Application of Dissimilarity Indices, Principal Coordinates Analysis, and Rank Tests to Peak Tables in Metabolomics of the Gas Chromatography/Mass Spectrometry of Human Sweat

ARTICLE *in* ANALYTICAL CHEMISTRY · SEPTEMBER 2007

Impact Factor: 5.64 · DOI: 10.1021/ac070134w · Source: PubMed

9 AUTHORS, INCLUDING:

Yun Xu
The University of Manchester
62 PUBLICATIONS **976** CITATIONS
SEE PROFILE

Elisabeth Oberzaucher
Alpen-Adria-Universität Klagenfurt
54 PUBLICATIONS **475** CITATIONS
SEE PROFILE

Karl Grammer
University of Vienna
161 PUBLICATIONS **5,071** CITATIONS
SEE PROFILE

Dustin J. Penn
University of Veterinary Medicine in Vienna
97 PUBLICATIONS **4,692** CITATIONS
SEE PROFILE

Anal. Chem. **2007,** *79,* 5633−5641

Application of Dissimilarity Indices, Principal Coordinates Analysis, and Rank Tests to Peak Tables in Metabolomics of the Gas Chromatography/Mass Spectrometry of Human Sweat

Yun Xu, Fan Gong, Sarah J. Dixon, and Richard G. Brereton*

Centre for Chemometrics, School of Chemistry, University of Bristol, Cantocks Close Bristol BS8 1TS, United Kingdom

Helena A. Soini and Milos V. Novotny

Institute for Pheromone Research and Department of Chemistry, Indiana University, 800 East Kirkwood Avenue, Bloomington, Indiana 47405

Elisabeth Oberzaucher and Karl Grammer

Department for Anthropology, Ludwig Boltzmann Institute for Urban Ethology, Althanstrasse 14, A-1090 Vienna, Austria

Dustin J. Penn

Konrad Lorenz Institute for Ethology, Austrian Academy of Sciences, Savoyenstr. 1a, A-1160 Vienna, Austria

The majority of works in metabolomics employ approaches based on principal components analysis (PCA) and partial least-squares, primarily to determine whether samples fall within large groups. However, analytical chemists rarely tackle the problem of individual fingerprinting, and in order to do this effectively, it is necessary to study a large number of small groups rather than a small number of large groups and different approaches are required, as described in this paper. Furthermore, many metabolomic studies on mammals and humans involve analyzing compounds (or peaks) that are present in only a certain portion of samples, and conventional approaches of PCA do not cope well with sparse matrices where there may be many 0s. There is, however, a large number of qualitative similarity measures available for this purpose that can be exploited via principal coordinates analysis (PCO). It can be shown that PCA scores are a specific case of PCO scores, using a quantitative similarity measure. A large-scale study of human sweat consisting of nearly 1000 gas chromatography/mass spectrometry analyses from the sweat of an isolated population of 200 individuals in Carinthia (Southern Austria) sampled once per fortnight over 10 weeks was employed in this study and grouped into families. The first step was to produce a peak table requiring peak detection, alignment, and integration. Peaks were reduced from 5080 to 373 that occurred in at least 1 individual over 4 out of 5 fortnights. Both qualitative (presence/absence) and quantitative (equivalent to PCA) similarity measures can be computed. PCO and the Kolomorogov-Smirnoff (KS) rank test are applied to these similarity matrices. It is shown that for this data set there is a reproducible individual fingerprint, which is best represented using the qualitative similarity measure as assessed both by the Hotelling t^2 statistic as applied to PCO scores and the probabilities associated with the KS rank test.

Knowledge discovery from large volumes of data is a prevalent theme in modern scientific research. In modern metabolomics,[1−4] it is common to produce large quantities of data, such as chromatograms of several samples consisting of a large number of compounds. The quantities of data are often too large for manual interpretation, an example being 1000 gas chromatography/mass spectrometry (GC/MS) chromatograms with up to 500 peaks or half a million peaks: to manually inspect, characterize, and quantify each peak at 5 min work per peak would take over 18 years working at 50 h a week and 45 weeks a year. Hence, data mining is required to handle these large data sets. Over the past few years, substantial public domain software has been

* To whom correspondence should be addressed. E-mail: r.g.brereton@bris.ac.uk.

(1) Shellie, R. A.; Welthagen, W.; Zrostliková, J.; Spranger, J.; Ristow, M.; Fiehn, O.; Zimmermann, R. *J. Chromatogr., A* **2005**, *1086*, 83−90.
(2) Schauer, N.; Steinhauser, D.; Strelkov, S.; Schomburg, D.; Allison, G.; Moritz, T.; Lundgren, K.; Tunali, U. R.; Forbes, M. G.; Willmitzer, L. A.; Fernie, R.; Kopka, J. *FEBS Lett.* **2005**, *579*, 1332−1337.
(3) Hope, J. L.; Prazen, B. J.; Nilsson, E. J.; Lidstrom, M. E.; Synovec, R. E. *Talanta* **2005**, *65*, 380−388.
(4) Gullberg, J.; Jonsson, P.; Nordström, A.; Sjöström, M.; Moritz, T., *Anal. Biochem.* **2004**, *331*, 283−295.

