Analytical Biochemistry 399 (2010) 58–63

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Analysis of metabolomic PCA data using tree diagrams

Mark T. Werth [a], Steven Halouska [b], Matthew D. Shortridge [b], Bo Zhang [b], Robert Powers [b,*]

[a] Department of Chemistry, Nebraska Wesleyan University, Lincoln, NE 68504, USA
[b] Department of Chemistry, University of Nebraska–Lincoln, Lincoln, NE 68588, USA

ARTICLE INFO

Article history:
Received 14 September 2009
Received in revised form 14 December 2009
Accepted 14 December 2009
Available online 21 December 2009

Keywords:
Metabolomics
Tree diagrams
Principal component analysis
Bootstrap analysis
NMR

ABSTRACT

Large amounts of data from high-throughput metabolomic experiments are commonly visualized using a principal component analysis (PCA) two-dimensional scores plot. The question of the similarity or difference between multiple metabolic states then becomes a question of the degree of overlap between their respective data point clusters in principal component (PC) scores space. A qualitative visual inspection of the clustering pattern in PCA scores plots is a common protocol. This article describes the application of tree diagrams and bootstrapping techniques for an improved quantitative analysis of metabolic PCA data clustering. Our PCAtoTree program creates a distance matrix with 100 bootstrap steps that describes the separation of all clusters in a metabolic data set. Using accepted phylogenetic software, the distance matrix resulting from the various metabolic states is organized into a phylogenetic-like tree format, where bootstrap values $\geqslant 50$ indicate a statistically relevant branch separation. PCAtoTree analysis of two previously published data sets demonstrates the improved resolution of metabolic state differences using tree diagrams. In addition, for metabolomic studies of large numbers of different metabolic states, the tree format provides a better description of similarities and differences between each metabolic state. The approach is also tolerant of sample size variations between different metabolic states.

The rapidly growing field of metabolomics seeks to describe and understand the different metabolic states of an organism. Recent reviews have described the application of metabolomic methods to diverse problems, including biomarker discovery, drug metabolism, nutrition, and environmental toxicology [1–5]. A fundamental question in many metabolomic studies is whether or not an altered metabolic state (e.g., disease, mutation, diet, drug) being studied is similar to or different from the reference state. The most common statistical approach for the analysis of metabolomic data is principal component analysis (PCA)[1] and partial least squares discriminant analysis (PLS–DA) [6–8]. As an illustration, more than 55 metabolomic or metabonomic articles have been published in the journal Analytical Biochemistry since 2001, with more than 45% of these articles using PCA, PLS–DA, or a comparable statistical tool. The focus of the remaining metabolomic articles has been metabolite identification or methodology development, where a statistical approach is not employed.

PCA or PLS–DA converts data obtained from high-throughput instrumental analysis into a qualitative visual presentation (scores plot) [9,10] showing the clustering of biological samples into either similar or different groupings. In some cases, sample data for different metabolic states are clearly separated into distinct clusters (e.g., wild-type cells vs. mutant cells). Other cases arise where the separation of data clusters is not so clearly defined. Even though the presentation of data in principal component (PC) scores space is the result of a statistical analysis, it is important to emphasize that the degree of separation between data clusters is not quantitatively addressed directly by the PCA approach. Recently, the MetaboAnalyst web server (http://metaboanalyst.ca) has been developed to provide a robust set of tools for the processing and analysis of metabolomic data [11]. PLS–DA and other supervised methods have a tendency to overfit the data and to identify nonexistent clustering patterns. MetaboAnalyst includes random forest [12] and support vector machine [13] methods to determine the reliability or significance of the PLS–DA discrimination. Similarly, a SIMCA Cooman's plot is used to predict class membership based on the distance to the model [14]. Alternatively, a simple visual inspection of the resulting scores plot does not provide a statistically meaningful answer to this basic question: are the clustering patterns in a scores plot significantly different?

Felsenstein encountered similar problems when attempting to assign confidence limits to phylogenetic trees [15] and resolved the problem by applying a bootstrap statistical approach [16,17]. This approach may also be applicable to the analysis of clustering patterns in scores plots for metabolomic data. The metabolome is complementary to the transcriptome and proteome, captures the

* Corresponding author. Fax: +1 402 472 9402.
E-mail address: rpowers3@unl.edu (R. Powers).

[1] *Abbreviations used:* PCA, principal component analysis; PLS–DA, partial least squares discriminant analysis; PC, principal component; 2D, two-dimensional; NMR, nuclear magnetic resonance; AZA, 8-azaxanthine; DCS, D-cycloserine; ANOVA, analysis of variance; MANOVA, multivariate ANOVA.

