Characterization of $^1$H NMR spectroscopic data and the
generation of synthetic validation sets

Paul E. Anderson$^1$, Michael L. Raymer$^1$, Benjamin J. Kelly$^1$, Nicholas V. Reo$^2$, Nicholas J. DiRaso$^3$ and T. E. Doom$^1$.

$^1$Department of Computer Science and Engineering, Dayton, OH 45435, $^2$Department of Biochemistry and Molecular Biology, Boonshoft School of Medicine, Cox Institute, Dayton, OH 45429 and $^3$Air Force Research Laboratory, Biosciences and Protection Division, Wright-Patterson AFB, OH 45433, USA

Received on June 3, 2009; revised on August 28, 2009; accepted on September 7, 2009

1 INTRODUCTION

The science of metabolomics (Fiehn, 2002)—the quantitative measurement of the metabolic response of biological systems to pathology or genetic modification—is a relatively young field that requires intensive signal processing and multivariate data analysis for interpretation of experimental results. Metabolomics techniques are used to identify biomarkers associated with: responses to toxin and pathophysiologic changes (Azmi et al., 2005; Lindon et al., 2001; Shockcor and Holmes, 2002), sample classification based on the type of toxic exposure (Beckonert, 2003), large scale human studies (Bijlsma et al., 2006), clinical diagnosis (Brindle et al., 2002; Griffin et al., 2001), differential gene expression (Bundy et al., 2002; Gavaghan et al., 2000), and the study of genetic disorders (Griffin et al., 2001).

Inherent to these data-driven applications is the need for statistical and computational techniques to facilitate the associated data analysis. As such, metabolomics is particularly subject to the proliferation of data preparation and analysis methods. The selection of the most appropriate data analysis techniques is a common problem for researchers working in the ‘omics’ fields (e.g. metabolomics, proteomics and genomics) (Robertson, 2005).

The interpretation of results requires in-depth knowledge of both the biological aspects and the analytical methods. As with other modern assays, there are a wide variety of potential data-transformation methods at each of the many data analysis steps (Davis et al., 2007; Stoyanova and Brown, 2002; van den Berg et al., 2006; Webb-Robertson et al., 2005). In current practice, selection methods are based upon the type of experiment, the specific hypothesis, expediency, and investigators’ background, experience and preference. The multivariate nature of these data can yield varied results dependent upon the choice of analytical method, and are highly subject to differing interpretations (Cloarec et al., 2005; Holmes et al., 2000).

Two techniques most often used to measure metabolite concentrations are nuclear magnetic resonance (NMR) (Lindon et al., 2001) and mass spectrometry (MS). Mass spectrometry includes an on-line separation step, such as high performance liquid chromatography (LC-MS) (Wilson et al., 2005) or gas chromatography (GC-MS) (Szopa et al., 2001). Both techniques provide complementary information and can be used to analyze urine, plasma and blood. Furthermore, NMR requires little sample preparation and is non-destructive, while MS provides higher sensitivity (Bezabeh et al., 2009; Geurtsen and Wang, 2008; Lewis et al., 2008; Reo, 2002; Robertson, 2005).

NMR-based metabolomics data processing and analysis is typically divided into five steps: (i) standard post-instrumental processing of spectroscopic data, (ii) quantification of spectral...
Characterization of 1H NMR spectroscopic data

The validation sets are developed by characterizing the salient features, (iii) normalization, (iv) scaling and (v) multivariate statistical modeling of data and pattern recognition. At each one of these steps, researchers must select among several algorithms for data processing and analysis. This task of selecting the ‘best’ technique for each step is complicated by several factors, including the limited number of direct comparisons of competing techniques, the ongoing creation of novel techniques, and the application-dependent nature of selecting a technique.

Common contemporary practice within the NMR-based metabolomics community is to evaluate and validate novel algorithms on empirical data or on simplified simulated data (Davis et al., 2007; Forsshed et al., 2005; Webb-Robertson et al., 2005). Empirical data captures the complex characteristics of experimental data, but the optimal or most correct analysis is unknown a priori; researchers are forced to rely on indirect performance metrics. For example, two spectral alignment algorithms might be compared based on their ability to enhance the class separation of data after principal component analysis and partial least squares discriminant analysis (Forsshed et al., 2005). Comparison of algorithms based on their indirect performance on empirical data is of limited value. More exacting performance metrics are necessary.

In order to achieve fair and complete analysis of competing techniques, a true or ‘most correct’ analysis of that data must be known. This is demonstrated for the assessment of alignment algorithms for LC-MS by establishing a ground truth by identifying peptides (Lange et al., 2008). As an alternative, metabolomics researchers often evaluate their algorithms on simplified simulated data with a known answer (Davis et al., 2007; Webb-Robertson et al., 2005). Unfortunately, the conclusions obtained on simulated data are only of value if the data sets are complex enough for results to generalize to true experimental data.

