Pacific Symposium on Biocomputing 12:115-126(2007)

LEVERAGING LATENT INFORMATION IN NMR SPECTRA FOR ROBUST PREDICTIVE MODELS

DAVID CHANG[1,2], AALIM WELJIE[1,3], JACK NEWTON[1]

[1]*Chenomx Inc., Suite 800, 10050 112 Street, Edmonton, Alberta, Canada*

[2]*Department of Chemical and Materials Engineering, University of Alberta, Edmonton, Alberta, Canada*

[3]*Metabolomics Research Centre, University of Calgary, Calgary, Alberta, Canada*

A significant challenge in metabolomics experiments is extracting biologically meaningful data from complex spectral information. In this paper we compare two techniques for representing 1D NMR spectra: "Spectral Binning" and "Targeted Profiling". We use simulated 1D NMR spectra with specific characteristics to assess the quality of predictive multivariate statistical models built using both data representations. We also assess the effect of different variable scaling techniques on the two data representations. We demonstrate that models built using Targeted Profiling are not only more interpretable than Spectral Binning models, but are more robust with respect to compound overlap, and variability in solution conditions (such as pH and ionic strength). Our findings from the synthetic dataset were validated using a real-world dataset.

1. Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy is a widely-used tool in the rapidly growing field of metabolomics, where the measurement of small molecule metabolites provides a chemical "snapshot" of an organism's metabolic state [1]. NMR is inherently quantitative and non-selective, thus a wealth of chemical information can be extracted from single NMR spectrum. Metabolomics studies often couple NMR spectral data with principal component analysis (PCA) and other pattern recognition techniques to uncover meaningful patterns in data sets [2]. Long-term goals of such computational model building include automation of data analysis as part of an integrated diagnostics platform [3] and personalized therapies [4]. Building statistical models from NMR spectra can be problematic however, as spectral distortions present potentially confounding artifacts to techniques such as PCA [5, 6].

These distortions have an origin in the hardware [7], the type and nature of the sample, and choice of acquisition and processing parameters [8]. For example, pre- and post-processing algorithms and the signal-to-noise (S/N) in the time domain impact data quality. Metabolite signals in complex mixtures often span several orders of magnitude, thus requiring a significant dynamic range in the receiver. Furthermore, aqueous samples such as urine or plasma require suppression of the water solvent peak which is 7-8x more concentrated than the metabolites of interest, resulting in distortions of the baseline and intensity of metabolite signals. Metabolites' resonance frequencies, lineshapes, and linewidths will vary between samples within an NMR metabolomics dataset irrespective of hardware considerations. Factors influencing these chemical modulations include sample pH, ionic composition, and inter-metabolite interactions [9]. As a result, statistical analyses require some form of pre-processing or data reduction to ensure that the variables of interest are representative of the underlying chemical data [10].

In this paper, the impact of spectral distortion on the quality of predictive statistical models built upon two alternative representations of NMR data is assessed. A simulated dataset is used to model various types of spectral distortion in a systematic manner, and two techniques for dimensionality reduction, spectral binning and targeted profiling, are used to represent these simulated spectra. The results are assessed using the regression/classification extension of PCA, partial least squares for discriminant analysis (PLS-DA) [11]. We validate our findings using a real-world data set of rat-brain extracts.

2. NMR Data Representations

An NMR spectrum is a linear combination of characteristic signals for each compound that is present in a given sample. As the concentration of a particular compound changes, the characteristic signal for that compound responds in a linear fashion. Thus, an NMR spectrum can be viewed from a theoretical perspective as follows:

$$\underset{[1\times n]}{d_{obs}} = \underset{[1\times k]}{c} \cdot \underset{[k\times n]}{[a \otimes s]} + \underset{[1\times n]}{u} + \underset{[1\times n]}{n} \quad (1)$$

where d_{obs} is a [1xn] vector of the observed NMR data, c is a [1xk] vector representing the concentrations of k known compounds in the mixture, and s represents a matrix of the spectral signatures present in the solution. a is a *spectrum calibration function* that is applied to each row of s to account for changes in the sample's pH, ionic strength, etc. u represents unknown contributions to the signal from unknown metabolites, lipoproteins, or any other contributions to the signal that are not explicitly modeled using s. Finally, the observed spectrum contains noise that is introduced by the NMR hardware and processing algorithms, n.

