2009

Characterization of 1H NMR Spectroscopic Data and the Generation of Synthetic Validation Sets

Paul E. Anderson
Wright State University - Main Campus

Michael L. Raymer
Wright State University - Main Campus, michael.raymer@wright.edu

Benjamin J. Kelly
Wright State University - Main Campus

Nicholas V. Reo
Wright State University - Main Campus, nicholas.reo@wright.edu

Nicholas J. DelRaso

See next page for additional authors

Authors
Paul E. Anderson, Michael L. Raymer, Benjamin J. Kelly, Nicholas V. Reo, Nicholas J. DelRaso, and Travis E. Doom

BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 22 2009, pages 2992–3000
doi:10.1093/bioinformatics/btp540

Data and text mining

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. DelRaso[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

Advance Access publication September 16, 2009

Associate Editor: John Quackenbush

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.

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 (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; 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

[*]To whom correspondence should be addressed.

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

The process of characterizing ^{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 (μ=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.

2.1 ^{1}H 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 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 ppm was zero. Then the peak intensities of each spectrum were normalized to a constant sum. The final data set consists of 22 ^{1}H 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 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 (σ), fraction Lorentzian (P), the center (x_c) and the width at half height of the Lorentzian (Γ):

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

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

$$G\big([M,\Gamma,x_c],x\big)=M\exp\Big(-(x-x_c)^2/(2\times\sigma^2)\Big), \quad (3)$$

where $\Gamma=2\sqrt{2\ln 2}\sigma$, 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 a flexible peak shape. The relationship between the width at half height of the Lorentzian peak and the SD of the Gaussian peak is fixed by assuming that both the height and the width at half height are the same

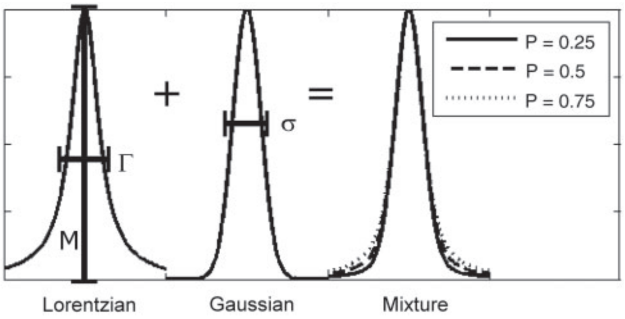

Fig. 1. Graphical representation of the construction of a Gaussian-Lorentzian peak and resulting mixture for different ratios of *P*.

for both peaks. This simplifies the model by avoiding a separate parameter for both the SD and width at half height. A graphical representation of the Gaussian–Lorentzian peak is shown in Figure 1.

The first step in decomposing a spectrum is to divide it into regions separated by signal that has been removed (e.g. the water signal). For the spectra considered in this paper, the signals that have been removed are the water, urea and TSP signals. This results in two independent regions divided by the water and urea signals. These regions are then divided into non-overlapping segments ranging from 0.05 to 0.15 ppm in size. To determine these segments, the spectrum is divided into uniform segments of size 0.05 ppm. Then the location of the minimum intensity (local minimum) in each odd numbered segment defines the adjusted segment boundaries. Thus, the width of a segment is varied to avoid placing a boundary in the middle of a 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:

$$\frac{1}{2}\sum_{i}^{m}\left(F(\beta,x_i)-y_i\right)^2, \tag{4}$$

where x_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, β 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:

$$F(\beta,x_i)=\sum_{j=1}^{N}S\left([M_j,\sigma_j,P_j,x_{cj}],x_i\right)+O, \tag{5}$$

where $[M_j,\sigma_j,P_j,x_{cj}]$ and the baseline offset O(constant for an entire segment) refer to parameters in the vector β. The parameters M_j, σ_j, P_j and x_{cj} refer to the height, SD, fraction of Lorentzian 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, x_{cj}, 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, Γ_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, σ_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.

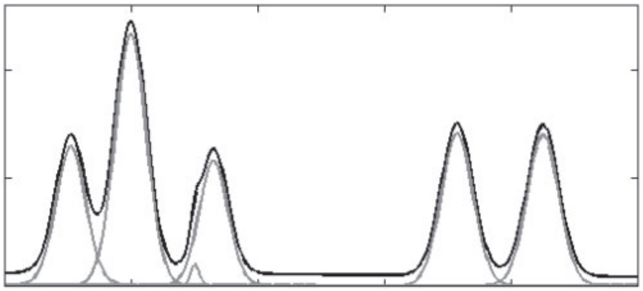

Fig. 2. Sample region of a spectrum decomposed into six peaks modeled as Gaussian–Lorentzian functions with a baseline offset.

