this baseline correction algorithm can create artificial positive intensities in the water region[26], a previous solvent suppression step is recommended and a subsequent step is dedicated to the removal of uninformative regions (*e.g.* water) to cope with this behaviour.

2.2.9 Negative values zeroing

Despite the application of baseline and phase corrections, the spectra may still have negative intensities at specific frequency values. This step simply sets all these negative values to zero since they cannot be properly interpreted.

In PepsNMR, this operation is done with the function `NegativeValuesZeroing()`.

Figure 3. The baseline is first estimated and then removed from the spectrum.

2.2.10 Warping

In metabolomic applications, due to the nature of biological samples and experimental conditions variations (*e.g.* pH, temperature or concentration), peak shifts or misalignment between identical features from different spectra, are observed. Peak alignment solutions commonly applied for NMR spectra in the literature are warping and bucketing (*i.e.* data reduction) and both methods are included in PepsNMR. The warping will indeed enhance the similarity between profiles by the way of shifts, stretches and/or compressions along their horizontal-axis. Recent reviews by Bloemberg *et al.*[2] and Vu and Laukens[31] report advantages and drawbacks of the most important warping procedures for signal alignment such as parametric time warping (PTW), semi-parametric time warping (SPTW), dynamic time warping (DTW), correlation optimised warping (COW), icoshift, CluPA, etc.

The warping method implemented in PepsNMR is inspired by different elements of time warping as developed by Eilers [8] and Van Nederkassel *et al.* [29]. It applies a warping function $w(v)$ to a normalised spectrum F , *i.e.* a distortion function of the ppm axis that combines a polynomial term and a penalised B-splines (P-splines) term. The warping function $w(v)$ is defined as:

$$w(v) = \sum_{k=0}^K \beta_k v^k + \sum_{l=1}^L \alpha_l B_l(v) \quad (5)$$

The first term is a polynomial of order K with β_k the corresponding polynomial coefficients. The second term is a weighted sum of B-splines, with L the number of B-splines and α_l the coefficient for the l^{th} B-spline $B_l(v)$. The B-spline curves are constructed from polynomial pieces and joined together smoothly. The normalised and now distorted spectral profile $F(w(v))$ is interpolated from the discrete warping function.

The warping principle consists of building $w(v)$ such that the distance between a warped spectrum $F(w(v))$ and the reference spectrum G is minimised. The parameters are found by the iterative minimisation of the following penalized least squares objective function:

$$Q = \sum_{j=1}^m (G_j - F_j(w(v)))^2 + \lambda \sum_{l=3}^L (\Delta^2 \alpha_l)^2 + \kappa \sum_{l=1}^L \alpha_l^2 \quad (6)$$

where the unknowns are β_k and α_l . Two penalty terms are applied to the B-splines coefficients[10]. The first one is based on second order differences $\Delta^2 \alpha_l$ and tunes the B-splines smoothness with parameter λ whereas the second one is a Ridge penalty applied to regularise the coefficients (*i.e.* control their variance) with parameter κ . The function `Warping()` warps the spectra with the following default parameter values : $K = 3$, $L = 40$, $\lambda = 0.01$ and $\kappa = 0.0001$. An illustration of warped spectra is given in Figure 4.

A mandatory prior normalisation is applied on each spectra that are de-normalised after warping. The concept of normalisation will be explained in a latter step. The choice of a reference spectrum among all profiles is not straightforward. With no prior knowledge and no pooled sample, a robust reference selection process, available in PepsNMR, selects the spectrum that minimises the sum of squared distances with all other spectra before or after warping. Note that when the samples come from different classes, using one reference sample per class can be misleading and is likely to induce an artificial between-class variability[2].

Figure 4. Spectra are horizontally warped to remove peak shifts and allow to accurately compare intensity differences.

2.2.11 Spectral window selection

In order to focus the data analysis on particularly informative spectral areas, the function `WindowSelection()` trims the spectral window between right and left chemical shift bounds (by default: 0.2 ppm and 10 ppm), leaving the spectra with a lowered number of descriptors m .

