January 2006

Negative impact of noise on the principal component analysis of NMR data

Steven M. Halouska
University of Nebraska - Lincoln

Robert Powers
University of Nebraska - Lincoln, rpowers3@unl.edu

Published in *Journal of Magnetic Resonance* 178:1 (January 2006), pp. 88-95.
doi:10.1016/j.jmr.2005.08.016

Negative impact of noise on the principal component analysis of NMR data

Steven Halouska and Robert Powers

Department of Chemistry, University of Nebraska–Lincoln, Lincoln, NE 68522, USA

Submitted May 2005; revised August 2005; published online 27 September 2005; in print January 2006.

Abstract: Principal component analysis (PCA) is routinely applied to the study of NMR based metabolomic data. PCA is used to simplify the examination of complex metabolite mixtures obtained from biological samples that may be composed of hundreds or thousands of chemical components. PCA is primarily used to identify relative changes in the concentration of metabolites to identify trends or characteristics within the NMR data that permits discrimination between various samples that differ in their source or treatment. A common concern with PCA of NMR data is the potential over emphasis of small changes in high concentration metabolites that would over-shadow significant and large changes in low-concentration components that may lead to a skewed or irrelevant clustering of the NMR data. We have identified an additional concern, very small and random fluctuations within the noise of the NMR spectrum can also result in large and irrelevant variations in the PCA clustering. Alleviation of this problem is obtained by simply excluding the noise region from the PCA by a judicious choice of a threshold above the spectral noise.

Keywords: Principal component analysis, Metabolomics, Impact of noise, NMR

1. Introduction

NMR is an extremely versatile analytical tool where the utility of NMR has recently been expanded to include the analysis of the metabolome [1]. Metabolomics is a natural extension of genomics and proteomics where the particular state or activity of a cell is monitored through the quantization of the low-molecular weight molecules present in the cell instead of directly following gene or protein expression levels [2]. Metabolomics has an intrinsic advantage over genomics and proteomics analysis since observed changes in the metabolome are directly coupled with changes in protein activity and cell function. A simple change in the expression level of a gene or protein does not necessarily correlate directly with a change in the activity level of a protein [3].

NMR is routinely being applied to monitor changes in the composition and concentration of metabolites found in biofluids and cell extracts to: (1) monitor drug toxicity [4], [5], [6] and [7], (2) identify disease markers [4], [8], [9], [10] and [11] and (3) explore in vivo protein function and activity [3], [12], [13] and [14]. ^{1}H NMR spectrum collected on the entire metabolome obtained from whole cell lysis or biofluids tend to be extremely complex due to the presence of hundreds of low-molecular weight compounds. Visual inspection or a spectral difference to identify metabolite concentration changes is relatively cumbersome if not generally impractical for large sample sizes. Instead, principal component analysis (PCA) is typically used to decipher changes in NMR based metabolomic data [15] and [16]. PCA is a well established statistical technique that determines the directions of largest variations in the data set, where a metabolomic data set is composed of a series of NMR spectra collected from numerous cell extracts or biofluid samples. The data are generally presented as a two or three-dimensional plot (scores plot) where the coordinate axis correspond to the principal components (representing the directions of the two or three largest variations in the data set). Effectively, each NMR spectrum is reduced to

a single point in the PC coordinate axis, where similar spectra will cluster together and variations along any of the PC axes will highlight experimental differences between the spectra.

The success of the application of PCA in the analysis of NMR metabolomic data is intrinsically dependent on the consistency of sample and data handling [17]. Any observed variations in the NMR data should be related to the state of the cell and organism, as opposed to subtle changes in chemical shifts, line-widths, baseline or artifacts from processing. To minimize these affects and to simplify data handling, NMR spectral data are usually divided into buckets with widths of 0.01–0.04 ppm [18] and [19]. This tends to smooth out errors from fluctuations in chemical shifts and line-shape between NMR spectra caused by sample handling or preparation. Another similar concern is the impact of changes in abundant metabolites relative to changes in the majority of low-concentration chemicals [20]. A relatively small random change in the concentration of an abundant metabolite would still result in an apparent large intensity change that may potentially mask a functionally relevant change in a low-concentration metabolite. The negative impact on the PCA scores plot would be an undesirable clustering of the NMR data that emphasized the irrelevant random changes of the abundant metabolite instead of the changes associated with the functionally relevant low-concentration metabolites. To minimize this issue, a transformation of the original data is performed that enhances the intensity of weak peaks relative to strong peaks and generates a constant variance in the data [10] and [21].

In this article, we describe the observation of another potential source of error in PCA of NMR metabolomic data that resulted in poor clustering of "ideal" NMR data with high similarity. The source of this error is the conceptual opposite of the random fluctuations of intense signals from abundant metabolites described above and as a result was completely unexpected. Extremely small variations within the noise of high signal-to-noise NMR spectra had a significantly and surprisingly negative impact in the quality of the clustering in PCA scores plot.

2. Materials and methods

2.1. NMR data collection

The NMR metabolomics test data sets consisted of three individual samples composed of either 500 mM or 1 mM of (i) ATP, (ii) glucose, and (iii) ATP and glucose. The compounds were dissolved in 99.8% D_2O with 50 mM phosphate buffer at pH 7.2 (uncorrected) and 5 mM of TMSP. The NMR spectra were collected on a Bruker 500 MHz Avance spectrometer equipped with a triple-resonance, *z*-axis gradient cryoprobe. ^{1}H NMR spectra were collected with 128 transients at 298 K with solvent presaturation of the residual HDO, a sweep-width of 5482 Hz and 32 K data points. Ten duplicate ^{1}H NMR spectra were collected sequentially for each of the three samples for a data set consisting of 30 NMR spectra for both the 1 and 500 mM set of samples.

2.2. Statistical analysis

The two sets of 30 NMR spectra were processed automatically using a macro in the ACD/1D NMR manager (Advanced Chemistry Development, Toronto, Ontario). The NMR data were Fourier transformed, zero-filled, phased and baseline corrected. The NMR spectra were processed using multiple protocols to eliminate the possible contribution of data processing to the observed spread in the PCA. The NMR spectra were processed with zero, one and four zero-fillings. The baseline was corrected using spectrum averaging or a polynomial fit of the noise. For spectrum averaging, the spectrum regions that do not contain signals are automatically defined by using a rectangular box (box half-width of 30 points). A peak is defined as having intensity 5-times greater than the noise standard deviation, where noise is defined as the minimal Root Mean Square error. The baseline is constructed by averaging the spectrum curve over these regions. Similarly, for polynomial fit the spectrum is equally divided into 64 regions. A polynomial of order 4 is fit to the regions that only contain noise. The polynomial is then subtracted from the entire spectrum.

The residual H_2O NMR resonance between 4.87 and 5.13 ppm was set to zero and excluded from the bucketing and PCA analysis. Each spectrum was referenced with the TMSP peak set to 0.0 ppm. A table of integral intensities bucketed into bins with a width of 0.025 or 0.04 ppm using the ACD intelligent or standard bucketing schemes were then exported to MS Excel. Instead of using a uniform bucket size of 0.025 or 0.04 ppm throughout the spectrum, the ACD intelligent bucketing protocol places the bucket divisions at local minima within the spectrum to avoid the splitting of peaks between buckets. The smaller bucketing size of 0.025 ppm resulted in a slightly better clustering of the data (see Supplemental Figs. 1S and 2S). There is a 1.48% improvement along P1, and 0.12% improvement along P2 in the variance using the 0.025 ppm bucketing size. An MS Excel macro was then used to combine the 30 spectra into a single file to normalize the binned intensities to a total integrated intensity of 1.0. The Excel spreadsheet was then imported into SIMCA (UMETRICS, Kinnelon, NJ) for PCA. Exclusion of the noise regions of the ^{1}H NMR spectra was accomplished by either limiting the bucket analysis in ACD/1D NMR manager to regions of the NMR spectrum that contained manually defined peaks or by an Excel macro that set the value of every bin below a certain intensity threshold to zero.