2.1. *Spectral Binning*

Spectral binning [2] is a widely-used technique where the spectrum is subdivided into a number of regions, and the total area within each bin is used as an abstracted representation of the original spectrum. The area encapsulated by a bin would ideally capture all of the area associated with a given resonance across all spectra in the dataset, thereby mitigating the effect of minor peak shift and line width variations for a compound across samples. A typical 64k NMR spectrum would be reduced using bin widths of 0.04 ppm, resulting in ~250 bin integral values. Spectral binning is agnostic of the underlying generative model described in Equation 1, however it is commonly used due to the ease of implementation and complete spectral coverage.

2.2. *Targeted Profiling*

Targeted profiling [8] is a technique that leverages a reference spectral database to directly recover the concentration matrix c from Equation 1, which is then used as the input to pattern recognition techniques such as PCA or PLS-DA. Targeted profiling can be viewed as a method of recovering the latent variables in the form of underlying metabolite concentrations that generated the observed spectral data. Because of its reliance on a spectral database s, targeted profiling does not directly model or deal with the unknown term u in Equation 1. Since u may contain potentially

important latent chemical information, it can be calculated directly as the residual from Equation 1, and spectral database-agnostic techniques such as spectral binning can be applied to u for subsequent analysis.

3. Methods

3.1. *Synthetic Study*

Several synthetic data sets were generated with specific characteristics to simulate, in a systematically controlled manner, some of the key challenges inherent in working with NMR data. The data for the synthetic study was generated using Chenomx NMR Suite 4.5 (Chenomx Inc., Edmonton, Alberta, Canada) compound database entries. Varying mixtures of twenty compounds, with the addition of DSS at 0.5 mM, were simulated. Compound concentrations for the following compounds were sampled randomly from a normal distribution: 2-oxoglutarate, acetate, acetone, alanine, betaine, carnitine, citrate, creatine, dimethylamine, fumarate, glucose, lactate, maleate, myo-inositol, taurine, tryptophan, tyrosine, urea, π-methylhistidine, τ-methylhistidine. Biologically viable population statistics of mean and standard deviation were used for each compound [Chang, Rankin, McGeer, Shah, Marrie, and Slupsky, submitted] and these concentrations remained fixed from simulation to simulation.

Random uncorrelated noise was added to each spectrum in the frequency domain. Each spectrum was generated to have an equivalent amount of noise by an approximate signal to noise ratio (SNR) of 100:1.

The effect of pH variability was simulated by randomly varying compound resonance frequencies within an empirically validated range. This range reflects the compound's NMR frequency response to pH levels ranging from pH 4 to 9 as determined from pH curves of pure reference spectra. The magnitude of this range was controlled to test the effects of pH variation via a transform fraction parameter. A fraction of 1.0 allowed clusters to be transformed over the entire pH 4 to 9 range, while a fraction of 0.1

would allow for clusters to be transformed over 10% of the range, centered at pH 7.0. The actual pH range that this represents will be different for each compound depending on the relative pH sensitivity of the compound near pH 7.0.

In order to generate two classes of spectra, the population statistics of one or more metabolites were changed for each simulation. The parameters used in each simulation are outlined in Table 1.

Table 1. Simulation Parameters for Synthetic Study.

