Preprocessing of NMR metabolomics data

Leslie R. Euceda[1], Guro F. Giskeødegård[1,2], Tone F. Bathen[1]

[1]Department of Circulation and Medical Imaging, Faculty of Medicine, The Norwegian University of Science and Technology (NTNU), Trondheim, Norway, [2]St. Olavs University Hospital, Trondheim, Norway

Corresponding Author: Leslie R. Euceda:

leslie.e.wood@ntnu.no

Abstract

Metabolomics involves the large scale analysis of metabolites and thus, provides information regarding cellular processes in a biological sample. Independently of the analytical technique used, a vast amount of data is always acquired when carrying out metabolomics studies; this results in complex datasets with large amounts of variables. This type of data requires multivariate statistical analysis for its proper biological interpretation. Prior to multivariate analysis, preprocessing of the data must be carried out to remove unwanted variation such as instrumental or experimental artifacts. This review aims to outline the steps in the preprocessing of NMR metabolomics data and describe some of the methods to perform these. Since using different preprocessing methods may produce different results, it is important that an appropriate pipeline exists for the selection of the optimal combination of methods in the preprocessing workflow.

Keywords: Baseline correction, biological marker, metabolomics, multivariate analysis, NMR spectroscopy, normalization, peak alignment, scaling, statistical data interpretation, variable selection

Introduction

Metabolism is the set of enzymatic biochemical reactions necessary for life processes to take place in living organisms[1]. Metabolomics involves the large scale analysis of metabolites, which are the products and/or intermediates of metabolism[2]. They include substances like carbohydrates, amino acids, nucleotides and vitamins[3]. Most metabolites are molecules of 1000 Da or lower in size; however, some large molecules, like lipoproteins and albumin, may fall into this group of substances[2].

Metabolites can be classified into different groups according to the field of study. In plant metabolomics, a metabolite can be either primary or secondary. Primary metabolites are involved in life-preserving functions related to growth, development and reproduction[4]. Although secondary metabolites are not directly crucial for short-term survival, they play a significant ecological role, usually related to plant defense[4]. Human metabolomics divides metabolites into endogenous and exogenous, the former being synthetized by the host organism while the latter originate from external substances like food or drugs. This review focuses on endogenous human metabolomics of biofluids and tissue. The collection of small-molecule metabolites present in a biological organism makes up its metabolome, which is a term parallel to the genome, transcriptome and proteome[5]. The metabolome gives insight on the physiological state of a cell or organism at a particular moment, under given conditions. It is dynamic: constantly changing due to the influence of extragenomic factors like environment, diet, toxins, disease processes, etc.

Daviss defines metabolomics as “the study of the unique chemical fingerprints that specific cellular processes leave behind”[6]. It is a relatively new field that can be considered the final step in the omics cascade, after genomics, transcriptomics and proteomics[7] (Figure 1). The omics cascade illustrates the biochemical information exchange between the different molecular levels in an organism, starting with the level of genes to the last level of metabolites. Because the metabolome is affected by post transcriptional and post translational events, such as environmental factors, when compared to these top three levels, metabolomic profiles provide information at a level that is closer to the phenotype[8]. Furthermore, these profiles can be useful in the early detection of a normal, healthy state becoming dysfunctional[9, 10]. The types and quantities of metabolites that make up these profiles depend on the condition of the living organism or cell they belong to, which makes it possible to establish metabolic differences between different states, gain insight on the biochemical pathways behind diseases and identify metabolites as potential biomarkers and targets for therapy.

The most common analytical techniques used to acquire metabolomics data are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS)[7, 11]. In NMR, a sample is submitted to a magnetic field and a radio frequency (RF) pulse is applied, causing the atomic nuclei to resonate at a specific frequency[12]. Some atomic nuclei, e.g. protons (^{1}H), possess a property called spin. When such nuclei are exposed to a strong magnetic field (B_0), their magnetic moments will align either parallel or anti-parallel of B_0, yielding a lower or higher energy spin state, respectively. The energies from nuclei in different states cancel each other out; however, more protons will exist in the lower energy state, causing a residual component of the magnetic moment. By applying an excitation radio frequency (RF) pulse, nuclei can be brought to a higher energy state of resonance. Following the excitation, the nuclei recover to the equilibrium state and the absorbed energies are released, emitting RF signals. Receiver coils detect the RF signals originating from the excited nuclei, and small differences in resonant frequencies can yield detailed information about the sample's molecular structure. The result of the analysis is an NMR spectrum, which is a plot of the intensity of absorption or emission as a function of the resonance frequency, which corresponds to the energy difference between the proton's two spin states (Figure 2). The chemical shift scale expresses the resonance frequency as a deviation from a reference signal in parts per million (ppm) and is thus independent of the magnetic field strength.

In MS, an ion source transforms the sample into volatile ions e.g. by bombarding it with electrons, photons, ions, molecules or thermal or electric energy[13]. The ions produced are then accelerated into a mass analyzer. Here, the energetically charged ions are continuosly detected and sorted according to their mass-to-charge ratio (*m/z*). Finally, the beam of ions is converted into an electric signal by a transducer to be processed and displayed at a further stage. The mass spectrometer's response is represented in a plot of relative intensity over *m/z*.

Both of the abovementioned analytical techniques have different strengths and weaknesses. Thus, although they can be used individually depending on the objectives of the metabolomics study, they complement each other. NMR provides more information on the chemical structure of metabolites making it easier to elucidate it[14]. Preparation of biofluid and tissue samples for NMR is simple and does not require time-consuming procedures as in MS. NMR sample preparation is also nondestructive, which enables further analyses of the same sample. When compared to MS, NMR is highly quantitative and reproducible[11, 15]. A drawback of NMR is its low sensitivity which does not make it possible to detect metabolites that are present in low abundance in a sample[11, 14, 15]. Limits of detection are in the micro- to millimolar range, while MS can detect compounds present as low as nanomolar or even picomolar and femtomolar concentrations depending on the method used[16, 17]. MS is usually combined with a chromatographic technique for prior separation of compounds in a sample. It provides high sensitivity and specificity, but is less useful for the elucidation of chemical structure, generating information only on the elemental composition and some structural fragments. In addition, sample preparation usually includes time-consuming pretreatment such as extraction or, when separating non-volatile compounds using gas chromatography, derivatization.

Whatever the method used to acquire metabolomics data, they all generate complex, high-dimensional data. In statistical terms, matrices consisting of much more variables (organized in columns) in relation to samples (organized in rows) will be produced. Many of the variables in these matrices are highly collinear, which means there will be an approximate linear relationship between at least some of them[18]. Classical multiple linear regression and similar techniques use the least squares approach (Equation 1) to fit models represented as $y=X\beta+E$, where β is a vector containing a regression coefficient for each variable in the matrix X, y is a vector containing a response or category for each sample and E is a residual matrix. Coefficients from collinear variables can compete and cancel each other out, making the inverse of X^TX in Equation 1 difficult to calculate or ill-conditioned[19]. In these cases, dimension reduction multivariate methods, such as principle component analysis (PCA) and partial least squares (PLS) – discriminant analysis (DA) should be used for modelling of the data.

$$\beta = (X^TX)^{-1}X^Ty \tag{1}$$

Prior to multivariate analysis, preprocessing of the data must be carried out to remove any variance not related to the property of interest. Several preprocessing procedures and methods exist that vary according to the characteristics of the data to be dealt with. Those characteristics are a result, to some extent, of the analytical technique used. This review outlines the main preprocessing steps focusing on NMR metabolomics datasets, particularly on those obtained from biofluids and tissue samples. A description of some of the methods available to carry out these procedures is also provided.

