3-2012

Localized Deconvolution: Characterizing NMR-Based Metabolomics Spectroscopic Data using Localized High-Throughput Deconvolution

Paul E. Anderson
Wright State University - Main Campus

Ajith H. Ranabahu
Wright State University - Main Campus

Deirdre A. Mahle

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

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

Localized Deconvolution: Characterizing NMR-based Metabolomics Spectroscopic Data using Localized High-throughput Deconvolution

Paul E. Anderson[1], Ajith H. Ranabahu[2], Deirdre A. Mahle[3], Nicholas V. Reo[4], Michael L. Raymer[2], Amit P. Sheth[2], and Nicholas J. DelRaso[3]

[1]Department of Computer Science, College of Charleston, Charleston, SC 29424

[2]Department of Computer Science, Wright State University, Dayton, OH 45435

[3]711 Human Performance Wing, Air Force Research Laboratory, Wright-Patterson AFB, OH

[4]Department of Biochemistry & Molecular Biology, Wright State University, Dayton, OH

Abstract— *The interpretation of nuclear magnetic resonance (NMR) experimental results for metabolomics studies requires intensive signal processing and multivariate data analysis techniques. Standard quantification techniques attempt to minimize effects from variations in peak positions caused by sample pH, ionic strength, and composition. These techniques fail to account for adjacent signals which can lead to drastic quantification errors. Attempts at full spectrum deconvolution have been limited in adoption and development due to the computational resources required. Herein, we develop a novel localized deconvolution algorithm for general purpose quantification of NMR-based metabolomics studies. Localized deconvolution decreases average absolute quantification error by 97% and average relative quantification error by 88%. When applied to a ^1H metabolomics study, the cross-validation metric, Q^2, improved 16% by reducing within group variability. This increase in accuracy leads to additional computing costs that are overcome by translating the algorithm to the map-reduce design paradigm.*

Keywords: Metabolomics, quantification, map-reduce, deconvolution

Web: http://ds.cs.cofc.edu/index.php/Localized_Deconvolution

Contact: andersonp@cs.cofc.edu

1. Introduction

Metabolomics, the measurement of metabolite concentrations and fluxes in various biological systems, is one of the most comprehensive of all bionomics [1]. Unlike proteomics and genomics that assess intermediate products, metabolomics assesses the end product of cellular function, metabolites. Changes occurring at the level of genes and proteins (assessed by genomics and proteomics) may or may not influence a variety of cellular functions. But metabolomics, by contrast, assesses the end products of cellular metabolic function, such that the measured metabolite profile reflects the cellular metabolic status. For instance, a disease process or exposure to a xenobiotic may interfere at the genomic or proteomic level, while it will always manifest itself at the metabolomic level. Further, nuclear magnetic resonance (NMR) spectroscopy of biofluids has been shown to be an effective method in metabolomics to identify variations in biological states [2], [3]. In contrast to various other proteomic, genomic, and metabolomic analyses, NMR spectroscopy is non-invasive, non-destructive, and requires little sample preparation [1].

Typically, NMR metabolic spectroscopic data are analyzed as follows: (1) standard post-instrumental processing of spectroscopic data, such as the Fourier transformation, phase adjustment, and baseline correction; (2) quantification of spectral signals commonly implemented via binning; (3) normalization and scaling; and (4) multivariate statistical modeling of data. Quantification of spectral signals, step (2), is a key step in the development of classification algorithms and biomarker identification (i.e., pattern recognition). A common method of quantification employed by the NMR community is known as binning or bucketing, which divides a NMR spectrum into several hundred regions. This technique is performed to (1) minimize effects from variations in peak positions caused by sample pH, ionic strength, and composition (Spraul et al. 1994); and (2) reduce the dimensionality for multivariate statistical analyses. The result is a data set with fewer features, thereby, increasing the tractability of pattern recognition techniques, such as principal component analysis (PCA) [4] and partial least squares discriminant analysis (PLS-DA) [5].

The standard quantification method is to divide a spectrum into several hundred non-overlapping regions or bins of equal size. This simple technique has been shown to be effective in the field of metabolomics [6], [7]. While standard quantification mitigates the effects from variations in peak positions, shifts occurring near the boundaries can result in dramatic quantitative changes in the adjacent bins due to the non-overlapping boundaries. This problem can be countered by incorporating a kernel-based binning method that weights the contribution of peaks by their distance from the center of the bin [8] or by dynamically determining the size and location of each bin [9], [10]; however, these techniques fail to remove irrelevant adjacent signals.