2.2.12 Bucketing

The high dimensionality of the data and small residual peak shifts can impede the future multivariate data analyses[21]. Bucketing integrates by intervals the m original spectral intensities into m^b predefined intervals. Since the warping already corrects for chemical shift misalignments, a *soft* bucketing technique is sufficient to smooth and correct for small deviations left provided that shifts of a same peak are small enough to be included in the same interval. Also called data reduction, the bucketing therefore reduces the number of points usually from $m > 15000$ to $m^b < 1000$.

Figure 5. Illustration of the target area to be integrated and the two bucketing methods (B. and C.) in PepsNMR.

In this step, the optimal trade-off lies between keeping the spectral information and removing the peaks shifts as well as decreasing the total number of variables. Among possible binning methods, PepsNMR's bucketing proposes two integration options, either trapezoidal or rectangular, with equally sized buckets and generalised to cut the original x -axis at any chosen location. Figure 5 illustrates the two possibilities. The function `Bucketing(Spectrum_data, mb = 500, ...)` reduces the spectra to m^b buckets.

2.2.13 Region removal

Included by default in the selected spectral window, resonances without interest can eventually alter the data analysis by the introduction of unwanted variation. With biological samples, one major problem is the presence of water resonance residuals and the positive values created by the baseline correction step, even with a previous application of pre-saturation and solvent signal suppression techniques[28]. By default, the function `RegionRemoval()` removes urea, maleic acid and water (4.5 – 6.1 ppm) peaks for urine spectra whereas the water resonance area (4.5 – 5.1 ppm) is only removed by default for serum spectra.

2.2.14 Region aggregation

Strong variations of peak positions are particularly present for ionised metabolites such as citrate, taurine, succinate or lactate[2] and cannot be totally avoided either by specific actions during samples preparation (*e.g.* pH buffer), constrained warping methods or soft bucketing.

Targeted data reduction can then be applied to avoid this unwanted source of variation with the function `ZoneAggregation()` in PepsNMR. The peak is appropriately reshaped with the aggregation technique suggested in Rousseau[26]: the total intensity in a $[a, b]$ ppm interval $T = \sum_{j=a}^{j=b} F_j$ is distributed across this area for the metabolite to have a unique symmetrical triangular peak and null intensities at the border. A further version of the library will include the possibility to aggregate intensities into a Lorentzian curve.

2.2.15 Normalization

Normalisation is defined as a row operation [5] $F_i \times c_i$ where each spectrum i is multiplied by a constant term c_i . It addresses the dilution factor problem: variations in absolute concentrations can be observed between biofluid samples whereas data analysis focuses only on relative peak differences between them. This variation is especially true between urine samples that do not possess the homeostasis properties such as serum samples do, but other technical issues can also create this unwanted variation. Various methods of normalisation coexist in the literature (see for example Craig *et al.*[5] or Smolinska *et al.*[28]).

The default method implemented in PepsNMR is the constant sum normalisation (CSN) technique (function `Normalization(Spectrum_data, type.norm = "mean", ...)`). With this method, $c_i = \frac{m^b}{\sum_{j=1}^{m^b} F_{ij}}$. The mean intensity for each spectrum becomes m_b . The probabilistic quotient normalisation (PQN)[6] — first developed for urine spectra and distinguishing the effects of overall and local metabolites concentration changes on the spectrum — is another method included in PepsNMR. Other normalisation techniques available are: median, quartile or by the intensity an endogenous metabolite, usually the creatinine that has a constant excretion rate, whose concentration is then assumed to be constant across samples.

2.3 Other PepsNMR features and functions

For the sake of reproducibility, PepsNMR gives the possibility to modify, view and save spectra together with their acquisition parameters as well as methods parameters during or after the whole pre-processing workflow. These different features are explained below.

The `PreprocessingChain()` function A wrapper for data import and pre-processing steps. When working with this function, the pre-processed spectral intensities and acquisition parameters matrices can be exported (.RData; .txt or .csv) as well as a text file with all the pre-processing input parameters to ensure traceability.

