OMICS A Journal of Integrative Biology
Volume 8, Number 2, 2004

Discrimination Models Using Variance-Stabilizing Transformation of Metabolomic NMR Data

PARUL V. PUROHIT,[1] DAVID M. ROCKE,[1,2] MARK R. VIANT,[3] and DAVID L. WOODRUFF[1,4]

ABSTRACT

After the extensive work that is being done in the areas of genomics, proteomics, and metabolomics, the study of metabolites has come of interest in its own right. Metabolites in biological systems give an understanding of the state of the system and provide a powerful tool for the study of disease and other maladies. Several analytical techniques such as mass spectrometry and high-resolution NMR spectroscopy have been used to study metabolites. The data, however, from these techniques remains quite complex. Traditionally, multivariate analyses have been used for such data. These methods however have an underlying assumption that the data is multivariate normal with a constant variance. This is not necessarily the case. It has been shown that a generalized log transformation renders the variance of the data constant effectively making the data more suitable for multivariate analysis. We demonstrate the effectiveness of these transformations on NMR data taken on a set of 18 abalone that were categorized as either being healthy, stunted, or diseased. We show how the transformation makes multivariate classification of the abalone into the healthy, stunted and diseased categories much more effective and gives a tool for identifying potential metabolic biomarkers for disease.

INTRODUCTION

To characterize the complete state of a cell requires, at a minimum, knowledge of the genome, the transcriptome, the proteome, and the metabolome, the latter being the complement of small molecular weight metabolites. Several analytical techniques such as mass spectrometry and high resolution nuclear magnetic resonance (NMR) spectroscopy have been used to measure metabolic profiles of biofluids and tissue extracts. This paper concerns the application of NMR spectroscopy to metabolomics, an area of recent substantial activity (Nicholson et al., 2002; Holmes et al., 2001; Azmi et al., 2002; Beckwith-Hall et al., 2002; Viant et al., 2003; Viant, 2003).

The data, however, from these methods remain quite complex. The measurements show non-constant variance of the two-component model type (Rocke and Lorenzato, 1995; Rocke and Durbin, 2001). This

[1]Center for Image Processing and Integrated Computing, University of California, Davis, California.
[2]Division of Biostatistics, University of California, Davis, California.
[3]School of Biosciences, University of Birmingham, Edgbaston, Birmingham, United Kingdom.
[4]Graduate School of Management, University of California, Davis, California.

variance in homogeneity can be ameliorated by use of a data transformation such as the generalized logarithm (glog) (Box and Cox, 1964; Rocke and Durbin, 2003; Durbin and Rocke, 2003; Durbin and Rocke, 2004; Durbin et al., 2002; Purohit and Rocke, 2003; Geller et al., 2003). Also, one typically faces the task of dealing with hundreds or thousands of measurements per sample (NMR data points) with a much smaller number of samples. Often, multivariate analysis methods such as principal components analysis (PCA) and partial least squares (PLS) are used in NMR metabolomic studies that have focused on classification of subjects (Lindon et al., 2001). These analyses are performed on a data matrix comprising of a matrix whose i^{th} row vector represents NMR data of the i^{th} abalone and the j^{th} column represents the NMR intensity for the j^{th} frequency shift. These shifts measured in parts per million (ppm) remain the same for each abalone and can be considered to be the variables describing each spectrum. In this context, these variables can be referred to as the predictors of the spectra.

This paper investigates the idea that improving the variance homogeneity of the data via the glog data transformation also makes classification and clustering more effective. We do this via an example data set consisting of a set of 18 red abalone (*Haliotis rufescens*), an important aquaculture species, that were categorized as either healthy, stunted, or suffering from withering syndrome (diseased). This classification was based on morphological data, as described previously (Viant et al., 2003). NMR spectra were obtained from extracts of muscle tissue from each animal. We show that PCA and cluster analysis obtain better classification of diseased and healthy samples and identification of disease related predictors after data transformation than before. We performed the glog transformation on this set of data along with other pre-processing techniques and performed PCA and cluster analysis on the data before and after the transformation. The transformation makes the classification of the abalone into the healthy, stunted and diseased categories much more effective for both multivariate methods.

MATERIALS AND METHODS

Sample preparation and NMR spectroscopy

The abalone used for the study were spawned in captivity and kept either in an upstream tank that received fresh seawater at ambient temperature or a downstream tank that received poorer quality water that had passed through several upstream tanks. As a result of the inferior quality of the downstream tank, the abalone in that tank became either diseased or stunted as measured by the shell length, which was considerably shorter than the abalone from the upstream tank. The abalone with shorter shells were labeled stunted if they otherwise appeared healthy. The abalone considered diseased were the ones that had considerable atrophy of the foot muscle associated with withering syndrome, a fatal disease in abalone.

A set of 18 abalone were used for this study: eight of which were healthy, five stunted, and five diseased. For each abalone, tissue samples were obtained from the foot muscle of the abalone. In addition, six samples were obtained from one other abalone (healthy) which were used as replicates to estimate the parameter associated with the generalized log transformation.

One dimensional ^{1}H NMR spectra of abalone muscle tissue extracts were measured at 500.11 MHz using an Avance DRX-500 spectrometer (Bruker, Fremont, CA), as described previously (Viant et al., 2003). The resulting NMR spectra were processed using standard methods yielding a dataset comprising 65,536 points from approximately 12 to -2 ppm (Viant et al., 2003). They were phase and baseline corrected and then calibrated (TMSP peak at 0.0 ppm) using XWINNMR software (version 3.1, Bruker).

Data pre-processing

Several pre-processing steps had to be undertaken before the transformation and multivariate analysis of the data could begin. We restricted ourselves to the frequency range of 0.2–10.0 ppm, which covered the regions of interest in each spectrum. The range outside this region presented a relatively flat region with no significant peaks. Data between 4.72 and 4.96 ppm containing the residual water peak was removed.

Then we normalized each spectrum to a total integrated intensity of 1.0. This allowed for the peak intensities to be independent of the quantity of the analyte. We also noticed that in the diseased abalone, the

overall intensity of the spectrum was lower than in the case of healthy and stunted abalone due to the withered nature of the diseased abalone. Normalization of these spectra facilitated comparison with the healthy and stunted spectra allowing for the identification of predictors that would be unique between the three cases rather than predictors that are different due to intensity only. The identification of these predictors can be used to specify metabolite biomarkers in the healthy versus diseased scenario.

