Fixed-Width Binning, Variable-Width Binning or No Binning: A Study of Different Binning Methods in NMR-based Metabolomics Analysis

Gregory Banik, Ph.D.,[1] Chen Peng, Ph.D*.,[1] Omoshile Clement, Ph.D, [1] Ty Abshear, [1] Scott Ramos, Ph.D.,[2] Brian Rohrback, Ph.D.,[2] Ian Lewis,[3] and John Markley, Ph.D[3]

[1]Bio-Rad Laboratories, Inc., Informatics Division, 3316 Spring Garden Street, Philadelphia, PA 19104 USA
[2]Infometrix, Inc., Suite 250, 10634 E. Riverside Drive, Bothell, WA 98011
[3] University of Wisconsin, Madison, USA

Introduction

In the multivariate analysis of NMR-based metabolomics data[1,2], small variations in the resonance position of the individual peaks caused by experimental and instrument-induced variations can adversely impact the PCA results [3,4]. Techniques to address the NMR peak misalignment issue include the commonly used binning or bucketing [5], and the more advanced algorithms that aim at shifting the individual peaks to reach a better alignment across the spectra [6-8].

Fixed-width binning (usually at 0.04 ppm) is commonly used to alleviate the impact of the peak misalignment by averaging up the data points falling inside the bin width. However, since it drastically reduces the data resolution, it makes it more difficult to interpret the PCA results such as identifying the changed metabolites from the loadings plot. Lately, multivariate analysis at full spectral resolution becomes practical as more integrated spectral processing and MVP software package becomes available [9] and its advantages become recognized [10], hence retaining the spectral resolution while aligning the shifted peaks becomes a more demanded technique. Theoretically, many of the reported methods that move local peaks to reach better peak alignment [6-8] should serve this goal. However the performance of such algorithms is inevitably hampered by peak overlaps, which are common in biofluid spectra.

We have developed the following tools intended to tackle the global and local peak misalignment problems:

- Global spectra alignment for multivariate analysis using full-resolution without binning.
- Fixed width binning, with graphical interface for the researcher to visualize and manually adjust the bins on top of the original spectra.
- IntelliBucket™, variable-width binning where the width of a bin is automatically adjusted in order to place bin boundaries at the local minima in the overlap density consensus spectrum.
- Automatic Filtering of NMR Spectra (AFNS), a novel method that selects spectral features based on their statistical significance and then smoothes the spectral points using their optimized filter widths.

The effects of the various options are demonstrated using two spectral datasets, with and without local chemical shift variations. The pros and cons of these methods are compared and discussed.

Methods and Algorithms

Global Spectral Alignment and PCA at full spectral resolution

The KnowItAll Metabolomics Edition is a fully integrated software package that includes applications for interactive and macro-based batch processing and database management of large quantity of NMR spectra, PCA, and identifying changed metabolites by metabolite database search [11]. The workflow, shown in Fig. 1, covers almost every aspect of the NMR-based metabolomics analysis in a single software package environment. Since there is no file transfer between different applications, there is virtually no limitation of the size of the X-matrix for PCA, and hence it is very efficient to do PCA on large spectral datasets typically of 32 or 64K points each spectrum.

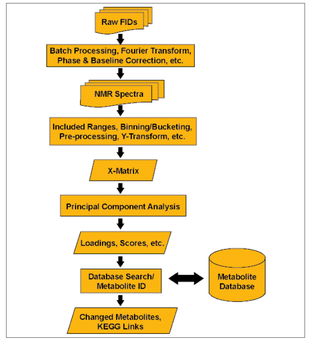

Figure 1. *The workflow of KnowItAll Metabolomics Edition for NMR-based metabolomics data analysis.*

Normally the chemical shift scales of the spectra are calibrated by setting the chemical shift of the reference peak of DSS to 0 ppm. In our experience, the resonance location of DSS usually varies in a range of around ±0.004-0.01 ppm, leading to obvious systematic misalignment for other peaks across the dataset (Fig. 2). We developed a tool for global spectral alignment using selected part of the spectra. Typically the user can select a narrow spectral area where the peaks are not supposed to vary across the dataset (called the 'focus area'), and the program aligns all spectra by matching the focus areas. The matching is based on either peak tops or the shape of the spectral curves. For a dataset without significant local peak shifts, such an alignment makes the dataset well-aligned and ready for the subsequent PCA using the full resolution spectral data points.