In order for comparisons of technique performance on simulated data to be of value, the data must emulate the salient features of experimental data. Identifying the pertinent characteristics is the most critical step in generating realistic synthetic data. Ideally, synthetic data should be indistinguishable from empirical data, yet retain a ‘known’ best analysis.

Herein, we propose a technique for creating realistic synthetic metabolomics validation sets based on NMR spectroscopic data. The validation sets are developed by characterizing the salient distributions in sets of empirical spectroscopic data. Each spectrum is modeled as a combination of Gaussian-Lorentzian peaks and a piecewise cubic interpolated baseline. Using this technique, several validation sets are constructed with a variety of characteristics present in ‘real’ data. A case study is presented to compare the relative accuracy of several alignment algorithms using the increased precision afforded by these synthetic data sets (Wong et al., 2005a and b).

2 SYSTEM AND METHODS

The process of characterizing 1H NMR spectroscopic data is divided into seven general steps: (i) Collect experimental data. (ii) Divide each spectrum into segments that are individually modeled by a set of Gaussian-Lorentzian peaks and a baseline offset, where the initial locations of the peaks are manually selected. The location and other peak parameters are adjusted by a non-linear curve-fitting routine. The manual selection of the initial locations is necessary due to the level of congestion typical of a 1H NMR spectrum. The full automatic deconvolution of an entire 1H NMR spectrum is an open research problem. (iii) Combine the segments to form a global model for each spectrum that is optimized by non-linear curve-fitting. (iv) Optimize the global model until the residual can be decomposed into normally distributed regions \(\mu\neq 0\). (v) Characterize the within-peak variability by matching peaks between spectra. (vi) Characterize the baseline variability by comparing baseline intensities between spectra. (vii) Extract the distributions for the peak parameters and baseline intensities.

After the characterization of the spectroscopic data, the process of generating synthetic data begins with the three general steps: (i) Generate a reference spectrum that will serve as the base for the entire data set. This spectrum contains the parameters for each peak (e.g. height, width and location) in addition to a reference baseline. These parameters are selected from the distributions extracted in the final step of spectral characterization. (ii) Generate individual spectra by varying the peak parameters and baseline intensities from the reference spectrum according to the extracted distributions. (iii) Add Gaussian distributed noise to each spectrum.

2.1 1H Spectroscopic data

The identification of biomarkers for a specific toxin is a common research area in metabolomics (Beckwith-Hall et al., 2002; Holmes et al., 2000). Here a biomarker is defined as a set of NMR signals that change after exposure to the toxin. Such an experiment consists of at least two groups (e.g. pre- and post-dose) for which spectroscopic data is compiled. Often, this experimental data is obtained by analyzing animal urine before and after acute toxic exposure.

The synthetic data sets developed in this manuscript are analogous to a set of control samples for a typical urinary metabolomics study using a rat animal model. The NMR spectral data were processed using Varian software and employed exponential multiplication (0.3 Hz line-broadening), Fourier transformation and baseline flattening (fifth-order polynomial and cubic spline fitting routines). The TSP signal was used as an internal chemical shift reference, and the regions surrounding the residual water signal (~4.8 ppm) and the urea signal (~5.8 ppm) were excluded from the analyses. The vertical shift of the entire spectrum was adjusted such that the mean of the intensities between 11.6 and 10.0 ppm was zero. Then the peak intensities of each spectrum were normalized to a constant sum. The final data set consists of 22 1H spectra from individual normal healthy rats. Additional information on the experimental techniques is given in the Supplementary Material.

2.2 Spectra characterization

2.2.1 Modeling the spectra Each spectrum is characterized by decomposing it into its constituent components: peaks, noise and baseline.

The observable NMR free induction decay (FID) signal is an exponential function for the peak parameters and baseline intensities from the reference spectrum according to the extracted distributions. (iii) Add Gaussian distributed noise to each spectrum.

The identification of biomarkers for a specific toxin is a common research area in metabolomics (Beckwith-Hall et al., 2002; Holmes et al., 2000). Here a biomarker is defined as a set of NMR signals that change after exposure to the toxin. Such an experiment consists of at least two groups (e.g. pre- and post-dose) for which spectroscopic data is compiled. Often, this experimental data is obtained by analyzing animal urine before and after acute toxic exposure.