A spectrum from Abalone 1 which is characteristic of a typical spectrum after baseline correction and normalization is shown in Figure 1a. In the range of interest, there were approximately 45,000 data points, a very large number of variables compared to the 18 observations in the study. Data reduction had to be performed in the predictor space. Binning techniques have been used in previous studies as a preliminary dimension reduction technique. In this approach, the predictors within a window of a defined width in ppm

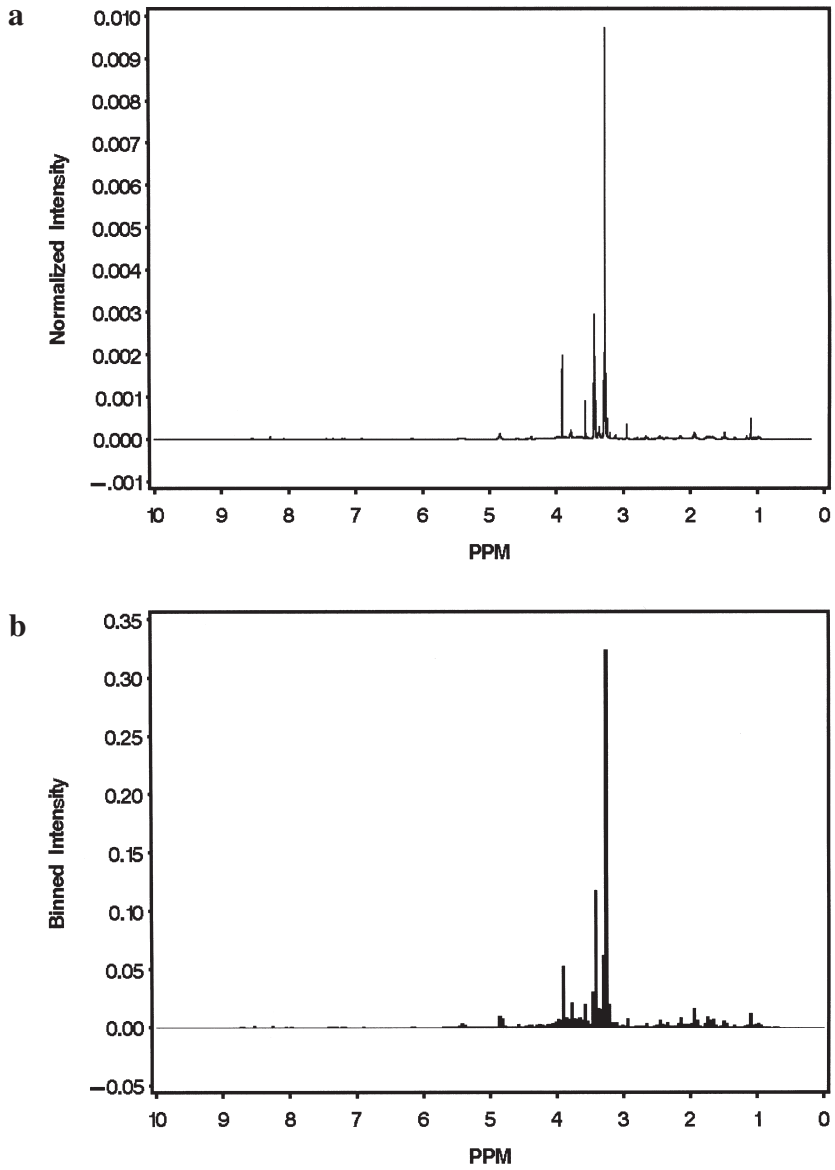


FIG. 1. (**a**) Typical ^{1}H NMR spectrum of muscle extract from a healthy abalone after baseline correction and normalization. (**b**) NMR spectrum after binning. The bin size here is 0.04 ppm and reduces the spectrum to 239 bins. These spectra correspond to Abalone 1.

is replaced by a single value corresponding to the sum of the intensities in the window. The window width, also known as the bin size, that has been most often used has been 0.04 ppm (Nicholson et al., 2002; Holmes et al., 2001). We employed this bin size for the study presented here. This meant that we collected 187 or 188 data points in a single bin. The width of a typical peak in the 500 MHz NMR spectra is ~0.0025 ppm and so the bin size is sufficiently large to encompass several peaks. However, the overall character of the spectrum is not affected by the size of this window as seen in Figure 1b as compared to the original spectrum. After binning, the spectrum dimension was reduced to 239 with the mid ppm value of each bin being considered the new predictor, a significant reduction from the original 45,000. Our data matrix that is used for multivariate analysis now has 18 rows and 239 columns.

Transformation

Many dimension reduction, clustering, and multivariate statistical techniques are based on the assumption that interpretation of variability does not need to depend on the measurement level. For many high-throughput assays, the data does not necessarily follow such an assumption and the variance in these data tends to rise with the intensities. One method that can be used to stabilize the variance is to perform a generalized log transformation of the form

$$g(\lambda) = ln\ (y^2 + \sqrt{y^2 + \lambda}) \tag{1}$$

where y is the original spectroscopic intensity, g the transformed intensity and λ is a transformation parameter. Similar transformation methods have been used for micro-array data but not for processing NMR data. It has been shown that the parameter λ can be estimated using a maximum likelihood method using a set of replicate measurements (Box and Cox, 1964; Rocke and Durbin, 2003; Durbin and Rocke, 2003, 2004). We employed a set of six replicate samples from one abalone for the estimation of the parameter λ. In this case, foot muscle tissue from one abalone was divided into six parts and independent NMR spectra obtained for each of the samples. As outlined in Durbin and Rocke, 2003, the likelihood can be maximized by minimizing the sum of squares of errors (SSE) of the transformed, Jacobian-corrected data. This optimization can be done by, for example, Newton's method, but in this case, we used a simple grid-search technique. Here, we estimated SSE for values of λ monotonically increasing by $0.1E^{-7}$. A plot of SSE as a function of λ for the 0.04 ppm bin size is shown in Figure 2 using the normalized data of the six replicates. For this bin size, the minimum in SSE is achieved for a value of λ equal to $2.7E^{-7}$.