0003-2697/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2009.12.022

functional or physiological state of the cell, and provides a link between genotypes and phenotypes [18]. Clearly, the range and quantity of metabolites observed are dependent on both the organism's proteome and genome, but direct correlations between gene expressions and the metabolome are low [19]. Nevertheless, metabolites have been associated with species evolution [20] and have been used to differentiate between different fungal species [21], to differentiate between different *Escherichia coli* species [22], and to monitor the adaptive evolution of yeast [23]. Phylogenetic trees have also been generated from the analysis of metabolic networks [24] and reproduce phylogenetic relationships between species derived from 16S RNA sequences [25]. Given that metabolomics maps reasonably well with phylogeny, it seemed appropriate to explore the application of tree diagrams and the bootstrap method to determine the significance of clustering patterns in scores plots.

A software program named PCAtoTree was developed to quantitatively analyze clusters of PC values. The program converts metabolomic data expressed as PC scores into a series of Euclidean distance matrices that can be used to generate metabolic trees and the corresponding bootstrap values. The resulting tree diagrams are intended to be used in combination with the original scores plot to decipher the significance of cluster similarity or differences. Importantly, the tree diagrams should not be interpreted as a hierarchal representation of the original metabolomic data [26].

Materials and methods

The PCAtoTree program (available on request) was written in the Awk scripting language running under the Linux operating system. The PCAtoTree program uses data from a PC scores plot generated by SIMCA (Umetrics, Kinnelon, NJ, USA). For each separate metabolic state, the PCAtoTree program calculates the average of each PC and the related standard deviations. Next, any data points having a PC value that is more than 2 standard deviations from the respective average are removed. The average PC values are then recalculated, and these average values define the cluster center for each metabolic state. To be consistent with the typical two-dimensional (2D) PC scores plots commonly found in the literature, the results presented here were calculated using only the first two PCs [27,28]. However, metabolomic data analysis often requires the use of additional components. The PCAtoTree program was written to accommodate eight PCs and can be expanded if necessary to include more PCs.

Distances between the average PC positions for each metabolic state are calculated using the standard equation for the Euclidean distance between two points:

$$\text{Distance} = \sqrt{(\Delta \text{PC}_1)^2 + (\Delta \text{PC}_2)^2 + \ \ldots} \qquad (1)$$

Because this distance matrix is calculated from the average PC position for each metabolic state, there is a single unique distance between any given pair of metabolomic states. Although it is possible to calculate the distances between every possible combination of data point pairs, the analysis of these distances is complicated primarily by the rapid progression in the total number of distances to be calculated with each successive data point. Also, it will not fundamentally change the definition of the cluster center but will unnecessarily complicate the bootstrapping analysis. Furthermore, a Euclidean distance does not capture direction, so an average distance from every possible combination of data points will overestimate the separation of closely spaced states. A similar circumstance occurs for a data set with a relatively large standard deviation, where the larger distances will skew the average separation.

Most important, the PCAtoTree program randomly resamples the data for each metabolic state to apply standard bootstrapping methods to assess the significance of the similarity (overlap) or difference (separation) observed between pairs of data clusters. For each metabolic state (e.g., wild-type state), the program randomly draws data points from the data set to determine an average PC score. It is important to note that the PCA analysis was performed only once and each round of bootstrapping uses the same PC scores data set. Data points previously flagged as outliers remain excluded. A particular data point may be chosen more than once or completely excluded in the new average PC calculation, but the total number of data points randomly drawn is equivalent to the original size of the data set. The new PC average values are then used to calculate a new distance matrix. Resampling is repeated until a total of 100 distance matrices have been created. The resulting distance matrix file is transferred to version 3.68 of the PHYLIP suite of software programs (http://www.phylip.com) [29,30] for completion of the bootstrapping and visualization of the results as metabolic tree diagrams.

Specifically, the Fitch program in PHYLIP performs a weighted least-squares analysis of the 100 Euclidean distance matrices from PCAtoTree to minimize the distances between the nodes [31]. The minimized distances are used to generate a summary metabolic tree. Virtually all metabolomic studies incorporate a reference or control group (e.g., wild-type cells, healthy subjects). This group is defined as the out-group for the purposes of the Fitch distance program. Fitch also produces a family of individual metabolic tree diagrams that are analyzed using the PHYLIP Consense package with the majority rule extended setting. Consense counts the number of times an identical branch is found. When the number of bootstrap steps (and therefore the number of individual trees) is 100, these values represent the percentage that a given branch appears. Using the majority rule extended setting [32], bootstrap values below 50% indicate that none of the possible branch patterns appears even 50% of the time and implies a statistically insignificant separation. Similarly, portions of the metabolic tree without a bootstrap value should be treated as unresolved. As bootstrap values rise above 50%, the confidence described by the tree branch also increases. Obviously, bootstrap values of 100% indicate complete statistical confidence in the overlap or separation of the metabolic states shown by the tree. Trees were visualized using the PHYLIP DrawGram program.

