Published in final edited form as:

Anal Chim Acta. 2020 January 25; 1095: 38–47. doi:10.1016/j.aca.2019.10.029.

Composite score analysis for unsupervised comparison and network visualization of metabolomics data

Joshua J. Kellogg^{†,§,*}, Olav M. Kvalheim[‡], Nadja B. Cech[†]

[†] Department of Chemistry & Biochemistry, University of North Carolina at Greensboro, Greensboro, NC, 27402, United States

[§] Department of Veterinary & Biomedical Sciences, Pennsylvania State University, University Park, PA, 16801, United States

[‡] Department of Chemistry, University of Bergen, Bergen 5020, Norway

Abstract

Metabolomics-based approaches are becoming increasingly popular to interrogate the chemical basis for phenotypic differences in biological systems. Successful metabolomics studies employ multivariate data analysis to compare large and highly complex datasets. A primary tool for unsupervised statistical analyses, principal component analysis (PCA), relies on the selection of a subsection of a maximum of three components from a larger model to visually represent similarity. The use of only three principal components limits the comprehensiveness of the model and can mask discrimination between samples. We have developed a new statistical metric, the composite score (CS), as a univariate statistic that incorporates multiple principal components to calculate a correlation matrix that enables quantitative comparisons of sample similarity between samples within one dataset based upon measured metabolome profiles. Composite score values were tabulated using profiles of complex extracts of dietary supplements from the plant *Hydrastis canadensis* (goldenseal) as a case study. Several outliers were unambiguously identified, and a PCA composite score network was developed to provide a graphical representation of the composite score matrix. Comparison with visualization using PCA score plots or dendrograms from hierarchical clustering analysis (HCA) demonstrates the utility of the composite score to as a

*Corresponding Author Telephone: 814-865-2287, Fax: 814-865-6140, jjk6146@psu.edu.

Associated Content

Supporting Information

The R script developed to calculate the composite score, plot the resulting heatmap, and prepare the data for network analysis is available for download at <https://github.com/jjkello/Composite-score>.

Supporting Information is available.

Author Information

The authors declare no competing financial interest.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

tool for metabolomics studies that seek to quantify similarity among samples. An R-script for the calculation of composite score has been made available.

Graphical Abstract

Keywords

Metabolomics; mass spectrometry; untargeted; natural products; multivariate statistical analysis; PCA; R

1. Introduction

Metabolomics analysis seeks to measure a holistic metabolome profile of a complex chemical or biological system, with the goal of correlating that profile to observable or quantifiable conditions in the system. Metabolomics has evolved into a defining analytical tool for chemical, biochemical, and biological analysis, and has been employed to study disease pathology [1], microbiome alterations [2], antibacterial resistance development [3], toxicology [4], and natural products drug discovery [5–7]. Some suggest that metabolomics is more responsive to endogenous and exogenous perturbations than other –omics approaches (e.g., transcriptomics or proteomics) [8]. Metabolomics is particularly useful in comparative analyses, which seek to correlate chemical profiles to differences between sample sets [9–10]. An untargeted metabolomics approach, where the entire measurable metabolome is analyzed without defined biomarker compounds to guide analysis, is especially useful when the study has no a priori mechanistic hypotheses.

Successful untargeted metabolomics studies rely on effective statistical analysis of the metabolomic dataset to guide interpretation and inform conclusions. Analysis of metabolomics datasets employs a variety of statistical methods, unsupervised methods such as hierarchical cluster analysis (HCA) [11–12], k-means clustering [11, 13], and principal component analysis (PCA) as well as supervised methods, for instance, soft independent modeling of class analogy (SIMCA) [14–15]. HCA, k-means clustering, and PCA are unsupervised representations of the data that can be used to ascribe and delineate clusters of

similar samples [11, 16–17]. While PCA was not originally designed as a clustering technique, the analysis of paired principal components in a scores plot have become ubiquitous in the analysis of metabolomics studies [12, 18–19], most frequently as graphical representations of pairwise comparisons of two principal components. However, visualization of only two components from a multi-component analysis inherently limits the amount of variability that is incorporated into the scores plot, and reduces the certainty of conclusions drawn [20].

For the k-means algorithm, one necessary pre-processing step is normalizing and scaling the data matrix, which can introduce errors into the data calculations [17]. In addition, k-means clustering relies on a pre-determined number of clusters to organize the dataset and k-means clustering can introduce artefacts when clusters are unevenly sized [21]. For untargeted analysis of natural products or other complex mixtures, the cluster size is often uneven (a few reference materials versus multiple unknown samples) or unknown, and k-means clustering is not an ideal approach in such scenarios.

SIMCA is a supervised statistical method, in which a PCA matrix is generated using one or more specified classes of data in order to determine their similarity against each other or to classify samples with unknown class-belonging [14]. This supervised approach is highly effective when a priori knowledge is available as to the sub-classes present in the sample (i.e. presence of known reference materials). However, for untargeted discovery analyses without prior knowledge about the sample set, SIMCA is not a preferred method [22].

When heat maps are employed to visually represent similarity, they frequently rely on raw spectral data, which has not been corrected to account for noise prior to visualization. To address this limitation, we recently introduced a metric that can provide a quantitative comparison mechanism employing multiple principle components from the PCA [19]. As described herein, we sought to utilize this metric, which we refer to as the “composite score” (CS), to provide a more complete comparison between samples compared to other traditional unsupervised analytical methods by capturing more of the systematic variation and information in the data into the metric. Using a case study of commercial goldenseal (*Hydrastis canadensis*) products, the composite scores metric was evaluated and compared with HCA, heat maps and traditional scores plots for untargeted metabolomics data analysis. With the work presented here, we also ventured to develop an accessible graphical representation of the correlation matrix using network analysis, similar in nature to a PCA scores plot in showing visual clusters of similar samples but overcoming the visualization problem posed by systematic variation spread across more than three principal components.

