**Abstract**

**Motivation:** Common contemporary practice within the nuclear magnetic resonance (NMR) metabolomics community is to evaluate and validate novel algorithms on empirical data or simplified simulated data. Empirical data captures the complex characteristics of experimental data, but the optimal or most correct analysis is unknown a priori; therefore, researchers are forced to rely on indirect performance metrics, which are of limited value. In order to achieve fair and complete analysis of competing techniques more exacting metrics are required. Thus, metabolomics researchers often evaluate their algorithms on simplified simulated data with a known answer. 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. Ideally, synthetic data should be indistinguishable from empirical data, yet retain a known best analysis.

**Results:** We have developed 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. Using this technique, several validation sets are constructed with a variety of characteristics present in ‘real’ data. A case study is then presented to compare the relative accuracy of several alignment algorithms using the increased precision afforded by these synthetic data sets.

**Availability:** These data sets are available for download at [http://birg.cs.wright.edu/nmr_synthetic_data_sets](http://birg.cs.wright.edu/nmr_synthetic_data_sets).

**Contact:** travis.doom@wright.edu

**Supplementary information:** Supplementary data are available at Bioinformatics online.

---

**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 (Bridle 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 (Choare 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; Gerszten 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 $^1$H NMR spectroscopic data

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; Forshed 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 (Forshed 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

2.1 Spectroscopic data

The characterization of $^1$H 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 $^1$H NMR spectrum. The full automatic deconvolution of an entire $^1$H 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 ($\mu$=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 validation sets is divided into 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.

2.2 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 decaying sinusoid leading to an approximate Lorentzian peak shape after Fourier transformation. Noise and baseline distortions arise from congested areas of the spectrum with multiple overlapping peaks, naturally broad signals from proteins or lipids, and the amplifier of a quadrature detection magnet system (Grage and Akke, 2003). The peaks in this analysis are modeled by Gaussian-Lorentzian functions that are defined by the magnitude ($M$), SD of the Gaussian ($\sigma$), fraction Lorentzian ($P$), the center ($x_c$) and the width at half height of the Lorentzian ($\Gamma$):

$$S([M,\sigma,P,x_c,\Gamma])=P \times L([M,\Gamma,x_c,\sigma])+(1-P) \times G([M,\sigma,x_c,\Gamma]),$$  
(1)

$$L([M,\Gamma,x_c,\sigma])=\frac{M \times \sqrt{\pi} \times \Gamma}{4 \sigma \times (x-x_c)^2 + \sigma^2},$$  
(2)

$$G([M,\sigma,x_c,\Gamma])=\text{Max} \phi((-x-x_c)^2/(2\sigma^2)), $$  
(3)

where \(\Gamma=2\sqrt{\ln2}\sigma\) and \(P\) is a real value between 0.0 and 1.0 that weights the contribution of the Lorentzian \([L(\ldots)]\) and Gaussian \([G(\ldots)]\) functions. The mixture of the Gaussian and Lorentzian peaks is selected to provide the contribution of the Lorentzian peaks and the Gaussian peaks is set by assuming that both the height and the width at half height are the same.
Thus, the width of a segment is varied to avoid placing a boundary in the region and the location of the peak’s half height (i.e. initial height) divided by two). The initial SD, \( \sigma \), 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 initial offset, \( O \), is defined as the minimum intensity in the segment. The lower and upper bounds for parameters are defined as:

\[
0 < M_j \leq \text{MAX}_j,
\]
\[
0 < \sigma_j \leq |x_i - x_{i-1}|,
\]
\[
0 \leq P_j \leq 1.0,
\]
\[
a_j < x_i < b_j,
\]
\[
0 \leq O \leq \text{MAX}_j,
\]

where \( \text{MAX}_j \) is the maximum height in the \( j \)-th segment, and \( a_j \) and \( b_j \) are the left and right boundaries of the segment. The boundaries for location of each peak, \([a_j, b_j]\), are defined as the locations corresponding to the minimum intensities between the current peak and the adjacent peaks. In the special cases of the first and last peaks of each segment, the segment boundary is used to define the region.

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 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 \( x_i \). 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; after that, the final results are then determined by editing the algorithm’s ability to find the global optimum. The results 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 \( x_i \). 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; after that, the final results are then determined by editing the algorithm’s ability to find the global optimum.
must be large enough to prevent the baseline from modeling individual peaks, while remaining small enough to accurately model the baseline. These intensities are interpolated onto a smooth baseline. Second, with the baseline held constant, the peaks are fit 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 (Graje 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 ($\mu = 0.01$) (Stevens, 1974, 1976, 1977, 1979), and the t-test is used to determine if a normally distributed region has a mean of zero ($\mu = 0.01$). 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 small (<0.01 ppm) non-normal segments are allowed between the normally distributed regions. The number of non-normal segments is determined by the following formula:

$$\frac{(x_{\text{max}} - x_{\text{min}})}{0.01} \times \alpha$$  \hspace{1cm} (8)

where $\alpha$ is the significance level, and $x_{\text{max}}$ and $x_{\text{min}}$ 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 also 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 residual no longer satisfies the stopping condition. 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.