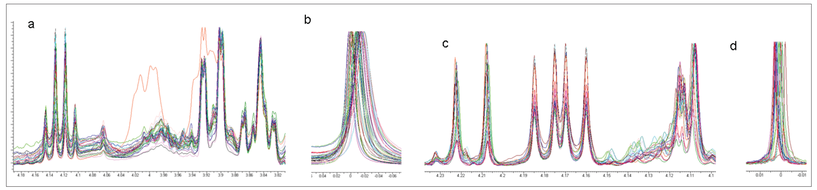

Figure 2. *Illustration of the chemical shift variation of DSS. (a) : Part of the diabetes spectra after globally aligned to peaks in 4.16-4.10 ppm. Note that the DSS peaks (b) show obvious misalignment. (c) : Part of the ATHK1 dataset after globally aligned to the peaks in 4.20-4.15 ppm. Note that other peaks still show misalignment, including the DSS peaks (d).*

Fixed-width Binning Within the Included Spectral Ranges

KnowItAll provides a graphical interface for the researcher to define the spectral ranges to be included for PCA, and once the bins are generated using the defined bin width, one can zoom in to interactively add, delete or edit the bins on top of all the spectra, either displayed as stacked plot or the novel Overlap Density Heatmap (ODH) mode.

IntelliBucket™ Based on Overlap Density Consensus Spectrum

In order not to split a single peak into two neighboring bins, the fixed width bins can be automatically tuned to accomodate a whole peak in a single bin in a user-defined variation range. This is a simple method, yet the key is how to locate the peak boundaries for a large set of spectra, especially when individual peaks may have different degrees of cross spectra misalignment.

In addition to the conventional overlay or stack plot where a set of multiple 1D spectra are plotted using different colors, we have developed a novel spectral visualization technology called "Overlap Density Heatmap (ODH)" [12], which is a 2D heatmap using different color to code the spectral areas with different degree of overlap. An ODH can be limited to show only the spectral areas with up to a certain degree of commonality or uniqueness, and by tracing the outline of the ODH, an 1D OD consensus spectrum can be reconstructed. The local minima in the OD consensus spectrum are used to adjust the bin widths (Fig. 3).

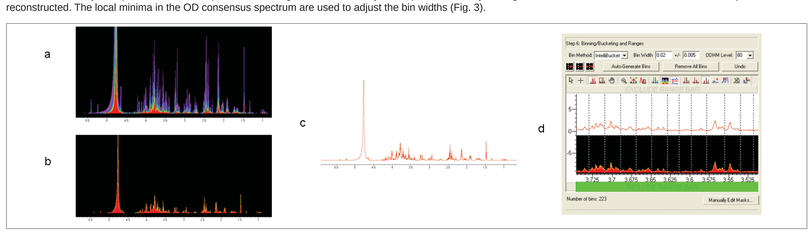

Figure 3. *Overlap Density Heatmap (a) OD level = 0, showing all spectra. (b) OD level = 80, showing areas of 80-100% overlap. (c) The OD consensus spectrum generated from b. (d). Bins generated with bin width automatically adjusted to accommodate local minima in the same OD consensus spectrum.*

Automatic Filtering of NMR Spectra (AFNS)

AFNS uses a rolling binning algorithm with multiple binwidths, and ANOVA (or t-test) based filtering as a means of identifying significant features in complex spectra. It is derived from the "kernel density estimation method" [13]. Instead of using a single binning width, it tries different bin widths (e.g. from 0 to 0.02 ppm in a step value of 0.002 ppm). For each trial bin width, w, the spectral points are smoothed according to (1) :

$$ss(i) = \sum_{j=-w}^{w} s(i+j)/(2w+1) \qquad (1)$$

where $1 \le (i+j) \le N$, the count of points in a spectrum. For each column in the matrix consisting of all the smoothed spectra, the M points (M is the number of experiments) are subject to one-way analysis of variance (ANOVA) using their corresponding classification infomation. If these data points show significant variance regarding the classes, i.e., F0 > Fc, (Fc is the critical F-distribution value, or can be a user-defined threshold), this column is retained and the used filter width, w, is retained. If a column passes when different trial w were used, the one corresponding to the maximum F_0 is retained.