2. Materials and Methods

2.1. General Reagents

All solvents and reagents were of ACS or spectroscopic grade, as necessary, and obtained from Thermo Fisher Scientific (Waltham, MA, USA).

2.2. Sample Acquisition and Preparation

Commercial *Hydrastis canadensis* (goldenseal) products (35 total) were purchased and prepared as described previously (coded GS-01 – GS-34 and GS-44 to preserve supplier anonymity) [23]. *Hydrastis canadensis* leaf (GS-35) and root (GS-36) reference materials that were authenticated and validated as originating from the species in question were purchased from ChromaDex (Irvine, CA, USA) (Table 1). All materials were obtained as dried powders and extracted using the same method [23–24]. Briefly, 200 mg of each sample were loaded into a scintillation vial, shaken overnight with 20 mL methanol, then filtered and dried under N₂. Samples were reconstituted to 1 mg/mL prior to ultraperformance liquid chromatography-mass spectrometry (UPLC-MS) analysis.

2.3. UPLC-MS Data Acquisition

Ultraperformance (UP) LC-MS data were acquired utilizing a Q Exactive Plus quadrupole-Orbitrap mass spectrometer (ThermoFisher Scientific) with an electrospray ionization (ESI) source coupled to an Acquity UPLC system (Waters, Milford, MA, USA). Triplicate injections of 3 μ L were performed. An Acquity UPLC BEH C18 column (1.7 μ m, 2.1 $\times$ 50 mm, Waters) was employed with a flow rate of 0.3 mL/min using a binary solvent gradient of H₂O (with 0.1% formic acid added) and CH₃CN (with 0.1% formic acid added): initial isocratic composition of 95:5 (H₂O:CH₃CN) for 1.0 min, increasing linearly to 0:100 over 7 min, followed by an isocratic hold at 0:100 for 1 min, gradient returned to starting conditions of 95:5 and held isocratic again for 2 min. The positive/negative switching ionization mode of the mass spectrometer was utilized over a full scan range of m/z 150–2000 with the following settings: capillary voltage, 5 V; capillary temperature, 300 °C; tube lens offset, 35 V; spray voltage, 3.80 kV; sheath gas flow and auxiliary gas flow, 35 and 20 units, respectively.

2.4. Raw Data Alignment

The UPLC-MS scan data were processed, aligned, and filtered using MZmine 2.31 software (<http://mzmine.github.io/>) with a modified version of a previously reported method [19, 23]. The following parameters were used for peak detection: noise level (absolute value), 1×10^5 counts; minimum peak duration, 0.05 min; tolerance for m/z intensity variation, 20%, which had been determined based upon inspection of the peak height, baseline, as well as previous studies [23]. Peak list filtering and retention time alignment algorithms were performed to refine peak detection. The join algorithm was used to integrate all the chromatograms into a single data matrix using the following parameters: the balance between m/z and retention time was set at 10.0 each, m/z tolerance was set at 0.001, and retention time tolerance was defined as 0.5 min. The peak areas for individual ions detected in triplicate extractions were exported from the data matrix for further analysis. Samples in which a particular marker ion was below the limit of detection were coded with a peak area of 0 to maintain a consistent number of variables throughout the dataset.

Chemometric analysis was completed using Sirius version 10.0 (Pattern Recognition Systems AS, Bergen, Norway). Data transformation to reduce heteroscedastic noise was carried out via a fourth root transform of raw spectral data [25]. Principal component analysis (PCA) and hierarchical clustering analysis were used for untargeted metabolomics

profiling of the samples with Sirius software. Other statistical analyses were performed in R using custom scripts (available for download, <https://github.com/jjkellogg/Composite-score>).

2.5. Composite Score

The composite score (CS) was developed and implemented as an accompanying R script (Supplemental Information, <https://github.com/jjkellogg/Composite-score>). In our previous work, the term “reproduced correlation coefficient (RCC)” was used to describe the metric derived in equations 1–5 below [19]. In the current manuscript, the name of this metric has been changed to composite score for clarity and to better reflect the nature of the statistical metric applied.

The decomposition of any data matrix can be represented as:

$$\mathbb{X} = \bar{\mathbb{X}} + \sum_{a=1}^A \mathbf{t}_a \mathbf{p}_a^T + \mathbb{E}_{PCA} \quad (1)$$

where $\mathbb{X}$ is of size $m \times n$, where n is the number of spectral variables, $\bar{\mathbb{X}}$ is the mean response across the model, $\mathbf{t}_a$ is a vector of size $m \times 1$, known as the scores vector, $\mathbf{p}_a$ is a vector of size $n \times 1$, known as the loadings vector, and $\mathbb{E}$ represents the residual information remaining in $\mathbb{X}$ after the projection onto the latent variable space. The score and loading vectors can be collected in score T and loading P matrices with dimensions, $n \times A$ and $m \times A$, respectively, for A latent variables (i.e., principal components).