There are several alternatives to spectral binning that still provide data dimension reduction [11]. Examples of these

include PARS [12], direct quantification [13], peak alignment tools in HiRes [14], and targeted profiling [15]. These techniques identify peaks or specific peak patterns in the spectra that are conserved across spectra. After the patterns have been identified, they are quantified by determining the peak area or amplitude. The accuracy of these algorithms is dependent on the spectral resolution, the quality of the peak alignment, and the breadth of spectroscopic pattern databases. Since spectral resolution is dependent upon the magnetic field strength (i.e., instrument specific), the spectral patterns in complex mixtures (e.g., urine and plasma) are also field dependent. This adds another level of complexity to targeted profiling techniques that attempt to match spectral patterns against standard spectra acquired at a specific magnetic field.

Despite the development of these alternative quantification techniques, binning remains a common technique for the NMR community owing to high throughput quantification technique [16], [11]. The wide spread use of advanced quantification algorithms has been hindered by the additional computing resources and manual intervention required to incorporate them into general metabolomics workflows. Herein, we propose a novel localized deconvolution algorithm for NMR spectroscopic data that removes adjacent and convoluting signal for significantly improved full spectrum quantification that does not rely on the breadth of annotated spectral databases. By pursuing a localized strategy for deconvolution, the algorithm is suited for implementation in the map-reduce paradigm that will allow for web-scale high-throughput availability. We show this technique is superior to alternative high-throughput quantification techniques by comparing the improvement in quantification accuracy on complex ^{1}H NMR spectroscopic data and realistic synthetic spectra.

2. Approach

The variability and complexity inherent in ^{1}H NMR spectra of biofluids requires sensitive signal processing and pattern recognition techniques to discover novel patterns in the data. The technique of spectral quantification is a general signal processing technique that reduces the dimensionality of spectroscopic data by transforming full resolution spectra into a feature vector for subsequent pattern recognition. The goals of which are to retain pertinent information and mitigate quantitative effects of peak misalignment. Biomarker identification can then be defined as finding a set of features that describe a pattern between groups, thus the success of biomarker identification is directly related to the quality of the feature vectors. Here a biomarker is defined as a set of NMR signals that change relative to some reference (i.e., before and after exposure to a toxin). Such an experiment will have at least two groups (e.g., pre-dose and post-dose) for which spectroscopic data is compiled. A significant step prior to biomarker identification is spectral quantification, our method, localized deconvolution, is comprised of three steps:

1) Solve the peak registration (correspondence) problem using an adaptive binning approach
2) Model the signals in each region using a Gauss-Lorentzian peak construct
3) Deconvolve the localized subproblem by removing adjacent and baseline signals

This technique is applied to a metabolomics study of toxicology for the identification of biomarkers associated with a kidney toxin (α-naphthylisothiocyanate) response.

3. Methods

3.1 Peak registration

The first step in localized deconvolution is to define the subproblems of interest, which are defined as regions containing a signal of interest across spectra. This problem, also known as the peak registration or correspondence problem, is solved by applying an adaptive binning technique: dynamic adaptive binning [9]. Peak registration is necessary to overcome the variability in signals between subjects (or samples). Our localized deconvolution technique leverages an adaptiving binning technique to generate the regions of interest, which can subsequently be solved in parallel; however, our method can be easily adapted to other methods of registration, including peak alignment and targeted approaches.

Dynamic adaptive binning determines the optimal bin configuration of n observed peaks as measured by an objective function. This process is divided into two steps: (1) determining the location of the observed peaks in each spectra and (2) finding the optimal bin boundaries with respect to the objective function. The identification of the observed peaks in each spectrum is accomplished by identifying local maxima after smoothing via a wavelet transform [17], [18], [19], [20], [21]. After the observed peaks of each spectrum have been determined, the algorithm determines the optimal bin configuration using a dynamic programming strategy. A detailed description of dynamic adaptive binning and proofs verifying optimal substructure can be found in [9].