After all the tial filter widths have been tested, the retained points comprises a reduced dataset, and for each column of the dataset, an optimized filter width w is recorded. Next all the points in the reduced dataset are smoothed according to (1) using the optimized filter width of the individual points.

Datasets

I. Diabetes Samples

Thirty-seven blood samples were collected from seventeen diabetic patients and twenty healthy people, then they were allowed to clot in plastic tubes for about two hours at room temperature. Aliquots of serum were collected from the blood and stored at -80°C until assayed. Before the NMR experiment, each sample (150ul) was diluted with solvent solution (300ul H_2O, 50ul D_2O and 3ul DSS). All spectra were measured at a temperature of 298K on a BRUKER Avance-500 spectrometer operating at the proton frequency of 500.13 MHz using pulse sequence ZGPR. (RD-90-t1-90- acquisition, RD being a relaxation delay of 1.5s during which the water resonance is selectively irradiated). For each sample, 64 scans were collected into 8K complex data points with a spectral width of 8012.8Hz.

II. ATHK1 and Regulation of the Osmotic Stress Response

This dataset was from our study to understand the role of ATHK1, a putative membrane histidine kinase, as the osmolyte sensor for the plant HOG1 pathway. We hypothesized that ATHK1 mutants would have altered steady state concentrations of the established osmolytes when the plants were exposed to saline media. The purpose of this study is to test this hypothesis in metabolite extracts of At (ATHK1 knowkouts), 35s (ATHK1 overexpressors) and wt (wild type) using a NMR based metabolomics approach.

Wild type (wt), ATHK1 knockout (At) and ATHK1 over expresser (35s) plants were germinated and grown in sterile liquid cultures of Murashige and Skoog medium. Plants were kept under continuous illumination under a shaker platform for one week. After one week either 1) a sterile sham of Murashig and Skoog medium or 2) NaCl (100mM final concentration) was added to each culture and allowed to incubate for 32 hours. The experiment was designed with four biological replicates for each of the 6 conditions. After the incubation, plants were removed from their media, washed, and flash frozen in liquid nitrogen. The frozen plant material was then lyophilized for 48 hours and ground to a fine powder in a coffee grinder. 300mg of each powdered homogenate was then added to a 22ml screw top vial with 16ml boiling water. Samples were then sealed and incubated at 100°C for 15 minutes. Extracts were then centrifuged at 4,000g for 30 minutes to pellet the cellular debris and the supernatant was filtered with 1) glass wool to remove suspended particulate matter and then 2) with a 5,000 MWCO vivaspin concentrator. The filtrate was then frozen and lyophilized to a dry powder. The resulting metabolite powder was then resuspended on a weight to volume basis with buffered NMR solvent at a ratio of 17.5uL of buffer per mg of dry extract. The NMR buffer was composed of D_2O with 50mM $NaPO_4$, 500uM sodium azide (to minimize microbial growth), and 500uM DSS (as a NMR chemical shift indicator and internal concentration reference). Samples were then titrated to an observed pH of 7.400 (+,- .004) and stored at -80°C until NMR spectral analysis. 1D ^{1}H spectra were collected on a Varian 600mHz cold probe equipped spectrometer. Spectra were collected in 4 scans with 4 silent scans using a 2 second acquisition time and a one second initial delay. All samples were hand shimmed until the DSS half height line widths were less that 1Hz.

Results and Discussion

For both spectral datasets, the raw spectra (FIDs) were automatically processed and saved into a database using the macro-based batch processing function of the ProcessIt™ NMR and MineIt™ applications. The macro included correction of DC offset, zero-filling (to 16K for dataset I and to 64K for dataset II), Lorentzian apodization of 0.5 Hz, Fourier Transform, automated phase correction (GoodLook™ algorithm), baseline correction with automatic base point detection and Spline fitting (for dataset I only. No baseline correction done for dataset II), and referencing (with the DSS peak set to 0 ppm). For dataset I, several spectra were imperfectly phased and were manually re-phased. Class information was added manually for each sample into the databases. Both datasets were globally aligned by focusing on a narrow spectral range (4.16-4.10 ppm and 4.20-4.15 ppm for datasets I and II, respectively). It's noticed that dataset I does not have obivous local peak shift, while dataset II shows significant local peak variation (Fig. 2).