The summed products of scores and loadings vectors in Equation 1 can be rewritten as $\hat{\mathbb{X}}_{PCA}$, which represents the estimated data from the latent variables (principal components) accounting for the major variation in $\mathbb{X}$ and yielding the simplified equation:

$$\mathbb{X} = \bar{\mathbb{X}} + \hat{\mathbb{X}}_{PCA} + \mathbb{E}_{PCA} \quad (2)$$

The product of the estimated PCA matrix $\hat{\mathbb{X}}_{PCA}$ and its transpose ($\hat{\mathbb{X}}_{PCA}^T$) divided by the norm results in a correlation matrix that can comparison between variables (equation 3) or objects (equation 4)

$$\hat{\mathbb{X}}_{PCA}^T \hat{\mathbb{X}}_{PCA} / \|\hat{\mathbb{X}}_{PCA}^T \hat{\mathbb{X}}_{PCA}\| \quad (3)$$

$$\hat{\mathbb{X}}_{PCA} \hat{\mathbb{X}}_{PCA}^T / \|\hat{\mathbb{X}}_{PCA} \hat{\mathbb{X}}_{PCA}^T\| \quad (4)$$

Translating this correlation into a composite score (CS) between two objects, ℓ and ℓ'

$$CS = \hat{\mathbf{x}}_{PCA \ell}^T \cdot \hat{\mathbf{x}}_{PCA \ell'} / \|\hat{\mathbf{x}}_{PCA \ell}\| \|\hat{\mathbf{x}}_{PCA \ell'}\| \quad (5)$$

The greater the composite score (CS) for two samples, the more similar their overall spectra are, and thus the closer the two samples are in metabolite profile.

2.6. PCA Composite Score Network Visualization

The pair-wise comparisons of the composite scores were translated into a network for visualization of the similarities between samples. All pair-wise comparisons were imported into Cytoscape 3.6.1 and displayed as a network of nodes, with edges described by the composite score value [26–27]. To simplify the network, duplicated edges and self-loops were removed from the network. Remaining nodes were organized using the edge-weighted spring embedded layout, based upon the “force-directed” paradigm of Kamada and Kawai [28], in which nodes possessing greater similarity are spatially closer to one another, while nodes that have a lesser similarity are placed further apart. Node colors were mapped based upon the source of the sample, and the edge thickness attribute was defined to reflect the composite score, with thicker lines indicating greater similarity between the two nodes (samples). Sub-networks were generated in Cytoscape by defining a minimum significant similarity score delineating similar samples. This score would be application-specific, enabling the user to tune their specificity to their needs. For the purposes of this study, 0.65 was selected as a significant similarity score due to its prevalence in other untargeted metabolomics workflows designed to associate two chromatograms or spectra [27].

3. Results & Discussion

3.1. Ambiguity of the Principal Component Analysis Scores Plots

The combined positive ion mode and negative ion mode metabolomics dataset yielded 5,423 features (unique *m/z*-retention time pairings) across 35 commercial goldenseal (*Hydrastis canadensis*) products and two standard reference materials. The sample set used for this study was chosen because it had previously been profiled in another metabolomics study [23]. Prior knowledge about the sample set enabled evaluation of the effectiveness and limitations of PCA to quantify differences among samples. The commercial goldenseal products are of two main types, those prepared from goldenseal roots and those prepared from goldenseal leaves (Table 1). Based on previous analyses, we were aware that there are chemical differences among the root and leaf products [24], and expected to see this reflected in PCA plots. Additionally, our previous study showed that three of the commercial products (GS-07, GS-20, and GS-33) [23], were adulterated, i.e. the label claims that these samples were *Hydrastis canadensis* were false, and the samples were actually prepared from plant material of a different species. We expected that PCA plots would identify these adulterated samples as outliers.

Plotting the model's first four principal components (PCs) in a pairwise fashion yielded six distinct PCA scores plots (Figure 1). Upon visual inspection, the spatial location of the three outlier samples (GS-07, GS-20, and GS-33), and their relative position compared to the rest of the dataset, shifted considerably depending on which pair of components were plotted. For instance, PC1 vs PC2 (Figure 1A) indicated the expected position of the three outlier samples outside the cluster of other commercial goldenseal supplements. However, the PC1

vs. PC2 projection was not effective for distinguishing the chemically different root and leaf samples.

Plotting other combinations of principal components, e.g., PC1 vs PC3 (Figure 1B), PC2 vs PC3 (Figure 1D), or PC3 vs PC4 (Figure 1F), enabled the root and leaf samples to be distinguished from each other, but the outlier samples were now observed to cluster amongst the rest of the samples. For PC1 vs PC3, only GS-20 could be visually identified as an outlier, as GS-33 was found within the goldenseal root cluster (within the 95% Hotelling's confidence ellipse), and GS-07 just outside of that grouping. PC2 vs PC3 identified GS-33 as an outlier, with GS-20 lying fully within the root sample cluster, and GS-07 located just outside this cluster. And PC3 vs PC4 yielded a single significant outlier, GS-20. The variability in results (Figure 1), depending on which two components are plotted, demonstrated the limitation of the common practice of reporting PCA plots with just two principal components. While a careful examination of the principal components individually could suggest that PC2 highlights the outlier samples and PC 3 differentiates the two disparate groups of goldenseal samples, there is no combination of principal components (Figure 1) that unambiguously identified both the groupings and the outliers. Even when those two components are taken together (PC2 vs PC3, Figure 1D), one of the outliers (GS-20) is still overlapping with the goldenseal cluster. Furthermore, this analysis was performed with already determined groupings. Without a priori knowledge of the sample identification or constituents, the ad hoc choice of paired PCA components to generate hypotheses can potentially mask one or more outliers, leading to specious results and conclusions.