Simulation #	Parameters	Value
1	Number of Files	200 (100 of each class)
	SNR	100
	Transform Fraction	0.1
	Group 1 Citrate/Tryptophan Mean ± Stdev (μmol)	2318 ± 1496 / 5 ± 2
	Group 2 Citrate/Tryptophan Mean ± Stdev (μmol)	1031 ± 945 / 10 ± 2
2	Number of Files	200 (100 of each class)
	SNR	100
	Transform Fraction	0.1
	Group 1 Maleate Mean ± Stdev (μmol)	30 ± 15
	Group 2 Maleate Mean ± Stdev (μmol)	60 ± 20
3	Number of Files	200 (100 of each class)
	SNR	100
	Transform Fraction	1
	Group 1 Citrate/Tryptophan Mean ± Stdev (μmol)	2318 ± 1496 / 5 ± 2
	Group 2 Citrate/Tryptophan Mean ± Stdev (μmol)	1031 ± 945 / 10 ± 2

3.2. *Rat Brain Extracts*

This real-world dataset is based on a previously published [12] dataset and was kindly provided by Dr. Brent McGrath and Dr. Peter Silverstone (Department of Psychiatry, University of Alberta). Twelve adult male Sprague-Dawley rats brains were dissected into frontal (fcx) cortex, temporal cortex (tcx), occipital cortex (ocx) and hippocampus (hipp) regions according to stereotaxic demarcation [12]. For spectral binning, bins widths of 0.04 ppm were used, with the following dark regions defined: DSS (the internal standard): -0.1-0.1ppm, 0.6-0.7 ppm; methanol (a byproduct of the extraction process): 3.33-3.37 ppm; water: 4.5-5.5ppm; imidazole (the pH indicator): 7.13-7.5, 7.82-8.68 ppm.

The following compounds were identified and quantified using the targeted profiling technique [8] as implemented in Chenomx NMR Suite 4.5: 4-aminobutyrate, acetate, adenosine, alanine, aspartate, betaine, choline, citrate, creatine, creatinine, formate, fumarate, glutamate, glutamine, glycerol, glycine, hypoxanthine, isoleucine,

lactate, leucine, lysine, methanol , N-acetylaspartate, serine, succinate, taurine, threonine, tyrosine, valine, xanthine, and myo-inositol.

3.3. *Multivariate Statistical Modeling*

All multivariate modeling was performed using SIMPCA-P+ 11.0 from Umetrics Inc. Permutations tests were performed using 100 permutations. R^2_X and R^2_Y are calculated as the fraction of the sum of squares of all X and Y that the model can explain using the latent variables. Q^2 is the fraction of the total variation in Y that can be predicted using the model via seven-fold cross-validation.

4. Results

4.1. *Synthetic Data*

By systematically varying key properties of the synthetic data sets, several aspects of building statistical models on NMR data representations were assessed. The first issue assessed was the effect of noise on the spectra. Specifically, noise was added to the spectrum to see how robust both spectral binning and targeted profiling methods were at being able to recover the latent information in the data in the presence of noise. What was observed was that if the noise was completely uncorrelated, then both methods are very robust to varying noise levels. (Data is available from supplementary materials.)

The next issue we examined was the choice of variable scaling and normalization methods, since this can have a large impact on the quality of results obtained from multivariate statistical methods such as PLS-DA. Normalization for all spectral binning data was to the total area of the NMR spectrum. No normalization was necessary for the targeted profiling results, since direct quantification can be obtained with the addition of an internal standard. Both the spectral binning data and targeted profiling data were mean centered and were scaled using unit variance (UV) or Pareto scaling. UV scaling involves weighting each of the variables by the variables' group standard deviation, and has the advantage of not biasing statistical models towards large concentration

compounds or high area bins. Pareto scaling involves the weighting each of the variables by the variables' group variance, which minimizes the impact of noise. Data from simulation #1 was used to evaluate the effects of these two scaling procedures. This simulation encoded class differentiation through citrate, present at relatively high concentrations, and tryptophan, present at relatively low concentrations. Figure 1a demonstrates that PLS-DA on UV scaled data can recover differences in both tryptophan and citrate, while the loadings plot of Pareto-scaled data (Figure 1b) is only able to distinguish the intense citrate signal. UV scaling was superior to Pareto scaling in recovering a model that accurately reflected the variables of interest (both low- and high-concentration metabolites) for both targeted profiling and spectral binning data.