3. Results and discussion

3.1. Principal component analysis of simulated metabolomic data

As a starting point to familiarize ourselves with the application of principal component analysis (PCA) of NMR based metabolomics data, we initiated a pilot study of simulated metabolomic data. The simulated metabolomic data simply consisted of three NMR samples composed of ATP, glucose and an ATP–glucose mixture. To optimize the similarity in the experimental data, duplicate NMR spectra were collected using the same sample. This provided us with a data set that was expected to yield tight clustering among the repeat data sets and known variances between the three unique samples.

The first comparison made was between the ATP and the ATP–glucose mixture samples. Again, we anticipated the major variance observed along PC1 would be attributed to the glucose NMR signals. Similarly, the variance along PC2 was expected to be attributed to instrument instability. We expected to see a relatively large variance along the PC1 axis and a tighter cluster along PC2. The PCA scores plot of the ATP and ATP–glucose mixture samples is illustrated in Fig. 1.To our surprise, we observed a relatively large scattering along PC2, equivalent in magnitude to the separation in PC1, but even more troubling was the observation that one of the ATP spectrum (#2) fell outside the 95% confidence level in the PCA plot. This observation was clearly a point of confusion. If this was a "normal" experimental data set, the PCA would flag this data point as an outlier and raise concerns of the origin of this sample, but in our simulated data set this is not a possibility. The samples are all identical. An alternative explanation that may have lead to this outlier would be a failure in either the data collection or the processing of the NMR spectrum.

The success of PCA of NMR metabolomics data is intrinsically tied to the consistency in the handling, preparation, collection, and processing of the NMR data [17]. Problems in phasing, referencing, baseline correction or instrument stability would easily lead to the observed scatter and the outlier seen along PC2. But, if any of these problems were present it would also result in a similar scatter along PC1. This is clearly not the case. It is also apparent that these processing or acquisition problems are not present by visually inspecting the NMR spectra. Fig. 2 compares the outlier ATP spectrum (#2) against the ATP spectrum (#9), which has a minimal variation along PC2. There is no visual difference between these two spectra that would easily justify the large difference along PC2.

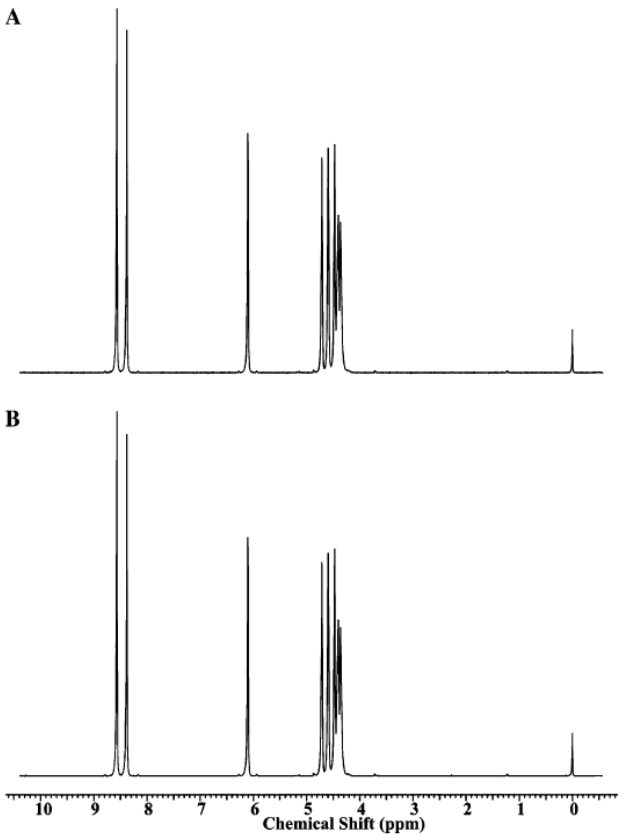

Fig. 2. ^{1}H NMR spectra of the (A) outlier ATP (#2) spectrum and (B) ATP (#9) spectrum with minimal variation along PC2.

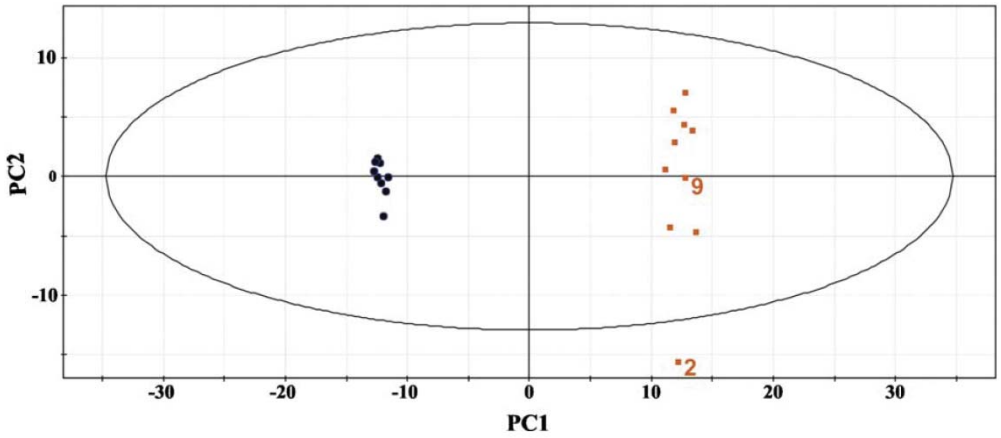

Fig. 1. PCA scoring plots of the set of 10 ATP (■) and ATP–glucose (•) NMR spectra.

To further verify that the processing protocol did not contribute to the large variations along PC2, the data were processed by varying the type of baseline correction, the number of zero-fillings, bucket width, and the method of binning (see Supplemental Figs. 1S–6S). Reducing the bucket width from 0.04 to 0.025 ppm did result in a small improvement in the scattering, with a 1.48% improvement along PC1 and a 0.12% improvement along PC2. Changing the baseline correction from a polynomial fit to spectral averaging or changing the number of zero-filling from zero to four or changing the binning method from intelligent bucketing to standard bucketing either had no beneficial effect or increased the PC2 variation. Interestingly, the specific characteristics of the PCA scores plots (absolute position of data points along the PC1 and PC2 axis) were sensitive to the details of the processing parameters even though the general appearance of each NMR spectrum was unchanged.

It is also conceivable that the relatively high sample concentration of 500 mM may have inadvertently contributed to the PC2 variation. To address this issue, the experiments were repeated exactly as before where the ATP, glucose and ATP–glucose concentrations were reduced to 1 mM. Essentially identical results were observed with the lower concentration samples (see Supplemental Figs. 7S–10S). This clearly indicates that sample concentration is not the source of the PC2 variation.

It is also interesting to note that the relatively more complicated ATP–glucose NMR spectra experiences a significantly smaller PC2 fluctuation compared to the ATP NMR spectra. This suggests that the observed PC2 variation is not primarily related to instrument stability since an opposite result would be expected. Simply, the larger number of NMR peaks present in the ATP–glucose spectrum increases the probability that a random fluctuation in peak intensity caused by instrument instability would occur between sequential data collection. Effectively, the ATP–glucose sample contains more probes to monitor instrument stability.

3.2. Difference loadings plot analysis of simulated metabolomic data

Comparison of the PC2 loading plots between the outlier ATP (#2) spectrum and the ATP spectrum (#9), which has a minimal variation along PC2, identifies the surprising source of the spread along PC2 (Fig. 3). The NMR bins that are responsible for most of the differentiation along PC2 are primarily associated with noise regions in the spectrum. Even more startling is the fact that the relative fluctuation in the intensity of these noise bins is extremely small compared to the intensity of real peaks in the spectrum. Fig. 4 illustrates an expanded view of one of the noise regions of the outlier ATP spectrum (#2) compared against the ATP spectrum (#9), which has a minimal variation along PC2. This expanded noise region of the NMR spectrum contains a large positive variation in the difference loadings plot (bin 363, 7.82–7.83 ppm), boxed area in Fig. 3. The expanded noise region does illustrate some random spikes in the noise that exhibit intensities greater than the average noise bands. These noise spikes are consistent with normal and expected variations in the instrument noise, and appear to correlate with the large variations observed in the PC2 loading plot. Nevertheless, the magnitude of the noise spikes and PC2 loadings do not appear to correlate. The largest PC2 loading for bin 363 (7.82–7.83 ppm) is 15, but the noise spike is lower in intensity compared to the spikes at 8.00 and 8.03 ppm, which have corresponding PC2 loadings that range from ~1 to 4. It is also important to keep the relative magnitude of these noise spikes in perspective with the remainder of the NMR spectrum. The relative intensity of the noise compared to real peaks, including ^{13}C satellites, is effectively zero (Fig. 3). On this scale, the relative intensity of the noise spikes compared to the typical noise band would be expected to be inconsequential and irrelevant. Table 1 lists some of the intensity values in the NMR noise bins that are responsible for the outlier ATP spectrum (#2) with corresponding values for other ATP spectra. Again, the in-