3.2. Selection of Model Components

To obtain a discriminating metric for comparing samples in complex datasets, relying on pairwise PCA components is not sufficient (Figure 1). To comprehensively compare the samples, we sought to combine the principal components to form a single metric of quantitative similarity between all the samples, the composite score (Equation 5). However, one of the overriding factors in developing a global unsupervised PC model is to determine an adequate number of principal components on which to base the model. If too few components are chosen, the model could lack pertinent information and provide an incomplete representation of the dataset. On the other hand, if too many principal components are included, the model will be overfit with noise, leading to an overparameterization (i.e., ascribing meaning to random noise in the dataset). In determining the number of relevant components, previous studies have suggested looking for a "consensus dimension" derived from the combined analysis of principal components by multiple statistical methods [29–30]. The PCA analysis of the goldenseal dataset produced a 38-component PCA model, but preliminary correlations based on a full 38-component model were overfit, and unable to highlight similarities between samples (Supplemental Information, Table S2). To avoid over- or under-fitting of the PCA model, it was first necessary to evaluate which components of the model were significantly contributing to the model.

Four main methods were employed to estimate the adequate number of principal components to retain for the final model. One of the most common stopping criteria in PCA is the Guttman-Kaiser rule, in which eigenvalues that are larger in magnitude than the average are retained. An extension of this, which takes into account effect of sample variance, is Jolliffe's modification of the Guttman-Kaiser rule, defined as 0.7 times the average of the variance [20]. From the PCA model of the data used herein, the conservative Jolliffe's modification to the Kaiser-Guttman rule retained nine principal components and was used as an upper bound for the number of components to include in the model, while the unmodified Kaiser-Guttman rule yielded seven principal components (Figure 2).

The broken stick approach is an apportionment model using the total variance at the data and comparing the contribution of each eigenvalue to the total model variance. As such, a component is retained if its associated eigenvalue is larger than the broken-stick distribution value [31]. Using the broken stick approach, four significant components were considered significant contributors (Figure 2).

Finally, a non-graphical approach to evaluating the scree plot was utilized, the optimal coordinate model. Determining the location of the inflection point from the scree plot relies on projecting each eigenvalue through the preceding coordinates, and the component is retained if its associated eigenvalue is greater than or equal to the estimated eigenvalue from the projection [32]. The optimal coordinate model, applied to the goldenseal commercial product data, recommended five principal components to retain for the final analysis (Figure 2). The general consensus between the broken-stick approach and the optimal coordinate model (four and five principal components, respectively) provided the final determination of the dimensionality of the PCA model, and five components were retained to calculate the composite score matrix. These four approaches to determining the number of principal components were included in the final R script (Supplemental Information, <https://github.com/jjkellogg/Composite-score>).

3.3. Hierarchical Clustering Analysis of Goldenseal Metabolome

Hierarchical clustering analysis (HCA) is another popular unsupervised statistical tool for comparing multiple samples. HCA does not provide a single partition of the dataset, but instead allows the user to visualize and decide the clusters present [33]. Hierarchical clustering employs a similarity metric between pairs of subjects to produce a dendrogram of nested clusters – Euclidean distance, Manhattan distance, Mahalanobis distance and maximum distance are frequently used metrics [33–35].

Hierarchical clustering analysis of the commercial goldenseal products employed an average-linkage algorithm to group objects based on Euclidean distance [36], and was performed using the first five principal components that were retrained from the overall model. The analysis revealed three dominant clusters (Figure 3). The three known outlier samples, GS-07, GS-20, and GS-33, were sequestered into a single cluster, distinct from the other samples. The two other groupings contained members from leaf and root product samples, with good discrimination between the two; four of the six leaf samples were located in a subcluster together. Two leaf samples were intermingled within the root material samples' cluster (Figure 3). Thus, employing HCA with the reduced principal components

yielded a more definite clustering solution that delineated different sample classes with a few mis-identifications present (Figure 3).

3.4. Similarity Comparisons Using the Composite Score matrix

The composite scores comprising the model data from the first five PCA components collectively encapsulated 74.4 % of the variation in the data. The corresponding scores and loadings served as input into the correlation calculations, which was achieved via a custom R script (Supplemental Information, <https://github.com/jkellogg/Composite-score>). The composite score represents a quantitative measure of the similarity of the systematic variation patterns between any two samples in the dataset. The data can be represented as an $n \times n$ matrix and visualized as a heatmap (Figure 4A). The full matrix with composite score values is presented in the Supplemental Information for reference (Supplemental Information, Table S1). Composite scores range from -1 (perfectly negatively correlated) to $+1$ (perfect positive correlation) between any two samples. The heatmap (Figure 5A) is shown with darker color density (deeper blue) to indicate a higher value of the composite score (and stronger correlation between two samples). For example, comparing one sample of commercial goldenseal root (GS-17) to the goldenseal root reference material (GS-36), the composite score provided a similarity score of 0.870 (Figure 4B). The LC-MS chromatograms for the two samples, both in the positive and negative ionization mode, demonstrated strong visual similarities. However, comparing commercial goldenseal sample GS-33 to the GS-36 reference material yielded a relatively poor correlation value of -0.334 , suggesting chemical differences between the two samples. Visual inspection of the chromatographic data for these samples (Figure 4C) revealed substantial differences in peak pattern, intensity, and retention time.

The close agreement between the chemical data from the LC-MS chromatograms supports the use of the composite score to describe chemical similarity in a pair-wise manner across a large dataset. The composite score also offers the benefit of comparing every sample against every other sample simultaneously, in one single calculation. This is a marked distinction from the HCA analysis (Figure 3) based upon the same PCA principal components, as comparisons of similarity beyond next neighbors is difficult to quantify.