3.2 Model the signals

While peak registration provides a mechanism for matching corresponding signals between spectra, quantification is still impaired by adjacent signal and baseline distortions. This problem is mitigated by removing adjacent signals that affect the true value of the signal of interest. The observable NMR free induction decay (FID) signal is an exponential decaying sinusoid leading to an approximate Lorentzian peak shape after Fourier transformation. These individual signals, S, are modeled by a Gaussian-Lorentzian function that is defined by the standard deviation of the Gaussian (σ), the

center (x_c), the width at half height of the Lorentzian (Γ), and the magnitude (M):

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

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

$$G([M\sigma,P,x_c],x) = Mexp(-(x-x_c)^2/(2\sigma^2)) \quad (3)$$

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 ($L(...)$) and Gaussian ($G(...)$) 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 standard deviation of the Gaussian peak is fixed by assuming that both the height and the width at half height are the same for both peaks. This simplifies the model by avoiding a separate parameter for both the standard deviation and width at half height.

3.3 Deconvolve

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 [22]. With the previously described model for the underlying signals, our algorithm removes unwanted signals from the region of interest. This deconvolution procedure divides each spectral subproblem into it's constituent signals (baseline, noise, and individual signal). These predefined regions and subproblems are adapted from the results of dynamic adaptive binning. If a targeted or peak alignment approach is taken, the regions can be defined as fixed width regions containing the targeted or aligned peaks of interest.

The solution to each subproblem is obtained by breaking each region into signal of interest, adjacent signal, and baseline. The baseline and adjacent signals are then removed, leaving the signal of interest. This construction of subproblems allows the problem to be transformed into the map-reduce paradigm (described later). As part of this work, two alternative definitions of the subproblems were explored:

1) Region of interest
2) Region of interest with adjacent buffer regions

By including adjacent buffer regions, it is hypothesized that better estimates of adjacent signals are obtained, thus, improving the accuracy of the quantification. Solutions to subproblems for both definitions are constructed by combining a model of baseline and a set of Gauss-Lorentzian peaks:

$$\Theta(\beta,x) = \sum_{j=1}^{N} S([M_j,\sigma_j,P_j,x_{cj}],x) + baseline([b_1,...,b_k],x) \quad (4)$$

$$\beta = [M_j,\sigma_j,P_j,x_{cj},b_1,...,b_k] \quad (5)$$

where $\Theta(\beta,x_i)$ is the model for each region with the model parameters, β. Further, N is the number of peaks in the subproblem, thus, M_j, σ_j, P_j, and x_{cj} refer to the height, standard deviation, fraction of Lorentzian, and the center of the j-th peak. $baseline(...)$ is a piecewise baseline linear function, where $b_1,...,b_k$ are the heights of the piecewise segments.

The final locations of the peaks and their parameters (e.g., width, height) are determined algorithmically by solving the corresponding nonlinear curve-fitting problem. The parameters of the nonlinear 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:

$$1/2 \sum_{i}^{m} (\Theta(\beta,x_i) - y_i)^2, \quad (6)$$

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 segment, β is a vector of parameters, and Θ is the model of each subproblem that will be fit.

The nonlinear 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 2). The initial standard deviation, 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. The initial baseline heights, b_i , is defined as the minimum intensity in the segment. The lower and upper bounds for parameters are defined as:

$$\begin{aligned} 0 < M_j &\leq MAX_i, \\ 0 < \sigma_j &\leq |s_L - s_R|, \\ 0 \leq P_j &\leq 1.0, \\ \alpha_i \leq x_{cj} &\leq \omega_i, \\ 0 \leq b_k &\leq MAX_i, \end{aligned}$$

where MAX_i is the maximum height in the i-th segment, and s_L and s_R are the left and right boundaries of the segment. The boundaries for location of each peak, $[\alpha_j,\omega_j]$,

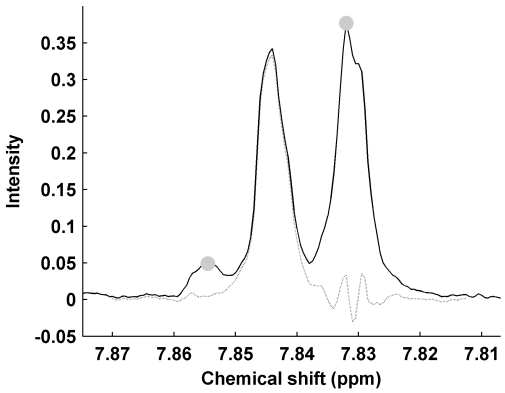


Fig. 1: Removal of adjacent signals (1st and 3rd peak) to target signal of interest (2nd peak) in overlapping regions

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.