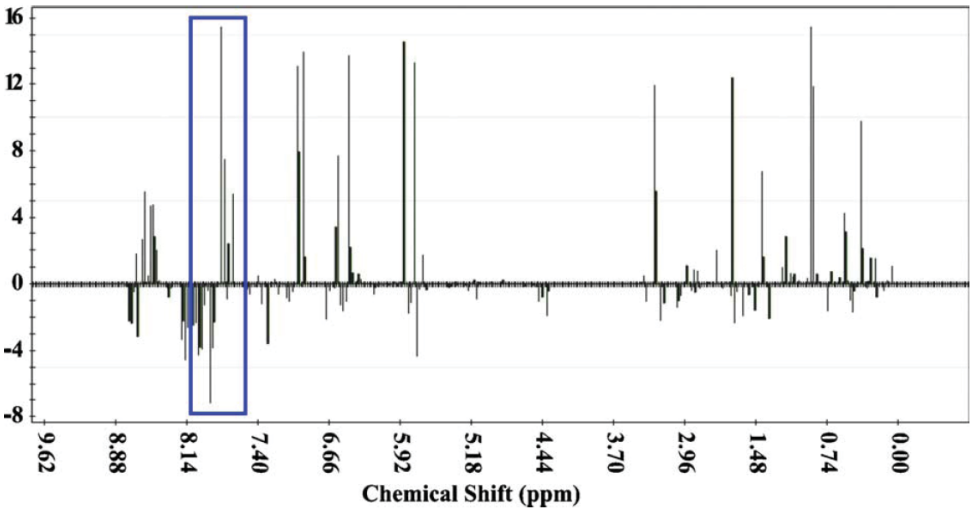

Fig. 3. PCA loading plots difference from the comparison of the outlier ATP (#2) spectrum and an ATP (#9) spectrum with minimal variation along PC2. The boxed area corresponds to the expanded noise regions illustrated in Fig. 4.

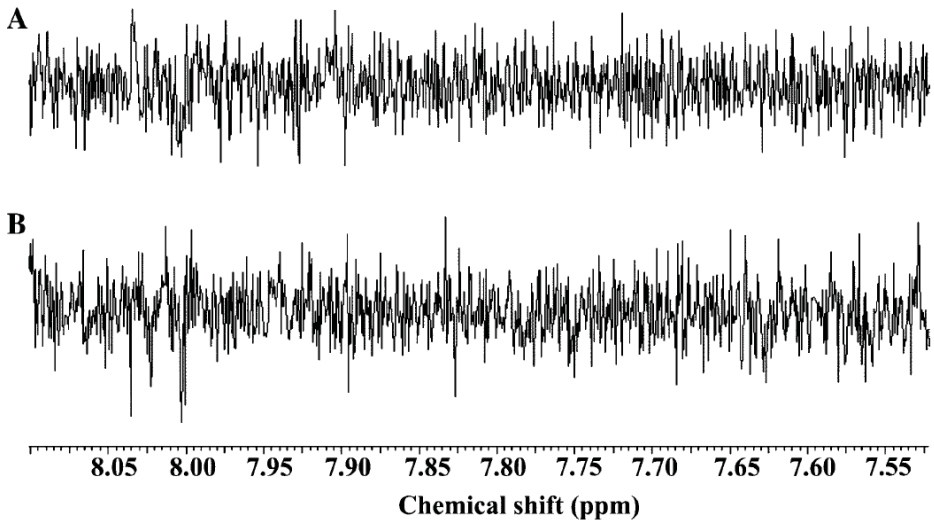

Fig. 4. Expanded view of the ^{1}H NMR noise region for the (A) outlier ATP (#2) spectrum and (B) ATP (#9) spectrum with minimal variation along PC2.

Table 1. Select intensity values of NMR noise after binning and normalization

Bin	(ppm)[a]	1	2[b]	3	4	5	6	7	8	9[c]	10
ATP spectra number											
358	[7.69...7.71]	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	3.08E−07	0.00E+00	6.89E−08
359	[7.71...7.74]	0.00E+00	6.14E−07	0.00E+00	3.17E−07	0.00E+00	6.67E−08	0.00E+00	0.00E+00	0.00E+00	1.31E−07
360	[7.74...7.76]	0.00E+00	0.00E+00	1.36E−06	0.00E+00	2.80E−07	0.00E+00	1.68E−08	0.00E+00	1.89E−07	0.00E+00
361	[7.76...7.79]	0.00E+00	2.33E−06	7.26E−07	1.68E−06	1.69E−07	0.00E+00	0.00E+00	0.00E+00	0.00E+00	6.25E−08
362	[7.79...7.82]	1.79E−06	0.00E+00	0.00E+00	1.40E−06	0.00E+00	1.61E−06	0.00E+00	0.00E+00	0.00E+00	1.11E−06
363	[7.82...7.83]	0.00E+00	1.04E−05	1.43E−07	3.50E−07	2.69E−08	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
364	[7.83...7.86]	0.00E+00	0.00E+00	0.00E+00	3.58E−07	0.00E+00	0.00E+00	1.08E−06	0.00E+00	0.00E+00	0.00E+00
365	[7.86...7.88]	1.15E−06	0.00E+00	3.95E−06	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.44E−06	7.72E−07
366	[7.88...7.91]	2.44E−05	8.51E−06	3.83E−05	2.34E−05	1.32E−05	3.42E−05	3.07E−05	1.05E−05	2.93E−05	9.75E−06
367	[7.91...7.92]	1.30E−05	0.00E+00	1.71E−05	1.08E−05	3.68E−07	1.63E−05	1.86E−05	0.00E+00	1.33E−05	8.64E−07
368	[7.92...7.94]	2.04E−05	0.00E+00	2.83E−05	2.33E−05	9.09E−06	2.63E−05	2.46E−05	2.39E−06	1.94E−05	4.19E−06

[a] Subset of the noise displayed in Fig. 4. The list of binned noise is centered around the largest positive peak (bin 363, 7.82–7.83 ppm) in the boxed region of Fig. 3.

[b] Binned noise for the outlier ATP spectrum number 2.

[c] Binned noise for the ATP spectrum with minimal variation along PC2.

tensity of these noise bins is effectively zero with small random fluctuations about 10^{-5}–10^{-7}, where some values are exactly zero. A large PC2 loading value was observed for bin 363 in ATP spectrum #2, where the intensity of this bin is 1.04 e^{-5} in spectrum #2 but varies from 0 to 1.43 e^{-7}. Apparently, since bin 363 for most of the ATP spectra is 0, a large PC2 loading value is attributed to ATP spectrum #2 because of a large *relative* difference even though the absolute difference is infinitesimal.

The contribution of noise to the difference loadings plot was not unique to the comparison of ATP spectra #2 and #9. Similar results were observed when other spectra were compared (see Supplementary Fig. 11S). The major differences in the difference loading plots were associated with noise regions, but the specific characteristics of the difference loadings plots varied randomly. The position and intensity of the spikes varied between the difference loadings plots. Again, this is consistent with the variability observed along PC2 in the PCA scores plot (Fig. 1) and the expected variability of noise. Clearly, these observations imply that the presence of noise may be detrimental to accurate clustering in NMR PCA scores plot.

3.3. Principal component analysis with a noise threshold

Assuming that the difference loadings plot analysis correctly identified that the PC2 variation is due to these extremely small fluctuations in noise regions of the spectrum and not another artifact of the PCA, the ATP and ATP–glucose NMR spectra were re-analyzed with the exclusion of noise from the binning. This was accomplished by either binning regions of the NMR spectrum that only contained peaks or by setting all bins that are below a certain intensity threshold to exactly zero. The PCA improved with the exclusion of the noise. None of the spectra fall outside the 95% confidence level and the relative range of variation along the PC2 axis have been reduced by a factor of 4–5 for the ATP spectra (Fig. 5). Similarly, the percent variance significantly increased for PC1 from 33.12 ± 10.31 to 83.37 ± 7.44% with the exclusion of noise. The contribution of noise to the scores plot was also evident by comparing 1 mM ATP NMR spectra with 500 mM ATP–glucose spectra. The same variance along PC2 was present for 1 mM ATP spectra that was similarly reduced by a ~4- to 5-fold by the exclusion of the noise.