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 distinctness between background and foreground facilitates the characterization of within-peak variation. Such real spectral features are well distinguished. Peaks of biofluids are very congested with multiple 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.

### 2.2.2 Characterizing the variability of the spectra

The model of each spectrum is comprised of a set of Gaussian–Lorentzian peaks and a piecewise cubic interpolated baseline. Each model is constructed such that the residual can be broken into independent normally distributed regions ($\mu = 0$). All of the peaks are further divided into foreground, background and baseline. The foreground (i.e. clearly observable) peaks provide a mechanism to estimate the within-peak variability of the 22 $^1$H spectra.

The peak parameters ($M_i, \sigma_i$, and $P_i$) for the signal peaks (combination of foreground and background peaks) and the baseline peaks are tested using the Anderson–Darling statistical test ($\alpha = 0.05$) to determine if they follow one of the following parametric distributions: Weibull, exponential (specific case of the Weibull distribution), normal, lognormal and Gumbel (also known as the extreme value type 1 distribution) (Krishnamoorthy, 2006). These common distributions are tested to discover if the parameters follow any of the aforementioned underlying distributions.

The peak parameters are common to both the signal and baseline peaks; however, the signal and baseline peaks are analyzed independently. Furthermore, there are parameters that are specific to each group. This is a result of the process that will be used to create a synthetic spectrum, where the signal peaks are placed first, followed by the piecewise interpolated baseline and peak parameters. The distance between adjacent peaks, the distance from the start of the spectrum to the first peak, and the distance from the end of the spectrum to the last peak are calculated to characterize the signal peaks.

The baseline intensities for the piecewise cubic interpolated baseline are calculated in relationship to the number of signal peaks per ppm, and the previous baseline intensity. The baseline peaks are then determined in relation
to the number of signal peaks per ppm and the baseline intensity. These values are calculated for each baseline segment of size 0.05 ppm. In addition, the distance to the first baseline peak and the distance to the last baseline peak is measured. Finally, the normalized sum of squared error is calculated to capture the within-baseline variability using the following formula:

$$\text{NSSE} = \sum_i \left( \frac{\sigma_i - \mu_i}{\mu_i} \right)^2$$

where $\mu_i$ is the mean of the $i$-th baseline intensity and $\sigma_i$ is the corresponding SD.

The residual is characterized by employing a sliding window of size 0.1 ppm with a step size of 0.05 ppm to calculate the SD of the residual along the spectrum. The number of signal peaks and the number of baseline peaks per ppm are calculated for each window.

For all of the components (peaks and baseline), the relationships between the parameters must be determined to emulate an accurate synthetic spectrum. This relationship is evaluated using the Spearman rank correlation ($\alpha = 0.05$) (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 variation of the signal peaks and the piecewise cubic interpolated baseline between spectra must be estimated. The degree 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 distance from center (capturing misalignment and pH effects), difference from average height, difference from average width and the difference from average fraction Lorentzian. The Anderson–Darling statistical test ($\alpha = 0.01$) is repeated for each peak and each of the aforementioned distributions. If $<1\%$ of the tests are significant (i.e. does not follow the distribution), then the parameter is assumed to follow the test distribution.

2.3 Generating a synthetic spectral data set

Any number of synthetic data sets can be generated from the characteristics of the experimental 1H NMR spectroscopic data set. A synthetic data set is based on a single base spectrum. The base spectrum is constructed in two stages: (i) generation of the signal peaks; and (ii) generation of the piecewise cubic interpolated baseline. The data set is then constructed by modifying the base spectrum to introduce between spectra variability to emulate the 1H spectral data set. Specifically, the height, width, fraction Lorentzian and location of the peaks are altered from the base spectrum to simulate real experiments. In addition, the piecewise cubic interpolated baseline is varied between spectra. Finally, the baseline peaks and Gaussian noise are independently generated for each spectrum.

2.3.1 Signal peaks

During the first stage, the signal peaks are generated by sampling the corresponding characteristic parameter distributions for the

![Fig. 5. Generation of a new peak by sampling the distributions for the height, width at half height, fraction Lorentzian and distance between adjacent peaks.](image)

![Fig. 6. Process of generating piecewise baseline ($\Delta = 0.5$ ppm) by applying a weighted mean to the intermediate baseline.](image)
The intensities are modified according to their SD until the sum of squared error from the baseline of the base spectrum reaches the target. In the synthetic data set, a target sum of squared error is estimated from the mean baseline of the empirical data. For each segment, the variability of the baseline is modeled by the within-peak variability. The variability is introduced by adjusting the peak parameters based on the local average. The local average is defined as 1/100th the size of the entire spectrum. For each signal peak, a matched peak is randomly selected as a model for its within-peak variability.