Simulated metabolic PCA data were used to illustrate the bootstrapping process (Fig. 1). Data points were randomly generated in Excel to produce four clusters of 10 points each: (i) circles had average PC1 and PC2 scores corresponding to −5.94 and −2.72, respectively, and standard deviations of 0.89 (PC1 axis) and 1.00 (PC2 axis); (ii) triangles had average PC1 and PC2 scores corresponding to −5.94 and 0.28, respectively, and standard deviations of 0.99 (PC1 axis) and 1.00 (PC2 axis); (iii) diamonds had average PC1 and PC2 scores corresponding to −3.91 and 2.28, respectively, and standard deviations of 0.98 (PC1 axis) and 1.00 (PC2 axis); and (iv) squares had average PC1 and PC2 scores corresponding to 4.93 and 0.28, respectively, and standard deviations of 0.74 (PC1 axis) and 1.00 (PC2 axis). The squares cluster is placed to the far right side of the PC scores plot to represent the reference or control metabolic state. The circles, triangles, and diamonds clusters are placed near each other on the left side of the PC scores plot and represent different metabolic states of interest. Open symbols represent the original data set and remain fixed.

Closed symbols represent the average center position obtained from random resampling of the data. These points move to reflect different possible resampling results. Fig. 1A represents the case where the resampling has resulted in the average center of the diamond data set being closest to the average center of the triangle data set. The Euclidean distance between the triangles and dia-

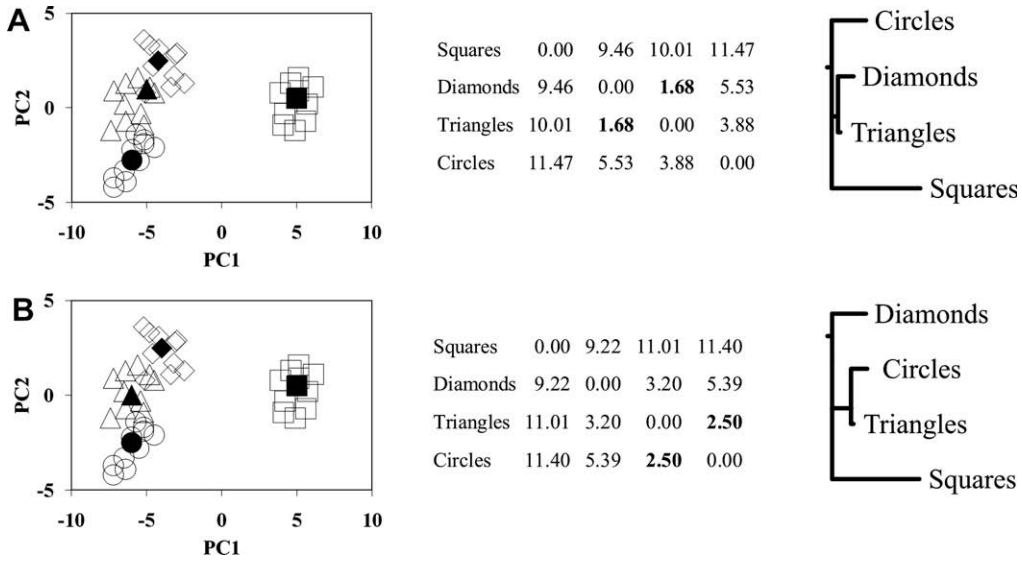

Fig. 1. Examples of PC scores plots with the corresponding Euclidean distance matrices and metabolic trees. Open symbols represent simulated experimental data points. Closed symbols represent the cluster center positions calculated after random sampling of the respective data sets. Panel (A) presents the case where the centers of triangles and diamonds are closest together. Panel (B) presents the case where the triangles and circles have the smallest separation.

monds is 1.68 U, and the metabolic tree shows the diamonds and triangles grouped together. Fig. 1B represents the case where resampling of the same experimental data has produced a new situation where the triangle center is now closest to the center for the circle metabolic state. The distance between the triangles and circles is 2.50 U, but the distance between the triangles and diamonds has now increased to 3.20 U. This scenario produces a metabolic tree that now pairs the triangles and circles. After repeating the bootstrap process 100 times, a single consensus distance for each pair of cluster centers should emerge. These consensus distances could be expressed as a distance matrix, although it would be difficult to readily discern relationships among the values. The tree format provides a more readily interpretable format in which to view the groupings, relative distances between groups, and (most important) the bootstrap values.