Preprocessing of NMR metabolomics data

Preprocessing is the transformationof raw data to make it apt for subsequent data analysis by adjusting the variability of each variable as well as that of the relationships between them[20]. It involves different procedures, each one directed to correct for different artifacts present in the data. Furthermore, different methods are available to carry out these procedures. The main steps necessary in the preprocessing of NMR data are described below.

Baseline correction

The baseline in NMR spectra consists of broad regions between peaks that do not contain signal of interest. Due to noise, i.e. the constant, random variation of positive and negative signals assumed to average to zero, the baseline will never be completely flat, but attempts should be made to make it as straight as possible. However, most NMR spectra display baseline distortion which can originate from different sources. This may be macromolecules that decay much faster in the time domain than those arising from the low molecular weight metabolites [21]. At short echo time [12] these signals from macromolecules are broad in the frequency domain and may alter the baseline. Other sources of baseline distortion include alterations in the free induction decay (FID), particularly of the first few data points[22], and instability in the instrument[23]. A distorted baseline alters the real intensity values in the spectra, resulting in less accurate peak assignment and integration. Furthermore, because most multivariate techniques cannot separate baselines and real signals[24], statistical results are affected when the baseline is not flat. Thus, appropriate baseline correction is an important step for removing unwanted variation.

Baseline correction of NMR-acquired data can be carried out in either the frequency or time domains[22, 25]. Frequency domain baseline correction is the typical approach and it involves baseline estimation and its subsequent subtraction directly from the measured spectrum[20] (Figure 3). Alternatively, the distorted points at the beginning of the FID in the time domain can be reconstructed by extrapolation[26], oversampling[25] or proper data acquisition timing to remove dead time[27] (Figure 4), for example. A combination of both methods can also be used. Frequency domain baseline correction is easier to adapt for specific purposes and spectra, such as metabolomics spectra[22]. We will discuss three baseline estimation methods and a strategy to select an optimal one for a given study. Some of the methods described below make use of a prior or combined step of noise removal, or smoothing, when modelling the baseline.

Iterative polynomial fitting[29] aims to eliminate the true signal or peaks of interest from the spectrum to estimate the baseline. The total spectrum, which will be referred to as the measured signal (*y*), is fitted with a low order polynomial. The signal is considered a function whose equation of polynomial fitting in the concise matrix form is shown in Equation 2, where *X* is a matrix of size $m \times (n+1)$ whose rows are each of the $j= 1, 2,..., m$ spectral points elevated from the zero to the *n*th power: $[1\ x_j\ x_j^2\ ...\ x_j^n]$. The column vector *a* represents the coefficients to be determined: $a_0, a_1, ..., a_n$. The polynomial fitting is obtained through Equation 3.

$$y = Xa \tag{2}$$

$$b = X(X^T X)^{-1} X^T y \tag{3}$$

A comparison between corresponding spectral points in the original (y_0) and estimated signal (b_k) is carried out, where *k* is the iteration number. Spectral points that have a higher value in y_0 are replaced by those in b_k, to obtain a new signal (y_k). The process is repeated taking y_k as the new signal until the criterion ρ (Equation 4) is met, resolving with b_k as the estimated baseline. There is no straightforward approach to determine the optimal polynomial order, and so the procedure may need to be repeated with different values for the power parameter *n* before a baseline that matches the original signal well is estimated. Another disadvantage is that fitting a different polynomial to each spectrum may result in variance in the dataset not related to the metabolic states under study[30]. Furthermore, in theory, the shape of NMR drifts is not that of polynomial functions[31].

$$\rho = \frac{\|b_k - b_{k-1}\|}{b_{k-1}} < 0.001 \tag{4}$$

Asymmetric least squares smoothing (AsLS)[32] uses the least squares algorithm to fit the baseline by minimizing the residuals through a vector of weights while adding a penalty to smooth the estimated baseline (Equation 5). The method is asymmetric as negative residuals are given more weight than positive ones.

$$S = \sum_i w_i \, (x_i - z_i)^2{}_i + \lambda \sum_i (\Delta^2 z_i)^2 \tag{5}$$

The first term in Equation 5 measures the fit to the original spectrum, where *w* is a vector of prior weights, *x* is the original spectrum and *z* is the estimation of the baseline in the current iteration (*i*). The second term is the smoothing penalty on *z*, where *λ* is a parameter that tunes the smoothing and *Δ^2* z_i is calculated according to Equation 6.

$$\Delta^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) = z_i - 2z_{i-1} + z_{i-2} \tag{6}$$

A parameter *p* tunes the asymmetry as it is used to compute the weights according to whether the residuals are positive or negative as shown in Equations 7 and 8, respectively.

$$w_i = 1 - p \tag{7}$$

$$w_i = p \tag{8}$$

Residuals calculated at each iteration can have different signs, leading to different weights (Equations 7 and 8). The process is iterated until the signs, and hence the weights, are unchanged. Using a vector of all ones for *w* as a starting point leads to a solution in about 10 iterations[33].

Advantages of AsLS include fast computation even for large signals and the tuning of the flexibility and position of the baseline with the abovementioned parameters, whose use make results completely reproducible. However, there is no automatic way to select these parameters and thus it must be done subjectively based on human judgment.

Dietrich et al. [34] introduced a derivative method involving automatic peak recognition followed by fitting a fifth degree polynomial to the remaining baseline and its subsequent subtraction from the spectrum. The noise is reduced prior to derivatization using a moving average filter of width *k* which replaces every spectral point (x_j) with the average of the interval including x_{j-k} to x_{j+k}. A first derivative of the spectrum is built by replacing each spectral data point with the difference between its value and the value of the next point. This allows us to distinguish the peaks from the baseline, as differences between their neighboring points are generally higher than those of the baseline[35]. The power spectrum is calculated and accentuates the peaks from the baseline. Anything below the threshold calculated as the sum of the mean of the entire power spectrum and three times the standard deviation is considered the baseline while the rest are peaks and are discarded. The threshold is calculated again until all data points are below it. For all points considered peaks, if both of its neighbors do not exceed the final threshold, it is put back into the baseline. Similarly, baseline points having both neighbors above the threshold are considered peaks. The final step is the polynomial fitting and subtraction of the baseline. This method is fast and easily automized but can be unstable when the spectrum has overlapping or broad peaks [35].

Liland et al. have proposed a procedure for the optimal selection of spectral baseline correction methods and parameters[36]. They apply this procedure on two different types of spectra: Raman and matrix-assisted laser desorption/ionization (MALDI)-Time of flight

(TOF). The procedure is said to be applicable for multivariate spectra regardless of the analytical technique used for acquisition. They recommend four steps: 1) selection of a limited number of parameter values to test; 2) baseline correction employing different methods and combinations of parameter values, statistical analysis and quality assessment; 3) visual inspection of the baseline resulting from the combination of parameter values achieving the best quality measure; 4) selection of the best performing method based on the quality measure and its subsequent application on spectra not used for modelling. A prediction is made for these corrected spectra using the statistical models obtained in step 2 to validate the performance of the selected baseline correction method.