Functions parameters For most pre-processing steps, different input parameters can be modified to adjust the methods for every dataset. Default values are automatically assigned to those parameters and can be modified according to the guidelines in the related *help* support in R. Parameters that are mentioned in this article need careful considerations to ensure an efficient pre-processing. They can be optimised by an appropriate experimental design and/or objective quality criteria as explained below in Section 3.

Functions order Some pre-processing blocks can be switched or bypassed with respect to the advised order while some corrections may be mandatory or highly recommended to precede other steps (*e.g.* solvent suppression before baseline correction or warping before bucketing) and several of them are compulsory (*e.g.* Fourier transform or phase corrections). As for their parameters, quality criteria can guide to decide on the appropriate functions order.

Visualisation tools After each step, data can quickly be visualised to check the effects of the applied pre-processing method(s) and detect unexpected or abnormal results that could mislead the data analysis, but also to spot outliers and global trends[21]. PepsNMR provides different ggplot2-themed plotting possibilities with the `Draw()` function to represent either the signal/spectral intensities or PCA scores and loadings. They can be directly displayed in the R interface or saved as png/pdf to be recovered later.

3. Quality criteria for PepsNMR evaluation

In the same perspective but beyond the scope of this article is the importance to rely on objective quality criteria to assess and quantify the pre-processing efficiency that leads to the improvement of informative content in the spectra. The development and wide use of such quality parameters, as named by Engel *et al.*[11], are still lacking and should be investigated more intensively across the scientific literature. Indeed, sub-optimal pre-processing strategies can deteriorate relevant spectral information and affect study conclusions. However there is currently no global consensus on which selection criteria other than visual inspection should be used to test for their efficiencies. In this paper, the evaluation of PepsNMR performances is done as suggested by Engel *et al.*[11]: the decision relies on a robust chemometric approach consisting of different quantitative quality parameters that aggregate into a super-parameter. This approach can either be used to compare the pre-processing efficiency between competing strategies or to calibrate appropriately the PepsNMR functions parameters and order.

This approach requires for every dataset under study a validation dataset to be collected to properly compare and calibrate the pre-processing strategies based on quantitative

quality criteria. These criteria are indeed measuring the spectral repeatability from a designed validation dataset with technical and/or biological replicates (*e.g.* subsamples from the same biological unit analysed at different times or involving different dilutions). The spectral repeatability is defined as the ratio between the repeatable, structured group "signal" from the replicates and i.i.d. noise under the hypothesis that the sampled spectra are composed of both structured signal and random noise. Note that this validation dataset should be similar but independent of the dataset of interest in the study.

In the next Section, quality criteria are measured for distinct pre-processing strategies that are applied on an identical validation dataset with known replicates class to rank their efficiency. These criteria are derived from unsupervised as well as supervised chemometric tools. They are gathered under the Metabolomic Informative Content (MIC) concept[13] and include inertia measures, PCA, clustering and PLS-DA related criteria. The inertia analysis decomposes the total variance into two complementary parts: the variance between the groups and the variance within the groups of observations. For the clustering, the (adjusted) Rand indexes measure the true class recovery efficiency and should be maximised while the Dunn and Davies-Bouldin indexes measure the clustering homogeneity and they have to be respectively maximised and minimised (formula details can be found in Féraud *et al.*[13]). PLS-DA builds a latent structure with Latent Variables (LV) to extract the variance components from the data but instead of maximising the spectral data matrix variance as does PCA, it maximises the variance of both the spectral data matrix (the regressors) and the class labels (the response) and their correlation. Its cross-validation criterion Q^2 assesses the model efficiency in class prediction.

The incentive to use unsupervised methods (PCA and K-means clustering) is to contrast the amount of class information that we are able to recover in the data after different pre-processing pipelines, provided that the between-groups variation is the main source of variability in the data. This information recovery is a blind process: no prior information is provided about the class labels. The supervised classification algorithm (PLS-DA) completes the performance analysis since it constitutes a standard multivariate data analysis that a researcher would use to discriminate between different classes of spectra.