The application of tree diagrams to analyze metabolomic PCA data was demonstrated using previously published results, where experimental procedures for cell culture growth, sample preparation, PCA protocols, and nuclear magnetic resonance (NMR) data collection, processing, and analysis were described in detail [27,28]. Specifically, tree diagrams were generated from the analysis of 8-azaxanthine activity in *Aspergillus nidulans* [27] and D-cycloserine activity in *Mycobacterium smegmatis* [28].

Results

PC scores data points were manually generated in Excel to explore the relationship between cluster overlap and the resulting bootstrap value. Again, four data sets were used with the cluster on the far right, representing the reference or control metabolic state. Data sets representing the three remaining metabolic states were adjusted to have the same average PC1 value and then overlapped along the PC2 axis. The spread (i.e., standard deviation) of the clusters in the PC2 dimension was empirically adjusted so that the standard deviations of all three clusters differed by no more than 0.02. One of these clusters (circles) was then incrementally separated along the PC2 axis, and the resulting data sets were analyzed. The results of these simulations are presented in Fig. 2.

The bootstrap value increases asymptotically as a function of the relative cluster separation distance (Fig. 2, top panel), which is defined as the distance between the cluster centers divided by the standard deviation of the cluster spread. Cluster standard deviations of 1 and 2 were used during the simulations. When the results were normalized to the relative distance, the curves overlapped (data not shown). The solid line represents the best-fit curve to simulations done using 10 data points per cluster. Point a in the top panel of Fig. 2 represents the case where the three clusters, with standard deviations of 2, were completely overlapped as shown in Fig. 2A below. The corresponding metabolic tree has one branch representing the squares and a second branch where the diamonds, triangles, and circles are grouped together.

Point b in the top panel of Fig. 2 represents a separation of the circles from the diamonds and triangles by 1 relative distance unit (see Fig. 2B). The spread of the data clusters was kept constant at 2 standard deviations. Visual analysis of Fig. 2B shows that a high degree of cluster overlap remains at this relative distance. However, the average bootstrap value for point b was 81%. The corresponding consensus metabolic tree, obtained from 81 of 100 random samplings of the data set, now has three branches. Squares and circles each form separate branches, and the third branch is the pairing of the triangles and diamonds. When the relative distance is increased to 3 (point c in top panel of Fig. 2), the bootstrap value has clearly reached the 100% upper limit. Each of the 100 samplings of the data set produced the same three-branched tree described for point b. Even at a relative distance of 3, the circles do not appear to be completely separated from the diamonds and triangles. This result illustrates that the tree diagrams can identify statistically distinct clusters that are not readily apparent by visually inspecting PCA scores plots.

The dashed line in Fig. 2 (top panel) represents the best-fit curve for results obtained when the number of data points per cluster was reduced from 10 to 6. As the number of data points per cluster decreases, even greater separation of the clusters is required to achieve the same bootstrap value. Even at a relative distance of 5, the bootstrap value has not yet reached 100%. This clearly demonstrates the inherent value in obtaining replicates of 10 or more to statistically differentiate between multiple metabolic states [33]. Assuming a normal distribution, the accuracy of the average position for the cluster will improve with the number of data points or, alternatively, an outlier may have a diminished impact on distorting the true center for the cluster. As the number