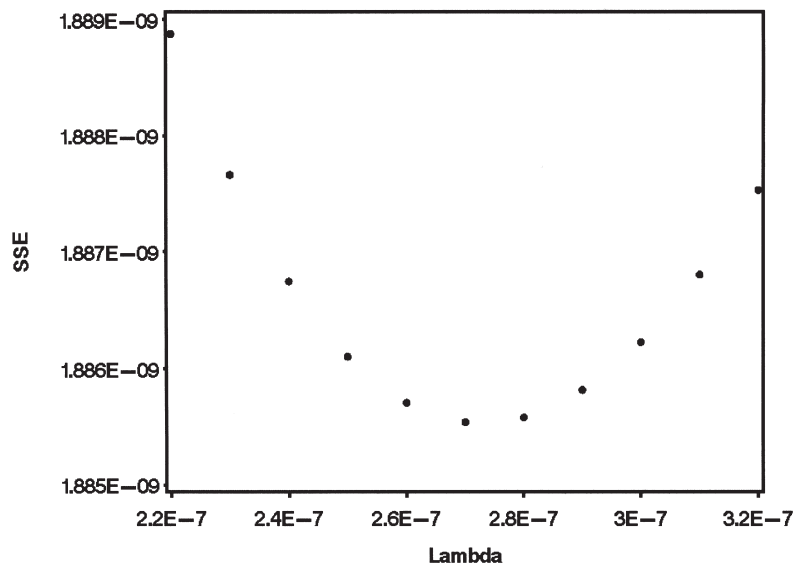


FIG. 2. A plot of the error sum of squares (SSE) versus λ using the maximum likelihood method. The value of λ at the minimum SSE is used for data transformation.

After transformation, we used S-Plus with default parameters to compute the robust M-estimate of location (median) and bisquare A-estimate of scale (standard deviation) of the 0.04 ppm binned data. Figure 3a,b show the relationship between the median, and the standard deviation for the untransformed data in two different plotting formats. These plots indicate that the standard deviation is approximately proportional to the robust median. Similar plots in Figure 3c,d show the results for the transformed data. Here, the standard deviation is relatively constant with respect to the median.

Further advantages of the transformation are evident from Figure 4, which shows the transformed spectrum for Abalone number 1. The untransformed data for this same abalone was shown in Figure 1a, with the corresponding binned data in Figure 1b. It is clear that the transformation enhances some of the very low intensity peaks barely visible in the untransformed case. These enhancements allow for the determi-

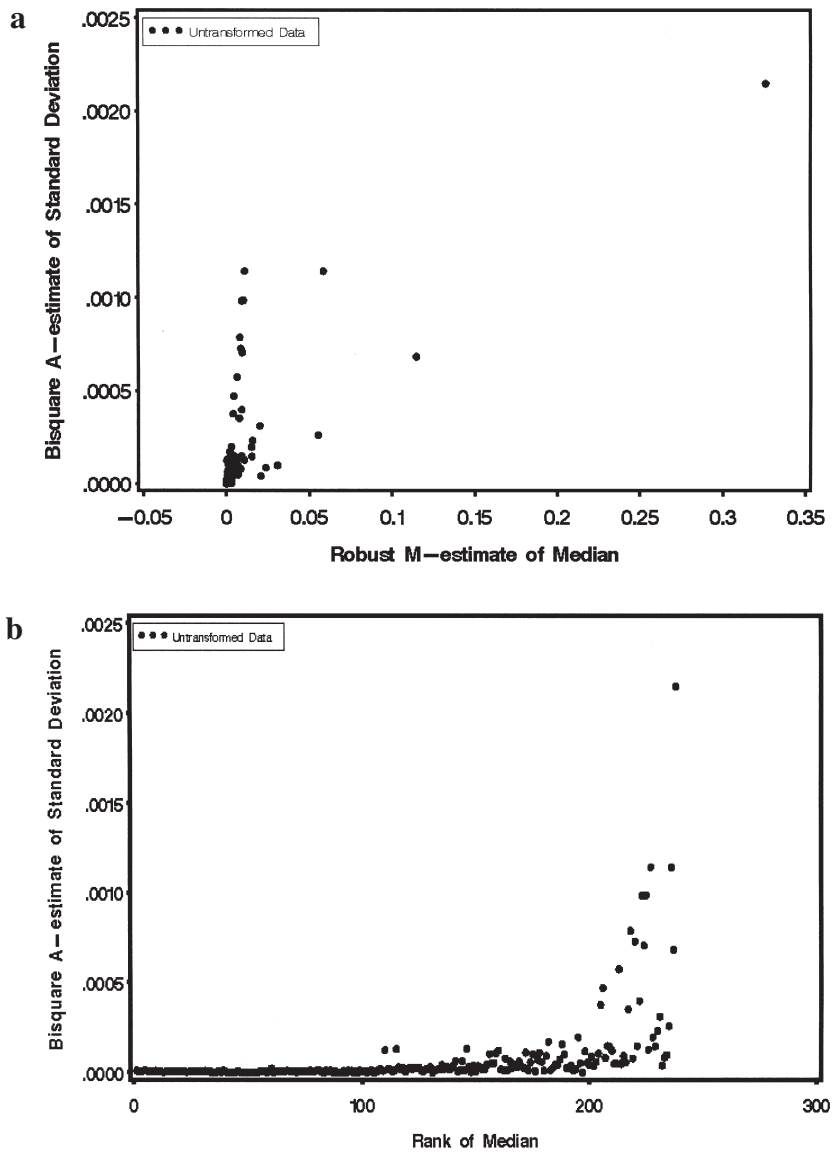


FIG. 3. (**a**) Robust standard deviation versus robust median for the replicate data before transformation. (**b**) Robust standard deviation versus the rank of the robust median for the replicate data before transformation. (**c,d**) The same as (**a,b**) after transformation.