4. Results and discussion

In this section, the advanced pre-processing methods in PepsNMR are compared to a manual data treatment with TopspinTM 3.1 and AMIX 3.9.14 on the grounds of two typical biofluids media: a human serum (HSerum) and a human urine (HUrine) datasets. To that purpose, the data tables are fully pre-processed either with PepsNMR or with Topspin/AMIX. The HSerum dataset comes from 4 different blood donors whereas the HUrine dataset is measured from 3 different urine donors. Both of them are designed in a way that the biological

information of between-donors variability should be the main factor affecting the spectral shape. Noisy parameters that were intentionally added to the data to reproduce standard analytical conditions are: day of measurement, sample replicates and sample dilution. More details on these data, their acquisition and pre-processing are available in Section 7.

4.1 Variance structure

If the pre-processing steps accurately manage to transform the data and correct for technical sources of variation, the main cause of variability in the two case studies would correspond to the biological differences between the donors.

As explained in Section 3, inertia analysis and PCA are first applied to characterise and quantify the amount of biological information recovery. Variance decomposition results are given in Table 1. For the two datasets, the inertia between the groups of samples is larger with PepsNMR than with manual pre-processing. Even if this difference in values for HUrine is small (+9.63 %), a real improvement can be observed for the HSerum data (+65.55 %): a larger part of inertia is captured by the variability among the groups rather than between them. Since a larger between-groups inertia implies a better group separation, PepsNMR performs better than the manual alternative.

Table 1 shows that the sum of Principal Components (PC) 1 and 2 is similar for the two pre-processing approaches for the HSerum dataset but the groups structure is only captured by PC2 for the manual pre-processing, as it can be seen in the scores plot in Figure 6. The main source of variation (79.53 %) has another origin: significant peak shifts are observed in the PC1 loadings plot (Figure 7). They are hiding and disseminating the information of interest. Scores plot of PC1 and PC2 for the HUrine dataset shares the same structure, regardless of the pre-processing mode. Spectral profiles are closely tied together and groups are clearly separated. PC1 discriminates between donor 3 and the two others whereas PC2 discriminates patient 1 from donor 2. Moreover, the PepsNMR workflow for this dataset recovers more informative variability in the two first components (+9.23 %) than the manual pre-processing.

Dataset	Pre-processing	Inertia		PCA variance
		Between (%)	Within (%)	PC1+PC2 (%)
HSerum	manual	21.444	78.556	90.483
HSerum	PepsNMR	86.99	13.01	89.751
HUrine	manual	82.185	17.815	82.817
HUrine	PepsNMR	91.819	8.181	92.049

Table 1. Repeatability is partly assessed by variance decomposition methods: inertia decomposition (% of between and within inertia) and PCA (cumulated % of variance for the two first PCs).

4.2 Clustering and classification performances

In this Sub-Section, the unsupervised and supervised classification algorithms (K-means clustering and PLS-DA) explained in Section 2.5 are applied to both datasets. Parameter

Figure 6. PCA scores plot of PC1 and PC2 for HSerum (A-B) and HUrine (C-D) datasets and the two pre-processing strategies (manual and PepsNMR). The colours represent the donors IDs.

Figure 7. First PCA loadings plot of HSerum dataset manually pre-processed.

K for the clustering is the number of donors. The number of LVs for PLS-DA is chosen based on the cross-validated Root-Mean-Square Error (RMSE): 4 for HSerum and 2 for HUrine. Accurate classification and homogeneity of groups are examined to evaluate the unsupervised tool while the predictive quality criterion Q^2 is used to rate the PLS-DA performance. Results from these two approaches can be visualised in Table 2.

Clustering-related indexes quantify either the similarity between two partitions (the true class and the one obtained with the blind clustering) or the clusters homogeneity. For the HSerum, higher Rand and Adjusted Rand indexes for PepsNMR clearly indicate a better recovery of the true classes of samples. Less importantly, even if the Dunn and Davies-Bouldin indexes show a better groups homogeneity for the manual pre-processing, the grouping does not represent correctly the classes. As for the variance structure analysis, the two pre-processing techniques for the HUrine dataset lead to analogous clustering results illustrated here by a perfect categorization of samples. Nonetheless, clusters with the PepsNMR pre-processed datasets do have better homogeneity criteria.