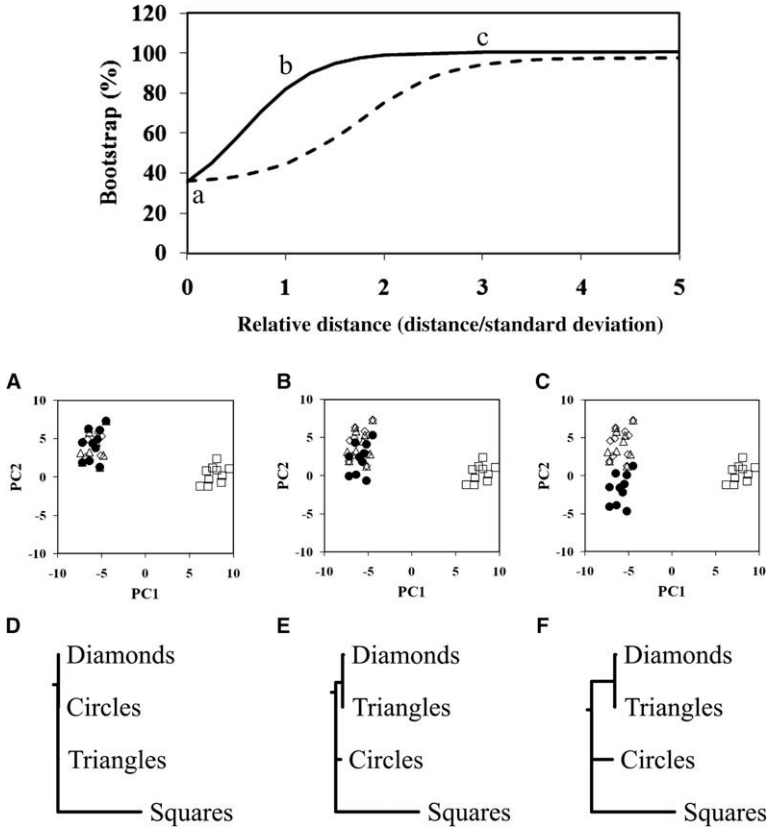

Fig. 2. Top panel: Best-fit curves obtained from simulations of bootstrap value as a function of relative cluster separation distance. The solid line was calculated using 10 data points per cluster. The dashed line was calculated using 6 data points per cluster. Middle panels: Selected PC scores plots used for simulations (10 data points per cluster, cluster standard deviation = 2): (A) relative distance = 0; (B) relative distance = 1; (C) relative distance = 3. Bottom panels (D–F): Metabolic trees corresponding to the PC scores plot shown above.

of data points increases, the probability of selecting any given point during the bootstrapping step is diminished and the distribution of data points away from the average also decreases. As a result, the cluster's center is more consistently defined and the range of distances calculated during the bootstrapping step will diminish with increasing data points, which in turn will decrease the variability in the number of distinct nodes and increase the bootstrapping number. Importantly, shifting of the curves shown in Fig. 2 (top panel) depends only on cluster size and not the standard deviation because the graphs overlap perfectly for different standard deviation values. Of course, the graphs do indicate that if the standard deviation for a cluster does increase, the separation between clusters would need to increase proportionally to maintain a similar bootstrap percentage.

For simulations with four metabolic states, the minimum bootstrap value was consistently around 35–38%. Presumably, the lower limit bootstrap value reflects clustering occurring at a higher probability than random chance would allow. For three overlapped clusters, a pairing would need to occur more than 33% of the time. When the number of metabolic states is increased to five, the lower limit of the bootstrap value decreases proportionally to the number of distinct metabolic states. However, the precise relationship between the lower limit of the bootstrap value and the number of metabolic states remains unclear. But the absolute value of the minimal bootstrap number is basically irrelevant because a bootstrap number below 50% is generally considered as insignificant.

It should be noted that the simulations presented in Fig. 2 were designed to focus on the separation of one data set (circles) from two overlapped clusters (diamonds and triangles). In these simulations, one group (squares) was defined as the reference or control data set and deliberately placed far away from the remaining three data sets (metabolic states). In addition, two groups (diamonds and triangles) were generated with essentially identical center positions for the clusters. Both of these constraints likely favored increased statistical significance for the intentional separation examined in the simulations.

To illustrate the application of this approach to actual NMR PCA scores data, we reexamined previously published data for a study of the interactions between the fungal cells of *A. nidulans* and the drug 8-azaxanthine (AZA) [27]. There are four metabolic states to be considered: the wild-type *A. nidulans,* the wild-type cells treated with AZA (which inhibits the enzyme urate oxidase), the urate oxidase deletion mutant of *A. nidulans,* and the urate oxidase deletion mutant treated with AZA. PCA of this system suggested that three of these metabolic states (wild-type treated with AZA, mutant, and mutant treated with AZA) were similar to each other and distinctly different from the wild-type metabolic state (see Fig. 3A).

Further analysis of the PC 2D scores data from the original study using the PCAtoTree program are also presented in Fig. 3A. As expected, the resulting tree diagram shows that the wild-type data cluster is clearly separated from the other samples. However, in the original data set, the wild-type with drug data appeared to be adjacent to the mutant and mutant with drug data in PC space. This small separation was attributed to some residual urate oxidase activity remaining in the presence of the drug [27]. Based on a qualitative visual analysis, there did not appear to be sufficient justification for separation of these groups. With development of the PCAtoTree program, this issue can be addressed more quantitatively. In 100 of 100 cases, the mutant and mutant with drug data were clustered together. Likewise, in 100% of the cases, the separa-