Peak Alignment

During spectral acquisition using NMR, peaks can shift left or right from their expected chemical shift due to changes in pH, temperature, inhomogeneous magnetic field and molecular interactions[37, 38]. This will create a variation in ppm values between spectra from different samples. For each spectrum included in a dataset, the signal for every metabolite should be aligned to the exact chemical shift to make the data comparable in multivariate analysis. If this is not the case, information corresponding to the same metabolite will be positioned in different columns of the data matrix hindering a true comparison.

Alignment methods are used to correct differences in chemical shift. Most of these methods position the signals in a spectrum to fit the corresponding chemical shift in a reference spectrum. A type of alignment methods known as warping does this by shifting, compressing or stretching the data until every signal matches the location, or chemical shift, of the same signal in the reference without changing the order of peaks[39].

Alignment can be carried out in two ways. The first is a global alignment in which the entire spectrum is corrected at once. Another approach is to divide the data into segments and align each segment separately. This is a more local approach and is more applicable to NMR spectra because peaks in a spectrum may not shift uniformly in direction or magnitude[20]. Different regions in a spectrum may exhibit peaks shifting in opposite directions or to different extents. However, misalignment can result if peaks are assigned into a wrong segment or are divided into two segments[24].

An example of global alignment methods is parametric time warping (PTW)[33]. PTW proposes a polynomial model for the warping function which is easy to interpret. The term of zero order, or the constant in the function, represents a shift. The first order term represents a compression or stretch, and higher order terms represent compression or stretches that depend on the variation of the spectra variable [39].

Correlation optimized warping (COW) was introduced by Nielsen et al.[40] for chromatographic data. This method uses two parameters to align spectra to match the reference: segment size and flexibility or slack[24, 37]. Based on the segment size, the spectrum is divided into local sections which will be aligned individually by compression or stretching. The flexibility indicates the maximum amount of data points that the segment can be stretched or compressed via interpolation. This process is carried out to maximize the overall correlation between sample and reference.

Another local approach which segments spectra is icoshift, which was developed by Savorani et al.[41] for the alignment of NMR acquired metabolomics data. This method employs fast Fourier transforms to maximize the correlation between a spectrum and a reference, and here correlation is optimized for each segment as opposed to overall correlation as in COW. Another difference is that the segments can be user defined and do not have to be equal in size (Figure 5). In addition, icoshift only shifts the segments and hence does not affect integration for quantification purposes by variation caused by stretching or compressing the spectra. This algorithm has shown to be faster than COW[24, 38, 41] because it aligns all spectra simultaneously. Giskeødegård et al.[42] showed that icoshift and COW performed better when comparing five different alignment methods for high resolution (HR) magic angle spinning (MAS) MR spectra based on similarity measures between spectra and reference, quantification of change due to alignment and performance of PLS- DA classification models resulting from aligned data. Only a few of the existing alignment algorithms have been presented here. A more extensive list of methods has been described by Bloemberg et al[39].

All of the abovementioned methods require a reference spectrum to align the rest of the spectra towards. Choosing or constructing an appropriate reference is therefore a critical step, and may affect the subsequent data analysis results. The reference should be the best possible representation of the group of spectra in the dataset in terms of number of peaks[37]. It can be either selected from the dataset of interest or an entirely new spectrum built by calculating a measure of spread of the whole dataset. Different criteria exist to select an individual spectrum from a dataset as a reference. It can be selected randomly as done by Pierce et al. for chromatographic data[43]. In cases where the spectrum that contains the most metabolite signals is known, it can be used as a reference. Another approach is to use the spectrum which is most similar to the rest by calculating for each spectrum, its correlation with the rest and then averaging to obtain a single score in the way of a mean correlation per spectra[44].

A reference as a measure of spread of an entire dataset can also be created in different ways. The mean spectrum usually includes all peaks appearing throughout the entire group of spectra. However, using this as a reference may also introduce atypical artifacts that may cause peak distortion after alignment[45], in addition to the mean spectrum having broad peaks. To avoid this, median or trimmed mean spectra can be used.

An alternative approach typically used for NMR data to correct small misalignments is binning, also known as bucketing. Spectra are segmented into a desired number of bins and all measurements inside each bin are summed by means of the integral of the signal or area under the curve [20, 24, 45]. This method makes the data less complex and more manageable by reducing the number of variables, which may facilitate data analysis. Moreover, the spectra are implicitly smoothed with this procedure. However, resolution will be reduced and, as is the case for local alignment methods, care must be taken to correctly place bins so as to not have peaks fall into the wrong bin or to split peaks in two (Figure 6). This can lead to loss of information or generation of artifacts.

Normalization

Metabolomic responses are reflected in differences in concentration of specific metabolites. Biological samples from which datasets are acquired regularly exhibit differences in overall metabolite concentrations. In the case of biofluids, dilution factors will not be exactly the same for each sample. For instance, the concentration of the constituents in urine differs according to the amount of water ingested or urinary volume excreted. In the same way, when analyzing tissue, the weight of prepared material may not be constant for all samples. This creates variations in signal intensities attributed to amount of material analyzed and dilution rather than changes in metabolic responses. Normalization methods aim to remove this effect to make spectra comparable. Different normalization approaches include area normalization, probabilistic quotient normalization (PQN), range normalization and normalization to a reference metabolite.

Area or integral normalization can be considered the standard for NMR-acquired metabolomics data[46]. The algorithm divides each data point by an equal total area, integral or mean of the spectrum. Figure 7 shows a hypothetical case in which three different sets of spectral data points were acquired from different amounts of the same tissue sample. For two spectra in which one was acquired from double the amount of sample, the signal intensity would be expected to be double for each data point. Similarly, three times the amount of sample will yield three times the intensity. Figure 7 shows how these tissue size differences are eliminated when using area normalization by dividing by the sum of intensities. However, this method has been reported to have limitations in robustness and accuracy when extreme quantities of single metabolites exist in samples[46, 47].

PQN[46] was developed to address the issues of area normalization in NMR-acquired metabolomics data of urine. It assumes that changes in overall metabolite concentrations affect the whole spectrum, while concentration differences in individual metabolites are reflected in only parts of the spectrum. Those individual metabolite peak differences do not depend on dilution only, as assumed for all peak differences by area normalization. The method follows four steps: 1) select or calculate a reference spectrum; 2) calculate the quotient between each of the data points of the spectrum to be normalized and the corresponding points in the reference spectrum; 3) find the median of these quotients and divide all points in the spectrum by this median; 4) repeat steps 2 and 3 for all spectra. Care must be taken in using a representative reference spectrum without a large portion of specific metabolite concentration changes in the total signal; it may be necessary to calculate it from control samples.

Min-max or range normalization[48] rescales the spectral data points to a new range, usually from zero to one or from one to minus one[49]. Equation 9 shows the linear transformation through which each of the *i* number of data points *x* in the spectrum, with values that range from x_{min} to x_{max}, is changed to range from new_{max} to new_{min}.

$$x_i \equiv \frac{x_{i-}x_{min}}{x_{max}-x_{min}} \times (new_{max} - new_{min}) + \ new_{min} \tag{9}$$

A fourth method used for urine spectra is to express the intensity of all signals relative to the peak areas of the normally-occurring reference metabolite creatinine.[20]. . The rationale for the use of creatinine is that its excretion rate or clearance is considered to remain constant [50]. However, creatinine excretion may be affected by metabolic changes caused by certain diseases or in children and the elderly due to muscle mass differences [47]. Other drawbacks include possible interference of overlapping peaks and the pH dependence of the intensity of the creatinine signal[46].