Through the solutions obtained for each subproblem, the frequency domain spectral data can be transformed into a feature vector by specifying a set of regions $\mathbf{R} = \{R_1, R_2, ...R_n\}$, where each region is identified by its chemical shift boundaries and the adjacent signals to remove from that region. The baseline is automatically removed from each region. By design, regions are allowed to overlap to filter out alternative sets of adjacent signals. This is demonstrated in Figure 1. The characterization of the metabolomics study for algorithm evaluation employs spectra binning to solve the correspondence problem; however, localized deconvolution can filter unwanted signals for the enhancement of targeted quantification, alignment algorithms, and other alternative quantification techniques.

3.4 Map-Reduce

A map-reduce architecture is employed to enable high-throughput spectral deconvolution. This architecture exposes cloud-based services using the web application framework Ruby on Rails. The algorithm is implemented as a Hadoop based map-reduce program using Hadoop streaming, a technique that allows one to use non Java based programs in the Hadoop architecture. This implementation uses a MATLAB implementation of the numerical optimization algorithm, in a similar fashion as experimented by [23], [24].

The Hadoop streaming mechanism processes data in lines. Hence the data format used as the input to the process is an independent deconvolution problem on each line. This is also important to maintain clear record boundaries for the record splitter. Given that this task is map centric, i.e. the critical process is performed during a map and reduce is merely a combine operation, the number of mappers is a sensitive operator. The map phase consists of solving the aforementioned non-linear optimization subproblem. The reduce step is the recombination and ordering of these results. In order to expose the Hadoop functions in a convenient way to the biologists and also for better integration with existing workflow engines, a web service is implemented. The web service follows the REST paradigm and can be accessed by an HTTP POST operation. The web service is deliberately made into an asynchronous service due to the longer processing time for larger jobs. The processing time varies depending on the complexity of the spectra, and therefore, could not be incorporated into a synchronous web service.

3.5 Cluster Setup

The Hadoop cluster consists of 15 dedicated server computers, each having 16GB of RAM and Quad core AMD processor and connected via Gigabit Ethernet. The Hadoop software version is 0.20.1. The cluster was configured to have a total map task capacity of 120 and reduce task capacity of 90. Jobs were submitted in groups of 5, 10, 15, and 20 (e.g., 5 spectra at a time).

3.6 Synthetic Data

Both empirical and synthetic spectroscopic data are employed to show the application of localized deconvolution. The synthetic spectroscopic data sets are based on urine ^{1}H spectra and were developed by characterizing the salient distributions in empirical spectroscopic data (Anderson et al., 2009). These synthetic data sets enable the use of exacting performance metrics because the true location and size of each peak is known *a priori*.

A synthetic data set of 20 complex ^{1}H spectra was generated, and it was analyzed by two direct measures of the spectral quantification accuracy for each algorithm: absolute quantification error (AQE) and relative quantification error (RQE):

$$AQE = \frac{100}{N * M} \sum_{b=1}^{M} \sum_{s=1}^{N} \left| \frac{predicted_{b,s} - true_{b,s}}{true_{b,s}} \right| \quad (7)$$

$$RQE = \frac{100}{M} \sum_{b=1}^{M} \left| \frac{std(predicted_b) - std(true_b)}{std(true_b)} \right| \quad (8)$$

where $predicted_{b,s}$ is the localized deconvolution results for bin b and spectrum s, $true_{b,s}$ is the true deconvolution results, M is the total number of bins, N is the total number of spectra, and $std(predicted_b)$ is the standard deviation of the set of all localized deconvolution results for bin b, and $std(true_b)$ is the standard deviation of the set of all true deconvolution results for bin b.