Findings with the PLS-DA are consistent with the previous

results: the classification algorithm has a very bad cross-validation performance for the human serum dataset manually pre-processed since the Q^2 is negative. With the PepsNMR strategy, the performance criterion is substantially improved. For the urine dataset, Q^2 is almost 1, reflecting the excellent class prediction for both pre-processing strategies.

Dataset	Pre-processing	K-Means clustering				PLS-DA Q^2
		Rand	Adj. Rand	Dunn	D-B	
HSerum	manual	0.638	0.136	0.615	0.563	-0.474
HSerum	PepsNMR	0.882	0.715	0.609	0.700	0.667
HUrine	manual	1.000	1.000	0.830	0.671	0.987
HUrine	PepsNMR	1.000	1.000	1.726	0.535	0.992

Table 2. Clustering/classification results for repeatability assessment: indexes derived from unsupervised K-means clustering ((Adjusted) Rand, Dunn and Davies-Bouldin (D-B) indexes) and a supervised PLS-DA validation criterion (the cross-validated coefficient of determination Q^2).

5. Conclusions

Spectroscopic data pre-processing is a keystone procedure for an enhanced information recovery in metabolomics and PepsNMR is the only existing R library able to comprehensively address a complex series of pre-processing steps dedicated to 1D ¹H-NMR data. The two designed experiments comparing PepsNMR to a classic pre-processing strategy confirm its practical interest: its use can increase the spectral repeatability, suggesting an overall better information recovery, and it can enhance the predictive power in a classification problem. Besides, the library has in-built assets: open-source, routines flexibility, automation and objectivity, structured documentation, results reproducibility and reporting. Although it is not possible to ensure that the package will always perform better than the classic pre-processing, its flexibility enables to combine blocks of it to other available tools for a greater efficiency. This library should furthermore help for a continuous improvement in the treatment of complex metabolomic data and complementary/alternative methods will be added in the future to this first version. Methods from PepsNMR are already available in Workflow4Metabolomics[15], a web-interface dedicated to metabolomic data processing, analysis and annotation where it works complementarily with other tools. Lastly and although essential, the selection of optimal parameters as well as the set up of an optimal procedure for methods selection is beyond the scope of this article. Methodological needs in this field[11] open up vast research possibilities.

6. Software availability

PepsNMR is available on GitHub (<https://github.com/ManonMartin/PepsNMR>). It shares commons R package characteristics: it can be easily installed via the function `devtools::install_github()`, called via the R console and has *help* pages for all user

functions and internal datasets. Datasets used in this article are available in PepsNMR at various processing stages. The additional R routines used to validate the pre-processing workflow can also be accessed in the MBXUCL R package (<https://github.com/ManonMartin/MBXUCL>).

7. Datasets

The package is applied on two real datasets from well-known human biofluids: serum (HSerum) and urine (HUrine) to demonstrate its performances. They were both designed to collect multiple measures from the same experimental unit and capture other technical sources of variation, allowing to compare the inter- and intra-unit variability. These datasets are available upon request

7.1 Human Serum

In the HSerum dataset, a blood sample was collected for 4 different donors. For each sample, 8 sub-samples were measured across 8 days with one sub-sample of each donor per day and permutations according to a latin hypercube sampling method. The total number of available FID signals is then $4 \times 8 = 32$. Data were acquired with a 500 MHz Bruker Avance spectrometer equipped with a TCI cryoprobe and using a CPMG relaxation-editing sequence with pre-saturation. Spectra are labelled as: JxDx where Jx is the day of measurement and Dx is the donor label. More details about this dataset can be found in Féraud *et al.* [13].