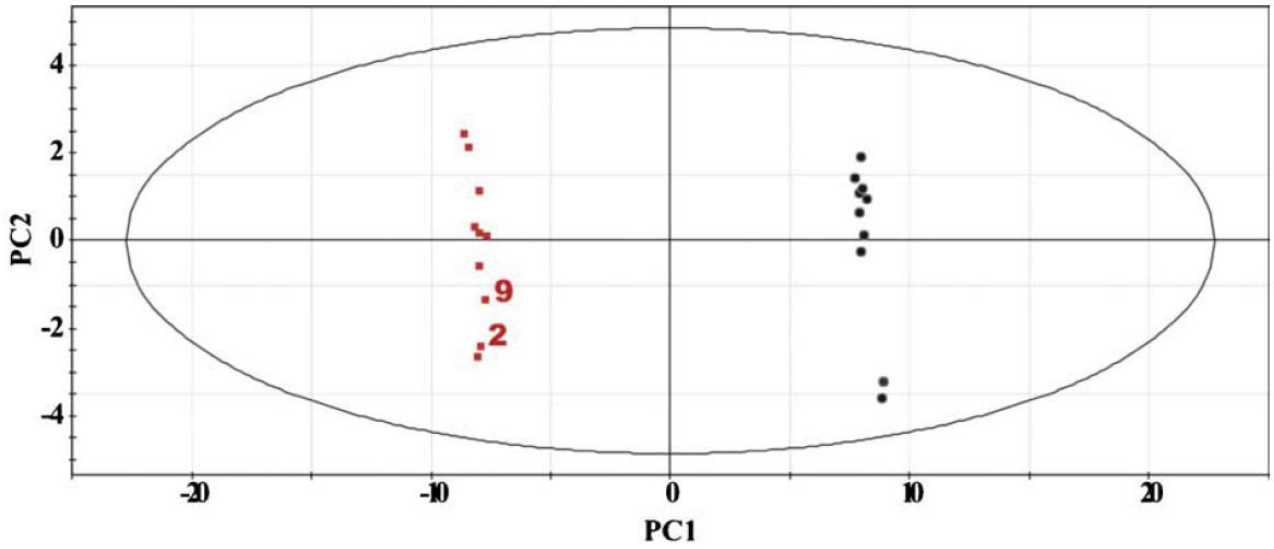

Fig. 5. PCA scoring plots of the set of 10 ATP (■) and ATP–glucose (•) NMR spectra after removal of the spectral noise by only binning NMR resonances.

Conversely, 500 mM ATP–glucose spectra were tightly clustered even with the inclusion of noise because of the relatively high signal-to-noise of 500 mM ATP–glucose NMR spectra compared to 1 mM ATP spectra (see Supplemental Figs. 12S–13S and Table 1S). Again, the variance along PC2 is directly correlated to the presence of noise in the NMR spectrum.

The noise component of an NMR spectrum does not convey any valuable information in the analysis of metabolomic data, but it is routinely included to simplify the data handling. This was based on the reasonable assumption that the inclusion of noise in the binning of NMR spectra would have a neutral impact on the PCA, where the binning process itself would minimize the noise intensity and its variation. The largest variations expected to be identified in PCA would be changes in the intensity of various metabolite NMR resonances. Unfortunately, our analysis indicates that small random changes in spectral noise may contribute to large incorrectly perceived variations in NMR spectra.

3.4. PCA including the glucose NMR data with and without a noise threshold

To determine if the observed large variation along the PC2 axis was an artifact created by comparing just two distinct and tightly clustered data sets, we added a third related NMR sample to the analysis. The third NMR sample only contains glucose and is expected to induce a significant PC2 variation in the PCA score plot. The PCA scores plot of the ATP, glucose and ATP–glucose NMR samples is illustrated in Fig. 6.As expected, large PC1 and PC2 variations result from the different composition of the three NMR spectra, effectively forming an equilateral triangle in the scores plot. The separation along either PC1 or PC2 is considerably larger than the variability among any members in the three distinct clusters. Nevertheless, the inclusion of the NMR noise region still results in a noticeable spread among the repeat NMR spectra within each cluster, especially for the ATP and glucose samples (Fig. 6A). The larger spread for the ATP and glucose samples is consistent with the fact that these NMR spectra

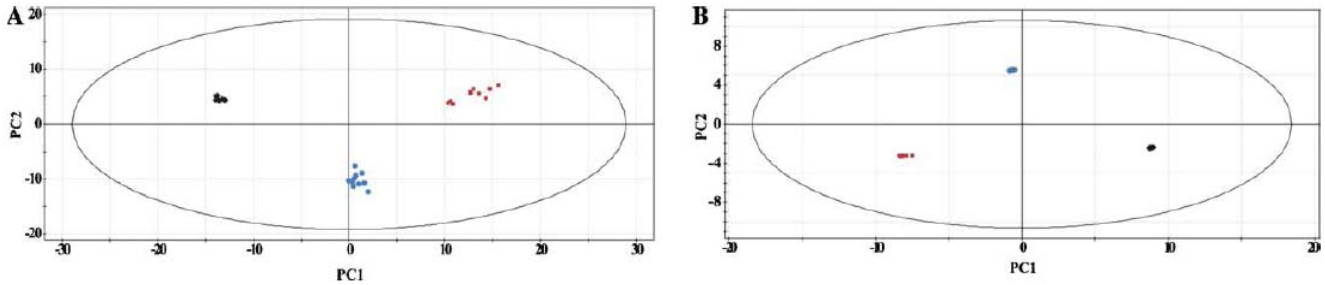

Fig. 6. PCA scoring plots of the 10 ATP (■) ATP–glucose (•) and glucose (●) NMR spectra with the (A) inclusion and (B) exclusion of noise.

would have more noise regions relative to the ATP–glucose spectra. Removal of the NMR noise regions results in a significant improvement in the clustering pattern in the PCA scores plot (Fig. 6B). As initially expected, the repeat NMR spectra are essentially on top of each other in each of the three clusters. Clearly, the inclusion of NMR noise regions results in a significant spread in the clustering of the PCA scores, where the noise does not correlate with any relevant sample characteristics.

In this "ideal" NMR metabolomic data, the large separation present in PC1 and PC2 permits easy discrimination of the ATP, glucose and ATP–glucose spectra despite the observed spread within each cluster caused by the presence of noise. Generally, this would not be the case when dealing with "real" biological data obtained from numerous cell lysis or biofluid samples. Thus, the level of discrimination expected from a set of typical NMR based metabolomic data may be compromised by the inclusion of NMR noise. The irrelevant spread in clustering induced by NMR noise may actually obscure the underlying features in the data resulting in the loss of any informative clustering in the PCA scores plot. Therefore, the standard protocol for processing NMR data for PCA should include the exclusion of noise especially since the noise provides no valuable information while potentially distorting the proper analysis of the NMR data.

Since, biofluid or cell extract data may contain weak NMR resonances that may be associated with functionally important metabolites, the choice of an appropriate noise threshold is critical to avoid the inadvertent elimination of these potentially valuable peaks. An iterative approach that adjusts the noise threshold to minimize the spread between repeat data points while simultaneously maximizing the separation between data collected under various cellular conditions may provide a mechanism to remove the negative impact of noise without compromising the data. A threshold corresponding to one standard deviation of the noise would be a reasonable starting point for the iterative approach where an upper-limit less than 2–3 times the noise would avoid eliminating peaks that can be reliably differentiated from the noise band.