The synthetic data sets developed in this manuscript are analogous to a set of control samples for a typical urinary metabolomics study using a rat animal model. The NMR spectral data were processed using Varian software and employed exponential multiplication (0.3 Hz line-broadening), Fourier transformation and baseline flattening (fifth-order polynomial and cubic spline fitting routines). The TSP signal was used as an internal chemical shift reference, and the regions surrounding the residual water signal (~4.8 ppm) and the urea signal (~5.8 ppm) were excluded from the analyses. The vertical shift of the entire spectrum was adjusted such that the mean of the intensities between 11.6 and 10.0 ppm was zero. Then the peak intensities of each spectrum were normalized to a constant sum. The final data set consists of 22 1H spectra from individual normal healthy rats. Additional information on the experimental techniques is given in the Supplementary Material.

2.2 Spectra characterization

2.2.1 Modeling the spectra Each spectrum is characterized by decomposing it into its constituent components: peaks, noise and baseline.
Thus, the width of a segment is varied to avoid placing a boundary in parameters, and the initial location, $x_{cj}$, is defined as double the distance (ppm) between the maximum intensity and the middle of the peak. In congested areas of the spectrum, each segment encompasses several peaks while remaining small to allow the initial fitting routines to be performed interactively.

Following the creation of the segments, the initial locations of the peaks are interactively selected. The final locations of the peaks and their parameters (e.g. width, height) are determined algorithmically by solving the corresponding non-linear curve-fitting problem. The parameters of the non-linear curve-fitting problem are estimated by a subspace trust-region method based on the interior-reflective Newton method (Coleman and Li, 1994, 1996). The parameters are adjusted to minimize the function:

$$
\sum_{j=1}^{N} \sum_{i=1}^{m} \left( \Gamma_i \cdot \frac{\beta, i}{F(\beta, x)} - y_i \right)^2. \tag{4}
$$

where $\beta, i$ and $y_i$ are the chemical shift and intensity of the $i$-th point in the segment, $m$ is the number of data points in the segment, $\beta$ is a vector of the parameters, and $F$ is the model that will be fit by the algorithm, which is composed of Gaussian–Lorentzian peaks and a baseline offset:

$$
F(\beta, x_i) = \sum_{j=1}^{N} \left( \frac{M_j \cdot \sigma_j \cdot P_j \cdot \omega_j}{\beta} \cdot \frac{x_i - \omega_j}{\beta} \right) + O. \tag{5}
$$

where $M_j, \sigma_j, P_j$, and $\omega_j$ refer to the height, width, and center of the $j$-th peak, respectively. An illustration of this model is shown in Figure 2, where a region of a spectrum is modeled as a combination of six peaks.

The non-linear curve-fitting algorithm estimates the optimal model parameters using their initial values and bounds. The initial location, $\omega_j$, of each peak is manually selected. The initial height, $M_j$, of each peak is defined as the difference between the maximum and minimum intensities in the region surrounding the peak. The initial value of the width at half height, $\Gamma_j$, is defined as double the distance (ppm) between the maximum intensity in the region and the location of the peak’s half height (i.e. initial height divided by two). The initial SD, $\sigma_j$, can then be computed from the width at half height. The initial fraction Lorentzian, $P_j$, of each peak is defined as 0.5.

The resulting parameters are then used as inputs to a second iteration of the non-linear curve-fitting algorithm. Additionally, the newly optimized peak locations are used to update the lower and upper bounds of $x_j$. This second iteration enhances the non-linear curve-fitting algorithm’s ability to find the global optimum. Following this second iteration, the results are visually inspected as a preliminary review; a statistically based stopping criterion is introduced later. Each segment is then adjusted by adding, removing and modifying the locations of the peaks. This procedure is repeated until the model passes a visual inspection. At this point in the characterization, the goal is an approximate model for each segment. These segments will be combined to form a global model, which will be adjusted until the residual can be decomposed into independent normally distributed regions, each with a mean of zero.

After defining the initial values and bounds for the parameters, the non-linear curve-fitting algorithm optimizes the parameters to minimize the difference between the model and the original data measured by Equation (4). The fitting parameters are then used as inputs to a second iteration of the non-linear curve-fitting algorithm. Additionally, the newly optimized peak locations are used to update the lower and upper bounds of $x_j$ for each segment. This second iteration enhances the non-linear curve-fitting algorithm’s ability to find the global optimum. Following this second iteration, the results are visually inspected as a preliminary review; a statistically based stopping criterion is introduced later. Each segment is then adjusted by adding, removing and modifying the locations of the peaks. This procedure is repeated until the model passes a visual inspection. At this point in the characterization, the goal is an approximate model for each segment. These segments will be combined to form a global model, which will be adjusted until the residual can be decomposed into independent normally distributed regions, each with a mean of zero.