3.7 Experimental Data

In addition to comparing spectral binning algorithms on synthetic data sets, this manuscript demonstrates the application of high-throughput localized deconvolution on empirical

Table 1: Mean/median absolute and relative quantification error for standard binning (Standard), localized deconvolution with positive baseline constraint (Region (+)), localized deconvolution with additional buffer and positive baseline constraint (Region & Buffer (+)), localized deconvolution (Region (+/-)), and localized deconvolution with additional buffer (Region & Buffer (+/-))

	Absolute Quantification Error		Relative Quantification Error	
	MEAN	MEDIAN	MEAN	MEDIAN
Standard	1405	125	409	45
Region (+)	36	25	50	25
Region & Buffer (+)	50	19	178	21
Region (+/-)	37	24	48	24
Region & Buffer (+/-)	55	18	172	21

data from a ^{1}H NMR-based experiment to monitor rat urinary metabolites after exposure to α-naphthylisothiocyanate (ANIT) [16]. A subset of this data set was used to compare the quantification algorithms. Specifically, an ANIT dose of 20 mg/kg at 2 days post-exposure was selected, and the performance of the algorithms were analyzed by studying the results of a standard supervised learning procedure, Orthogonal Projection onto Latent Structures (O-PLS) [25].

The O-PLS model was evaluated on its predictive ability, using the Q^2 (coefficient of prediction) metric. Q^2 was calculated as follows:

$$Q^2 = 1 - \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{9}$$

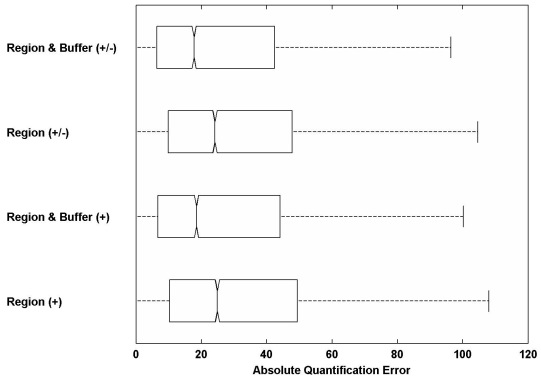
where $PRESS$ is the Predicted REsidual Sum of Squares calculated as the residual e between the predicted and actual Y during leave-one-out cross-validation, SSY is the Sum of Squares for y, $\bar{y}$ is the y mean across all samples, and y_i is the y value for sample i. As Q^2 approaches 1, the more predictive capability the model exhibits. A Q^2 value less than 0 shows the model has no predictive power.

4. Results and Discussion

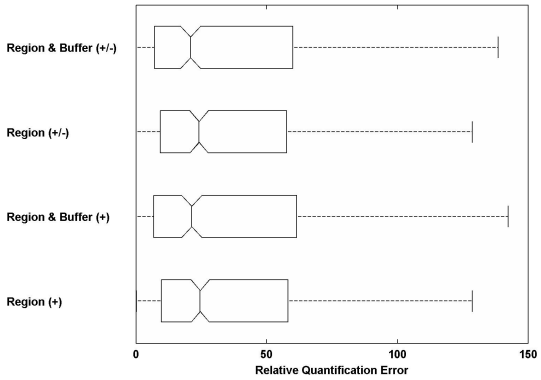
Standard high throughput quantification techniques, such as uniform binning or bucketing, have shown to be effective in reducing the dimensionality and mitigating spectral misalignment; however, these techniques often introduce erroneous quantification errors due to overlapping and adjacent signals. To illustrate the advantages of localized deconvolution, we analyzed synthetic and empirical data. The absolute and relative accuracy of quantification was measured on realistic ^{1}H synthetic spectroscopic data, which were modeled after a traditional urine NMR-based metabolomics study. These results are summarized in Table 1. The difference in performance by including a buffer region and constraining the baseline to positive offsets are shown in Figures 2(a) and 2(b).

As determined by a one-way ANOVA (α = 0.05 assumed for all subsequent statistical tests), the means absolute quantification error for all quantification methods are significantly different. Comparing pairs of methods shows that the standard quantification mean absolute quantification error is significantly different than all localized deconvolution methods using the Tukey-Kramer multiple test correction. To evaluate the median absolute error, the Kruskal-Wallis test was applied to the performance data; the results of which showed that there is a difference between quantification methods as measured by the median absolute quantification error. Specifically, the standard quantification median absolute quantification error is significantly different from all localized deconvolution methods. The mean relative quantification error is significantly different for all methods (one-way ANOVA). The standard quantification mean relative quantification error is significantly different from all localized deconvolution methods (Tukey-Kramer multiple test correction).