4. Conclusion

The principal component analysis of NMR metabolomic data is proving to be a powerful tool for the evaluation of toxicity, protein function, and the identification of disease markers. A fundamental benefit of PCA is the identification of distinct clusters in a scores plot that highlights discriminating characteristics reflecting the source or treatment of the NMR samples. Essential to the successful interpretation of NMR PCA data is a requirement that the observed variations identified by PCA are related to features of the biological sample and not an artifact of data manipulation or sample handling. Processing NMR data for PCA generally includes binning the entire spectrum, which also incorporates all the noise regions. Our analysis of "ideal" metabolomic data indicates that this inclusion of noise may result in significant and irrelevant spreading of the PCA scores clusters that may inhibit proper interpretation of the data. A simple solution is a routine application of a filter to exclude the noise region below a defined peak intensity threshold.

Acknowledgments

This work was supported by grants from the Protein Structure Initiative of the National Institutes of Health (P50 GM62413), Nebraska Tobacco Settlement Biomedical Research Development Funds and the Maude Hammond Fling Faculty Research Fellowship.

References

[1] R. Goodacre, S. Vaidyanathan, W.B. Dunn, G.G. Harrigan and D.B. Kell, Metabolomics by numbers: acquiring and understanding global metabolite data, *Trends in Biotechnology* 22 (2004), pp. 245–252.

[2] 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* 29 (1999), pp. 1181–1189.

[3] B.H. ter Kuile and H.V. Westerhoff, Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway, *FEBS Letters* 500 (2001), pp. 169–171.

[4]J.P.Shockcor, and E.Holmes, Metabonomic applications in toxicity screening and disease diagnosis. Current Topics in Medicinal Chemistry (Hilversum, Netherlands) 2 (2002) 35–51.

[5] D.G. Robertson, M.D. Reily, R.E. Sigler, D.F. Wells, D.A. Paterson and T.K. Braden, Metabonomics: evaluation of nuclear magnetic resonance (NMR) and pattern recognition technology for rapid in vivo screening of liver and kidney toxicants, *Toxicological Sciences* 57 (2000), pp. 326–337.

[6] J. Van der Greef, E. Davidov, E. Verheij, J. Vogels, R. Van der Heijden, A.S. Adourian, M. Oresic, E.W. Marple and S. Naylor, The role of metabolomics in systems biology: a new vision for drug discovery and development, *Metabolic Profiling* (2003), pp. 171–198.

[7] K. Nicholson Jeremy, J. Connelly, C. Lindon John and E. Holmes, Metabonomics: a platform for studying drug toxicity and gene function, nature reviews, *Drug discovery* 1 (2002), pp. 153–161.

[8] G. Navon, H. Burrows and J.S. Cohen, Differences in metabolite levels upon differentiation of intact neuroblastoma × glioma cells observed by proton NMR spectroscopy, *FEBS Letters* 162 (1983), pp. 320–323.

[9] J. Pfeuffer, I. Tkac, S.W. Provencher and R. Gruetter, Toward an in vivo neurochemical profile: quantification of 18 metabolites in short-echo-time 1H NMR spectra of the rat brain, *Journal of Magnetic Resonance* 141 (1999), pp. 104–120.

[10] H.C. Keun, T.M.D. Ebbels, H. Antti, M.E. Bollard, O. Beckonert, E. Holmes, J.C. Lindon and J.K. Nicholson, Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling, *Analytica Chimica Acta* 490 (2003), pp. 265–276.

[11] R.-J.A.N. Lamers, J. DeGroot, E.J. Spies-Faber, R.H. Jellema, V.B. Kraus, N. Verzijl, J.M. TeKoppele, G.K. Spijksma, J.T.W.E. Vogels, J. van der Greef and J.H.J. van Nesselrooij, Identification of disease- and nutrient-related metabolic fingerprints in osteoarthritic guinea pigs, *Journal of Nutrition* 133 (2003), pp. 1776–1780.

[12] L.M. Raamsdonk, B. Teusink, D. Broadhurst, N. Zhang, A. Hayes, M.C. Walsh, J.A. Berden, K.M. Brindle, D.B. Kell, J.J. Rowland, H.V. Westerhoff, K. Van Dam and S.G. Oliver, A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations, *Nature Biotechnology* 19 (2001), pp. 45–50.

[13] S.G. Oliver, Functional genomics: lessons from yeast, *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 357 (2002), pp. 17–23.

[14] O. Fiehn, Metabolomics-the link between genotypes and phenotypes, *Plant Molecular Biology* 48 (2002), pp. 155–171.

[15] R. Stoyanova and T.R. Brown, NMR spectral quantitation by principal component analysis, *NMR in Biomedicine* 14 (2001), pp. 271–277.

[16] J.C. Lindon, E. Holmes and J.K. Nicholson, Pattern recognition methods and applications in biomedical magnetic resonance, *Progress in Nuclear Magnetic Resonance Spectroscopy* 39 (2001), pp. 1–40.

[17] B.C.M. Potts, A.J. Deese, G.J. Stevens, M.D. Reily, D.G. Robertson and J. Theiss, NMR of biofluids and pattern recognition: assessing the impact of NMR parameters on the principal component analysis of urine from rat and mouse, *Journal of Pharmaceutical and Biomedical Analysis* 26 (2001), pp. 463–476.

[18] A. Ross, G. Schlotterbeck, W. Klaus and H. Senn, Automation of NMR measurements and data evaluation for systematically screening interactions of small molecules with target proteins, *Journal of Biomolecular NMR* 16 (2000), pp. 139–146.

[19] E. Holmes, P.J.D. Foxall, J.K. Nicholson, G.H. Neild, S.M. Brown, C.R. Beddell, B.C. Sweatman, E. Rahr and J.C. Lindon *et al.*, Automatic data reduction and pattern recognition methods for analysis of 1H nuclear magnetic resonance spectra of human urine from normal and pathological states, *Analytical Biochemistry* 220 (1994), pp. 284–296.

[20] W. El-Deredy, Pattern recognition approaches in biomedical and clinical magnetic resonance spectroscopy: a review, *NMR in Biomedicine* 10 (1997), pp. 99–124.

[21] P.V. Purohit, D.M. Rocke, M.R. Viant and L.D. Woodruff, Discrimination models using variance-stabilizing transformation of metabolomic NMR data, *Omics* 8 (2004), pp. 118–130.

Supplementary data associated with this article follows.

Supplementary Material: "Negative Impact of Noise on the Principal Component Analysis of NMR Data" Steven Halouska and Robert Powers

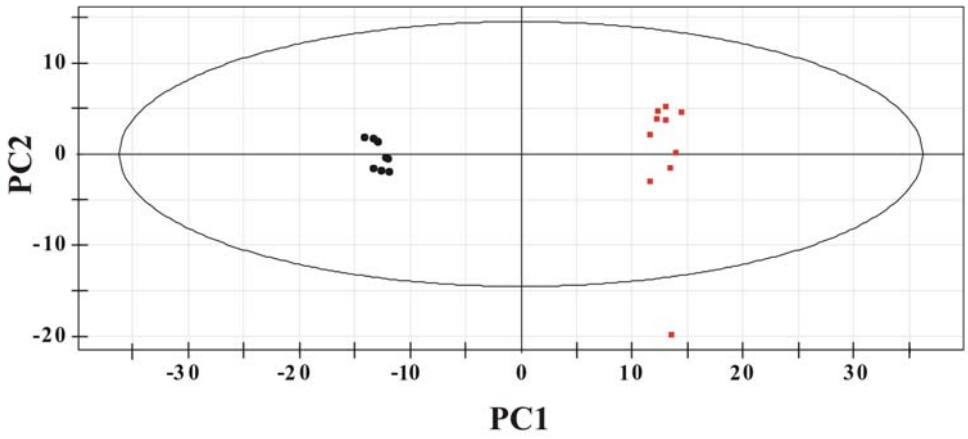

Figure 1S: PCA scoring plot of set of ten 500mM ATP ( ) and 500mMATP-glucose (●) NMR spectra using intelligent bucketing with bin size of .025 ppm.