7.2 Human Urine

In the second dataset, the morning urines of 3 different donors were collected and directly aliquoted in eppendorfs with the following procedure: for each donor, 4 aliquotes containing each 400 μ l of urine + 300 μ l of D2O buffer + 10 μ l of TMSP and 4 aliquotes containing 320 μ l of urine + 380 μ l of D2O buffer et 10 μ l of TMSP. Hence, each donor sample was divided into 2 sub-samples: one was kept pure and the other 25 % diluted. The 4 aliquotes of each dilution were analysed on 4 different days. For each day, the order of measurement was held constant across the 6 sub-samples. The total number of collected FID signals is then: $3 \times 2 \times 4 = 24$. Spectra are labelled as: SxDx where x is the donor label and D is the dilution (0: no dilution; 1: 25 % diluted).

7.3 Spectral pre-processing and outlier removal

Both datasets were pre-processed with PepsNMR and a combinaison of TopspinTM and Amix for manual pre-processing. With PepsNMR, the full advised workflow was applied with default parameters. The region removal and zone aggregation steps were adapted given the type of medium. For region removal, the water (4.5 – 5.1 ppm) area was removed for the serum while the water, urea and maleic acid (4.5 – 6.1 ppm) areas were removed for the urine medium. The zone aggregation step was only applied to the urine dataset for the citrate peak (2.5 – 2.7 ppm). The manual pre-processing included the following sequence: exponential apodization,

Fourier transform, first and zero order phase corrections, internal calibration, polynomial baseline correction, bucketing and spectral regions removal.

Outliers were detected with the ROBPCA[17] approach, a robust PCA used to detect spectra with a high score distance and/or with a high orthogonal distance to the PCA space. Two outlying observations were removed from the HSerum dataset (J6D4 and J6D3) and one from the HUrine dataset (S1D0E1). Note that a missing observation in the manual pre-processing dataset (J4D3) was completely removed for the data analysis.

Acknowledgments

The first author gratefully acknowledges funding from the Belgian Fund for Scientific Research (F.R.S.-FNRS) with a FRIA grant. The F.R.S.-FNRS further supported this work with P. de Tullio as a senior research associate. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) to the Institute of Statistics, Biostatistics and Actuarial Sciences is also gratefully acknowledged.

References

- [1] A. Alonso, S. Marsal, and A. Julià. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology*, 3:23, 2015.
- [2] T. G. Bloemberg, J. Gerretzen, A. Lunshof, R. Wehrens, and L. M. Buydens. Warping methods for spectroscopic and chromatographic signal alignment: a tutorial. *Analytica chimica acta*, 781:14–32, 2013.
- [3] L. Buydens. Towards tsunami-resistant chemometrics. *The analytical Scientist*, 813:24–30, 2013.
- [4] T. D. Claridge. *High-Resolution NMR Techniques in Organic Chemistry*, volume 27 of *Tetrahedron Organic Chemistry Series*. Elsevier, 2009.
- [5] A. Craig, O. Cloarec, E. Holmes, J. K. Nicholson, and J. C. Lindon. Scaling and normalization effects in nmr spectroscopic metabolomic data sets. *Analytical chemistry*, 78(7):2262–2267, 2006.
- [6] F. Dieterle, A. Ross, G. Schlotterbeck, , and H. Senn. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in ¹h nmr metabonomics. *Analytical Chemistry*, 78(13):4281–4290, 2006.
- [7] P. H. C. Eilers. A perfect smoother. *Analytical Chemistry*, 75:3631–3636, 2003.
- [8] P. H. C. Eilers. Parametric time warping. *Analytical Chemistry*, 76(2):404–411, 2004.
- [9] P. H. C. Eilers and H. F. Boelens. Baseline correction with asymmetric least squares smoothing. Leiden University Medical Centre Report, 2005.
- [10] P. H. C. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.
- [11] J. Engel, J. Gerretzen, E. Szymańska, J. J. Jansen, G. Downey, L. Blanchet, and L. M. Buydens. Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, 50:96–106, 2013.
- [12] L. R. Euceda, G. F. Giskeødegård, and T. F. Bathen. Pre-processing of nmr metabolomics data. *Scandinavian journal of clinical and laboratory investigation*, 75(3):193–203, 2015.
- [13] B. Féraud, B. Govaerts, M. Verleysen, and P. De Tullio. Statistical treatment of 2d nmr cosy spectra in metabolomics: data preparation, clustering-based evaluation of the metabolomic informative content and comparison with 1h-nmr. *Metabolomics*, 11(6):1756–1768, 2015.
- [14] T. Gebregiworgis and R. Powers. Application of nmr metabolomics to search for human disease biomarkers. *Combinatorial chemistry & high throughput screening*, 15(8):595–610, 2012.
- [15] F. Giacomoni, G. Le Corguillé, M. Monsoor, M. Landi, P. Pericard, M. Pétéra, C. Duperier, M. Tremblay-Franco, J.-F. Martin, D. Jacob, S. Goulitquer, E. A. Thévenot, and C. Caron. Workflow4metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics*, 31(9):1493, 2015.
- [16] R. Goodacre, D. Broadhurst, A. K. Smilde, B. S. Kristal, J. D. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, T. Ebbels, D. B. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjöström, J. Trygg, and F. Wulfert. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, 3(3):231–241, 2007.
- [17] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden. Robpca: a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.
- [18] J. Keeler. 4 fourier transformation and data processing. In *Understanding NMR Spectroscopy*. John Wiley & Sons, 2002.
- [19] A. K. Kosmides, K. Kamisoglu, S. E. Calvano, S. A. Corbett, and I. P. Androulakis. Metabolomic fingerprinting: challenges and opportunities. *Critical ReviewsTM in Biomedical Engineering*, 41(3), 2013.
- [20] E. M. Lenz and I. D. Wilson. Analytical strategies in metabonomics. *Journal of proteome research*, 6(2):443–458, 2007.