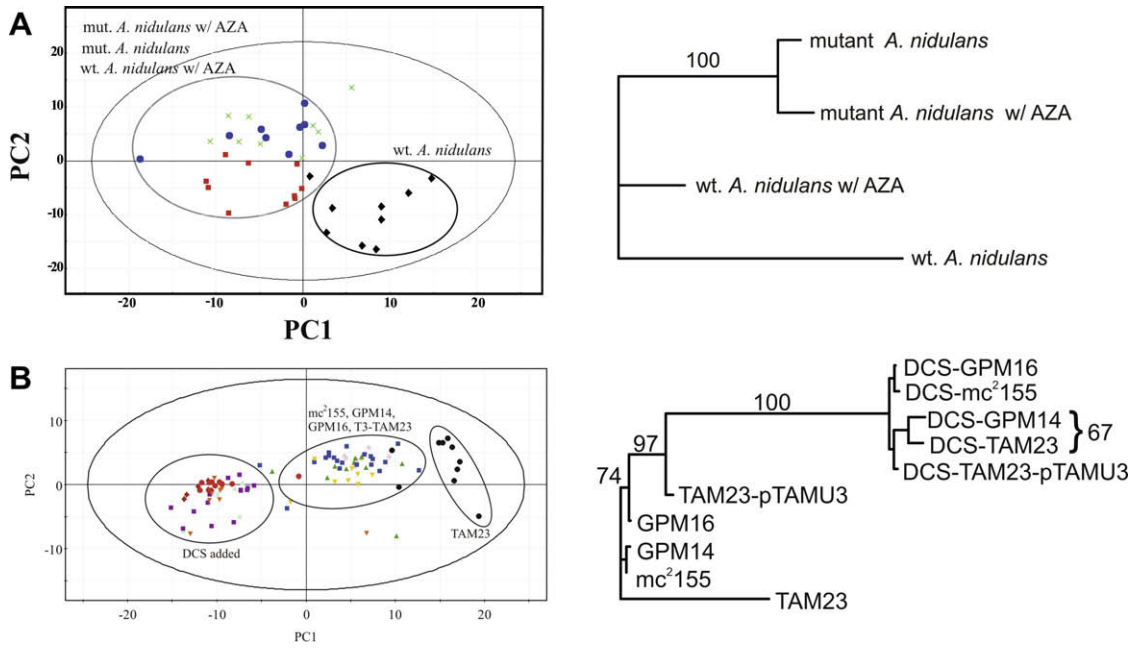

Fig. 3. (A) PC scores plot and corresponding metabolic tree for the *A. nidulans* fungal mutant and drug-treated samples. The labels correspond to *A. nidulans* urate oxidase mutant (green ×), wild-type with AZA (red ■), urate oxidase mutant with AZA (blue ●), and wild-type cells (black ♦). wt, wild-type. (B) PC scores plot and corresponding metabolic tree for the *M. smegmatis* mutants and drug-treated samples. The labels correspond to wild-type (mc^2155, blue ■), D-alanine racemase null mutant (TAM23, black ●), D-alanine racemase overproducing DCS-resistant mutant (GPM14, pink ♦), undefined DCS-resistant mutant (GPM16, green ▲), and TAM23 complemented with wild-type d-alanine racemase gene (TAM23–pTAMU3, yellow ▼). Cells treated with DCS to wild-type (purple ■), TAM23 (red ●), GPM14 (brown ♦), GPM16 (light blue ▲), and TAM23–pTAMU3 (orange ▼). Bootstrap values above 50% are indicated. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.) (Reprinted with permission from Refs. [27] and [28], Copyrights 2006 and 2007 by American Chemical Society).

tion between wild-type with drug and the two mutant data sets was maintained. This result demonstrates the improved resolution of data clusters made possible by using a metabolic tree approach.

A metabolomic study of the effects of D-cycloserine (DCS) on mycobacterial metabolism provides a more challenging system to easily analyze using a PCA 2D scores plot [28]. In particular, sample populations for the 10 different metabolic states range from 6 to 20. The original PCA scores plot and the resulting metabolic tree are shown in Fig. 3B. At the bottom of the tree, the D-alanine racemase null mutant (TAM23) cells are clearly separated from the remaining 9 metabolic states. Next, the wild-type (mc^2155), D-alanine racemase overproducing DCS-resistant mutant (GPM14), undefined DCS resistant mutant (GPM16), and TAM23 complemented with wild-type D-alanine racemase gene (TAM23–pTAMU3) cells are found on the left side. These were originally reported as clustering together. The metabolic tree analysis shows that the TAM23–pTAMU3 separates from this group 97% of the time. Clearly, these similar cell lines have small, but significant, variations from each other. The branch leading to the cells treated with DCS has a bootstrap value of 100, indicating a complete separation of the drug-treated samples from the non-drug-treated samples. Within the group of five cell lines treated with DCS, the DCS–GPM14 and DCS–TAM23 were associated with each other 67% of the time. Other associations within the DCS-treated samples occurred at even lower levels. This more quantitative analysis does not contradict any conclusions presented in the original report.