The heatmap for the original goldenseal dataset was scaled to highlight only positive correlations by setting the background color (white) for any correlations <0 (Figure 5). The outlier status of samples GS-07, GS-20, and GS-33 is recognizable in the composite score matrix by their lack of distinct correlations with the standard reference materials (GS-35 and GS-36) or the majority of commercial goldenseal products. GS-07 showed no strong correlations (composite score >0.65), and moderate correlations with three samples (GS-33, 0.868; GS-31, 0.357; GS-04, 0.218; GS-05, 0.262); however, there were no positive correlations between this sample and the goldenseal reference materials. Sample GS-20 showed strong correlations with GS-03 (0.742) and GS-19 (0.777), and moderate correlations with GS-27 (0.541) and GS-029 (0.540), but only mild associations with the root goldenseal reference material (GS-36, 0.331). GS-33 had only one moderate correlation with another sample (GS-44, 0.868), and no correlation with reference standards. The other goldenseal samples evidenced positive correlation ($CS >0$) with either one of the reference

materials (GS-35 and GS-36), with the exception of GS-04 and GS-12. These two samples were correlated with other goldenseal samples (e.g., GS-04 correlated with GS-12 (0.954), GS-25 (0.788) and GS-21 (0.780)) and evidenced low correlations with the outlier samples (e.g., GS-04's correlation with GS-33 = -0.088). While the reasoning of the lack of reference correlation is not currently known, it must be restated that this dataset was built on commercially supplied samples, and as such their botanical provenance could not be authenticated.

3.5. Composite Score Network Diagram

While the composite score matrix, either as a heatmap (Figures 4 and 5) or table of data (Supplemental Information, Table S1) provided excellent and efficient pair-wise comparisons between samples, a heatmap representation is not convenient to observe trends across the entire dataset. To address this, a network analysis approach was employed using PCA composite scores [26–27] (Figure 6). Significant correlations ($CS > 0.65$) were used to form the edges of the network, and edge thickness was directly proportional with the strength of the connection (correlation) between the two samples.

Three outlier samples (GS-07, GS-20, and GS-33) had previously demonstrated variable relation to the main dataset, depending on which two principal components were selected to graphically display the PCA model (Figure 1). The composite score network, however, clearly delineated these three samples as having few to no significant connections to other samples. A network including significant composite correlation scores ($CS > 0.65$, Figure 6) displayed two distinct clusters of samples, with the leaf material distinct from the root samples. In addition, GS-07 and GS-33 did not possess any significant connections to other samples and were visually observable as outliers, while GS-20 was a skewed datapoint, separate from the main cluster of samples even though it did have three connections to samples.

The lack of connectivity between GS-33 and other samples follows the lack of correlation evidenced in the composite score matrix (Figures 4 and 5). GS-33 was clearly designated as an outlier in the other statistical analyses (Figures 1, and 3), and is likely due to the presence of a non-*H. canadensis* botanical within the product [23]. GS-07 also did not manifest any correlations with goldenseal root or leaf samples, and appears to be a mixture of botanicals that would result in multiple weak correlations between samples [23]. Finally, the minor connections between GS-20 and four goldenseal root samples (GS-03 GS-27, GS-29, and GS-19) is likely to be due to some metabolite overlap between these botanicals, especially the main alkaloid present in both samples, berberine [37–38]. This similarity, derived from the composite correlation, was supported by comparing the chromatographic and mass spectral data of the three outlier samples to other goldenseal products, as was reported previously [23].

The network analysis incorporates a distinct advantage over the HCA calculation of similarity (Figure 3). First, using a network software such as Cytoscape allows for the degree of visualized similarity to be tuned and provide multiple viewpoints of the data at different correlation thresholds. For the current study, a composite score cut-off of 0.65 was chosen as it is a common mass spectrometry-based similarity limit, seen in other

metabolomics studies [REF]. This flexibility is not possible with HCA. Furthermore, for datasets with larger numbers of objects, the network analysis diagram could provide scalability to detect subnetworks. The ability to set different similarity thresholds is a distinct advantage of the composite score and the network analysis of facilitating deeper analysis of complex datasets.