After defining the initial values and bounds for the parameters, the non-linear curve-fitting algorithm optimizes the parameters to minimize the difference between the model and the original data measured by Equation (4). The fitting parameters are then used as inputs to a second iteration of the non-linear curve-fitting algorithm. Additionally, the newly optimized peak locations are used to update the lower and upper bounds of $x_j$. This second iteration enhances the non-linear curve-fitting algorithm’s ability to find the global optimum. Following this second iteration, the results are visually inspected as a preliminary review; a statistically based stopping criterion is introduced later. Each segment is then adjusted by adding, removing and modifying the locations of the peaks. This procedure is repeated until the model passes a visual inspection. At this point in the characterization, the goal is an approximate model for each segment. These segments will be combined to form a global model, which will be adjusted until the residual can be decomposed into independent normally distributed regions, each with a mean of zero.

After defining the initial values and bounds for the parameters, the non-linear curve-fitting algorithm optimizes the parameters to minimize the difference between the model and the original data measured by Equation (4). The fitting parameters are then used as inputs to a second iteration of the non-linear curve-fitting algorithm. Additionally, the newly optimized peak locations are used to update the lower and upper bounds of $x_j$. This second iteration enhances the non-linear curve-fitting algorithm’s ability to find the global optimum. Following this second iteration, the results are visually inspected as a preliminary review; a statistically based stopping criterion is introduced later. Each segment is then adjusted by adding, removing and modifying the locations of the peaks. This procedure is repeated until the model passes a visual inspection. At this point in the characterization, the goal is an approximate model for each segment. These segments will be combined to form a global model, which will be adjusted until the residual can be decomposed into independent normally distributed regions, each with a mean of zero.
The curve-fitting procedure is repeated until the residual can be decomposed into independent normally distributed regions.

must be large enough to prevent the baseline from modeling individual peaks, while remaining small enough to accurately model the baseline. These intervals are automatically decreased as each region is refined. Second, with the baseline held constant, the peak fitting procedure is repeated using a sliding window of width 0.04 ppm, encompassing several peaks. The window is used to select those peaks that will be fit during the current iteration. Those peaks outside of the window are held constant. A step size of 0.01 ppm is used to provide overlap between adjacent windows. Finally, after the sliding window has covered the entire spectrum, the baseline is updated again with the peaks held constant. This procedure results in the first global model.

The noise from the amplifier of a quadrature detection magnet system has been shown to follow a white and Gaussian distribution about the baseline (Grage and Akke, 2003). Thus, the global model is interactively modified until the residual can be decomposed into independent normally distributed regions ($\mu = 0$). The Anderson–Darling test is used to determine if each region follows a normal distribution ($\alpha = 0.01$) (Stevens, 1974, 1976, 1977, 1979), and the $t$-test is used to determine if a normally distributed region has a mean of zero ($\alpha = 0.01$). The minimum width of each region is 0.025 ppm (60 data points). Each region is extended until it no longer follows a normal distribution with a mean of zero. To provide flexibility, a number of ($\alpha = 0.01$) non-normal segments are allowed between the normally distributed regions.

The number of non-normal segments is determined by the following formula:

$$\frac{(\text{max} - \text{min})}{\text{max} + \text{min}} \times \alpha$$

where $\alpha$ is the significance level, and max and min are the maximum and minimum chemical shift values of the spectrum, respectively. An example region is shown in Figure 3.

In addition to defining a stopping condition for the interactive procedure described above, analyzing the residual can be used to refine the model for each spectrum. Where two models satisfy the requirement that the residual can be decomposed into independent normally distributed regions equally well, the more parsimonious model is preferred. To achieve this objective, each peak (smallest to largest) is tested for removal from the model until the peak can be decomposed into independent normally distributed regions with a mean of zero.

This process is repeated until no additional peaks can be removed. Once this is finished, a single peak is considered as a replacement for every pair of adjacent peaks. Two potential peaks are fit independently as a single Gaussian-Lorentzian peak. The two adjacent peaks are then replaced by the single peak, if the two stopping conditions are met and the $R^2$ value is above 0.98. This is repeated until no two peaks can be combined, and results in a global model for each spectrum consisting of Gaussian-Lorentzian peaks and a cubic interpolated baseline.

Once each spectrum is modeled by a set of Gaussian–Lorentzian peaks and a piecewise cubic interpolation baseline model, the peaks are separated into three groups: baseline, background and foreground. The distinction between background and foreground facilitates the characterization of within-peak variation. Such real spectral features arise since the $^1$H NMR spectra of biofluids are very congested with overlapping peaks, or can sometimes contain naturally broad signals from proteins or lipids (more prevalent in blood samples). In urinary spectra, these broad signal regions are mostly due to numerous overlapping metabolite signals that are at or near the limits of NMR detection (sometimes referred to as chemical noise). In practice, measurement of these signals is not possible because they are too weak and poorly resolved, but their presence tends to distort the baseline; therefore, our peak-fitting algorithm must address these spectral features.