Discussion

The PCAtoTree approach was developed to address the question of overlap versus separation of data clusters in PC scores plots produced by metabolomic studies. This question is critical to the interpretation of scores plots produced from PCA and similar statistical techniques. Overlapped data clusters suggest a similar metabolic process, whereas separated data clusters imply different underlying metabolic mechanisms. The overlap/separation issue is further complicated as the number of metabolic states included in the study increases. However, due to difficulties with the proper application of common statistical tests (e.g., univariate *T* test, analysis of variance [ANOVA], multivariate ANOVA [MANOVA]), this question has not been addressed until now.

The PCAtoTree approach recognizes that the overlap/separation issue can be rephrased more quantitatively in terms of distances between the centers of the data clusters. Another advantage of the bootstrap approach is that it is less sensitive to the number of data points per cluster. This facilitates comparisons when the control and sample populations are of different sizes, as was the case in the mycobacterial study of Halouska and coworkers [28]. Even though the approach can be applied to samples of variable size, one should keep in mind that larger sample populations are still desirable.

Tree branches with bootstrap values below 50% suggest overlap of the data clusters in PC scores space or an insignificant separation. In the metabolic studies, these metabolic states would be considered as essentially identical. Conversely, bootstrap values greater than 50% indicate that the corresponding branch of the metabolic tree occurs more often than alternative branching patterns. As seen in Fig. 2, bootstrap values greater than 50% are possible even when visual analysis of the data confirms that substantial overlap of the clusters remains. This illustrates a strength of the tree diagram approach: an unbiased analysis of cluster centers identifies separations that are not easily visualized.

The issue of cluster overlap or separation can now be expressed in more quantitative terms using tree diagrams with bootstrap values. However, in the simplest sense, the question of whether or not two clusters are overlapped has only two answers: overlapped or not overlapped. The bootstrap value does not provide a definitive

answer to the overlap question in all cases, but it does provide a useful measure of the confidence in the resulting interpretation of the experimental data. Our analysis also clearly demonstrates that increasing the number of replicate data points significantly improves the reliability of differentiating between two clusters. Fig. 2 illustrates a dramatic increase in the bootstrap number from a doubling in the number of data points while keeping all other factors constant. Effectively, a larger number of replicate data points better define the cluster's center, and this in turn allows an improved differentiation between clusters that are still visibly overlapped. This result also makes clear the challenge involved in attempting to assess the degree of cluster overlap or separation by visual inspection alone.

Application of the PCAtoTree approach to the drug metabolism results of Forgue and coworkers [27] demonstrated that the separation of the wild-type with drug metabolic state from the mutant metabolic states was small but significant by virtue of the fact that it was found in all 100 cases. A higher level of confidence can now be placed in a previously uncertain observation. In the case of the mycobacterial metabolism study by Halouska and coworkers [28], the clear overlap of the five DCS-treated samples can now be expressed as occurring 100 of 100 times. A claim of separation of TAM23–pTAMU3 from the three nearby groups (mc^2155, GPM14, and GPM16) is based on 97 of 100 results, whereas a claim of separation of GPM16 from mc^2155 and GPM14 is supported at the lower level of 74 of 100 results.

In summary, the PCAtoTree approach addresses the question of similarity, or difference, between metabolic states in a metabolomic experiment. Bootstrapping provides a more quantitative measure of the confidence in a claim of separation or overlap of experimental data clusters. The approach is applicable where sample populations between metabolic states vary significantly. Also, tree diagrams will greatly simplify the analysis of complex metabolic studies that involve numerous experimental conditions where the visual inspection of clustering patterns is extremely challenging.

Acknowledgments

This work was supported in part by funds from the American Heart Association (0860033Z). M.T.W. was supported by a sabbatical leave from Nebraska Wesleyan University. The research was performed in facilities renovated with support from the National Institutes of Health (NIH, RR015468-01).