The Principal Component Analysis (PCA) was run with the Analyzelt™ MVP application. For dataset I, the spectral regions of 10-5.15 ppm and 4.75 - 0.5 ppm were used in order to exclude the strong water peaks and other baseline regions. Prior to PCA, each spectrum was transformed by subtracting by its baseline value (the value of the first point in the region of 10-5.15 ppm) and dividing by sample 2-norm (i.e., vector length normalization). Mean-centering was used in pre-processing. For dataset II, the regions of 5.5-5.15 ppm and 4.5-0.5 ppm were used. Similar preprocessing and Y-transform were used except that the baseline subtraction was not done.

For dataset I, we used full-resolution data (i.e., 9,297 points from the included regions) and fixed-width binning (using a bin width of 0.04 ppm and generating 228 bins). As shown in Fig. 4, the scores plots show very similar classification.

However, the loading plots are of very different resolution (Fig. 5). Using the SearchIt application, peaks can be picked from such a loadings plot, and the peaks are searched against the ^{1}H NMR spectral library of 226 standard metabolites. With the loadings plot of Fig. 5a, D-glucose was reported as one of the top hits. With the one of Fig. 5b, however, much fewer useful peaks can be picked and hence the search failed to give sensible results. Even for human analysis, the low resolution loadings plot gives much less information and is less useful.

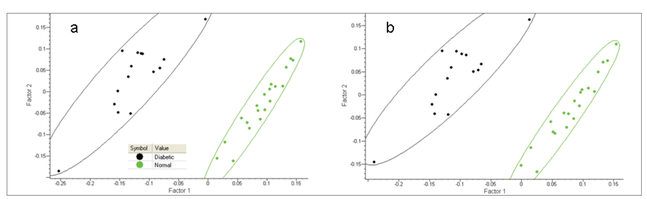

Figure 4. *Similar scores plots from the PCA of the diabetes dataset (I) (a) without binning and (b) with binning using a fixed bin width of 0.04 ppm.*

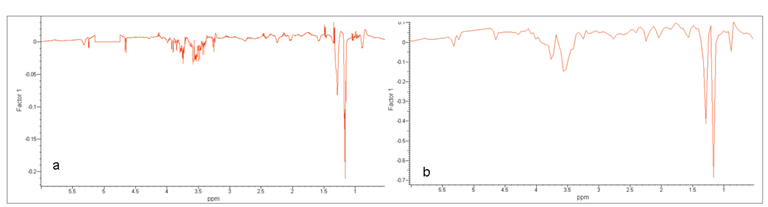

Figure 5. *The loadings plots from the PCA of the diabetes dataset (I) : (a) without binning and (b) with binning using a fixed bin width of 0.04 ppm.*

For dataset II, we used full-resolution data, fixed-width binning, IntelliBucket™, and AFNS. Fig. 6 shows the resulting scores plots (PC1 and PC2).

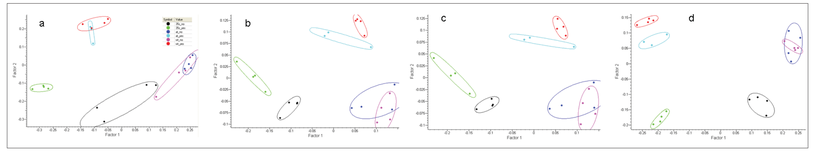

Figure 6. *Scores plots from the PCA of the ATHK1 dataset (II) : (a) No binning, using 17,088 points from the included regions. (b) fixed-width binning (using a bin width of 0.02ppm and generating 218 bins). (c) IntelliBucket binning (using a bin widht of 0.02ppm and variation range of ± 0.005 ppm, OD level =80, generating 223 bins), and (d) AFNS (bin width=0 - 0.02 ppm at step values of 0.002 ppm, critical F=4.6, leading to 10,645 points being used for PCA).*

It is noted that the AFNS algorithm gave the best separation of the classes than the other methods. As shown in Fig. 7, AFNS selected about 10 K significant points from the original 17K points. Although it applied various bin widths for PCA, the resulting loadings plot retains the same resolution as the original spectra, hence it is good for automatic or visual interpretation. We hypothesized that ATHK mutants would have distinct osmolyte profiles when exposed to high salt media but more similar metabolic profiles under low salt conditions. AFNS supported this prediction and demonstrated that the high and low salt condition for every genotype could be discriminated with a single principal component. AFNS also showed that there are significant differences in osmolyte concentrations between the three genotypes. These data support our observation of lower ATHK1 knockout viability and higher over-expresser viability in100mM NaCl in comparison to the small phenotypic variation observed in low salt media.