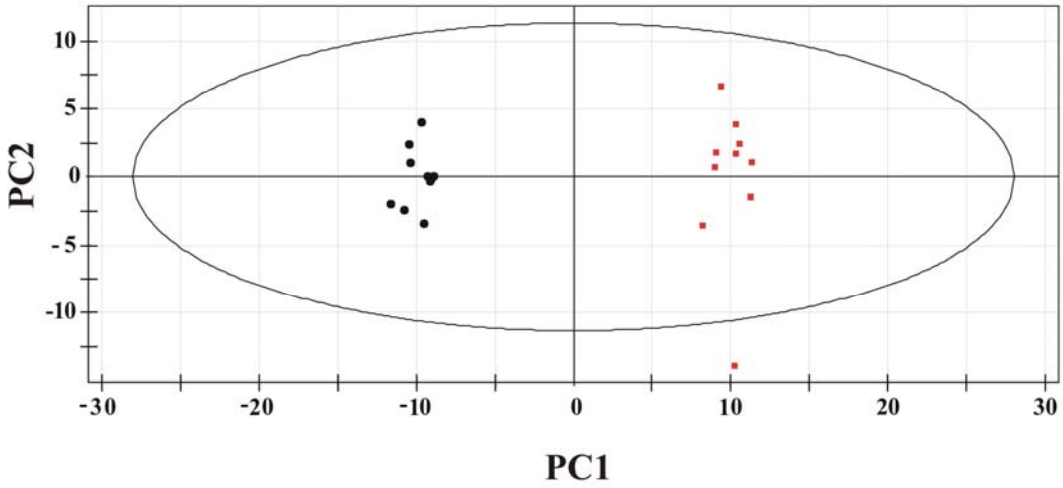

Figure 2S: PCA scoring plot of set of ten 500mM ATP ( ) and 500mM ATP-glucose (●) NMR spectra using intelligent bucketing with bin size of .040 ppm.

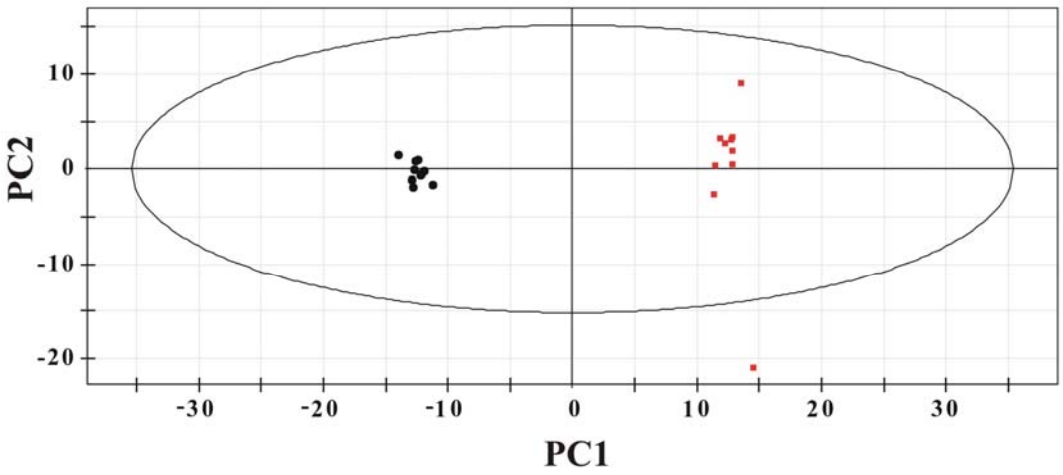

Figure 3S: PCA scoring plot of set of ten 500mM ATP ( ) and 500mM ATP-glucose (●) NMR spectra using standard bucketing with bin size of .025 ppm.

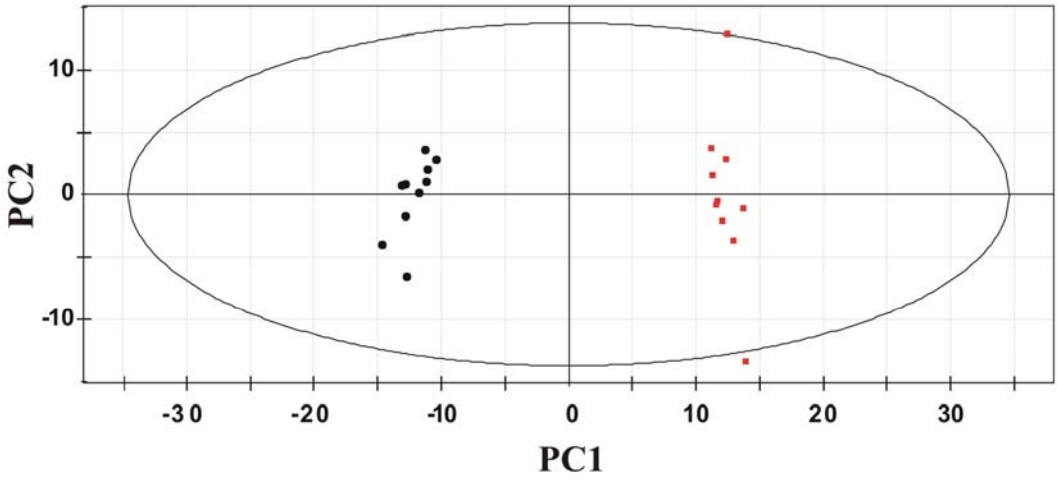

Figure 4S: PCA scoring plot of set of ten 500mM ATP ( ) and 500mM ATP-glucose (●) NMR spectra with 1X zero filling.

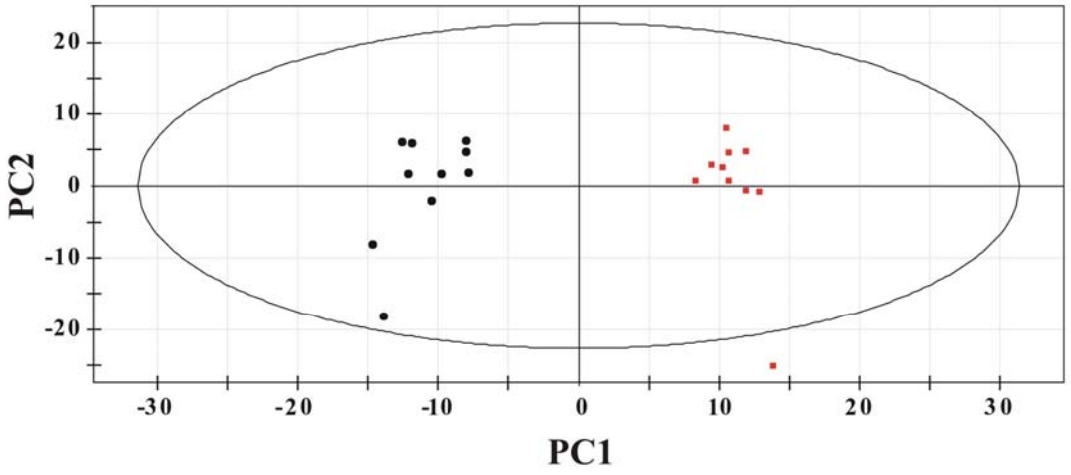

Figure 5S: PCA scoring plot of set of ten 500mM ATP ( ) and 500mM ATP-glucose (●) NMR spectra with 4X zero filling.

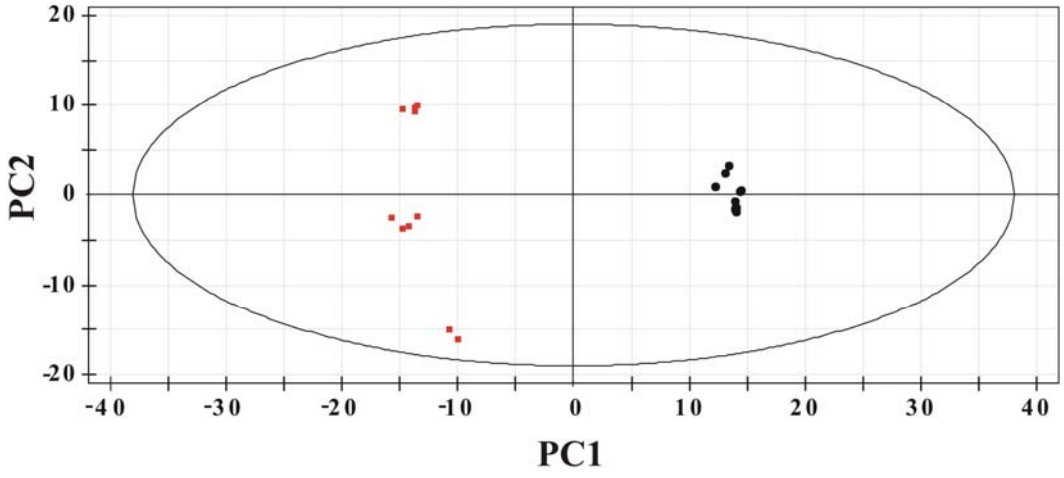

Figure 6S: PCA scoring plot of set of ten 500mM ATP ( ) and 500 mM ATP-glucose (●) NMR spectra with polynomial baseline correction