Among the four different versions of localized deconvolution, a one-way ANOVA showed that the means of the absolute quantification error are significantly different, and the mean absolute quantification error of Region & Buffer (+/-) is significantly different from the means of Region (+)



(a)



(b)

Fig. 2: Box and whisker plot of the absolute quantification error (a) and the relative quantification error (b)

and Region (+/-). Using the Kruskal-Wallis test, the medians are significantly different, and specifically, the medians of Region & Buffer (+) and Region & Buffer (+/-) are significantly different from the average rank of Region (+) and Region (+/-). A Tukey-Kramer correction was applied to correct for multiple tests. The one-way ANOVA on the mean relative quantification error and Kruskal-Wallis test on the median relative quantification error failed to reject their null hypotheses. i.e., there is not a significant difference among the localized deconvolution methods when examining relative quantification error.

These significant results demonstrate the error in approximating the underlying peak signals with standard binning. If two peaks are adjacent in a spectrum, the degree to which they influence each other will be proportional to their intensity and proximity. Adjacent peaks that are drastically smaller will be heavily influenced by the larger adjacent peaks. Quantifying these smaller peaks is of particular interest to the metabolomics community, as the magnitude of the peak does not determine it relevance in any given study. By modeling each peak individually while simultaneously providing high throughput quantification, localized deconvolution significantly improves the absolute and relative quantification accuracy in NMR-based metabolomics.

In addition to demonstrating the improvement gained through localized deconvolution on synthetic data, we analyzed its effect on quantifying a study of toxicity, as measured by subsequent pattern recognition methods. Specifically, we observed an improvement of 16% in the cross-validated measure Q^2. during the application of a standard supervised learning method, orthogonal projection onto latent structures (O-PLS). The Q^2 metric improved from 0.7569 to 0.8782 after applying localized deconvolution (Region (+/-)). The improved Q^2 metric can be attributed to removing within group variability. Figure 3 shows this improvement in the projected space used to separate the two groups (48 hrs, 20 mg/kg and 0 hrs, Control). The x-axis is representative of the signal responsible for the difference in the groups. The y-axis is signal uncorrelated to the difference in the groups. The tightening of the within group variability on the x-axis leads to the improvement of the Q^2 metric.

The adoption of a general purpose high-throughput quantification method by the metabolomics community is dependent on its ease of applicability. This can be broken into two parts: speed and flexibility. By providing access via RESTful web interface, we are providing a resource that can be incorporated in scientific workflows and other quantification methods. Using a map-reduce framework allows us to parallelize the deconvolution procedure and run the process at a rapid rate. The running time is dependent on the number of mappers, which is shown in Figure 4. On a moderately sized cluster with 15 nodes, it requires approximately 4 minutes to complete a detailed deconvolution of five congested ^{1}H spectra from the data using 20 mappers.

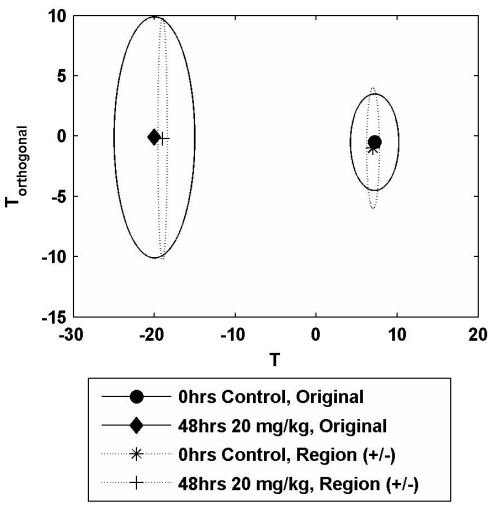


Fig. 3: O-PLS results showing the separation between 0 hrs, Control and 48 hours after a 20 mg/kg dose using 10 fold cross-validation