A heuristic identifies baseline peaks whose width at half height is greater than six times their height. The background and foreground peaks are distinguished by the minimum distance between a maximum and its corresponding minima, where maxima are matched to the nearest peak. The minimum distance from maximum to minimum is calculated from the model consisting of Gaussian-Lorentzian peaks and piecewise cubic interpolated baseline. If this distance is above four times the SD of the entire residual, then it is considered a foreground peak (i.e. observable). A sample illustration of a set of Gaussian-Lorentzian peaks divided into groups is shown in Figure 4.

The baseline intensities for the piecewise cubic interpolated baseline are determined in relation to the last peak. The baseline peaks are then determined in relation to the spectrum to the last peak are calculated to characterize the signal peaks. Furthermore, a second condition is added to check the local region around the selected peak, specifying that a region of 0.15 ppm (containing multiple peaks) centered on the peak can be decomposed into independent normally distributed regions with a mean of zero.

This process is repeated until no additional peaks can be removed. Once this is finished, a single peak is considered as a replacement for every pair of adjacent peaks. Two potential peaks are fit independently as a single Gaussian-Lorentzian peak. The two adjacent peaks are then replaced by the single peak, if the two stopping conditions are met and the $R^2$ value is above 0.98. This is repeated until no two peaks can be combined, and results in a global model for each spectrum consisting of Gaussian-Lorentzian peaks and a cubic interpolated baseline.
Any number of synthetic data sets can be generated from the characteristics of the data. The number of synthetic peaks and the number of baseline peaks per ppm are calculated for each spectrum. For all of the components (peaks and baseline), the relationships between the parameters must be determined to create an accurate synthetic spectrum. This relationship is evaluated using the Spearman rank correlation coefficient (\( \rho = 0.05 \) (Spearman, 1904), if the parameters do not follow a parametric distribution, otherwise, the correlation is evaluated using the Pearson correlation coefficient.

The distributions described above detail the components of a single spectrum. The baseline peaks and residual are independently generated for each spectrum, however, the variability of the signal peaks and the piecewise cubic interpolated baseline between spectra must be estimated. The magnitude of this variability can be modified when creating a validation set. The variability within each signal peak can be approximated from the foreground peaks, which can be matched between spectra. After the peaks are matched the task of modeling the within-peak variation is straightforward; however, the results of a peak-matching algorithm cannot be verified on the experimental data set. This type of evaluation will be available after the creation of a synthetic data set. The goal of characterizing the within-peak variation is to provide an approximation that will be used as a basis for the synthetic data sets. The resulting within-spectrum distributions can be varied to create several synthetic data sets to achieve a more robust validation.

The peak-matching algorithm begins by arbitrarily selecting one of the spectra to serve as a reference spectrum. The rest of the spectra are then matched to this spectrum by matching its foreground peaks to the nearest peak in the reference spectrum. If two or more peaks from the same spectrum are matched to the same reference peak, these ambiguous matches are removed from the data used to characterize within-peak variability. This algorithm will result in a set of peaks that have been matched between spectra that characterize within-peak variation. The within-peak distributions include the degree of variation of a baseline can be controlled by modifying the variability within each signal peak can be approximated from the foreground peaks, which can be matched between spectra. After the peaks are matched the variability of the within-peak variation is straight-forward; however, the results of a peak-matching algorithm cannot be verified on the experimental data set. This type of evaluation will be available after the creation of a synthetic data set. The goal of characterizing the within-peak variation is to provide an approximation that will be used as a basis for the synthetic data sets. The resulting within-spectrum distributions can be varied to create several synthetic data sets to achieve a more robust validation.

The distributions described above detail the components of a single spectrum. The baseline peaks and residual are independently generated for each spectrum, however, the variability of the signal peaks and the piecewise cubic interpolated baseline between spectra must be estimated. The magnitude of this variability can be modified when creating a validation set. The variability within each signal peak can be approximated from the foreground peaks, which can be matched between spectra. After the peaks are matched the task of modeling the within-peak variation is straightforward; however, the results of a peak-matching algorithm cannot be verified on the experimental data set. This type of evaluation will be available after the creation of a synthetic data set. The goal of characterizing the within-peak variation is to provide an approximation that will be used as a basis for the synthetic data sets. The resulting within-spectrum distributions can be varied to create several synthetic data sets to achieve a more robust validation.