References

[1] R. Powers, NMR metabolomics and drug discovery, Magn. Reson. Chem. 47 (2009) S2–S11.
[2] M. Coen, E. Holmes, J.C. Lindon, J.K. Nicholson, NMR-based metabolic profiling and metabonomic approaches to problems in molecular toxicology, Chem. Res. Toxicol. 21 (2008) 9–27.
[3] D.W. Nebert, E.S. Vesell, Can personalized drug therapy be achieved? A closer look at pharmaco-metabonomics, Trends Pharmacol. Sci. 27 (2006) 580–586.
[4] S. Rezzi, Z. Ramadan, L.B. Fay, S. Kochhar, Nutritional metabonomics: applications and perspectives, J. Proteome Res. 6 (2007) 513–525.
[5] D.G. Robertson, M.D. Reily, J.D. Baker, Metabonomics in pharmaceutical discovery and development, J. Proteome Res. 6 (2007) 526–539.
[6] E. Holmes, H. Antti, Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR spectra, Analyst 127 (2002) 1549–1557.
[7] J. Trygg, E. Holmes, T. Lundstedt, Chemometrics in metabonomics, J. Proteome Res. 6 (2007) 469–479.
[8] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2001) 109–130.
[9] R. Stoyanova, T.R. Brown, NMR spectral quantitation by principal component analysis, NMR Biomed. 14 (2001) 271–277.
[10] J.C. Lindon, E. Holmes, J.K. Nicholson, Pattern recognition methods and applications in biomedical magnetic resonance, Prog. Nuclear Magn. Reson. Spectrosc. 39 (2001) 1–40.
[11] J. Xia, N. Psychogios, N. Young, D.S. Wishart, MetaboAnalyst: A web server for metabolomic data analysis and interpretation, Nucleic Acids Res. 37 (2009) W652–W660.
[12] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.
[13] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.
[14] D. Coomans, I. Broeckaert, M.P. Derde, A. Tassin, D.L. Massart, S. Wold, Use of a microcomputer for the definition of multivariate confidence regions in medical diagnosis based on clinical laboratory profiles, Comput. Biomed. Res. 17 (1984) 1–14.
[15] J. Felsenstein, Confidence limits on phylogenies: An approach using the bootstrap, Evolution 39 (1985) 783–791.
[16] B. Efron, G. Gong, A leisurely look at the bootstrap, the jackknife, and cross-validation, Am. Stat. 37 (1983) 36–48.
[17] B. Efron, E. Halloran, S. Holmes, Bootstrap confidence levels for phylogenetic trees, Proc. Natl. Acad. Sci. USA 93 (1996) 13429–13434.
[18] O. Fiehn, Metabolomics: the link between genotypes and phenotypes, Plant Mol. Biol. 48 (2002) 155–171.
[19] U. Roessner, J. Bowne, What is metabolomics all about?, Biotechniques 46 (2009) 363–365
[20] F. Pietra, Evolution of the secondary metabolite versus evolution of the species, Pure Appl. Chem. 74 (2002) 2207–2211.
[21] J. Smedsgaard, J. Nielsen, Metabolite profiling of fungi and yeast: from phenotype to metabolome by MS and informatics, J. Exp. Bot. 56 (2005) 273–286.
[22] R.P. Maharjan, T. Ferenci, Metabolomic diversity in the species *Escherichia coli* and its relationship to genetic population structure, Metabolomics 1 (2005) 235–242.
[23] M.-Z. Ding, X. Zhou, Y.-J. Yuan, Metabolome profiling reveals adaptive evolution of *Saccharomyces cerevisiae* during repeated vacuum fermentations, Metabolomics, in press, available from doi:10.1007/s11306-009-0173-3.
[24] S.J. Oh, J.-G. Joung, J.-H. Chang, B.-T. Zhang, Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks, BMC Bioinform. 7 (2006) 284–296.
[25] A. Mazurie, D. Bonchev, B. Schwikowski, G.A. Buck, Phylogenetic distances are encoded in networks of interacting pathways, Bioinformatics 24 (2008) 2579–2585.
[26] O. Beckonert, M.E. Bollard, T.M.D. Ebbels, H.C. Keun, H. Antti, E. Holmes, J.C. Lindon, J.K. Nicholson, NMR-based metabonomic toxicity classification: Hierarchical cluster analysis and *k*-nearest-neighbour approaches, Anal. Chim. Acta 490 (2003) 3–15.
[27] P. Forgue, S. Halouska, M. Werth, K. Xu, S. Harris, R. Powers, NMR metabolic profiling of *Aspergillus nidulans* to monitor drug and protein activity, J. Proteome Res. 5 (2006) 1916–1923.
[28] S. Halouska, O. Chacon, R.J. Fenton, D.K. Zinniel, R.G. Barletta, R. Powers, Use of NMR metabolomics to analyze the targets of D-cycloserine in mycobacteria: role of D-alanine racemase, J. Proteome Res. 6 (2007) 4608–4614.
[29] J. Felsenstein, PHYLIP Phylogeny Inference Package (Version 3.2), Cladistics 5 (1989) 164–166.
[30] J.D. Retief, Phylogenetic analysis using PHYLIP, Methods Mol. Biol. 132 (2000) 243–258.
[31] W.M. Fitch, E. Margoliash, Construction of phylogenetic trees, Science 155 (1967) 279–284.
[32] A. Cotton James, M. Wilkinson, Majority-rule supertrees, Syst. Biol. 56 (2007) 445–452.
[33] M.P.H. Verouden, J.A. Westerhuis, M.J. van der Werf, A.K. Smilde, Exploring the analysis of structured metabolomics data, Chemometr. Intell. Lab. Syst. 98 (2009) 88–96.