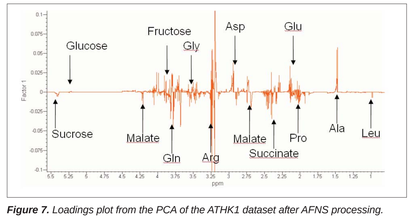

Figure 7. *Loadings plot from the PCA of the ATHK1 dataset after AFNS processing.*

Conclusion

The NMR spectral data used in NMR-based metabolomics studies can be of very different quality and nature. Using the diabetes and ATHK1 datasets, we demonstrated how the different binning tools, in combination with the seamlessly integrated software applications, can help researchers analyze such data effectively and efficiently. While more in-depth studies and evaluation of the ATHK1 dataset are in progress and will be published elsewhere, we can draw the following conclusion based on the current study:

- If the spectral peaks do not manifest significant local chemical shift variation, the global spectral alignment followed by multivariate analysis using the full-resolution dataset is the first choice. The resulting loadings plots, at the original spectral resolution, are well-suited for interpretation of changed metabolites.
- If the spectral peaks have significant local chemical shift variation, a global spectral alignment can be done first to align the majority of the peaks, and next AFNS can be applied to extract the significant spectral features and achieve better classification. The resulting loadings plots also have the original spectral resolution, and are well-suited for interpretation of changed metabolites.
- The traditional binning and bucketing methods, enhanced by using the OD consensus spectra, can still be useful in cases where the number of variable must be limited, or when the classification information is not available for AFNS processing.

References

1. Brown, T.R. and Stoyanova, R., NMR spectral quantiation by principal-component analysis. *J. Magn. Reson. B*. **112**, 32-43 (1996)
2. Brekke, T., Kvalheim and O.M., Sletten, E., Prediction of physical properties of hydrocarbon mixtures by partial-least squares calibration of carbon-13 nuclear magnetic resonance data. *Anal. Chim. Acta*, **223**, 123-134 (1989).
5. Holmes, E., Foxall, P.J.D., Nicholson, J.K., Neild, G.H., Brown, S.M., Beddel, R.R., Sweatman, B.C., Rahr, E., Lindon, J.C., Spraul, M. and Neidig, P., *Anal. Biochem.*, **220**, 284-296 (1994)
6. Vogels, J.T.W.E., Tas, A.C., Venekamp, J. and Van der Greef, J., Partial linear fit: A new NMR spectroscopy preprocessing tool for patern recognition applications. *J. Chemometrics*, **10**, 425-438 (1996)
7. Torgrip, R.J.O., Aberg, M., Karlberg, B. and Jacobsson, S.P., Peak alignment using reduced set mapping, *J. Chemometrics*, **17**, 573-582 (2003)
8. Wu., W., Daszykowski, M., Walczak, B., Sweatman, B.C., Connor, S.C., Haselden, J.N., Crowther, D.J., Gill, R.W. and Lutz, M. W., Peak alignment of urine NMR spectra using fuzzy warping, *J. Chem. Inf. Model.*, **46**, 863-875 (2006)
9. Peng, C., Banik, G., Wang, T., Xia, B., Ramos, S., Toward Diagnosis of Diabetes by NMR and Multivariate Analysis, *Bio-Rad Technical Note*, March 2006
10. Cloarec, O., Dumas, M.-E., Craig, A., Barton, R., H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J.C., Holmes, E. and Nicholson, J., Statistical Total Correlation Spectroscopy: An Exploratory Approach for Latent Biomarker Identification from Metabolic ^{1}H NMR Data Sets, *Anal., Chem.*, **77**,1282-1289 (2005)
11. The raw spectral data were adapted from Biological Magnetic Resonance Data Bank at the University of Wisconsin at Madison: http://www.bmrb.wisc.edu/.
12. Clement, O., Abshear, T., Banik, G. and Peng, C., Overlap Density Heatmap technology, a novel tool for spectroscopic and metabolomics study, *Scientific Computing*, Sept., 2006
13. Silverman, B.W., *Density Estimation for Statistics and Data Analysis*. Chapman and Hall. London, 1986.