The peak-matching algorithm begins by arbitrarily selecting one of the spectra to serve as a reference spectrum. The rest of the spectra are then matched to this spectrum by matching its foreground peaks to the nearest peak in the reference spectrum. If two or more peaks from the same spectrum are matched to the same reference peak, these ambiguous matches are removed from the data used to characterize within-peak variability. This algorithm will result in a set of peaks that have been matched between spectra that characterize within-peak variation. The within-peak distributions include the degree of variation of a baseline can be controlled by modifying the variability within each signal peak can be approximated from the foreground peaks, which can be matched between spectra. After the peaks are matched the task of modeling the within-peak variation is straightforward; however, the results of a peak-matching algorithm cannot be verified on the experimental data set. This type of evaluation will be available after the creation of a synthetic data set. The goal of characterizing the within-peak variation is to provide an approximation that will be used as a basis for the synthetic data sets. The resulting within-spectrum distributions can be varied to create several synthetic data sets to achieve a more robust validation.

The peak-matching algorithm begins by arbitrarily selecting one of the spectra to serve as a reference spectrum. The rest of the spectra are then matched to this spectrum by matching its foreground peaks to the nearest peak in the reference spectrum. If two or more peaks from the same spectrum are matched to the same reference peak, these ambiguous matches are removed from the data used to characterize within-peak variability. This algorithm will result in a set of peaks that have been matched between spectra that characterize within-peak variation. The within-peak distributions include the degree of variation of a baseline can be controlled by modifying the variability within each signal peak can be approximated from the foreground peaks, which can be matched between spectra. After the peaks are matched the task of modeling the within-peak variation is straightforward; however, the results of a peak-matching algorithm cannot be verified on the experimental data set. This type of evaluation will be available after the creation of a synthetic data set. The goal of characterizing the within-peak variation is to provide an approximation that will be used as a basis for the synthetic data sets. The resulting within-spectrum distributions can be varied to create several synthetic data sets to achieve a more robust validation.
This may be the result of a mixture of chemical noise in some regions and
Due to the large number of peaks (e.g. peak height and location) for the synthetic
peaks are normalized as fractional differences from their means. Then for
variability is introduced by adjusting the peak parameters based on the
The last step to creating a synthetic spectrum is to introduce variability
to the piecewise baseline. The variability of the baseline is modeled by the
2.3.4 Generating parameters Due to the large number of peaks (≈1500)
in each of the 22 spectra, sampling directly from the parameter values
and width of a peak are correlated, they must be selected from an appropriate
and width of a signal peak are correlated with the fraction Lorentzian, but that
the distance between adjacent peaks is correlated with the height and width
but not the fraction Lorentzian. To solve this problem, the height, width and
specifications for the alignment of SELDI and MALDI type clinical proteomics
data. Thus, this case study will also provide an evaluation of their
available download and have been organized according to their characteristics
2.4 Case study: comparing alignment algorithms Three preexisting alignment
advantages of using synthetic validation sets that accurately capture the
width of a signal peak are correlated with the fraction Lorentzian, but that
multivariate distribution. An example of the second case is if the height and
the underlying multivariate distribution. The final value is determined by
soriting the values for each parameter independently and then generating a
preexisting parameters, they constrain the range of values that can be selected
from the table.
2.3.6 Available data sets The procedure to generate spectral data sets can
These modifications include selecting a fraction of the peaks to create a sparser
true white thermal noise in other regions. This is modeled by estimating
by sampling the corresponding distributions.
characteristic parameter distributions for their height, width and fraction Lorentzian.
The locations of the baseline peaks are randomly selected within each segment.
The location of the first and last baseline peaks are selected by sampling the
The total number of baselines peaks for each spectrum is equivalent to the
2.3.2 Noise The SD of the noise is not constant throughout the spectrum.
This may be the result of a mixture of chemical noise in some regions and
true white thermal noise in other regions. This is modeled by estimating
the SD of the noise every 0.05 ppm with respect to the number of signal
peaks in the neighborhood (0.1 ppm). These estimates are then interpolated
to determine the SD of the noise along the entire spectrum.
2.3.4 Within-spectrum variability Each spectrum in the synthetic data set
is constructed by adding the spectrum independent components (baseline
peaks and noise) and by modifying the base spectrum. The within-peak
variability is introduced to the signal peaks, and finally, the piecewise
baseline in the synthetic data set, a target sum of squared error is estimated.
within-peak variability. The baseline peaks are introduced to each spectrum by selecting
the number of baseline peaks per segment with relation to the number of
signal peaks and the baseline intensity. The baseline peaks are generated
by sampling the characteristic parameter distributions for their height, width and fraction Lorentzian.
The locations of the baseline peaks are randomly selected within each segment.
The location of the first and last baseline peaks are selected by sampling the
corresponding distributions.
2.3.3 Noise The SD of the noise is not constant throughout the spectrum.
This may be the result of a mixture of chemical noise in some regions and
true white thermal noise in other regions. This is modeled by estimating
the SD of the noise every 0.05 ppm with respect to the number of signal
peaks in the neighborhood (0.1 ppm). These estimates are then interpolated
to determine the SD of the noise along the entire spectrum.
2.3.4 Within-spectrum variability Each spectrum in the synthetic data set
is constructed by adding the spectrum independent components (baseline
peaks and noise) and by modifying the base spectrum. The within-peak
variability is introduced to the signal peaks, and finally, the piecewise
baseline in the synthetic data set, a target sum of squared error is estimated.
within-peak variability. The baseline peaks are introduced to each spectrum by selecting
the number of baseline peaks per segment with relation to the number of
signal peaks and the baseline intensity. The baseline peaks are generated
by sampling the characteristic parameter distributions for their height, width and fraction Lorentzian.
The locations of the baseline peaks are randomly selected within each segment.
The location of the first and last baseline peaks are selected by sampling the
corresponding distributions.
2.3.3 Noise The SD of the noise is not constant throughout the spectrum.
This may be the result of a mixture of chemical noise in some regions and
true white thermal noise in other regions. This is modeled by estimating
the SD of the noise every 0.05 ppm with respect to the number of signal
peaks in the neighborhood (0.1 ppm). These estimates are then interpolated
to determine the SD of the noise along the entire spectrum.
2.3.4 Within-spectrum variability Each spectrum in the synthetic data set
is constructed by adding the spectrum independent components (baseline
peaks and noise) and by modifying the base spectrum. The within-peak
variability is introduced to the signal peaks, and finally, the piecewise
baseline in the synthetic data set, a target sum of squared error is estimated.
within-peak variability. The baseline peaks are introduced to each spectrum by selecting
the number of baseline peaks per segment with relation to the number of
signal peaks and the baseline intensity. The baseline peaks are generated
by sampling the characteristic parameter distributions for their height, width and fraction Lorentzian.
The locations of the baseline peaks are randomly selected within each segment.
The location of the first and last baseline peaks are selected by sampling the
corresponding distributions.
by recursively searching for the optimal minimum size to divide the spectra (i.e. segment size). Both algorithms require the maximum shift of a segment to be specified. This comparison used a maximum shift of 20 points (∼0.01 ppm).