4. Conclusions

Chemically-complex samples are ubiquitous in our environment, from natural products [39–40], food and nutraceutical matrices [19, 41], and environmental samples [42]. Analyzing variation and being able to ascribe (dis)similarity between samples is a challenge. Metabolomics remains a powerful analytical tool to ascertain differences between samples; however, the greatest obstacle for metabolomics experiments tends not to be in the instrumental analysis and data collection, but rather in meaningful, robust data interpretation and annotation. Multivariate analysis of a metabolomics study is key to modeling variance and providing a rigorous description of the dataset. This study illustrates some of the inherent limitations of using principal component analysis (PCA) for clustering analysis of high-dimensional metabolomics data. While PCA is not inherently a clustering metric, it has evolved into a primary tool for gauging clusters of complex data like that found in metabolomics studies. For PCA, the model is usually visualized on two axes, presented as a score plot, which inherently limits the data (and associated model variance) offered visually. Furthermore, depending on the selection of components to produce the score plot, there could be a shifting relationship between samples (Figure 1), which could lead to misguided discussions, conclusions, and further hypotheses, especially for unknown sample sets. The composite score represents a quantitative metric that incorporates the significant contributing PCA model components into a single value and provides a useful alternative to the ad hoc visual comparisons of PCA scores plots. The composite correlation allows for simultaneous pair-wise comparisons between all samples in the dataset (Figure 4), and the visual parameters of a heatmap depiction of the correlations enables facile comparison of samples (Figure 5). Furthermore, composite correlation data are ideally suited for transformation into a two-dimensional network diagram (Figure 6) for visual representation of the relative spatial distribution of samples based upon their correlations. The composite score has the potential to provide clear, quantitative scores as a means for measuring (dis)similarity between complex samples in a large data set.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Center for Complementary and Integrative Health (NCCIH) at the National Institutes of Health (NIH), specifically the Ruth L. Kirschstein Postdoctoral National Research Service Award (F32 AT009816) and the Center of Excellence for Natural Product Drug Interaction Research (NaPDI, U54 AT008909). Mass spectrometry analyses were conducted in the Triad Mass Spectrometry Facility at the University of North Carolina at Greensboro (<https://chem.uncg.edu/triadmslab/>). The authors would like to acknowledge Dr. Laura Sanchez (orcid.org/0000-0001-9223-7977) for her technical assistance with networking software, and Dr. Brian O'Connor for his assistance in the development of some R-encoded statistical comparisons.

6. References

- [1]. Dona AC; Coffey S; Figtree G, Translational and emerging clinical applications of metabolomics in cardiovascular disease diagnosis and treatment. *Eur. J. Prev. Cardiol* 2016, 23 (15), 1578–1589. [PubMed: 27103630]
- [2]. Liu R; Hong J; Xu X; Feng Q; Zhang D; Gu Y; Shi J; Zhao S; Liu W; Wang X, Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nat. Med* 2017, 23 (7), 859. [PubMed: 28628112]
- [3]. Zampieri M; Zimmermann M; Claassen M; Sauer U, Nontargeted metabolomics reveals the multilevel response to antibiotic perturbations. *Cell Rep.* 2017, 19 (6), 1214–1228. [PubMed: 28494870]
- [4]. Dinis-Oliveira RJ, Metabolomics of methadone: clinical and forensic toxicological implications and variability of dose response. *Drug Metab. Rev* 2016, 48 (4), 568–576. [PubMed: 27320437]
- [5]. Li G; Zhang Z; Quan Q; Jiang R; Szeto SS; Yuan S; Wong W.-t.; Lam HH; Lee SM-Y; Chu IK, Discovery, synthesis, and functional characterization of a novel neuroprotective natural product from the fruit of *Alpinia oxyphylla* for use in Parkinson's disease through LC/MS-based multivariate data analysis-guided fractionation. *J. Proteome Res.* 2016, 15 (8), 2595–2606. [PubMed: 27246451]
- [6]. Shang N; Saleem A; Musallam L; Walshe-Roussel B; Badawi A; Cuerrier A; Arnason JT; Haddad PS, Novel approach to identify potential bioactive plant metabolites: pharmacological and metabolomics analyses of ethanol and hot water extracts of several Canadian medicinal plants of the Cree of Eeyou Istchee. *PLOS One* 2015, 10 (8), e0135721. [PubMed: 26263160]
- [7]. Kellogg JJ; Todd DA; Egan JM; Raja HA; Oberlies NH; Kvalheim OM; Cech NB, Biochemometrics for Natural Products Research: Comparison of Data Analysis Approaches and Application to Identification of Bioactive Compounds. *J. Nat. Prod* 2016, 79 (2), 376–386. [PubMed: 26841051]
- [8]. Kell DB; Brown M; Davey HM; Dunn WB; Spasic I; Oliver SG, Metabolic footprinting and systems biology: the medium is the message. *Nat. Rev. Microbiol* 2005, 3, 557. [PubMed: 15953932]
- [9]. Hou Y; Braun DR; Michel CR; Klassen JL; Adnani N; Wyche TP; Bugni TS, Microbial Strain Prioritization Using Metabolomics Tools for the Discovery of Natural Products. *Anal. Chem* 2012, 84 (10), 4277–4283. [PubMed: 22519562]
- [10]. Stewart DA; Winnike JH; McRitchie SL; Clark RF; Pathmasiri WW; Sumner SJ, Metabolomics Analysis of Hormone-Responsive and Triple-Negative Breast Cancer Cell Responses to Paclitaxel Identify Key Metabolic Differences. *J. Proteome Res.* 2016, 15 (9), 3225–3240. [PubMed: 27447733]
- [11]. Beckonert O; E. Bollard M; Ebbels TMD; Keun HC; Antti H; Holmes E; Lindon JC; Nicholson JK, NMR-based metabolomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. *Anal. Chim. Acta* 2003, 490 (1), 3–15.
- [12]. Caesar LK; Kvalheim OM; Cech NB, Hierarchical cluster analysis of technical replicates to identify interferents in untargeted mass spectrometry metabolomics. *Anal. Chim. Acta* 2018, 1021, 69–77. [PubMed: 29681286]
- [13]. Wang W; Cheng K-K; Deng L; Xu J; Shen G; Griffin JL; Dong J, A clustering-based preprocessing method for the elimination of unwanted residuals in metabolomic data. *Metabolomics* 2017, 13 (1), 10.
- [14]. Tsugawa H; Tsujimoto Y; Arita M; Bamba T; Fukusaki E, GC/MS based metabolomics: development of a data mining system for metabolite identification by using soft independent modeling of class analogy (SIMCA). *BMC Bioinformatics* 2011, 12 (1), 131. [PubMed: 21542920]
- [15]. Wold S, Pattern recognition by means of disjoint principal components models. *Pattern Recogn.* 1976, 8 (3), 127–139.
- [16]. Abdi H; Williams LJ, Principal component analysis. *WIREs Comp. Stat* 2010, 2 (4), 433–459.
- [17]. Arora P; Deepali; Varshney S, Analysis of K-Means and K-Medoids Algorithm For Big Data. *Procedia Comput. Sci.* 2016, 78, 507–512.