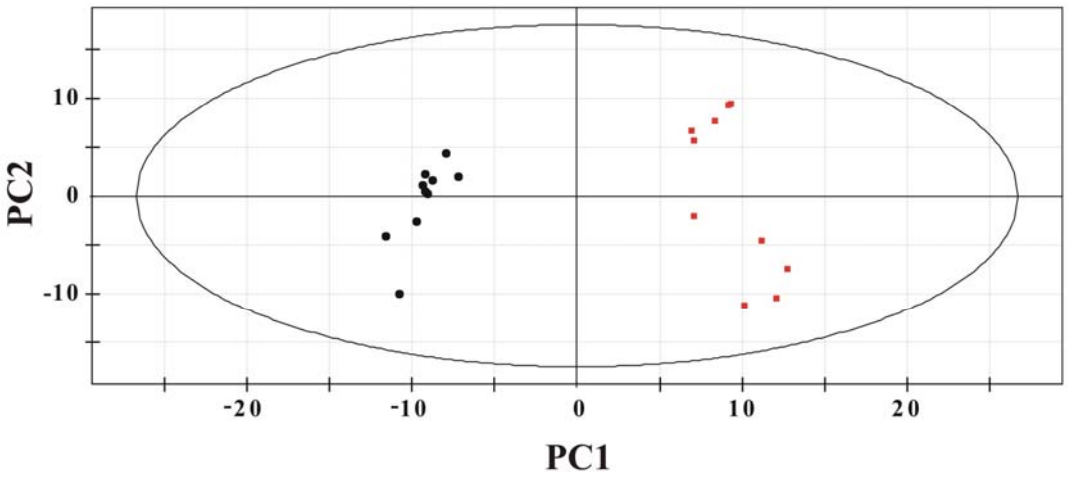

Figure 7S: PCA scoring plot of set of ten 1mM ATP ( ) and 1mM ATP-glucose (●) NMR spectra using intelligent bucketing with bin size of .025 ppm.

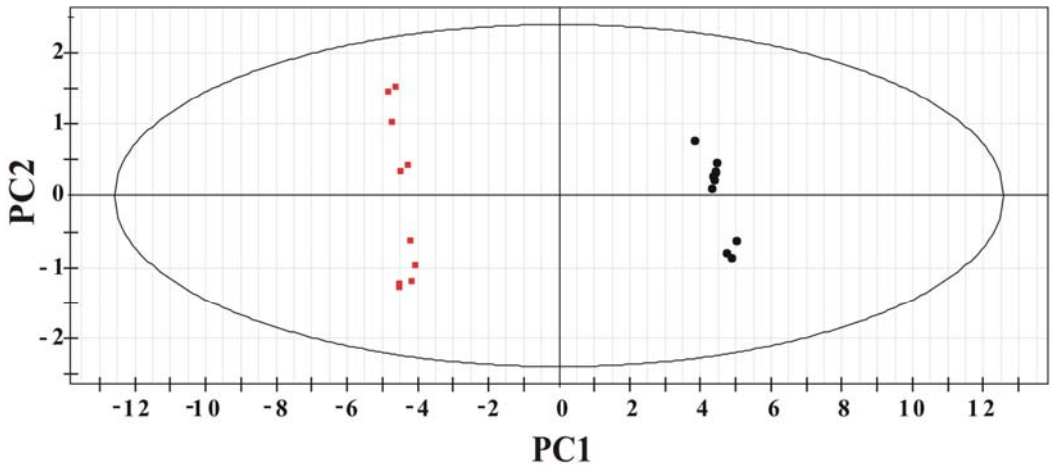

Figure 8S: PCA scoring plot of set of ten 1mM ATP ( ) and 1mM ATP-glucose (●) NMR, editing out the noise.

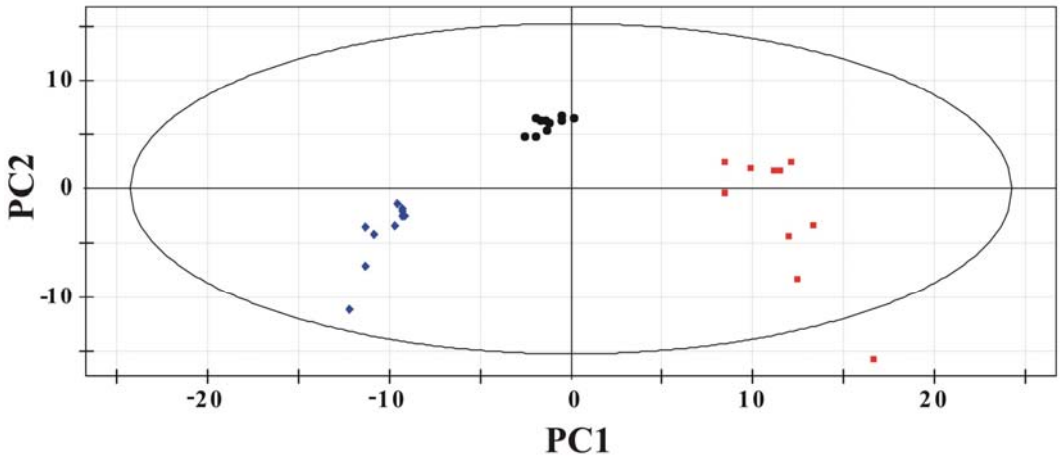

Figure 9S: PCA scoring plot of set of ten 1mM ATP ( ), 1mM ATP-glucose (●), and 1mM glucose (●) NMR spectra using intelligent bucketing with bin size of .025 ppm

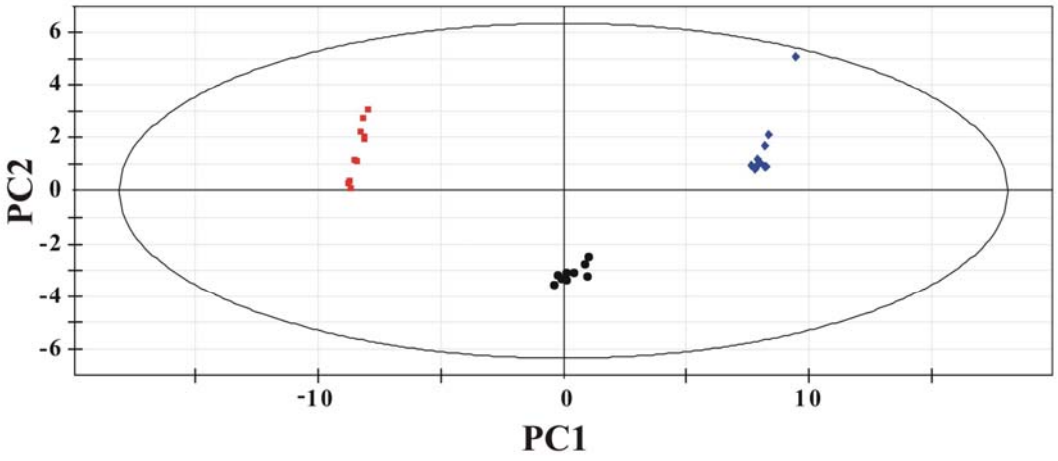

Figure 10S: PCA scoring plot of set of ten 1mM ATP ( ), 1mM ATP-glucose (●), and 1mM glucose (●) NMR spectra, editing out the noise