Before any normalization is attempted, it is important to consider the removal of spectral signals that are known to be unrelated with biological effects of interest, such as the residual water signal and lipid peaks arising from adipose tissue, whose intensities may interfere with this procedure.

Scaling

Metabolites that are more abundant will usually display larger differences among samples that can mask smaller, yet biologically important changes in metabolites of low abundance. Scaling aims to balance signal intensity variances that originate from difference in average abundance of metabolites. While normalization is a sample/row-wise operation, scaling is performed column-wise on the variables[47, 51]. Unlike normalization operations, which are performed independently on each sample, scaling operations depend on all samples in the dataset. Typically used scaling methods include autoscaling, pareto scaling and vast scaling[15, 47, 52].

Mean centering is a column-wise centering method typically used prior to scaling. Mean centering transforms all values so that they vary around zero instead of around the mean value[15]. This corrects for the displacement between metabolites existing in high and low concentrations, resulting in the simplest model possible[47] (Figure 8). Mean centering is carried out by subtracting the column mean intensity from each individual intensity value.

Autoscaling or standardization is a method which converts all metabolites or spectral data points to have unit variance. It is therefore also known as unit variance scaling. The algorithm divides each value in a column by the standard deviation after mean centering[20, 52].

However, variables containing only noise are given larger influence when applying this method[15, 20]. NMR spectra include many baseline regions with random noise, which makes autoscaling less suited for this type of data.

Pareto scaling is an approach which is similar to autoscaling except it uses the square root of the standard deviation as a scaling factor instead of the standard deviation alone[15]. The resulting values are more similar to the original, raw measurement. Pareto scaling is an intermediate between the extremes of no scaling and autoscaling; however, large biologically relevant variations in the data are decreased more than small variations, and so the influence of the former is reduced significantly when compared to the unprocessed data.

Variable stability, or Vast scaling, is a third method that accentuates relatively stable metabolites that do not show large fluctuations. It is an extension of autoscaling: after mean centering, the data is divided by both the column standard deviation and the coefficient of variation (cv). A group specific cv can be calculated if a priori knowledge about sample groups is available. This method is not suitable for data having large induced variation and lacking group structure[52].

Variable selection

Typically, when conducting metabolomics studies, previous knowledge exists about the origin of the analyzed samples and supervised methods are used to build models for the classification of samples into biological categories based on the acquired metabolic profiles. The biological category each sample falls into are represented using dummy variables[53]. For example, the code 1 can be assigned to one category and -1 to another. These codes are arranged in the form of a response variable (*y*) with rows corresponding to each sample (Figure 9). In most cases, not all of the chemical shift regions are informative or related to biological variances of interest. Variables that are uninformative may add noise to the model and if included may produce misleading results and interpretations[54].

Variable selection is the process of reducing the original number of variables or columns in the dataset by discriminating informative variables from the ones that are not related to the response *y*. Because each spectral data point region represents a metabolite, selecting informative variables in metabolomics datasets can be considered as selecting informative metabolites that are relevant to the property of interest, and can lead to the identification of potential biomarkers[55, 56].

Variable selection is not a preprocessing procedure per se, as it involves statistical modelling in order to determine the best subset of variables. It is described here because it may be carried out prior to the construction of the final statistical models. However, it is important to keep in mind that this procedure should be regarded as being part of statistical analyses rather than preprocessing and requires proper validation to avoid biased results.

The minimum redundancy-maximum relevance method uses a quantity called mutual information as a measure of variable importance[57]. If the values of a given variable are random or uniformly distributed between categories, the information that is mutual between the variable and the response *y* is zero. However, if the variable accounts for large differences between classes, then there is large amount of mutual information between the variable and the response. First developed for gene expression data, this method not only selects the variables most relevant to the property of interest, but it also excludes variables that provide the same information as others that have been selected, in this way minimizing redundancy. This is useful for dimensionality reduction of NMR spectra in which nearby data points are highly correlated. The risk in this is that different highly correlated metabolites could be discarded while still being involved in the pathway of a disease being studied.

Variable significance can also be evaluated using permutation tests. For a given variable *j*, two different models are built: one with permuted values of *j* and one without. The performance of both models is evaluated. To do so, the models are tested on data not used for training, or building, the model. The model built with the training set is used to predict the response of each spectrum in the test set with and without permuted *j*. A model quality measure is calculated based on the predicted value and the true value of the response for each spectrum. The entire process is repeated for each of the defined number of permutations. A mean of performance quality measures is obtained for both normal prediction and prediction with permuted *j*, which can then be compared. This method requires high computational power and can be very time-consuming.

A third approach involves the calculation of a selectivity ratio for each variable based on a target projection model[58]. While other statistical methods produce models consisting of many components, target projection models are composed by a single predictive component, which makes them easier to interpret. From the target projection model built, explained (v_{expl}) and residual (v_{res}) variances can be calculated for each of the *p* variables in the dataset. The selectivity ratio for each variable will then be the relationship between these variances (Equation 10) The higher the selectivity ratio, the more informative the variable is. However, a suitable threshold must be established to determine at what selectivity ratio a variable should be selected as important or informative.

$$SR_j = \frac{v_{expl,j}}{v_{res,j}}, j = 1,2,3, \dots, p \qquad (10)$$

When selecting a subset of variables and building statistical models, a proper validation scheme should be implemented to ensure the reliability of model performance and unbiased results[59, 60]. It is highly important that the scaling factors are calculated from training data and applied to test set data to avoid overfitting. Anderssen et al. have proposed a double-layer validation to avoid over-optimistic model performance [61].

Concluding remarks

Metabolomics data, whether acquired by MS or NMR, is high dimensional and complex and requires multivariate statistical techniques for interpretation. Prior to data analysis, preprocessing must be carried out to remove artifacts and biologically irrelevant variance. Preprocessing of NMR metabolomics data includes several procedures; the optimal method to carry out may vary depending on the type of samples analyzed. Here, we have described some of the methods that have been applied for the preprocessing of NMR metabolomics data. Using different methods for the same preprocessing procedure may produce different results, and the group that works most appropriately together when applied to the dataset in question should be selected. It may be useful for each laboratory to establish a pipeline to choose the combination of preprocessing procedures and methods that will optimize NMR metabolomics data analysis results based on the aim of the analysis.

The result of appropriate data preprocessing is the elimination of spectral artifacts and removal of irrelevant variation, making NMR spectra more comparable. It is then possible to proceed to analyze the data using multivariate statistical methods. The description of such methods are beyond the scope of this paper, and the authors refer to [24] for further descriptions of multivariate methods for analysis of NMR data.

Acknowledgements

This work was supported by grants from the Research Council of Norway (FRIMEDBIO221879).

References

[1] Manahan SE. Toxicological Chemistry and Biochemistry. Boca Raton, FL, USA: CRC Press; 2003.

[2] Nielsen J. Metabolomics in Functional Genomics and Systems Biology. Metabolome Analysis An Introduction. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2007.