Overlap of NMR resonances from different metabolites is another issue hampering the analysis of complex biofluid spectra. Further complications arise from compound overlap with dominant peaks such as urea, where low intensity peaks are often lost in traditional analyses due to the overwhelming magnitude of the urea signal. Simulation #2 generated a dataset in which a single metabolite, maleate, differentiates the two classes and overlaps with the high concentration urea signal, which varies randomly (i.e. urea does not encode class discrimination). Figure 2 shows the scores, loadings, and permutations tests for spectral binning and targeted profiling methods. One can see from the loadings plot in Figure 2b, that targeted profiling methods identify maleate as a significant metabolite even under severe overlap conditions, while spectral binning shown in Figure 2a fails. Spectral binning is also prone to generating highly overfit models as shown by the permutation test in Figure 2, whereas targeted profiling models show no signs of overfitting. Permutation tests help assess overfitting by randomly permuting class labels and refitting a new model with the same number of components as the original model. An overfit model will have similar R^2 and Q^2 to that of the randomly permuted data. Well fit models will have R^2 and Q^2 values that are always higher than that of the permuted data.

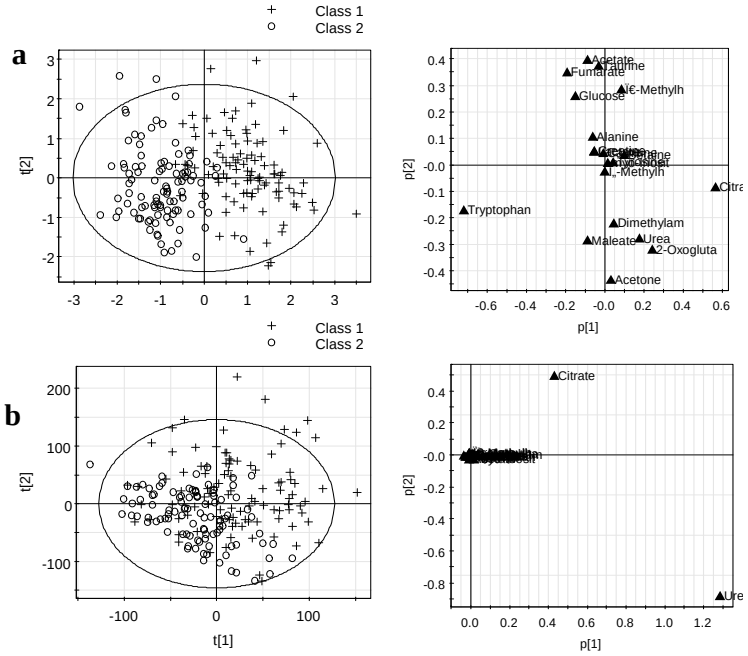

Figure 1. PLS-DA models (scores plot left, loadings plot right) of targeted profiling data using **a**) unit variance scaling **b**) Pareto scaling.

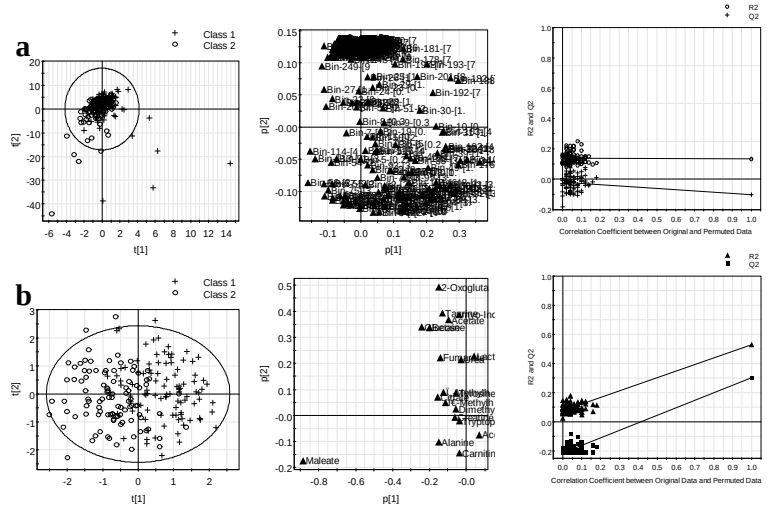

Figure 2. PLS-DA models (scores plot left, loadings plot center, permutation plot right) for **a**) spectral binning and **b**) targeted profiling methods under conditions of large overlap.