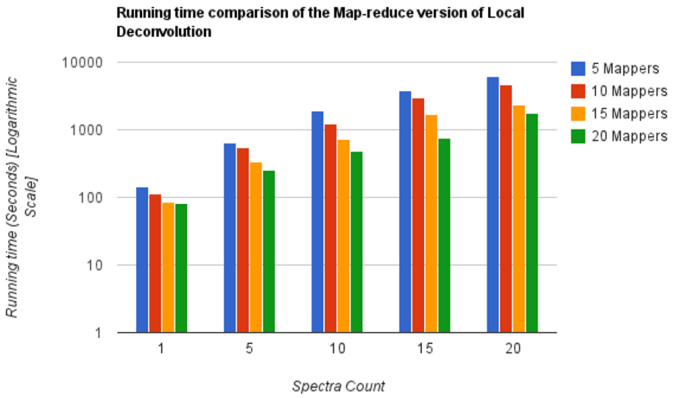


Fig. 4: The running time required to quantify different ^{1}H spectra as a function of the number of mappers

In our implementation, we set the default number of mappers at 20 since it seemed to provide reasonable running times for the typical file sizes encountered in our experimental set up; however, for larger files, higher number of mappers definitely makes an improvement and can be set accordingly by passing the relevant parameter.

5. Conclusion

In conclusion, we have shown that localized deconvolution is a robust method to process highly congested spectra that improves accuracy over standard high-throughput quantification methods. Our algorithm is naturally decomposed into concurrent tasks which are implemented in a map-reduce paradigm with a Web-service interface, thus, providing a scalable and accessible tool for the metabolomics community.

Our experiments have shown that the removal of adjacent, convoluting, and irrelevant signals results in significantly improved absolute and relative quantification, as demonstrated on realistic synthetic data. The performance metrics also demonstrate that including a buffer region does not improve overall accuracy, and allowing the baseline to be positive or negative results in the best accuracy. However, it was observed that specific spectral configurations did benefit from including a buffer region. Developing an algorithm to take advantage of the strengths of both methods is currently in process.

The advantages of our method were also observed on an experimental metabolomics data set of organ toxicity. Specifically, the within group scatter was reduced by localized deconvolution, resulting in an improved cross-validation score (Q^2); however, this increase in accuracy leads to additional computing costs. Such issues can easily be overcome by parallelizing the process with map-reduce and making use of cheaply available cloud resources. While our method provides a significant improvement over standard binning methods, alternative techniques that rely on annotated spectral databases, such as targeted and direct quantification methods, can also improve their accuracy by filtering and removing obfuscating signals with localized deconvolution.

6. Acknowledgement

We would like to acknowledge the Kno.e.sis Cloud Computing Collaboratory for providing computing resources: http://knoesis.wright.edu/aboutus/infrastructure/cloud.

References

[1] N. V. Reo, "NMR-based metabolomics," *Drug and Chemical Toxicology*, vol. 25, no. 4, pp. 375–382, 2002.

[2] J. C. Lindon, E. Holmes, and J. K. Nicholson, "Pattern recognition methods and applications in biomedical magnetic resonance," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 39, no. 1, p. 1, 2001.

[3] J. K. Nicholson, J. C. Lindon, and E. Holmes, "Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data," *Xenobiotica*, vol. 29, no. 11, p. 1181, 1999.

[4] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[5] H. Martens and T. Naes, *Multivariate Calibration*. London: Wiley, 1989.

[6] S. C. Connor, R. A. Gray, M. P. Hodson, N. M. Clayton, J. N. Haselden, I. P. Chessell, and C. Bountra, "An NMR-based metabolic profiling study of inflammatory pain using the rat FCA model," *Metabolomics*, vol. 3, no. 1, pp. 29–39, 2007.

[7] T. L. Whitehead, B. Monzavi-Karbassi, and T. Kieber-Emmons, "1H-NMR metabonomics analysis of sera differentiates between mammary tumor-bearing mice and healthy controls," *Metabolomics*, vol. 1, no. 3, pp. 269–278, 2005.

[8] P. E. Anderson, N. V. Reo, N. J. DelRaso, T. E. Doom, and M. L. Raymer, "Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics," *Metabolomics*, vol. 4, no. 3, pp. 261–272, 2008.

[9] P. Anderson, D. Mahle, T. Doom, N. Reo, N. DelRaso, and M. Raymer, "Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data," *Metabolomics*, pp. 1–12, 2011. [Online]. Available: http://www.springerlink.com/index/C5NP143U061K0585.pdf

[10] R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, and J. C. Wilson, "Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform," *Chemometrics & Intelligent Laboratory Systems*, vol. 85, no. 1, pp. 144–154, 2007.