[3] Christians U, Albuisson J, Klawitter J, Klawitter J. The Role of Metabolomics in the Study of Kidney Diseases and in the Development of Diagnostic Tools. In: Edelstein CL, editor. Biomarkers of Kidney Disease. London, UK: Academic Press, Elsevier Inc.; 2011.

[4] Hurtado-Fernandez E, Pacchiarotta T, Gomez-Romero M, Schoenmaker B, Derks R, Deelder AM, Mayboroda OA, Carrasco-Pancorbo A, Fernandez-Gutierrez A. Ultra high performance liquid chromatography-time of flight mass spectrometry for analysis of avocado fruit metabolites: method evaluation and applicability to the analysis of ripening degrees. J Chromatogr A 2011;1218:7723-38.

[5] Fiehn O. Metabolomics — the link between genotypes and phenotypes. Plant Mol Biol 2002;48:155-71.

[6] Daviss B. Growing Pains for Metabolomics. The Scientist. 2005:25-8.

[7] Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol 2012;13:263-9.

[8] Hollywood K, Brison DR, Goodacre R. Metabolomics: current technologies and future trends. Proteomics 2006;6:4716-23.

[9] Mickiewicz B, Vogel HJ, Wong HR, Winston BW. Metabolomics as a Novel Approach for Early Diagnosis of Pediatric Septic Shock and Its Mortality. Am J Respir Crit Care 2013;187:967-76.

[10] Boudonck KJ, Mitchell MW, Német L, Keresztes L, Nyska A, Shinar D, Rosenstock M. Discovery of Metabolomics Biomarkers for Early Detection of Nephrotoxicity. Toxicol Pathol 2009;37:280-92.

[11] Pan Z, Raftery D. Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics. Anal Bioanal Chem 2007;387:525-7.

[12] Jacobsen N. NMR Spectroscopy Explained: Simplified Theory, Applications and Examples for Organic Chemistry and Structural Biology. Hoboken, NJ, USA: John Wiley and Sons, Inc.; 2007.

[13] Skoog D, Holler J, Crouch S. Principles of Instrumental Analysis, Sixth Edition: Thomson Brooks/Cole; 2007.

[14] Peironcely JE, Reijmers T, Coulier L, Bender A, Hankemeier T. Understanding and classifying metabolite space and metabolite-likeness. PloS one 2011;6:e28966.

[15] Smolinska A, Blanchet L, Buydens LMC, Wijmenga SS. NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. Anal Chim Acta 2012;750:82-97.

[16] Hopkin K. The study of metabolomics is attracting a flurry of biotechs and academics, with research implications ranging from plant biology to drug discovery. 2004 [cited 2014 March 26]; Available from: http://www.bio-itworld.com/archive/071404/omics_sidebar_5593.html

[17] Kristal BS, Shurubor YI, Kaddurah-Daouk R, Matson WR. High-Performance Liquid Chromatography Separations Coupled With Coulometric Electrode Array Detectors: A Unique Approach to Metabolomics. In: Weckwerth W, editor. Metabolomics Methods and Protocols. Totowa, NJ, USA: Humana Press Inc.; 2007.

[18] Mandel J. Use of the Singular Value Decomposition in Regression Analysis. Am Stat 1982;36:15-24.

[19] Jackson JE. A User's Guide to Principal Components. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 1991.

[20] Engel J, Gerretzen J, Szymańska E, Jansen JJ, Downey G, Blanchet L, Buydens LMC. Breaking with trends in pre-processing? TrAC Trends Anal Chem 2013;50:96-106.

[21] Sarma MK, Nagarajan R, Thomas MA. Brain MR Spectroscopy In Vivo: Basics and Quantitation of Metabolites. In: Muftuler LT, editor. Quantifying Morphology and Physiology of the Human Body Using MRI. Boca Raton, FL, USA: CRC Press, Taylor & Francis Group; 2013.

[22] Xi Y, Rocke DM. Baseline correction for NMR spectroscopic metabolomics data analysis. BMC bioinformatics 2008;9:324.

[23] Wang KC, Wang SY, Kuo CH, Tseng YJ. Distribution-based classification method for baseline correction of metabolomic 1D proton nuclear magnetic resonance spectra. Anal Chem 2013;85:1231-9.

[24] Liland KH. Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis. TrAC Trends Anal Chem 2011;30:827-41.

[25] Wider G. Elimination of Baseline Artifacts in NMR Spectra by Oversampling. J Magn Reson 1990;89:406-9.

[26] Marion D, Bax A. Time domain linear extrapolation. J Magn Reson 1989;83:205-11.

[27] Heuer A, Haeberlen U. A new method for suppressing baseline distortions in FT NMR. J Magn Reson (1969) 1989;85:79-94.

[28] Schmidt-Rohr K, Spiess HW. Multidimensional Solid-State NMR and Polymers. San Diego, California, USA: Academic Press Limited; 1994.

[29] Gan F, Ruan G, Mo J. Baseline correction by improved iterative polynomial fitting with automatic threshold. Chemometr Intell Lab Syst 2006;82:59-65.

[30] Advanced Preprocessing: Noise, Offset, and Baseline Filtering. 2013 [cited 2013 November 7]; Available from: http://wiki.eigenvector.com/index.php?title=Advanced_Preprocessing:_Noise,_Offset,_and_Baseline_Filtering

[31] Cobas JC, Bernstein MA, Martín-Pastor M, Tahoces PG. A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. J Magn Reson 2006;183:145-51.

[32] Eilers PHC, Boelens FM. Baseline Correction with Asymmetric Least Squares Smoothing. 2005 [cited 2013 November 7]; Available from: http://www.science.uva.nl/~hboelens/publications/draftpub/Eilers_2005.pdf

[33] Eilers PHC. Parametric Time Warping. Anal Chem 2004;76:404-11.

[34] Dietrich W, Rüdel CH, Neumann M. Fast and precise automatic baseline correction of one- and two-dimensional nmr spectra. J Magn Reson (1969) 1991;91:1-11.

[35] Miao S, Koenders E, Knobbe A. Automatic baseline correction of strain gauge signals. Struct Control Health Monit 2014:n/a-n/a.

[36] Liland K, Almøy T, Mevik B-H. Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra. Appl Spectrosc 2010;64:1007-16.

[37] Jellema RH. 2.06 - Variable Shift and Alignment. In: Brown SD, Tauler R, Walczak B, editors. Comprehensive Chemometrics. Oxford: Elsevier; 2009. p. 85-108.

[38] Khakimov B, Bak S, Engelsen SB. High-throughput cereal metabolomics: Current analytical technologies, challenges and perspectives. J Cereal Sci.

[39] Bloemberg TG, Gerretzen J, Lunshof A, Wehrens R, Buydens LMC. Warping methods for spectroscopic and chromatographic signal alignment: A tutorial. Anal Chim Acta 2013;781:14-32.

[40] Nielsen N-PV, Carstensen JM, Smedsgaard J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. J Chromatogr A 1998;805:17-35.

[41] Savorani F, Tomasi G, Engelsen SB. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. J Magn Reson 2010;202:190-202.

[42] Giskeødegård GF, Bloemberg TG, Postma G, Sitter B, Tessem M-B, Gribbestad IS, Bathen TF, Buydens LMC. Alignment of high resolution magic angle spinning magnetic resonance spectra using warping methods. Anal Chim Acta 2010;683:1-11.