The initial offset, O, is defined as the minimum intensity in the segment. The lower and upper bounds for parameters are defined as:

$$\begin{array}{rcccl} 0 & < & M_j & \leq & MAX_j, \\ 0 & < & \sigma_j & \leq & |s_L - s_R|, \\ 0 & \leq & P_j & \leq & 1.0, \\ \alpha_j & < & x_{cj} & < & \omega_j, \\ 0 & \leq & O & \leq & MAX_j, \end{array} \tag{6}$$

where MAX_j is the maximum height in the j-th segment, and s_L and s_R are the left and right boundaries of the segment. The boundaries for location of each peak, $[\alpha_j,\omega_j]$, 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_{cj}. 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 the segments are modeled individually, all of the segments are combined to obtain a global model, which is defined as follows:

$$\Theta(\beta,x_i)=\sum_{j=1}^{N}S\left([M_j,\sigma_j,P_j,x_{cj}],x_i\right)+\text{baseline}(\beta,x_i), \tag{7}$$

where $\Theta(\beta,x_i)$ is the global model with the model parameters, β. Furthermore, N is the number of peaks in the entire spectrum, thus, M_j, σ_j, P_j and x_{cj} refer to the height, SD, fraction of Lorentzian, and center of the j-th peak. The baseline model, baseline(β,x_i), is the piecewise cubic interpolation of baseline intensities (i.e. height of the baseline) spaced 0.05 ppm apart (Fritsch and Carlson, 1980). The baseline intensities are parameters of the model (β), and thus, they are determined by the non-linear curve-fitting algorithm.

Due to the large number of peaks (i.e. parameters) in the rat urine spectra described above, the global model is fit in an iterative fashion. First, the peaks determined from independently fitting the segments are held constant as the baseline model is fit. The initial values of the baseline intensities are the offsets of the independent segments. These baseline intensities are uniformly spaced at an interval of 0.05 ppm. The interval between baseline intensities

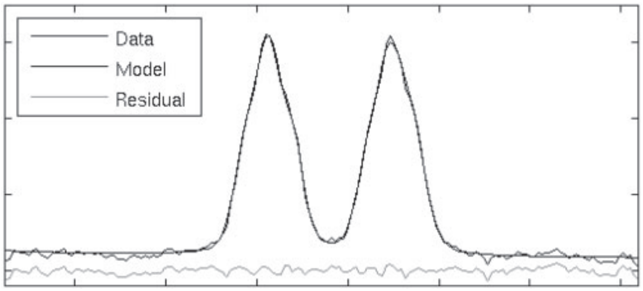

Fig. 3. Illustration of a region showing the residual (i.e. signal minus model). The curve-fitting procedure is repeated until the residual can be decomposed into independent normally distributed regions.

must be large enough to prevent the baseline from modeling individual peaks, while remaining small enough to accurately model the baseline. These intensities are interpolated to create 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 (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 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:

$$\frac{(x_{\max} - x_{\min})}{0.01} \times \alpha \qquad (8)$$

where α is the significance level, and $x_{\max}$ and $x_{\min}$ 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.

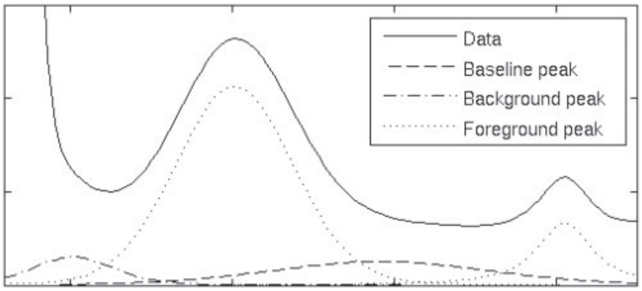

Fig. 4. Illustration of a set of Gaussian–Lorentzian peaks divided into three groups: foreground, background and baseline.

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 spectra 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_j, σ_j, and P_j) 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 baseline peaks. 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:

$$\text{NSSE} = \sum_i \left(\frac{\mu_i - \text{sigma}_i}{\mu_i} \right)^2, \tag{9}$$

where μ_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 create 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 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 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 ^{1}H 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 ^{1}H 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 height, width, fraction Lorentzian and location. For example, the positions of the peaks are determined by sampling the distance between adjacent peaks distribution, and the heights of each peak are selected by sampling the peak height distribution. The generation of a new peak is illustrated in Figure 5. The location of the first and last signal peaks are selected by sampling the corresponding distributions.