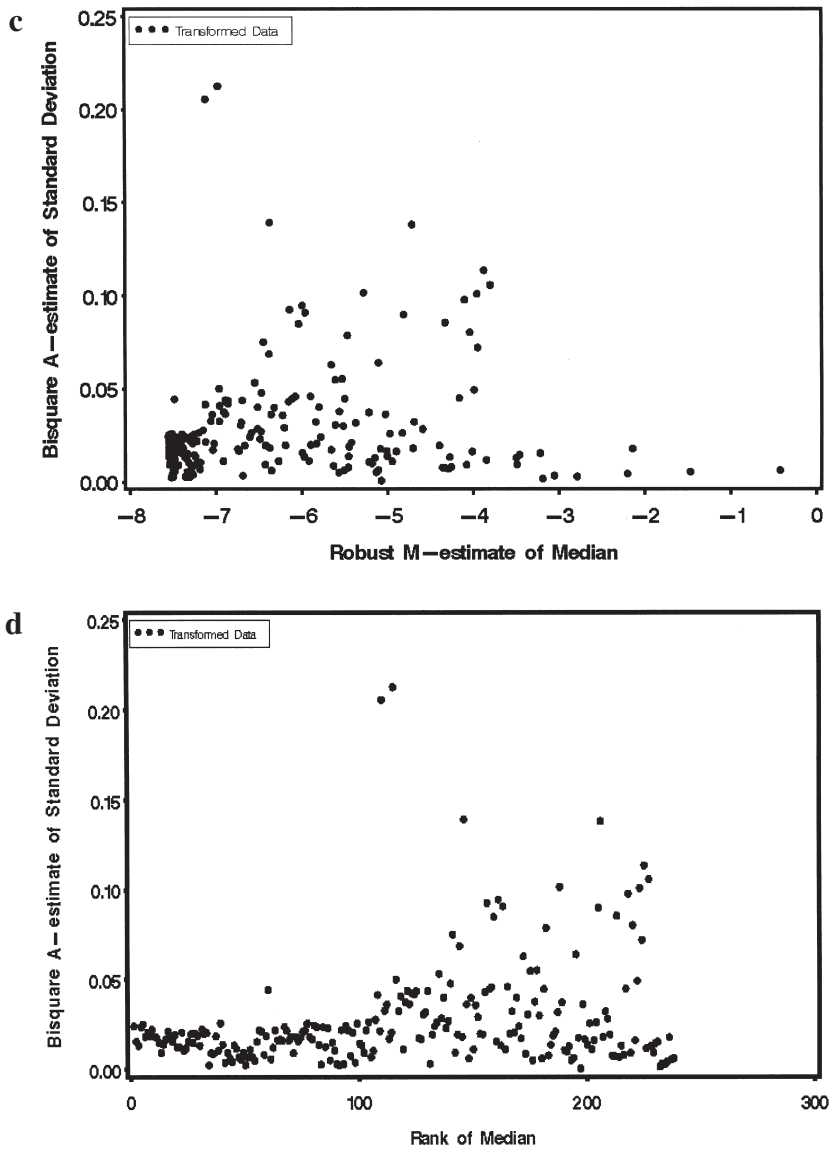


FIG. 3. *Continued.*

nation of the low intensity predictors that could be crucial in determining the metabolic profile biomarkers for different physiological states.

Multivariate methods

Our goals in multivariate analysis of the NMR data are to explore methods that divide abalone into their groups (without using the prior classification) as well as to determine the predictors that are significantly different between the groups. One method that has been used frequently for data of this nature is principal component analysis (PCA). In this method, new variables known as the principal components are formed that are linear combinations of the previous predictors (binned ones) in a way that the variance explained by each successive component is maximized. A plot of the projections of the sample positions along the first and second principal components gives the relationship between the observations (also known as the scores plot). Samples that cluster together on the plot indicate similar metabolic profiles. A plot of the first

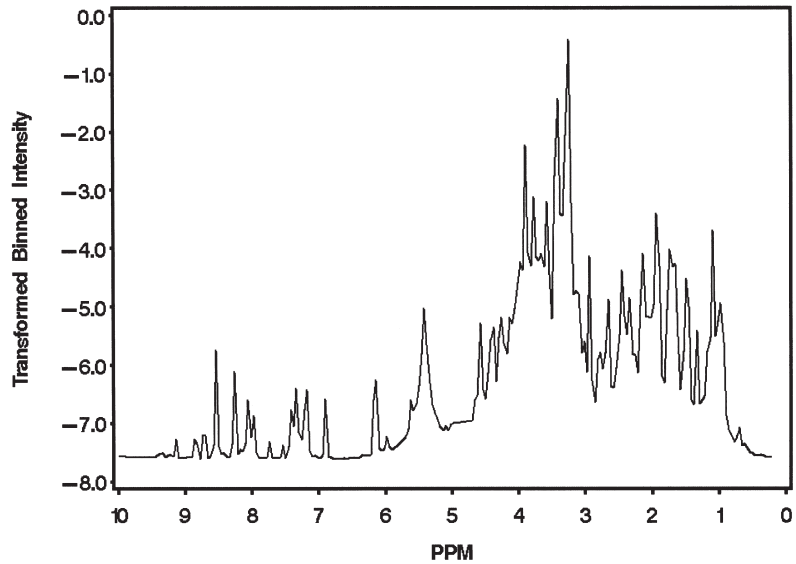


FIG. 4. Transformed NMR spectrum. This particular one corresponds to Abalone 1.