[43] Pierce KM, Hope JL, Johnson KJ, Wright BW, Synovec RE. Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis. J Chromatogr A 2005;1096:101-10.

[44] Wu W, Daszykowski M, Walczak B, Sweatman BC, Connor SC, Haselden JN, Crowther DJ, Gill RW, Lutz MW. Peak Alignment of Urine NMR Spectra Using Fuzzy Warping. J Chem Inf Model 2006;46:863-75.

[45] Vu TN, Laukens K. Getting Your Peaks in Line: A Review of Alignment Methods for NMR Spectral Data. Metabolites 2013;3:259-76.

[46] Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR metabonomics. Anal Chem 2006;78:4281-90.

[47] Craig A, Cloarec O, Holmes E, Nicholson JK, Lindon JC. Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets. Anal Chem 2006;78:2262-7.

[48] Shalabi LA, Shaaban Z, Kasasbeh B. Data Mining: A Preprocessing Engine. JCS 2006;2:735-9.

[49] Jayalakshmi T, Santhakumaran A. Statistical Normalization and Back Propagation for Classification. IJCTE 2011;3:1793-8201.

[50] Lindon JC, Nicholson JK, Holmes E. Handbook of Metabonomics and Metabolomics. Oxford, UK: Elsevier; 2007.

[51] Hendriks MMWB, van Eeuwijk FA, Jellema RH, Westerhuis JA, Reijmers TH, Hoefsloot HCJ, Smilde AK. Data-processing strategies for metabolomics studies. TrAC Trends Anal Chem 2011;30:1685-98.

[52] van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 2006;7:142.

[53] Whitfield PD, Noble PJM, Major H, Beynon RJ, Burrow R, Freeman AI, German AJ. Metabolomics as a diagnostic tool for hepatology: validation in a naturally occurring canine model. Metabolomics 2005;1:215-25.

[54] Xiaobo Z, Jiewen Z, Povey MJW, Holmes M, Hanpin M. Variables selection methods in near-infrared spectroscopy. Anal Chim Acta 2010;667:14-32.

[55] Rajalahti T, Arneberg R, Berven FS, Myhr K-M, Ulvik RJ, Kvalheim OM. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemometr Intell Lab 2009;95:35-48.

[56] Zhang W, Zhang L, Li H, Liang Y, Hu R, Liang N, Fan W, Cao D, Yi L, Xia J. GC-MS Based Serum Metabolomic Analysis of Isoflurane-Induced Postoperative Cognitive Dysfunctional Rats: Biomarker Screening and Insight into Possible Pathogenesis. Chromatographia 2012;75:799-808.

[57] Ding C, Peng H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. J Bionform Comput Biol 2005;3:185-205.

[58] Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr KJ, Kvalheim OM. Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation

and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles. Anal Chem 2009;81:2581-90.

[59] Lubbe A, Ali K, Verpoorte R, Choi YH. NMR-Based Metabolomics Analysis. In: Lammerhofer M, Weckwerth W, editors. Metabolomics in Practice: Successful Strategies to Generate and Analyze Metabolomic Data. Weinheim, Germany: Wiley-VCH; 2013.

[60] Rubingh C, Bijlsma S, Derks EPA, Bobeldijk I, Verheij E, Kochhar S, Smilde A. Assessing the performance of statistical validation tools for megavariate metabolomics data. Metabolomics 2006;2:53-61.

[61] Anderssen E, Dyrstad K, Westad F, Martens H. Reducing over-optimism in variable selection by cross-model validation. Chemometr Intell Lab Syst 2006;84:69-74.

Figures

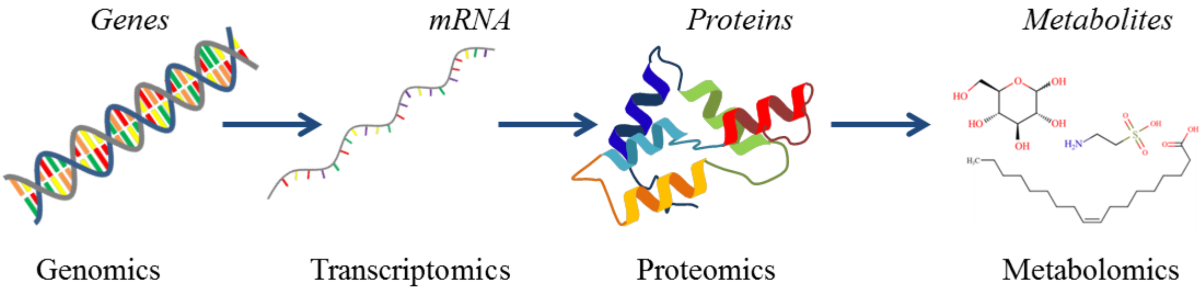


Figure 1. The Omics Cascade depicting the different molecular levels in an organism. The arrows represent biochemical processes through which information is transferred from one level to the next. Metabolomics, as the final step in the cascade, studies molecules that are closer to the phenotype.

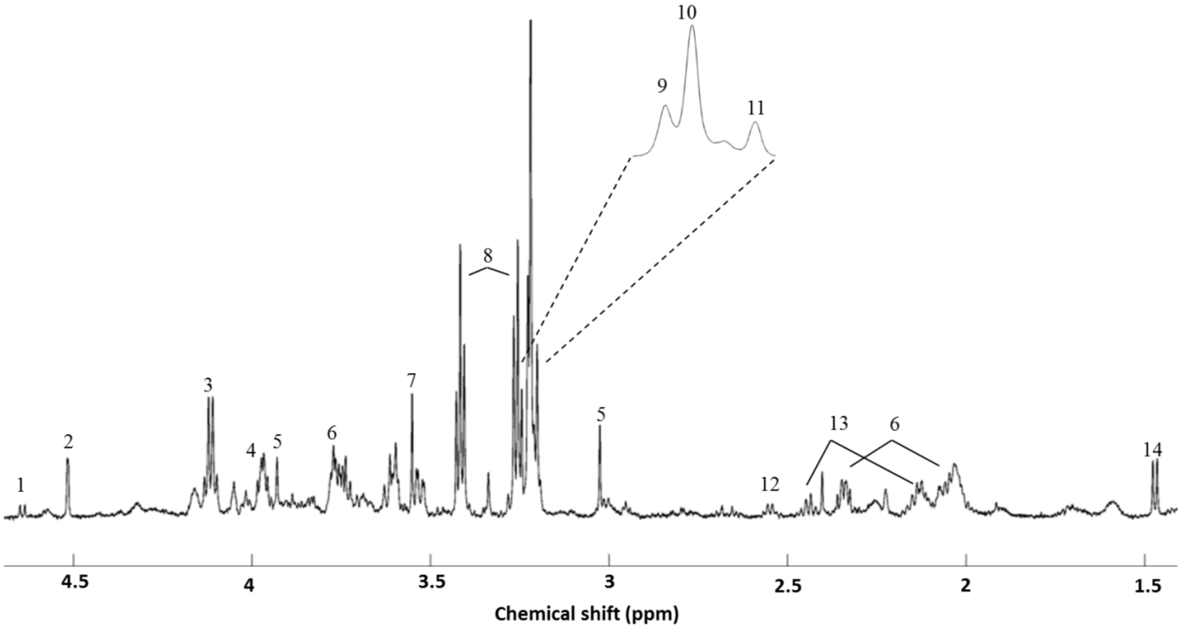


Figure 2. Proton high resolution (HR) magic angle spinning (MAS) magnetic resonance (MR) spectrum of a breast tumor biopsy. Metabolite peaks observed include 1) β-glucose, 2) ascorbate, 3) lactate, 4) tyrosine, 5) creatine, 6) glutamate, 7) glycine, 8) taurine, 9) glycerophosphocholine, 10) phosphocholine, 11) choline, 12) methionine, 13) glutamine and 14) alanine.