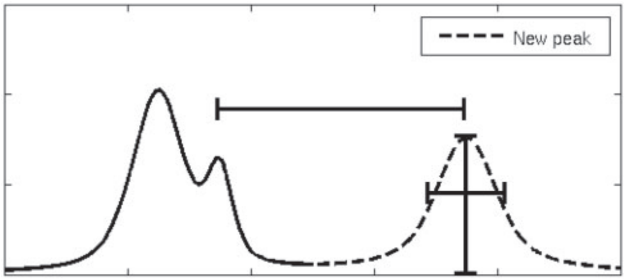

Fig. 5. Generation of a new peak by sampling the distributions for the height, width at half height, fraction Lorentzian and distance between adjacent peaks.

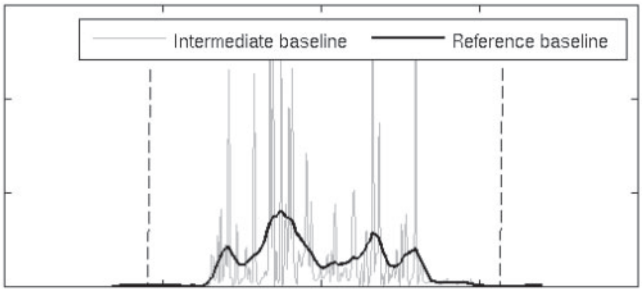

Fig. 6. Process of generating piecewise baseline ($\Delta = 0.5$ ppm) by applying a weighted mean to the intermediate baseline.

2.3.2 Baseline The second component, the baseline, is composed of a piecewise cubic interpolated baseline of uniform segments of size 0.05 ppm and baseline peaks. The baseline is divided into three regions to accurately model segments of the baseline with different characteristics. The first and third regions contain the baseline intensities (i.e. height of the baseline) from the beginning of the spectrum to the first peak and the baseline intensities from the end of the spectrum to the last peak, respectively. The second region consists of the intervening baseline intensities. The first and third regions remain relatively flat, while the third region contains the majority of the baseline distortion. The process of generating a synthetic baseline is shown in Figure 6, where the first step is to generate a reference baseline that will serve as a base for the baselines of the individual spectra.

The reference baseline is generated by smoothing an intermediate baseline that is created by sampling the baseline intensities distributions. The reference baseline intensities, s_i, are computed as the weighted average of the adjacent intermediate baseline intensities, u_j, within a minimum distance, Δ:

$$s_i = \sum_j w_{ij} u_j, \tag{10}$$

$$w_{ij} = \frac{1 - |x_i - x_j|}{\sum_{\forall k} 1 - |x_i - x_k|}, \tag{11}$$

where $|x_i - x_j|$ is the distance between the baseline intensities. Furthermore, the degree of variation of a baseline can be controlled by modifying the minimum distance, Δ (i.e. for a gradual baseline use a large minimum distance).

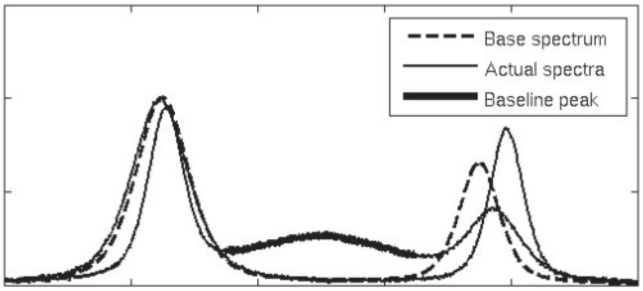

Fig. 7. Two simplified spectra with their associated base spectrum.

The individual spectrum baselines are generated to conform to the overall shape of the reference baseline. The amount of variation from the base spectrum is determined by generating a target NSSE. A specific baseline is then generated from the reference baseline by individually adjusting its intensities using their corresponding SDs. The SDs control the regions of the spectrum that have higher variability (i.e. the third region). These intermediate intensities are then smoothed according to Equation (11). The smoothed intensities are then iteratively adjusted until they reach the target NSSE.

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 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.

After adding the spectrum independent components, the within-peak variability is introduced by adjusting the peak parameters based on the matched foreground peak distributions. The parameter values of the matched peaks are normalized as fractional differences from their means. Then for each signal peak, a matched peak is randomly selected as a model for its within-peak variability.

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 sum of squared error from the mean baseline of the empirical data. For each baseline in the synthetic data set, a target sum of squared error is estimated. The intensities are modified according to their SD until the sum of squared error from the baseline of the base spectrum reaches the target.