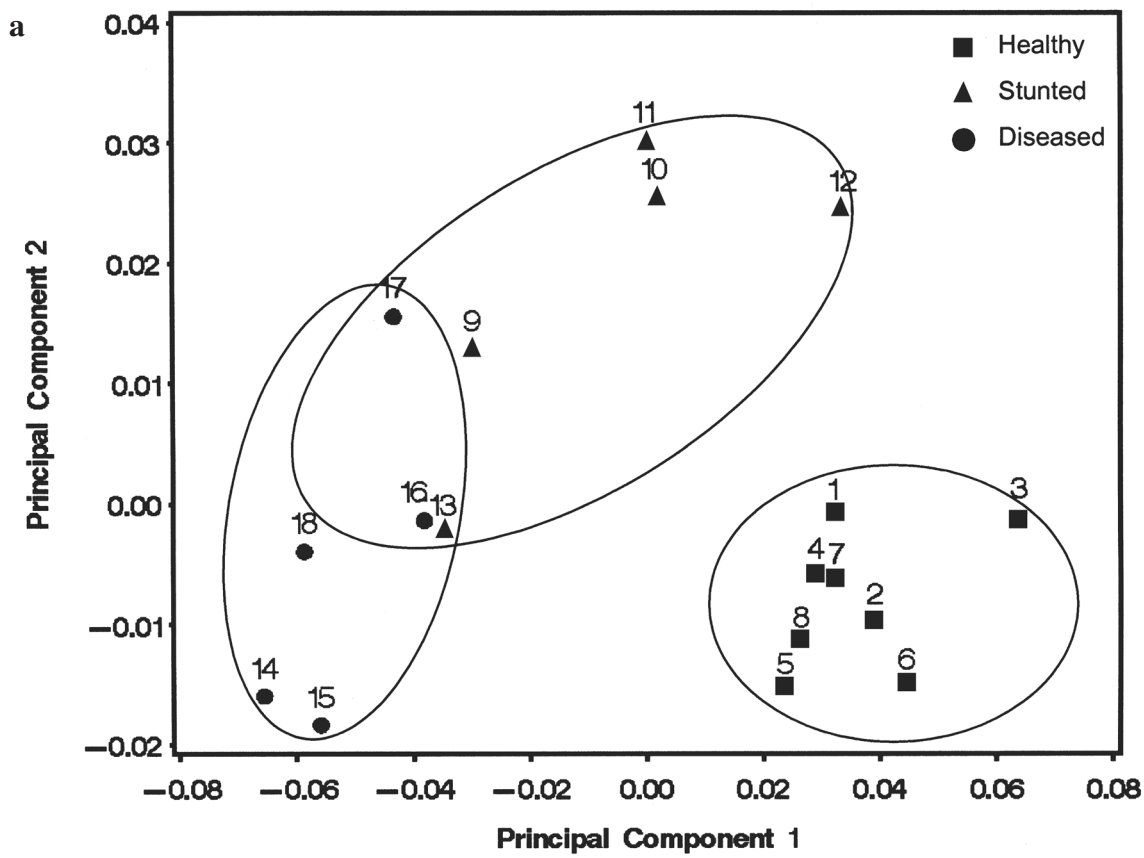
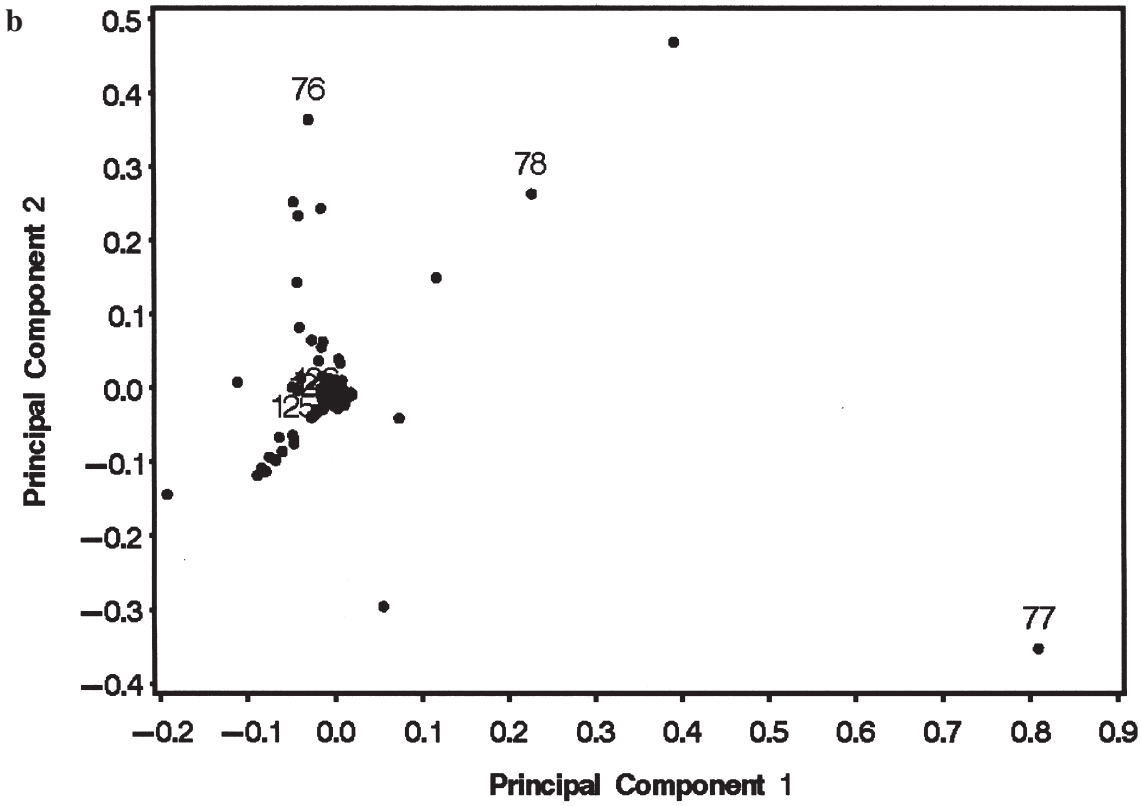


FIG. 5. PCA and cluster analysis results for the untransformed data for the 18 abalone. (**a**) The scores plot showing the relationships between abalone. (**b**) The loads plot showing the relationships between predictors. (**c**) Results from cluster analysis using the complete method and fixing the number of groups at 3.



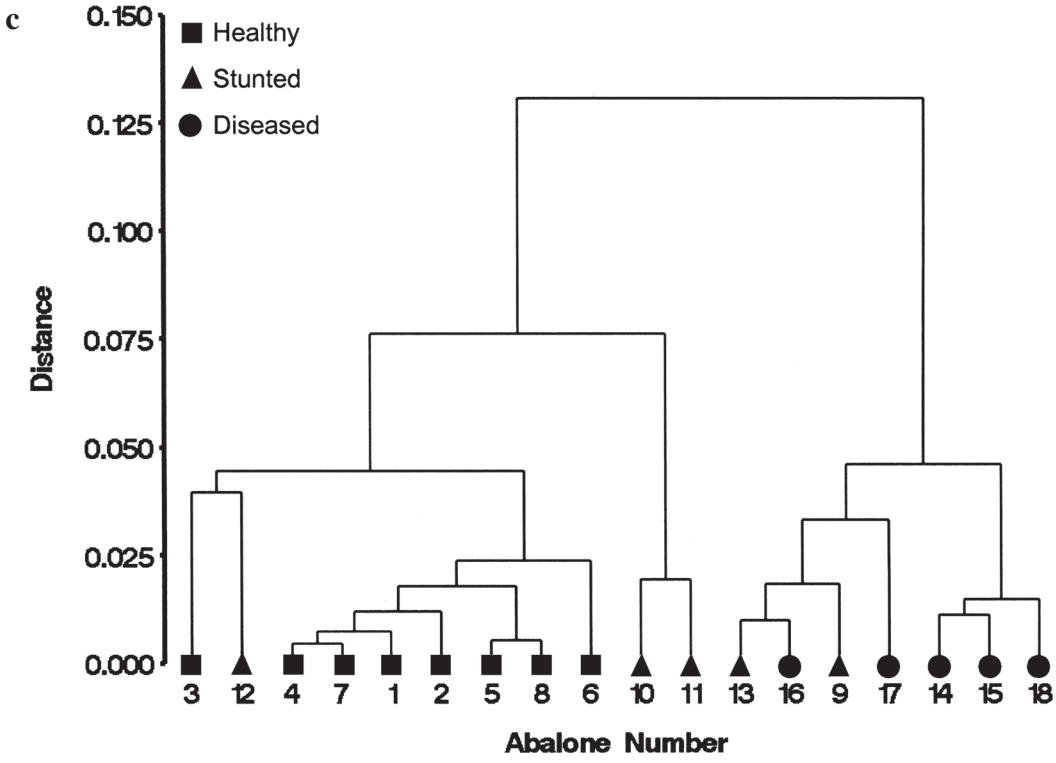


FIG. 5. *Continued.*

two eigenvectors of the covariance or the correlation matrix gives the relationship between the predictors (also known as the loadings plot). We used the covariance matrix and computed mean-centered principal components (i.e., each column of our data matrix had zero mean; we achieved this by subtracting off the mean of each column from the original data), the predictors close to (0, 0) in the loadings plot are the most unchanged among the abalone whereas the ones at the extremities are significantly changed (Khattree and Naik, 2000). One can attribute the changed predictors to the state of the abalone.