The use of a synthetic data set facilitates the development of metrics that can directly measure the relative performance of the algorithms. For the alignment algorithms, the optimal alignment is known a priori. To compute the optimal alignment the peak shift is removed from each spectrum to align the peaks with the target spectrum. This alignment can then be directly compared to the alignment results from the aforementioned algorithms. This is quantified by the average sum of squares error that is defined as follows:

\[
\text{ASSEu} = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N} (y_{j,i} - a_{j,i})^2,
\]

where \(y_{j,i}\) is the perfectly aligned value of the \(i\)-th data point in the \(j\)-th spectrum, and \(M\) and \(N\) are the number of spectra to align to the target and the number of data points in each spectrum, respectively. The ASSE of the unaligned spectra is compared to the ASSE after alignment. The relative increase (RI) in ASSE measures the ability of an alignment algorithm to correct for misalignment, where a positive increase indicates an improvement. The RI metric is calculated as follows:

\[
\text{RI} = \frac{\text{ASSEu} - \text{ASSE}}{\text{ASSE}},
\]

where \(\text{ASSEu}\) is the average sum of squares error for unaligned spectra and \(\text{ASSE}\) is the average sum of squares error for aligned spectra.

### 3 RESULTS

#### 3.1 Parameters

The creation of synthetic spectral data sets begins by characterizing the underlying parameter distributions. These distributions are extracted using the procedure described in Section 2.2.2. The components of a synthetic spectrum are the signal peaks, baseline peaks, baseline intensities that define the cubic interpolated baseline and the noise. Furthermore, each spectrum is decomposed into ∼1700 peaks. Most of the parameters do not follow one of the parametric distributions listed in Section 2.2.2; therefore, they are treated as non-parametric. The exceptions include the baseline intensities and sum of squared error from the mean baseline; these parameters follow a normal distribution (\(\alpha = 0.05\)).

After analyzing each parameter individually, the relationship between parameters for each component was tested using the Spearman rank correlation. These relationships will determine the details of how a synthetic spectrum is constructed. For example, if the peak height and width are not correlated, then they can be selected independently. This procedure is described in detail in Section 2.3. The significant correlations (\(\alpha = 0.05\)) for each component are shown in Table 1.