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 approximates the actual distribution. The method for selecting a parameter (e.g. peak height and location) for the synthetic spectrum depends on whether that parameter is correlated with one or more parameters, and whether the values of any of these parameters are 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, then 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 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:

$$\mathrm{ASSE} = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N}\left(y_{j,i}-a_{j,i}\right)^2}{M}, \qquad (12)$$

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:

$$\mathrm{RI} = \frac{\mathrm{ASSE}_u - \mathrm{ASSE}_a}{\mathrm{ASSE}_u}, \qquad (13)$$

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.

Table 1. Relationships between the parameters for each of the components: (a) signal peaks, (b) baseline peaks, (c) piecewise baseline and (d) noise

Height, M ↔ Width, σ
Height, M ↔ Fraction Lorentzian, P
Height, M ↔ Distance between adjacent signal peak
Width, σ ↔ Fraction Lorentzian, P
Width, σ ↔ Distance between adjacent signal peaks
(a) Signal peaks

Height, M ↔ Width, σ
Width, σ ↔ Fraction Lorentzian, P
Number of baseline peaks per ppm ↔ number of signal peaks per ppm
Number of baseline peaks per ppm ↔ baseline intensities
(b) Piecewise baseline

Baseline intensity ↔ number of signal peaks per ppm
Baseline intensity ↔ Previous baseline intensities
(c) Baseline peaks

Standard deviation,SD ↔ number of signal peaks per ppm
(d) Noise

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.

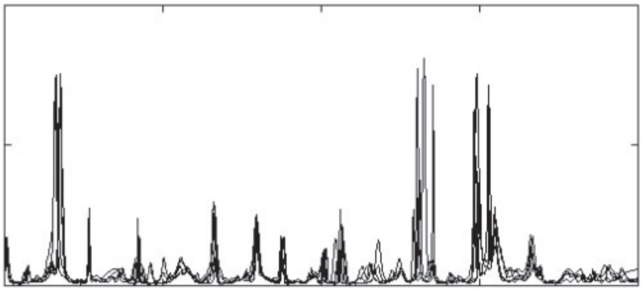

Fig. 8. Sample region of one of the synthetic data sets used in the evaluation of the alignment algorithms (∼4.5 ppm).

Table 2. Average and SD of the RI of the ASSE for 30 synthetic data sets

Algorithm	Average	Standard deviation	P-value
Peak matching method	3.59	12.33	0.0935
PAFFT correlation method	2.78	3.09	0.0002
RAFFT correlation method	5.16	7.30	0.0016

The P-value of applying the t-test to determine if the RI is >0.

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 ($\alpha = 0.05$).

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 ^{13}C 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 ($\alpha = 0.05$). The peak matching alignment algorithm fails to improve the alignment. According to the authors, this is a result of the congestion typical of ^{1}H 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 ($\alpha = 0.05$) 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/HEPB), The Henry M Jackson Foundation for Advancement of Military Medicine, Inc. (Contract No.132633).

Conflict of Interest: none declared.

REFERENCES

Anderson,P.E. *et al*. (2009) Nuclear magnetic resonance synthetic validation sets. Available from: http://birg.cs.wright.edu/nmr_synthetic_data_sets

Azmi,J. *et al*. (2005) Chemometric analysis of biofluids following toxicant induced hepatotoxicity: a metabonomic approach to distinguish the effects of 1-naphthylisothiocyanate from its products. *Xenobiotica: Fate Safety Eval. Foreign Comp. Biol. Syst.*, **35**, 839.

Beckonert, (2003) NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. *Anal. Chim. Acta*, **490**, 3.

Beckwith-Hall,B.M. *et al*. (2002) NMR-based metabonomic studies on the biochemical effects of commonly used drug carrier vehicles in the rat. *Chem. Res. Toxicol.*, **15**, 1136.

Bezabeh,T. *et al*. (2009) Detecting colorectal cancer by 1H magnetic resonance spectroscopy of fecal extracts. *NMR Biomed.*, **22**, 593–600.

Bijlsma,S. *et al*. (2006) Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anal. Chem.*, **78**, 567.

Brindle,J.T. *et al*. (2002) Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nature Med.*, **8**, 1439.

Bundy,J.G. *et al*. (2002) Earthworm species of the genus Eisenia can be phenotypically differentiated by metabolic profiling. *FEBS Lett.*, **521**, 115.

Cloarec,C. *et al*. (2005) Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of Biomarker changes in 1H NMR spectroscopic metabonomic studies. *Anal. Chem.*, **77**, 517–526.

Coleman,T.F. and Li,Y.Y. (1994) On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds. *Math. Prog.*, **67**, 189–224.

Coleman,T.F. and Li,Y.Y. (1996) An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.*, **6**, 418–445.