- [18]. Booker A; Suter A; Krnjic A; Strassel B; Zloh M; Said M; Heinrich M, A phytochemical comparison of saw palmetto products using gas chromatography and 1H nuclear magnetic resonance spectroscopy metabolomic profiling. *J. Pharm. Pharmacol* 2014, 66 (6), 811–822. [PubMed: 24417505]
- [19]. Kellogg JJ; Graf TN; Paine MF; McCune JS; Kvalheim OM; Oberlies NH; Cech NB, Comparison of Metabolomics Approaches for Evaluating the Variability of Complex Botanical Preparations: Green Tea (*Camellia sinensis*) as a Case Study. *J. Nat. Prod.* 2017, 80 (5), 1457–1466. [PubMed: 28453261]
- [20]. Jolliffe IT; Morgan BJT, Principal component analysis and exploratory factor analysis. *Stat. Methods Med. Res* 1992, 1 (1), 69–95. [PubMed: 1341653]
- [21]. Jain AK, Data clustering: 50 years beyond K-means. *Pattern Recogn. Lett* 2010, 31 (8), 651–666.
- [22]. Harnly J; Jabolonski J; Moore J, A Model for Nontargeted Detection of Adulterants In Botanicals: Methods and Techniques for Quality & Authenticity, Reynertson KA; Mahmood K, Eds. CRC Press: Boca Raton, FL, 2015; p 91.
- [23]. Wallace ED; Oberlies NH; Cech NB; Kellogg JJ, Detection of adulteration in *Hydrastis canadensis* (goldenseal) dietary supplements via untargeted mass spectrometry-based metabolomics. *Food Chem. Toxicol.* 2018, 120, 439–447. [PubMed: 30031041]
- [24]. Britton ER; Kellogg JJ; Kvalheim OM; Cech NB, Biochemometrics to Identify Synergists and Additives from Botanical Medicines: A Case Study with *Hydrastis canadensis* (Goldenseal). *J. Nat. Prod* 2018, 81 (3), 484–493. [PubMed: 29091439]
- [25]. Kvalheim OM; Brakstad F; Liang Y, Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. *Anal. Chem* 1994, 66 (1), 43–51.
- [26]. Smoot ME; Ono K; Ruscheinski J; Wang PL; Ideker T, Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 2011, 27 (3), 431–2. [PubMed: 21149340]
- [27]. Yang JY; Sanchez LM; Rath CM; Liu X; Boudreau PD; Bruns N; Glukhov E; Wodtke A; de Felicio R; Fenner A; Ruh Wong W; Linington RG; Zhang L; Debonsi HM; Gerwick WH; Dorrestein PC, Molecular Networking as a Dereplication Strategy. *J. Nat. Prod* 2013, 76 (9), 1686–1699. [PubMed: 24025162]
- [28]. Kamada T; Kawai S, A simple method for computing general position in displaying three-dimensional objects. *Lec. Notes. Comput. Sc* 1988, 41 (1), 43–56.
- [29]. Cangelosi R; Goriely A, Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct* 2007, 2 (1), 2. [PubMed: 17229320]
- [30]. Valle S; Li W; Qin SJ, Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods. *Ind. Eng. Chem. Res* 1999, 38, 4389–4401.
- [31]. Jackson DA, Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 1993, 74 (8), 2204–2214.
- [32]. Raïche G; Walls TA; Magis D; Riopel M; Blais J-G, Non-graphical solutions for Cattell's scree test. *Methodology* 2013, 9, 23–29.
- [33]. Boccard J; Veuthey J-L; Rudaz S, Knowledge discovery in metabolomics: An overview of MS data handling. *J. Sep. Sci* 2010, 33 (3), 290–304. [PubMed: 20087872]
- [34]. Jain AK; Murty MN; Flynn PJ, Data clustering: a review. *ACM Comp. Surv.* 1999, 31 (3), 264–323.
- [35]. Ren S; Hinzman AA; Kang EL; Szczesniak RD; Lu LJ, Computational and statistical analysis of metabolomics data. *Metabolomics* 2015, 11 (6), 1492–1513.
- [36]. Kaufman L; Rousseeuw PJ, Finding groups in data: an introduction to cluster analysis. John Wiley & Sons: 2009; Vol. 344.
- [37]. Rackova L; Majekova M; Kost'alo D; Stefek M, Antiradical and antioxidant activities of alkaloids isolated from *Mahonia aquifolium*. *Structural aspects. Bioorgan. Med. Chem* 2004, 12 (17), 4709–15.
- [38]. Weber HA; Zart MK; Hodges AE; Molloy HM; O'Brien BM; Moody LA; Clark AP; Harris RK; Overstreet JD; Smith CS, Chemical comparison of goldenseal (*Hydrastis canadensis* L.) root powder from three commercial suppliers. *J. Ag. Food Chem* 2003, 51 (25), 7352–8.