Performing PCA on the data also gives the opportunity to reduce the dimension of the matrix. Since a large portion of the variance is covered by the initial principal components, the data can be described in terms of these components. From the scree plot of the data (Khattree and Naik, 2000), we concluded that five components were sufficient to describe the data. The variance represented in these components was 90.3% in the untransformed data and 90.8% for the transformed data. This reduced the data matrix to a dimension of 18×5. This data was then subjected to cluster analysis, another technique for separating the groups of abalone. Cluster analysis using hierarchical clustering was used on both the untransformed and transformed data.

RESULTS

Figure 5a,b shows the results of PCA analysis as applied to the untransformed data. The scores plot shows the relationship between the 18 abalone metabolic profiles. The healthy abalone comprising samples 1–8

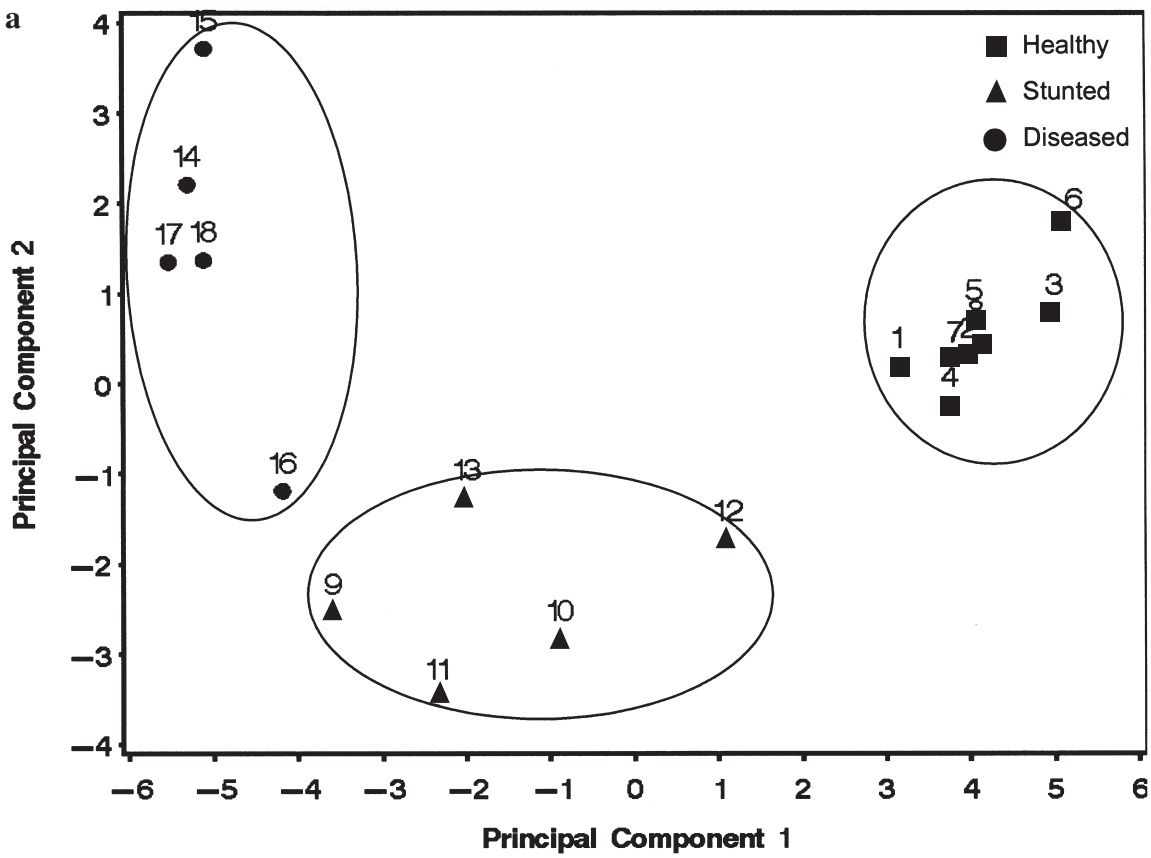


FIG. 6. PCA and cluster analysis results for the transformed data. (**a**) The scores plot showing the relationships between abalone. (**b**) The loads plot showing the relationships between predictors. (**c**) Results from cluster analysis using the complete method and fixing the number of groups at 3.