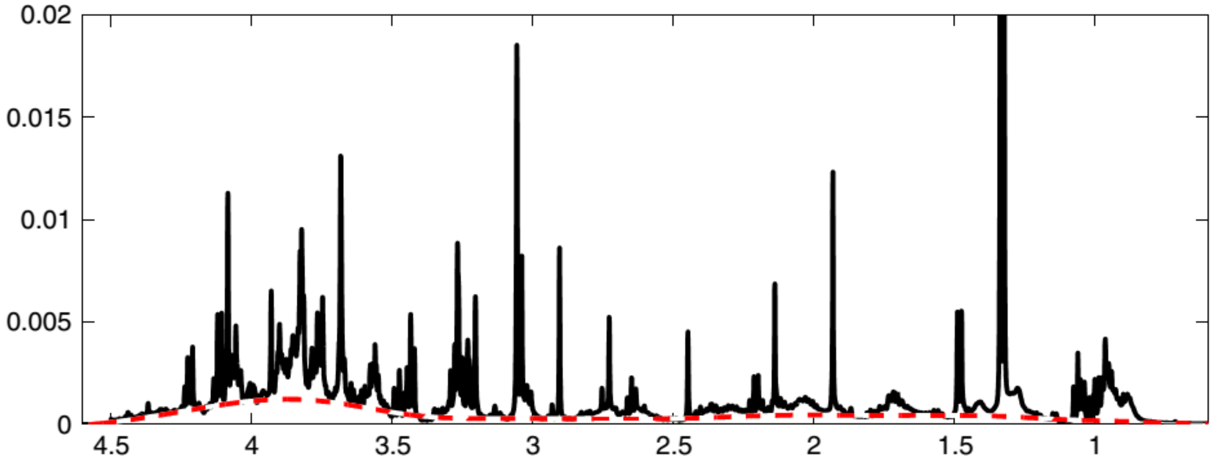


Figure 3. Baseline estimation of an NMR spectrum. Reprinted from Trends in Analytical Chemistry, Vol 30, Liland, KH. *Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis*, 827-841, Copyright 2011, with permission from Elsevier.

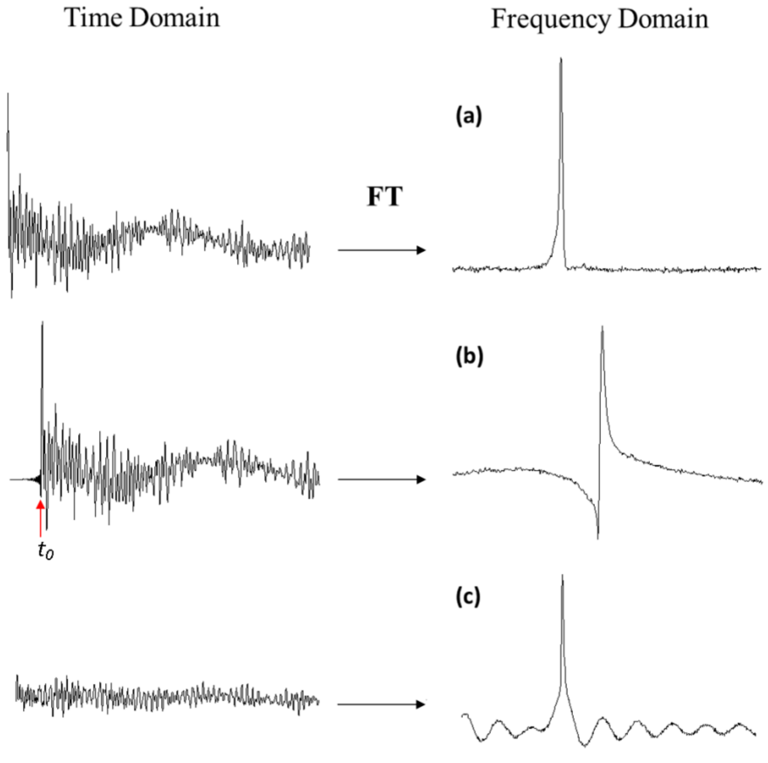


Figure 4. Baseline correction in the time domain based on proper data acquisition timing. (a) Absence of dead time leads to no baseline distortion. (b) Phase errors are observed with both dead time and signal detection starting at time zero. (c) Baseline distortion observed when signal detection starts after a dead time of t_0.

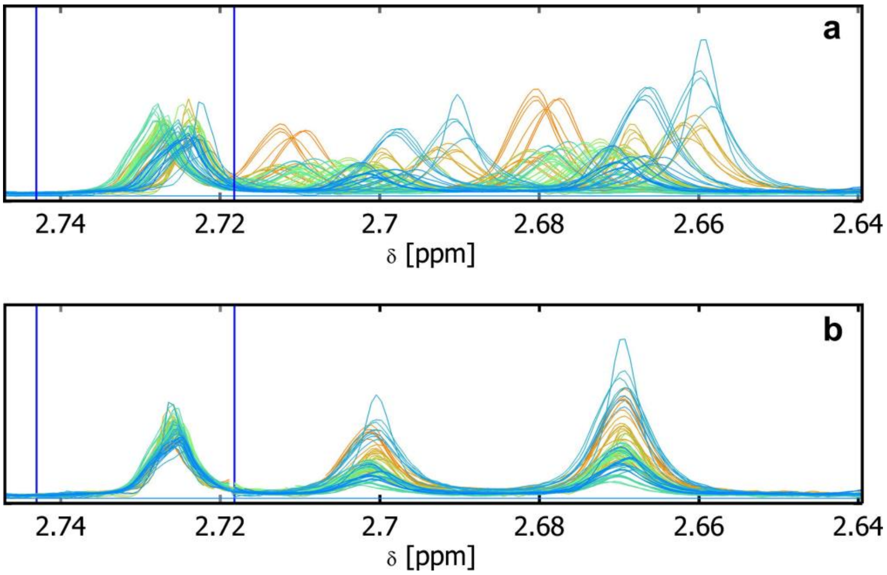


Figure 5. NMR data (a) before and (b) after alignment using the icoshift algorithm. Two intervals of different sizes defined by the user are shown here.

Reprinted from: Journal of Magnetic Resonance, Vol 202, Savorani F, Tomasi G, Engelsen SB. *icoshift: A versatile tool for the rapid alignment of 1D NMR spectra*, 190-202, Copyright 2010 with permission from Elsevier.