#### 3.2 Case study: comparing alignment algorithms

The case study illustrates the advantages of using the synthetic validation sets to directly compare algorithms. Three spectral alignment algorithms were selected to test their applicability to NMR spectral data. The algorithms were compared on 30 synthetic validation sets each containing five spectra. A sample region of one of these data sets is shown in Figure 8, and additional examples are available in the Supplemental Material.

Each alignment algorithm was applied to the data sets using the first spectrum in the data set as the reference. The algorithms are tested to determine if a statistically significant positive change in RI is observed. These results are shown in Table 2. The PAFFT and RAFFT correlation alignment algorithms show a significant positive change after alignment (\(\alpha = 0.05\)). The peak matching alignment algorithm fails to improve the alignment as measured by ASSE. This is most likely a result of the congestion typical of 1H spectra according to the authors (Wong et al., 2005b).

A second comparison between RAFFT and PAFFT using the two-sample t-test showed that RAFFT was significantly better than PAFFT (\(\alpha = 0.05\)).

### Table 1. Relationships between the parameters for each of the components:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Signal peaks</th>
<th>Baseline peaks</th>
<th>Piecewise baseline</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height, (M) ↔ Width, (\sigma)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height, (M) ↔ Fraction Lorentzian, (P)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height, (M) ↔ Distance between adjacent signal peak</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Width, (\sigma) ↔ Fraction Lorentzian, (P)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of baseline peaks per ppm ↔ Number of signal peaks per ppm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline intensity ↔ Previous baseline intensities</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) Signal peaks</td>
<td>(b) Piecewise baseline</td>
<td>(c) Baseline peaks</td>
<td>(d) Noise</td>
<td></td>
</tr>
</tbody>
</table>
4 DISCUSSION

Novel algorithms for metabolomics data analysis are commonly compared and evaluated on empirical and simulated data. While simulated data is attractive as it enables the quantification of direct performance metrics, its value is directly tied to its ability to capture the salient features of experimental data. In contrast, empirical data captures the complex characteristics of experimental data, but comparisons are often formed on indirect performance metrics because the optimal or correct output is difficult to obtain a priori.

In this manuscript, we develop a technique for creating synthetic validation sets that characterize the salient features based on NMR spectroscopic data of rat urine samples from a metabolomics experiment. The validation sets were developed by modeling each spectrum as a combination of Gaussian-Lorentzian peaks and a piecewise cubic interpolated baseline. Each spectrum was constructed such that the residual could be decomposed into regions that follow a normal distribution, each with a mean of zero. The characterization time on a typical desktop machine (Pentium 4, 2 GB of RAM) averages several hours/spectrum. Using the distributions resulting from the characterization, the validation sets are automatically generated, and their running time depends on the number of data sets requested and the number of spectra for each data set. For 100 data sets with 50 spectra each, the worst-case running time for this procedure is several hours. To provide instant access to these data sets, several validation sets were constructed with a variety of characteristics and are publicly available (Anderson et al., 2009). Furthermore, additional synthetic data sets are actively being developed, including 13C NMR synthetic data sets.

Three alignment algorithms are selected to illustrate the procedure of comparing algorithms on the novel validation sets. Two of the alignment algorithms based on the cross-correlation (RAFFT and RAFFT) showed a significant positive change after alignment as measured by ASSE (α = 0.05). The peak matching alignment algorithm fails to improve the alignment. According to the authors, this is a result of the congestion typical of 1H spectra (Wong et al., 2005b). Comparing the RAFFT and RAFFT alignment algorithms, the RAFFT algorithm was significantly better than RAFFT using the two-sample t-test (α = 0.05) as measured by ASSE. This is due to the ability of the RAFFT algorithm to optimize the minimum segment size employed during the alignment. This comparison illustrates the advantages of the synthetic validation sets. A more detailed analysis of the three aforementioned algorithms in addition to other alignment algorithms is an area of future research (Forshed et al., 2005; Torgrip et al., 2003).

The case study demonstrates the procedure of comparing and validating algorithms on the novel synthetic data sets. Specifically, a direct performance metric, ASSE, is calculated using the correct spectral alignment, which is unavailable for experimental data. In addition, the distributions associated with peak specific parameters may be employed by quantification techniques as a statistical basis. The data sets will facilitate the development of novel algorithms in addition to improving the quality of algorithm comparisons. The availability of this data significantly improves the ability of researchers to select the most appropriate algorithms for their experimental data analysis.

Funding: Air Force Research Laboratory, Human Effectiveness Directorate, Biotechnology Branch (AFRL/HEPB), The Henry M Jackson Foundation for Advancement of Military Medicine, Inc. (Contract No.132633).

Conflict of Interest: none declared.

REFERENCES


P.E. Anderson et al.