- [39]. Kingston DGI, A Natural Love of Natural Products. *J. Org. Chem* 2008, 73 (11), 3975–3984. [PubMed: 18459734]
- [40]. Wolfender J-L; Marti G; Thomas A; Bertrand S, Current approaches and challenges for the metabolite profiling of complex natural extracts. *J. Chromatogr. A* 2015, 1382, 136–164. [PubMed: 25464997]
- [41]. Sut S; Baldan V; Faggian M; Peron G; DallAcqua S, Nutraceuticals, a new challenge for medicinal chemistry. *Curr. Med. Chem* 2016, 23 (28), 3198–3223. [PubMed: 27319583]
- [42]. Schoenfuss HL; Furlong ET; Phillips PJ; Scott T-M; Kolpin DW; Cetkovic-Cvrlje M; Lesteberg KE; Rearick DC, Complex mixtures, complex responses: Assessing pharmaceutical mixtures using field and laboratory approaches. *Environ. Toxicol. Chem* 2016, 35 (4), 953–965. [PubMed: 26561986]

Highlights

- Combined multiple principal components into a single statistical metric
- Composite score unambiguously described outliers from the dataset
- Network analysis provided clear graphical interpretation of metabolomics data
- Efficient tool for ascribing similarity and differences among samples

Figure 1.

Principal component analysis (PCA) of data from commercial goldenseal (*Hydrastis canadensis*) samples. Different combinations of PCA components illustrate variability in the spatial distribution of the outlier (adulterated) samples (red diamonds), goldenseal leaf samples (green symbols and ellipses), and goldenseal root samples (yellow symbols and ellipses). Depending on which two PCA components are selected, the leaf samples are clustered independently of the root samples (1B, 1D, and 1F), or their confidence ellipses overlap suggesting no significant differences (1A, 1C, 1E). Similarly, the choice of component pairing affects whether or not the outlier samples are located within the 95% Hotelling's confidence ellipse. This exemplifies the difficulties in using pairs of two principal components as a basis for visualizing differences between samples.

Figure 2.

The number of significant principal components to include in the calculation of composite score from the PCA model was determined via a comparison of the variance (eigenvalue) for each components versus multiple statistical means of model evaluation: optimal coordinate model, broken stick model, Kaiser-Guttman criterion, and Joliffe's modification of the Kaiser-Guttman criterion. The Joliffe's modification of Kaiser-Guttman yielded 9 components as an upper bound on the significant number of principal components. Subsequently, the optimal coordinate approach yielded five adequate principal components, respectively, and thus those first five were retained to perform composite score calculations.

Figure 3.

Hierarchical cluster analysis (HCA) of goldenseal (*Hydrastis canadensis*) product samples based upon the same five principal components that were determined to be significant (see Figure 2). Samples are color coded to indicate their botanical origin: root material (yellow symbols), leaf material (green), or outlier sample (red) (Table 1). The outlier samples (GS007, GS-20, and GS-33) were all clustered as distinct from the other samples. The HCA clustering largely differentiated between leaf and root material, with a couple exceptions: the leaf samples GS-31 and GS-18 were interspersed amongst the root material.

Figure 4.

Composite score analysis of a goldenseal (*Hydrastis canadensis*) metabolomics dataset including 35 samples and two reference materials. A) Heatmap of the 37×37 composite score matrix displays darker hues as an indication of greater correlation between two samples. B) Comparison of a representative commercial goldenseal root sample (GS-17) and the authenticated goldenseal root reference material (GS-36), which reported high correlation, and visual inspection of the base peak LC-MS chromatograms in both positive and negative ionization mode showed close similarities. C) Comparison of the known outlier

(GS-33) and GS-36, which had a low composite score, evidenced chromatograms that were quite different in both ionization modes.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 5.

Composite score heat map of 35 commercial goldenseal (*Hydrastis canadensis*) products and two relevant authentic reference materials (leaf material, GS-35 and root material, GS-36).

The heatmap was scaled to show only positive correlations (composite score > 0).

Figure 6.

Composite Principal Component Analysis network diagram for goldenseal metabolomics dataset. Composite similarity scores > 0.65 were used to construct networks in Cytoscape [26]. All of the commercial goldenseal root products (yellow symbols), along with the relevant authentic *H. canadensis* reference material (GS-36; orange symbol) are clustered in a distinct sub-network. Similarly, all of the commercial goldenseal leaf products (light green symbols), along with the authenticated leaf reference material (GS-35; dark green symbol) were clustered together. Two outlier samples (GS-33 and GS-07; red symbols) were clearly

evidenced as being distinct from these two clusters. The third outlier, GS-20 (red symbol) had three weak connections to other root products.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1.Commercial *Hydrastis canadensis* (goldenseal) products chosen for the study

Product Type ^a	Product Codes	Figure Color ^b
Commercial goldenseal root	GS-01, GS-02, GS-03, GS-04, GS-06, GS-08, GS-09, GS-10, GS-12, GS-14, GS-15, GS-16, GS-17, GS-19, GS-21, GS-22, GS-23, GS-24, GS-25, GS-26, GS-27, GS-28, GS-29, GS-30, GS-32, GS-34, GS-44	Yellow
Commercial goldenseal leaf	GS-05, GS-11, GS-13, GS-18, GS-31	Light green
Authenticated goldenseal root reference material	GS-36	Orange
Authenticated goldenseal leaf reference material	GS-35	Dark green
Outlier (non-goldenseal or blend)	GS-07, GS-20, GS-33	Red

^aDifferent botanical sources and commercial versus authenticated reference material were included, as well as three products identified in a previous study [23] as outliers from the main dataset.

^bFigure color is the color used to represent specific sample class in Figures 1, 2, 3, and 7.