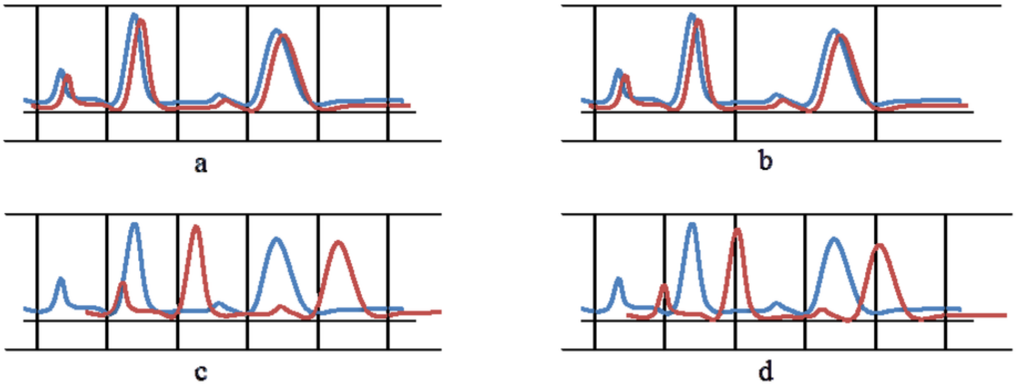
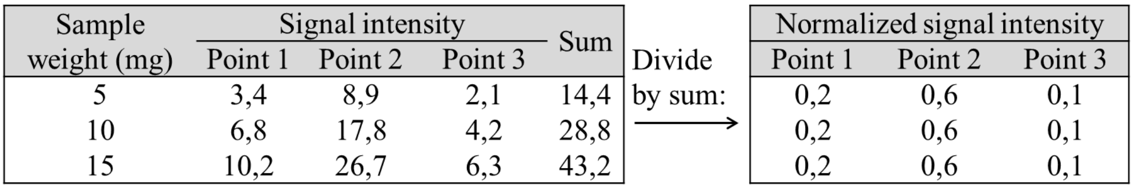


Figure 6. Hypothetical examples of how binning addresses peak shifts. (a) A good binning: each peak corresponding between the red and the blue spectra fall into the same bin. (b) Multiple peaks fall into the same bin. (c) The first peak of the red spectrum falls in the same bin of the second peak of the blue spectrum. (d) Peak shifts across the boundaries of bins. Reprinted from: Metabolites, Vol 3, Vu TN, Laukens K. *Getting Your Peaks in Line: A Review of Alignment Methods for NMR Spectral Data*, 259-276, Copyright 2013 (Open Access).



Sample weight (mg)	Signal intensity			Sum
	Point 1	Point 2	Point 3	
5	3,4	8,9	2,1	14,4
10	6,8	17,8	4,2	28,8
15	10,2	26,7	6,3	43,2

Divide by sum: →

Normalized signal intensity		
Point 1	Point 2	Point 3
0,2	0,6	0,1
0,2	0,6	0,1
0,2	0,6	0,1

Figure 7. Hypothetical area normalization to eliminate variance related to the amount of sample analyzed. The signal intensity of three sets of spectral data points acquired using different amount of the same tissue sample are divided by the sum of the intensity in each set to make each spectrum comparable.

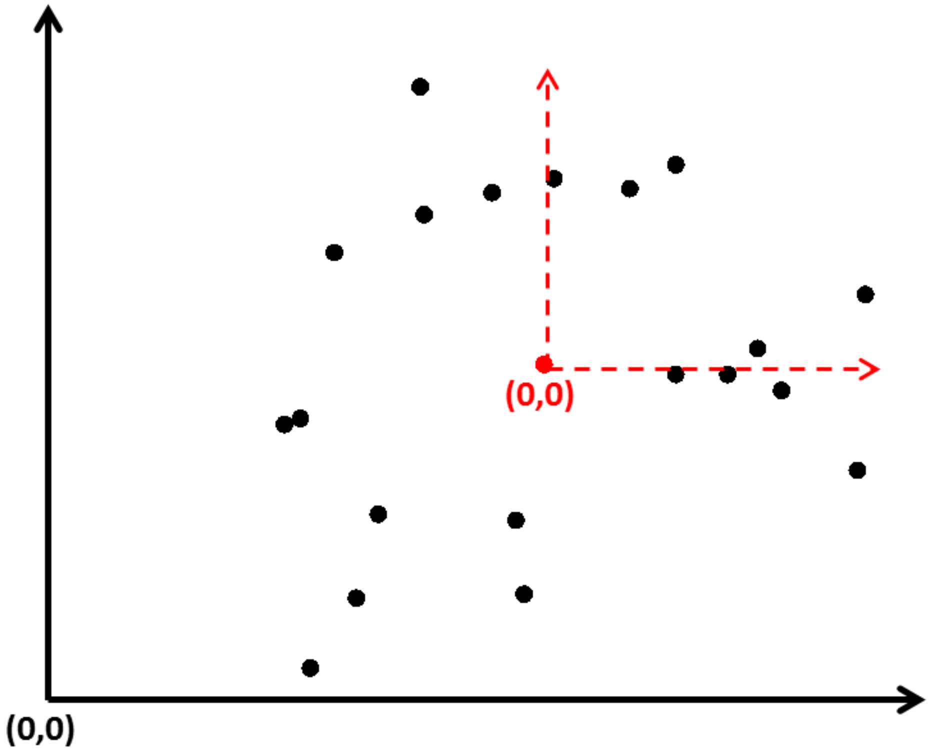
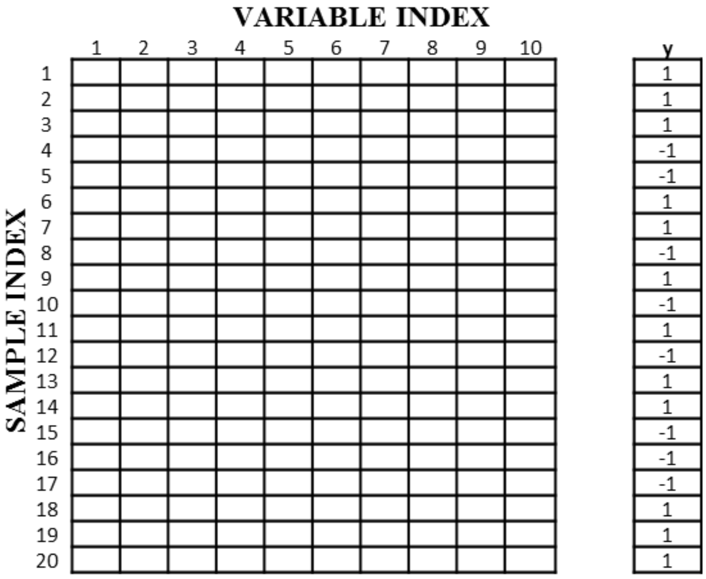


Figure 8. Depiction of data mean centering in two dimensions. Original data (black axis) vary around the mean. Mean centering shifts data left and down so that zero becomes the new centroid (red axis).



VARIABLE INDEX

SAMPLE INDEX

	1	2	3	4	5	6	7	8	9	10	y
1											1
2											1
3											1
4											-1
5											-1
6											1
7											1
8											-1
9											1
10											-1
11											1
12											-1
13											1
14											1
15											-1
16											-1
17											-1
18											1
19											1
20											1

Figure 9. Representation of a dataset consisting of 20 samples or spectra and 10 variables or spectral data points. The response variable *y* contains dummy variables where 1 represents a particular category and -1 another, for example presence and absence of a certain disease, respectively.