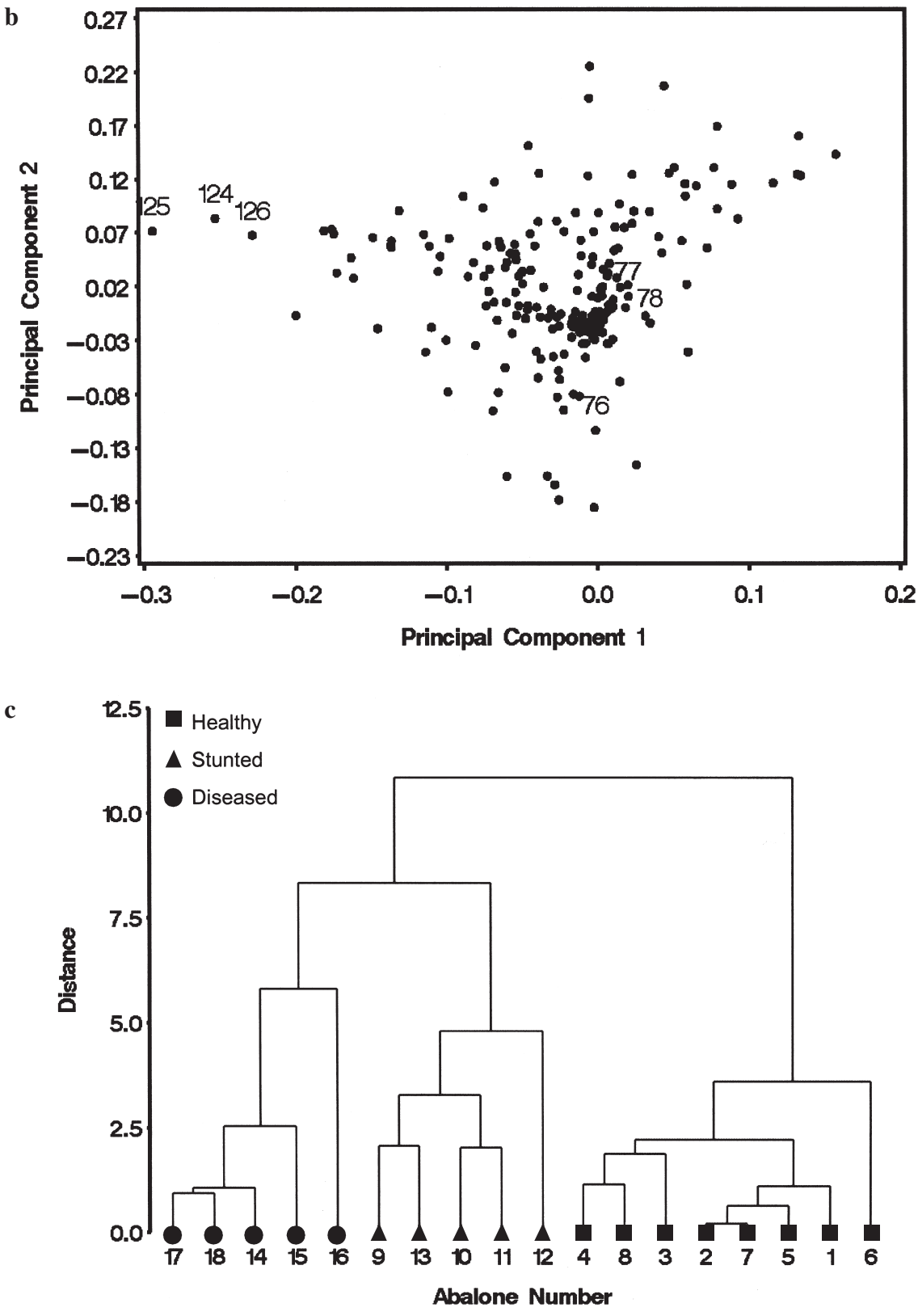


FIG. 6. *Continued.*

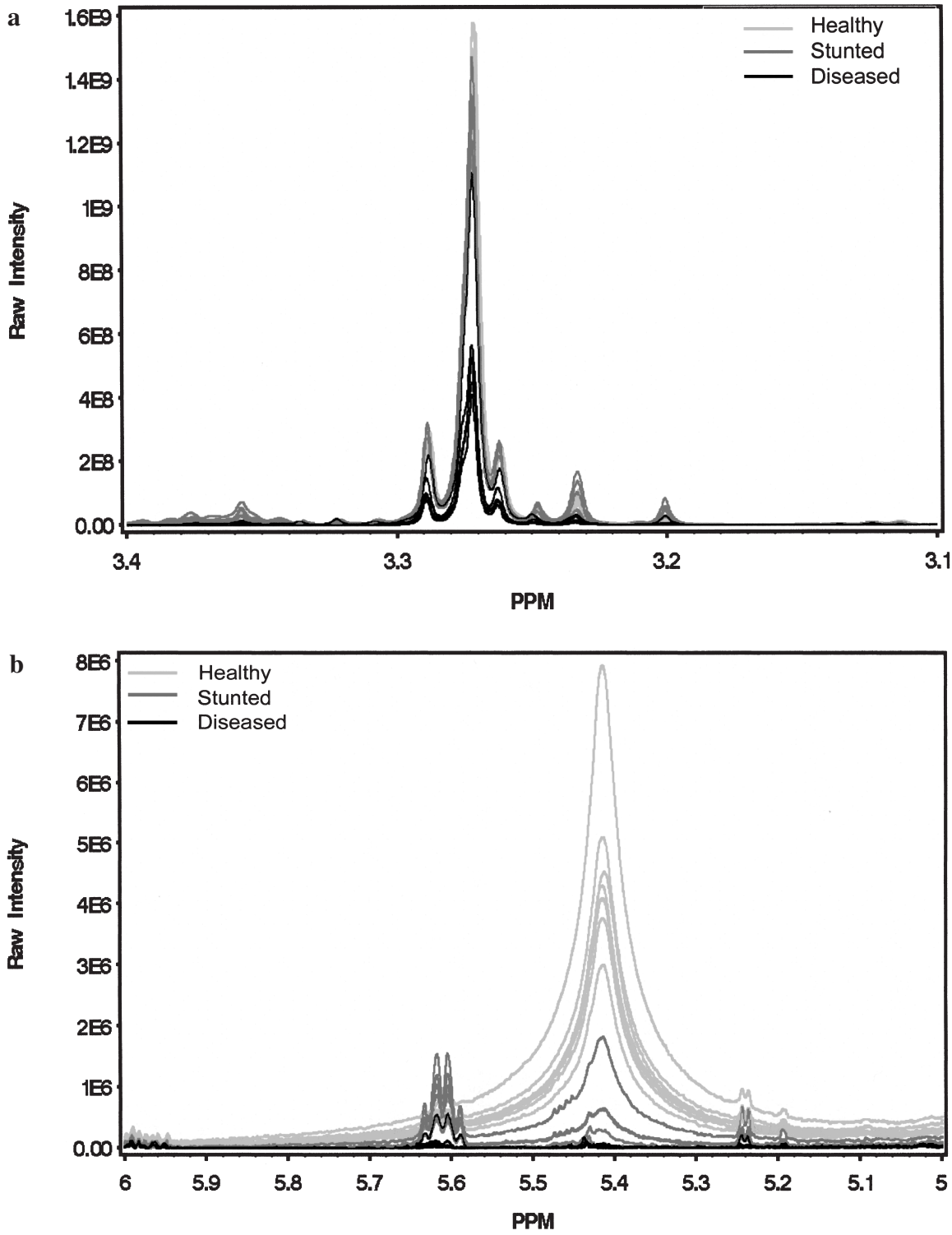


FIG. 7. (**a**) Raw NMR spectra after baseline subtraction for all abalone in the region 3.1–3.4 ppm, which includes bins 76–78. (**b**) The same as in **a**, except that the spectrum is in the region 5–6 ppm, which includes bins 124–126.

are clustered effectively, whereas there is an overlapping of the stunted group (samples 9–13) and the diseased group (samples 14–18). In particular, abalone 13 and 16 that belong to the stunted and diseased categories respectively, are not well separated. The loads plot shows that the bins are gathered together near the (0, 0) position. Bins 76–78 are among the most outlying in Figure 5b, and correspond to a frequency range of 3.22–3.30 ppm.