Davis,R.A. *et al*. (2007) Adaptive binning: an improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometr. Intell. Lab. Syst.*, **85**, 144–154.

Fiehn,O. (2002) Metabolomics—the link between genotypes and phenotypes. *Plant Mol. Biol.*, **48**, 155.

Forshed,J. *et al*. (2003) Peak alignment of NMR signals by means of a genetic algorithm. *Anal. Chim. Acta*, **487**, 189–199.

Forshed,J. *et al*. (2005) A comparison of methods for alignment of NMR peaks in the context of cluster analysis. *J. Pharm. Biomed. Anal.*, **38**, 824–832.

Fritsch,F.N. and Carlson,R.E. (1980) Monotone piecewise cubic interpolation. *SIAM J. Numer. Anal.*, **17**, 238–246.

Gavaghan,C.L. *et al*. (2000) An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: application to the C57BL10J and Alpk:ApfCD mouse. *FEBS Lett.*, **484**, 169.

Gerszten,R.E. and Wang,T.J. (2008) The search for new cardiovascular biomarkers. *Nature*, **451**, 949–952.

Grage,H. and Akke,M. (2003). A statistical analysis of NMR spectrometer noise. *J. Magn. Reson.*, **162**, 176–188.

Griffin,J.L. *et al*. (2001) Metabolic profiling of genetic disorders: a multitissue 1H nuclear magnetic resonance spectroscopic and pattern recognition study into dystrophic tissue. *Anal. Biochem.*, **293**, 16–21.

Holmes,E. *et al*. (2000) Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chem. Res. Toxicol.*, **13**, 471–478.

Krishnamoorthy,K. (2006) *Handbook of Statistical Distributions with Applications*. Boca Raton, Chapman & Hall.

Lange,E. *et al*. (2008) Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, **9**, 375.

Lewis,G.D. *et al*. (2008) Metabolite profiling of blood from individuals undergoing planned myocardial infarction reveals early markers of myocardial injury. *J. Clin. Invest.*, **118**, 3503–3512.

Lindon,J.C. *et al*. (2001) Pattern recognition methods and applications in biomedical magnetic resonance. *Progr. Nucl. Magn. Reson. Spectr.*, **39**, 1.

Reo,N.V. (2002) NMR-based metabolomics. *Drug Chem. Toxicol.*, **25**, 375–382.

Robertson,D.G. (2005) Metabonomics in toxicology: a review. *Toxicol. Sci.*, **85**, 809–822.

Shockcor,J.P. and Holmes,E. (2002) Metabonomic applications in toxicity screening and disease diagnosis. *Curr. Topics Med. Chem.*, **2**, 35.

Spearman,C. (1904) The proof and measurement of association between two things. *Amer. J. Psychol.*, **15**, 72–101.

Stevens,M.A. (1974) EDF statistics for goodness of fit and some comparisons. *J. Amer. Statist. Assoc.*, **69**, 730–737.

Stevens,M.A. (1976) Asymptotic results for goodness-of-fit statistics with unknown parameters. *Ann. Statist.*, **4**, 357–369.

Stevens,M.A. (1977) Goodness of fit for the extreme value distribution. *Biometrika*, **64**, 583–588.

Stevens,M.A. (1979) Tests of fit for the logistic distribution based on the empirical distribution function. *Biometrika*, **66**, 591–595.

Stoyanova,R. and Brown,T.R. (2002) NMR spectral quantitation by principal component analysis. *J. Magn. Reson.*, **154**, 163–175.

Szopa,J. *et al*. (2001) Identification and quantification of catecholamines in potato plants (Solanum tuberosum) by GC-MS. *Phytochemistry*, **58**, 315.

Torgrip,R.J.O. *et al*. (2003) Peak alignment using reduced set mapping. *J. Chemometr.*, **17**, 573–582.

van den Berg,R.A. *et al*. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, **7**, 142.

Webb-Robertson,B.-J.M. *et al*. (2005) A study of spectral integration and normalization in NMR-based metabonomic analyses. *J. Pharm. Biomed. Anal.*, **39**, 830–836.

Wilson,I.D. *et al*. (2005) HPLC-MS-based methods for the study of metabonomics. *J. Chromatogr. B*, **817**, 67.

Wong,J.W. *et al*. (2005a) SpecAlign—processing and alignment of mass spectra datasets. *Bioinformatics*, **21**, 2088–2090; Available at http://physchem.ox.ac.uk/∼jwong/specalign/.

Wong,J.W. *et al*. (2005b) Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Anal. Chem.*, **77**, 5655–5661.