- [21] K. H. Liland. Multivariate methods in metabolomics—from pre-processing to dimension reduction and statistical analysis. *TrAC Trends in Analytical Chemistry*, 30(6):827–841, 2011.
- [22] K. H. Liland, T. Almøy, and B.-H. Mevik. Optimal choice of baseline correction for multivariate calibration of spectra. *Applied spectroscopy*, 64(9):1007–1016, 2010.
- [23] A. D. Maher, S. F. Zirah, E. Holmes, and J. K. Nicholson. Experimental and analytical variation in human urine in ^1H nmr spectroscopy-based metabolic phenotyping studies. *Analytical Chemistry*, 79(14):5204–5211, 2007.
- [24] R. Marion. Pre-processing of nmr spectra: review and evaluation of baseline correction, normalization, scaling and transformation methods. Master thesis, Ecole de Statistique, Biostatistique et Sciences Actuarielles, Université catholique de Louvain, 2016.
- [25] S. Ren, A. A. Hinzman, E. L. Kang, R. D. Szczesniak, and L. J. Lu. Computational and statistical analysis of metabolomics data. *Metabolomics*, 11(6):1492–1513, 2015.
- [26] R. Rousseau. *Statistical contribution to the analysis of metabonomics data in ^1H NMR spectroscopy*. PhD thesis, Institut de Statistique, Biostatistique et Sciences Actuarielles, Université catholique de Louvain, 2011.
- [27] W. M. Siebert. *Circuits, signals, and systems*, volume 2. MIT press, 1986.
- [28] A. Smolinska, L. Blanchet, L. M. Buydens, and S. S. Wijmenga. Nmr and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Analytica chimica acta*, 750:82–97, 2012.
- [29] A. Van Nederkassel, M. Daszykowski, P. H. Eilers, and Y. Vander Heyden. A comparison of three algorithms for chromatograms alignment. *Journal of Chromatography A*, 1118(2):199–210, 2006.
- [30] J. Vanwinsberghe. Bubble: development of a matlab tool for automated ^1H -nmr data processing in metabonomics. Unpublished manuscript, Strasbourg University, 2005.
- [31] T. N. Vu and K. Laukens. Getting your peaks in line: A review of alignment methods for nmr spectral data. *Metabolites*, 3(2):259, 2013.
- [32] E. T. Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63–75, 1922.