In order to perform a formal unsupervised cluster analysis, we employed only the first five PCs from the PCA. We then performed hierarchical cluster analysis using the complete method fixing the number of clusters to 3. In this case, the observations do not cluster as expected into the healthy, stunted and diseased categories as shown in Figure 5c.

Figure 6a–c shows the results of similar analyses after the data have been transformed. In the scores plot (Fig. 6a), the three groups belonging to the healthy, stunted and diseased abalone are well separated according to the initial morphological assessment. The loads plot (Fig. 6b) is dramatically different from the loads plot based on untransformed data. Here, the set of bins 124–126 are significantly different whereas bins 76–78 that were significantly different in the untransformed case are not considered particularly significant here and lie close to (0, 0). Cluster analysis of the reduced matrix is demonstrated in Figure 6c. The result clearly indicates the improvement in discrimination capabilities due to the transformation as the groups are clearly separated, unlike the situation with the untransformed data.

The significance of the bin selection can be ascertained by plotting the original raw data in the region of the bins mentioned above. Such plots are shown in Figure 7a,b. Figure 7a shows the raw data for all the abalone in the region of the spectrum 3.1–3.4 ppm, which includes bins 76–78. These bins were indicated as significant for the untransformed data. As seen in Figure 7, the significance of the bins stems from an overall intensity change in the peaks. However, if a similar plot is drawn for the range of 5–6 ppm, which includes bins 124–126 (indicated as significant for the transformed data), the significance of the bins is realized by a unique peak at $\sim$5.42 ppm for the healthy abalone and to a lesser extent by the stunted abalone. The absence of this peak could thus be a bio-marker for the disease of withering syndrome. Of course, additional chemical analysis and experimentation is needed to confirm this.

The 3.1–3.4-ppm range in the spectra represent the strongest peaks and the untransformed data is identifying the small differences in the intensities of these peaks. Following transformation, the analysis is able to discern arguably more important differences between the weaker peaks.

CONCLUSION

In this paper, we have demonstrated that the use of a variance stabilizing transformation of NMR spectroscopic data can greatly improve the ability to classify samples and to interpret clusterings. We have demonstrated by example that it is possible to identify potential metabolite markers that may be associated with disease by using sophisticated data analysis techniques. Transformation of the data that enables constant variance provides a unique opportunity to identify markers that otherwise may be difficult to detect without the transformation. Transformation also allows for better separation of the classes of organisms and provides a unique opportunity to find biomarkers.

ACKNOWLEDGMENTS

The research reported in this paper was supported by grants from the National Science Foundation (ACI 96-19020) and the National Institute of Environmental Health Sciences, National Institutes of Health (P43-ES04699). We are grateful for helpful suggestions from two referees.

REFERENCES

AZMI, J., GRIFFIN, J.L., ANTTI, A., et al. (2002). Metabolic trajectory characterization of xenobiotic-induced hepatotoxic lesions using statistical batch processing of NMR data. Analyst **127,** 271–276.

BECKWITH-HALL, B.M., BRINDLE, T., BARTON, R.H., et al. (2002). Application of orthogonal signal correction to minimize the effects of physical and biological variation in high-resolution ^{1}H NMR spectra of biofluids. Analyst **127,** 1283–1287.

BOLLARD, M.E., HOLMES, E., LINDON, J.C., et al. (2001). Investigations into biochemical changes due to diurnal variation and estrus cycle in female rats using high-resolution ^{1}H NMR spectroscopy of urine and pattern recognition. Anal Biochem **295,** 194–202.

BOX, G.E.P., and COX. D.R. (1964). An analysis of transformations. Roy Stat Soc Series B **26,** 211–252.

DURBIN, B.P., HARDIN, J.S., HAWKINS, D.M., et al. (2002). A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics **18,** S105–S110.

DURBIN, B., and ROCKE, D. (2004). Variance-stabilizing transformations for two-color microarrays. Bioinformatics **20,** 660–667.

DURBIN, B., and ROCKE, D.M. (2003). Estimation of transformation parameters for microarray data. Bioinformatics **19,** 1360–1367.

GELLER, S., GREGG, J., HAGERMAN, P., et al. (2003). Transformation and normalization of oligonucleotide microarray data. Bioinformatics **19,** 1817–1823.

HOAGLIN, D.C., MOSTELLER, F., and TUKEY, J.W., eds. (1983). *Understanding Robust and Exploratory Data Analysis* (Wiley, New York).

HOLMES, E., NICHOLSON, J.K., and TRANTER, G. (2001). Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks. Chem Res Toxicol **14,** 182–191.

KHATTREE, R., and NAIK, D.N. (2000). *Multivariate Data Reduction and Discrimination* (SAS Institute Inc., Cary, NC).

LINDON, J.C., HOLMES, E., and NICHOLSON, J.K. (2001). Pattern recognition methods and applications in biomedical magnetic resonance. Prog Nucl Res Spec **39,** 1–40.

NICHOLSON, J.K., CONNELLY, J., LINDON, J.C., et al. (2002). Metabonomics: a platform for studying drug toxicity and gene function. Nat Rev Drug Discov **1,** 153–161.

PUROHIT, P.V., and ROCKE, D. (2003). Discriminant models for high-throughput proteomics mass spectrometer data. Proteomics **3,** 1699–1703.

ROCKE, D.M., and DURBIN, B. (2001). A model for measurement errors for gene expression arrays. J Comput Biol **8,** 557–569.

ROCKE, D.M., and LORENZATO, S. (1995). A two-component model for measurement error in analytical chemistry. Technometrics **37,** 176–184.

ROCKE, D., and DURBIN, B. (2003). Approximate variance-stabilizing transformations for gene-expression microarray data. Bioinformatics **19,** 966–972.

VIANT, M.R. (2003). Improved methods for the acquisition and interpretation of NMR metabolomic data. Biochem Biophys Res Commun **310,** 943–948.

VIANT, M.R., ROSENBLUM, E.S., and TJEERDEMA, R.S. (2003). NMR-based metabolomics: a powerful approach for characterizing the effects of environmental stressors on organism health. Environ Sci Technol **37,** 4982–4989.

Address reprint requests to:
Dr. Parul V. Purohit
Center for Image Processing and Integrated Computing
University of California
Davis, CA 95616

E-mail: pvpurohit@ucdavis.edu