After adding the signal peak, a matched peak is selected and used to determine the SD of the noise along the entire spectrum.

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 signal 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 of the base spectrum is modified for each spectrum. A simplified base spectrum and two synthetic spectra with peak variability are shown in Figure 7.

2.3.5 Generating parameters Due to the large number of peaks (~1500) in each of the 22 spectra, sampling directly from the parameter values is preexisting. For example, if the height and width of a peak are correlated, they must be selected from an appropriate multivariate distribution. An example of the second case is if the height 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 fraction Lorentzian are selected from a multivariate distribution. Then the height and width are used as preexisting values to select the distance between adjacent peaks.

The correlated parameters are drawn from a multivariate distribution represented as a table of values. If these parameters are not correlated to any preexisting parameters, they can be selected from a table that captures the underlying multivariate distribution. The final value is determined by sorting the values for each parameter independently and then generating a uniform random number between previous and next parameter. This resolves the problem of fixing the exact values of the parameters. When there are 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 be modified to produce validation sets with different characteristics. Some of these modifications include selecting a fraction of the peaks to create a sparser spectrum, selecting a subset of the peaks to be consistent across spectra and modifying the distributions via transformations (e.g. multiplication, addition, logarithm and exponential). In addition to generating control data sets, treatment data sets are also created with varying degrees of response. These data sets are available for download and have been organized according to their characteristics (Anderson et al., 2009).

2.4 Case study: comparing alignment algorithms Three preexisting alignment algorithms were chosen to illustrate the advantages of using synthetic validation sets that accurately capture the characteristics of empirical data (Wong et al., 2005a, b). These three alignment algorithms are available in the spectral processing software suite: SpecAlign (Wong et al., 2005a). These algorithms were developed specifically for the alignment of SELDI and MALDI type clinical proteomics data. Thus, this case study will also provide an evaluation of their applicability to NMR spectral data. The three algorithms include alignment algorithms based on peak matching or fast Fourier transform cross-correlation.

The first algorithm aligns peaks that have been automatically selected in each spectrum. Potential peaks are selected by sliding a window across the spectra to determine if there is a change in the gradient from positive to negative. These peaks are selected if they are above the baseline cutoff and also above the average intensity across the local region of the spectrum. The baseline cutoff is defined as the fraction of the baseline under the baseline intensity at which the algorithm should ignore picking peaks. The baseline is automatically determined via a restrained moving average, where only values less than the local average are added to the global moving average. The local average is defined as 1/100th the size of the entire spectrum. For the peak-picking algorithm, the default parameters were used (window: 21, baseline cutoff: 0.5, height ratio: 1.5). After the peaks have been identified, each spectrum is aligned to an arbitrarily chosen target spectrum. For each spectrum, the peaks are individually aligned to the closest peak in the target spectrum. The alignment is performed by adjusting the minima adjacent to the selected peaks, where points that are inserted are estimated by least-squares fitting about the neighboring points.

The next two alignment algorithms are based on the fast Fourier transform cross-correlation. These alignment algorithms are the peak alignment by fast Fourier transform (PAFFT) and the recursive alignment by fast Fourier transform (RAFFT). These two algorithms do not depend on peak picking and are therefore more suitable to highly congested spectra (Wong et al., 2005b). These algorithms divide the spectra into segments before the evaluation of the best shift via the fast Fourier transform cross-correlation. The recursive alignment by fast Fourier transform (RAFFT) extends PAFFT.
by recursively searching for the optimal minimum size to divide the spectra (i.e. segment size). Both comparisons used a maximum shift of 20 points (±0.01 ppm).

The creation of a synthetic spectral 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:

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

where \( y_{ij} \) 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:

\[ RI = \frac{ASSE_u - ASSE_a}{ASSE_u} \]

where ASSE\(_u\) is the average sum of squares error for unaligned spectra and ASSE\(_a\) is the average sum of squares error for aligned spectra.

3 RESULTS

3.1 Parameters

The creation of synthetic spectral data set 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 \(^1\)H 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 (\( \alpha = 0.05 \)).
Characterization of ¹H NMR spectroscopic data

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 empirical 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 ²³ 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 (PAFFT 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 PAFFT and RAFFT alignment algorithms, the RAFFT algorithm was significantly better than PAFFT 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., 2003; 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/HEBP), 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.