Sample matrix conditions such as pH and ionic strength can have profound effects on metabolites' NMR resonance frequencies. These shifts can directly influence the quality of the models that are generated using NMR data, and were modeled with simulation #3. Both spectral binning and targeted profiling gave rise to models that were able to separate the data in the latent variable space. However, the quality of the model generated with the spectral binning data was low and resulted in overfitting as shown in permutation plots (Supplementary Figures). This is due to the large number of variable weights used in the loadings. A large number of variables share similar weights because the same significant resonances are now migrating over adjacent bins due to pH/ionic strength variation. Models built on targeted profiling data, which accounts for the shifts in resonance locations directly in the modeling process, are able to separate the two groups and do not overfit the data.

The final effect studied is the impact of limited sample sizes on predictive capacity, a typical problem in metabolomics studies. The effect of sample size was shown using a subset from Simulation #3. The size of the dataset was reduced from 100 to 20 samples in each class. Even with a limited sample size, the targeted profiling approach resulted in well fit PLS-DA models, as assessed by the permutations tests. While the descriptive features of tryptophan and citrate are not as clearly distinguished in the loadings plot, the permutation plot indicates that even with a small number of samples the data is not overfit. The results for spectral binning, however, are quite deceptive, as the PLS-DA model shows very good separation of classes in the scores plot. However, the model generated has an extremely high degree of overfitting – the majority of the randomly permuted models generate Q^2 values higher than that of the non-permuted model (Supplementary Figures).

4.2. *Rat Brain Extract*

The rat brain extract dataset is a real-world dataset that exhibits many of the phenomena we have seen in the synthetic dataset. The spectra contain noise, have metabolite resonances that shift due to pH, and have low-concentration metabolites that are important in

differentiating the different brain regions, thus making it a suitable model dataset to validate our findings from the synthetic dataset. This dataset was acquired at high resolution (800MHz) and contains ~30 NMR-visible compounds. We did not find that the choice of variable scaling affected the quality of the generated models for this dataset. We therefore used unit variance scaling for the results shown below.

We found that using spectral binning generated a model with lower predictive accuracy than targeted profiling data: Q^2 for spectral binning was 0.468, whereas Q^2 for targeted profiling was 0.522.

As in our synthetic dataset, we found that spectral binning-based results were prone to overfitting. To test for overfitting, we randomly permuted the class labels for the PLS-DA analysis 100 times. With the spectral binning dataset, we found that some of the models generated with random permutations of the data had higher Q^2 and R^2 values than the non-permuted data. This is illustrated in Figure 3a. Internal validation of the model based on the targeted profiling representation of the NMR data do not exhibit any characteristics of an overfit model, as shown in Figure 3b. The targeted profiling representation uses only 27 variables to represent the latent information in the dataset, thereby restricting the degrees of freedom available in the construction of a model, and reducing the capacity of the model to overfit the data.

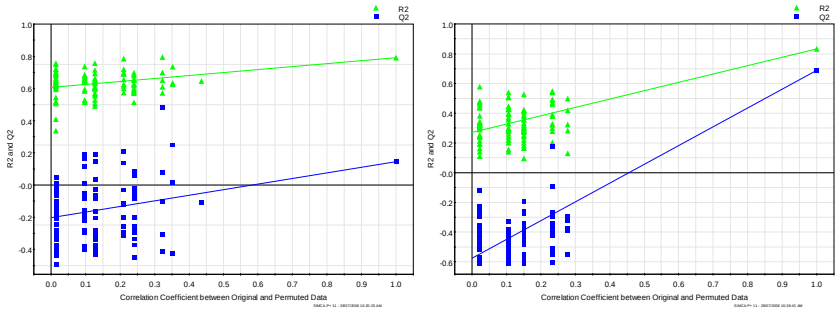

Figure 3. **a,** Internal validation of spectral binning, showing clear evidence of overfitting with random permutations of the data generating better R2 and Q2 values than the non-permuted data. **b**, Internal validation of targeted profiling, showing clear decrease in performance on permuted data.