[11] K. M. Å berg, E. Alm, and R. J. O. Torgrip, "The correspondence problem for metabonomics datasets," *Analytical and Bioanalytical Chemistry*, vol. 394, pp. 151–162, 2009.

[12] J. Forshed, R. J. Torgrip, K. M. Aberg, B. Karlberg, J. Lindberg, and S. P. Jacobsson, "A comparison of methods for alignment of NMR peaks in the context of cluster analysis," *J Pharm Biomed Anal*, vol. 38, no. 5, pp. 824–832, 2005.

[13] D. J. Crockford, H. C. Keun, L. M. Smith, E. Holmes, and J. K. Nicholson, "Curve-Fitting Method for Direct Quantitation of Compounds in Complex Biological Mixtures Using 1H NMR: Application in Metabonomic Toxicology Studies," *Analytical Chemistry*, vol. 77, no. 14, pp. 4556–4562, 2005.

[14] Q. Zhao, R. Stoyanova, S. Du, P. Sajda, and T. R. Brown, "HiRes - a tool for comprehensive assessment and interpretation of metabolomic data," *Bioinformatics*, vol. 22, no. 20, pp. 2562–2564, 2006.

[15] A. M. Weljie, J. Newton, P. Mercier, E. Carlson, and C. M. Slupsky, "Targeted Profiling: Quantitative Analysis of 1H NMR Metabolomics Data," *Analytical Chemistry*, vol. 78, no. 13, pp. 4430–4442, 2006.

[16] D. Mahle, P. Anderson, and N. DelRaso, "A generalized model for metabolomic analyses: application to dose and time dependent toxicity," *Metabolomics*, 2011. [Online]. Available: http://www.springerlink.com/index/H7861V6218327H10.pdf

[17] B. K. Alsberg, A. M. Woodward, and D. B. Kell, "An introduction to wavelet transforms for chemometricians: A time-frequency approach," *Chemometrics & Intelligent Laboratory Systems*, vol. 37, no. 2, p. 215, 1997.

[18] H. F. Cancino-De-Greiff, R. Ramos-Garcia, and J. V. Lorenzo-Ginori, "Signal de-noising in magnetic resonance spectroscopy using wavelet transforms," *Concepts in Magnetic Resonance*, vol. 14, no. 6, pp. 388–401, 2002.

[19] K. Kaczmarek, B. Walczak, S. de Jong, and B. G. Vandeginste, "Preprocessing of two-dimensional gel electrophoresis images," *Proteomics*, vol. 4, no. 8, p. 2377, 2004.

[20] C. Perrin, B. Walczak, and D. L. Massart, "The Use of Wavelets for Signal Denoising in Capillary Electrophoresis," *Anal. Chem.*, vol. 73, no. 20, pp. 4903–4917, 2001. [Online]. Available: http://pubs3.acs.org/acs/journals/doilookup?in_doi=10.1021/ac010416a

[21] X. G. Shao, A. K. Leung, and F. T. Chau, "Wavelet: a new trend in chemistry," *Accounts of Chemical Research*, vol. 36, no. 4, p. 276, 2003.

[22] H. Grage and M. Akke, "A statistical analysis of NMR spectrometer noise," *Journal of Magnetic Resonance*, vol. 162, no. 1, pp. 176–188, 2003.

[23] K. Gunaratna, P. Anderson, A. Ranabahu, and A. Sheth, "A study in Hadoop streaming with MATLAB for NMR data processing," in *Processing of 2nd IEEE International Conference on Cloud Computing (Cloudcom)*, Indianapols, IN, 2010.

[24] A. Manjunatha, P. Anderson, A. Ranabahu, and A. Sheth, "Identifying and Implementing the Underlying Operators for Nuclear Magnetic Resonance based Metabolomics Data Analysis," in *Proceedings of 3rd International Conference on Bioinformatics and Computational Biology (BICoB)*, New Orleans, LA, 2011.

[25] J. Trygg and S. Wold, "O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter," *Journal of Chemometrics*, vol. 17, no. 1, pp. 53–64, 2003.