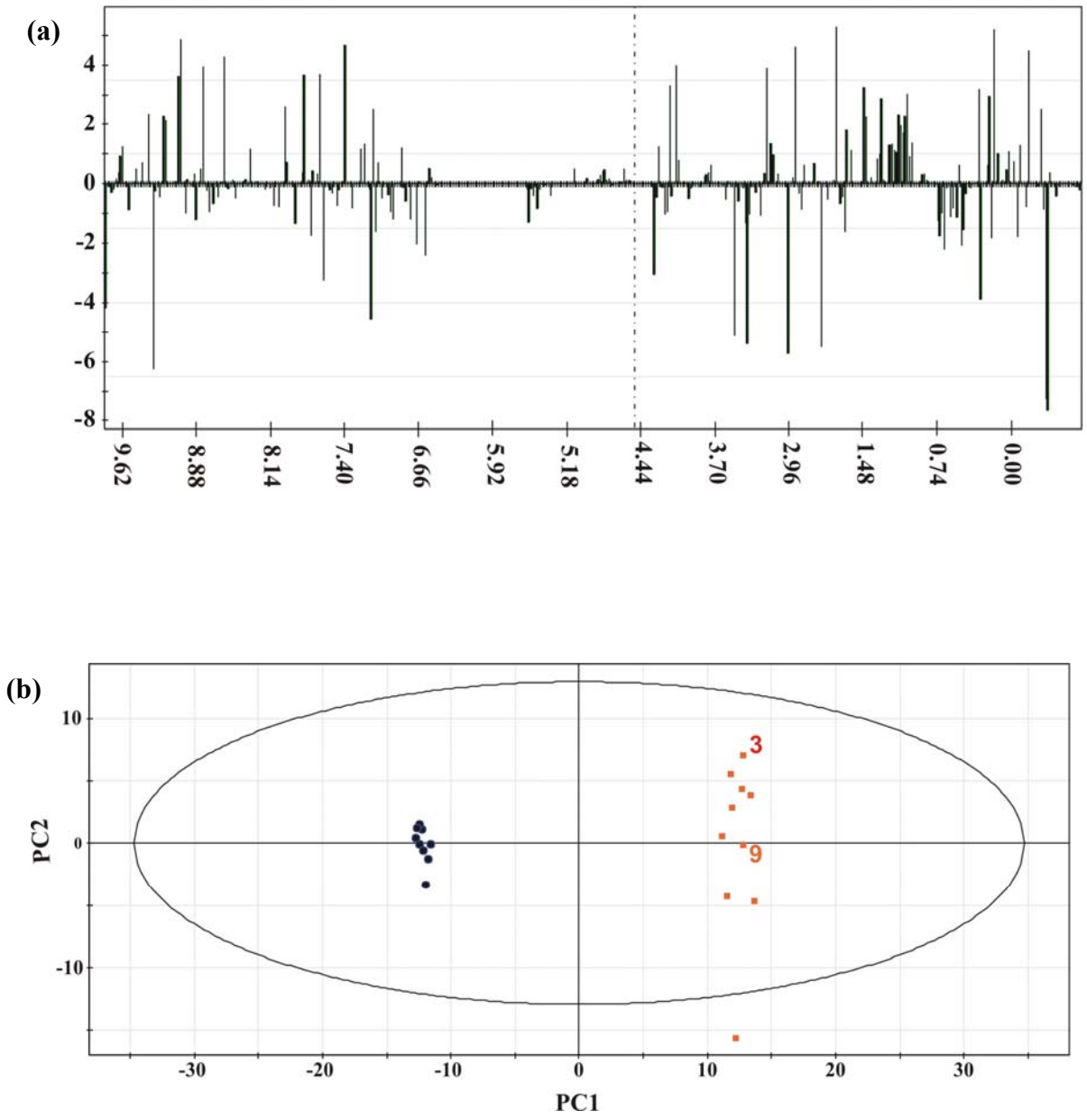

Figure 11S: (a) PCA loading plots difference from the comparison of the ATP (#3) spectrum and the ATP (#9) spectrum with minimal variation along PC2. (b) PCA scoring plots of the set of ten ATP ( ) and ATP-glucose (●) NMR spectra. ATP spectra #3 and #9 are labeled to indicated their relative positioning in the PCA scores plot.

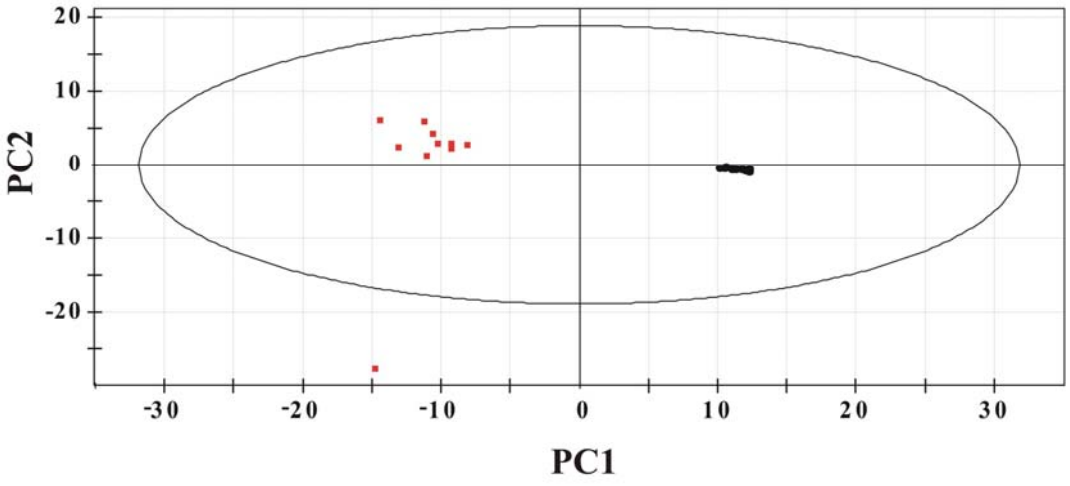

Figure 12S: PCA scoring plot of set of ten 1 mM ATP ( ) and 500 mM ATP-glucose (●).

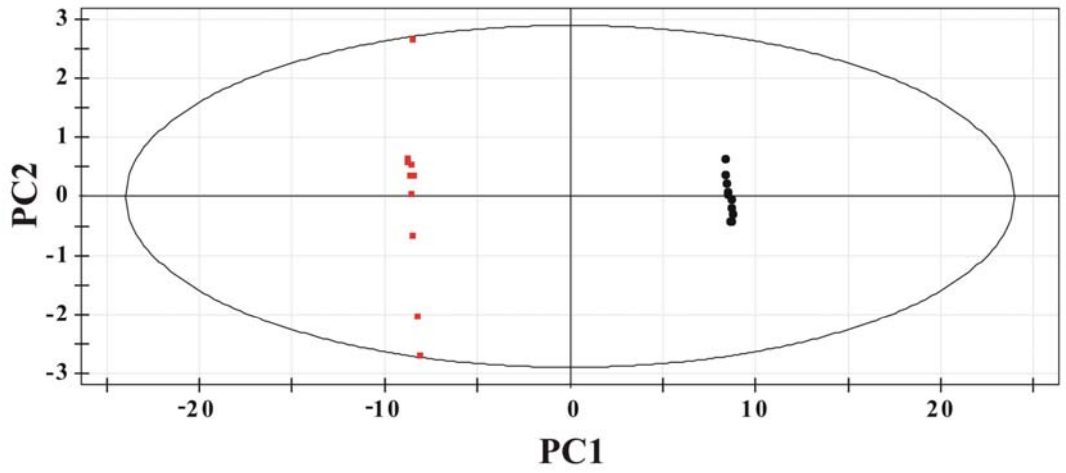

Figure 13S: PCA scoring plot of set of ten 1 mM ATP ( ) and 500 mM ATP-glucose (●), editing out the noise.

Table 1S: Percent variance of the principal component analysis for the various NMR data with the inclusion or exclusion of noise regions.

	Percent Variance	
NMR Sample Set	PC1	PC2
500mM ATP & ATP-Glucose with noise	42.89%	5.95%
500mM ATP & ATP-Glucose signal only	90.64%	4.15%
500mM ATP & Glucose & ATP-Glucose with noise	31.46%	13.84%
500mM ATP & Glucose & ATP-Glucose signal only	74.41%	24.86%
1mM ATP & ATP-Glucose with noise	38.74%	6.20%
1mM ATP & ATP-Glucose signal only	88.19%	10.96%
1mM ATP & Glucose & ATP-Glucose with noise	19.38%	7.73%
1mM ATP & Glucose & ATP-Glucose signal only	80.24%	9.87%