5. Conclusion

We have demonstrated how the inherent properties of NMR spectroscopy can impact the predictive ability of models built upon spectral binning and targeted profiling representations of NMR data by using a novel method for synthetically generating NMR spectra. The quality of predictive models built was quantitatively assessed, as was the relative robustness of these two methods. Under the experimental design chosen, both methods are very robust with respect to noise. In contrast, variable scaling methods can affect both the quality and interpretability of the models generated. W found for targeted profiling data, unit variance scaling generates a more robust data representation. Targeted profiling was also found to be an effective dimensionality reduction technique that, overall, is more robust with respect to spectral distortions and high dynamic range metabolites than spectral binning, and is less prone to overfitting than spectral binning models. These findings were validated on a real-world dataset of rat-brain extracts consisting of ~30 NMR detectable metabolites, in which statistical models were less prone to overfitting based on a spectral profiling representation of the data. Spectral binning is a common method for data reduction due to the speed of analysis, while current targeted profiling implementations require interactive input and are relatively time-intensive. While the rat-brain extract study represents a relatively simple dataset, targeted profiling has successfully been applied to extensive studies of serum [Weljie, Dowlatabadi, Miller, Vogel, Jirik, submitted] and urine [Chang, Rankin, McGeer, Shah, Marrie, and Slupsky, submitted]. As increasingly automated methods for quantitative profiling of NMR data become available, we expect database-driven targeted profiling to become the data-reduction method of choice.

6. Supplementary Information

Supplementary Figures and Data is available at http://www.chenomx.com/publications/PSB2007

References

[1] J. C. Lindon, E. Holmes and J. K. Nicholson, *Anal. Chem.* **75**, 384A (2003)

[2] E. Holmes, H. Antti, *Analyst* **127**, 1549 (2002)

[3] D. S. Wishart, L. M. M. Querengesser, B. A. Lefebvre, N. A. Epstein, R. Greiner and J. B. Newton, *Clinical Chemistry* **47**, 1918 (2001)

[4] T. A. Clayton, J. C. Lindon, O. Cloarec, H. Antti, C. Charuel, G. Hanton, J. P. Provost, J. L. Le Net, D. Baker, R. J. Walley, J. R. Everett and J. K. Nicholson, *Nature* **440**, 1073 (2006)

[5] M. Defernez, I. J. Colquhoun, *Phytochemistry* **62**, 1009 (2003)

[6] S. Halouska, R. Powers, *J. Magn Reson.* **178**, 88 (2006)

[7] R. Siuda, G. Balcerowska and D. Aberdam, *Chemometrics Intell. Lab. Systems* **40**, 193 (1998)

[8] A. M. Weljie, J. Newton, P. Mercier, E. Carlson and C. M. Slupsky, *Anal. Chem.* **78**, 4430 (2006)

[9] J. C. Lindon, J. K. Nicholson, E. Holmes and J. R. Everett, *Concepts in Magnetic Resonance* **12**, 289 (2000)

[10] B. J. Webb-Robertson, D. F. Lowry, K. H. Jarman, S. J. Harbo, Q. R. Meng, A. F. Fuciarelli, J. G. Pounds and K. M. Lee, *J Pharm. Biomed. Anal.* **39**, 830 (2005)

[11] Umetrics AB, *Multi- and Megavariate Data Analysis: Principles and Applications*, Umeå, (2001).

[12] B. M. McGrath, A. J. Greenshaw, R. McKay, A. M. Weljie, C. M. Slupsky and P. H. Silverstone, *Int. J. Neurosci.* **(In Press)** (2006)