10.1021/ac070134w CCC: $37.00
Published on Web 06/30/2007

developed to simplify these data sets, primarily based on principal components analysis (PCA)[5−7] and partial least-squares (PLS).[8,9] These approaches, while well established, have limitations under certain circumstances. A particular application involves studying methods where group sizes are small and so cannot easily be modeled using predictive classifiers such as PLS-DA.[10,11] However, if there are several small groups, an alternative is to see if the same pattern repeats itself over and again; whereas the predictive model from a single small group may not in itself be conclusive, if similar trends are seen many times, then this is good evidence of a trend, but different approaches are required to study repeated patterns. A second issue, common to many mammal and human metabolomic studies is that very few peaks are found in all or most of the samples;[12] this has advantages in that there are often characteristic marker compounds for individuals or small groups of individuals (which may be due to polymorphic genes that occur only in a fraction of a population) but which are absent in the majority of samples: biologists often compare organisms by looking at which features they have in common, and by analogy, in metabolomic studies we can compare samples according to which detectable peaks are in common, which overcomes a limitation of PCA that does not usually function effectively on sparse matrices. There are a large number of possible qualitative similarity measures[13] hence opening the opportunity for many approaches for the comparison of samples.

In biology, there are many methods based on determining similarity between objects, such as hierarchical cluster analysis.[14] The first step is to produce a pairwise distance matrix between the objects (corresponding in our case to GC/MS information). The aim of these approaches is to look at how features in a set of objects compare, and they can be employed in a wide variety of situations, for example, facial recognition, looking at dimensions of organisms, or looking at the presence or absence of features. In metabolomics, we can also look at similarities between samples based on characteristics of their chromatograms. Whereas it is common to present this information in the form of dendrograms, it is relatively unusual to use this similarity information for exploratory data analysis. The first step is to compute a dissimilarity matrix between objects. The next is to construct a new set of points in a low-dimensional space, the Euclidean distance between these objects corresponding to the distance between the points in the original space formed by the dissimilarity matrix. This technique is often called multidimensional scaling (MDS).[15−20] MDS can be divided into two types of methods depending on whether the mapping from the distance matrix to Euclidean distance space is linear or nonlinear. Linear mapping is often denoted principal coordinates analysis (PCO) while nonlinear mapping is often denoted metric/nonmetric MDS. It can be shown that PCO and PCA give the same solution when the Euclidean distance is used to construct the dissimilarity matrix to be approximated in the first step.[20] An important advantage of PCO over PCA is that there are many distance metrics other than the Euclidean distance can be used to represent the pattern of the data: hence, it can give different views of the pattern of the data set according to the criteria chosen for the dissimilarity measure. For example, the Euclidean distance is often not best suited to model categorical data, especially when there are multiple zeroes in the matrix and some special distance metrics are required (e.g., Hamming distance, Jaccard distance, etc).[22−25] In this paper, we apply PCO to peak table data, obtained as described elsewhere.[12] Based on the peak table, two distance matrices are constructed: one using the Euclidean distance to measure quantitative difference between samples and another using the Jaccard distance[21] to measure the qualitative difference between the samples, relating to whether compounds are detected in a sample or not. The higher the proportion of compounds detected in two samples the more similar the samples. PCO is applied on these two distance matrices to visualize the pattern of the data points. In addition to visualization, PCO can also be used as a form of data reduction and statistics as to the tightness and similarity of clusters such as Hotelling's t^2 test can be applied[26] to the PCO scores.

Similarity measures have an additional advantage in that it is possible to rank how similar two samples are, the lower the dissimilarity, the more similar the samples. If there are repeat samples from a single individual, it is possible to ask whether repeats from the same individual are more similar than samples between different individuals and, hence, whether there is a statistically significant individual biometric fingerprint as determined by GC/MS. The similarity between samples can be computed and separated into two groups, the first involving the ranks of repeats from within the same individual and another is the ranks of samples between different individuals. The distributions of these two rank lists can be compared by using statistical tests. Both the χ^2 and the Kolmorogov−Smirnov (K−S) tests[27−29] can be used to compare the observed with theoretical distributions as well as two experimental distributions measured from two

(5) Brereton, R. G. *Analyst* **2000**, *125*, 2125−2154.
(6) Wold, S; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab.* **1987**, *2*, 37−52.
(7) Brereton, R. G. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*; Wiley: Chichester, 2003.
(8) Wold, S.; Ruhe, A.; Wold H.; Dunn W. *J. SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735−743.
(9) Geladi, P.; Kowalski B. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(10) Brereton, R. G. *Trends Anal. Chem.* **2006**, *25*, 1103−1111.
(11) Dixon, S. J.; Xu, Y.; Brereton, R.G.; Soini, H. A.; Novotny M. V.; Oberzaucher, E.; Grammer, K.; Penn, D. J. *Chemom. Intell. Lab. Syst.* **2007**, *87*, 161−172.
(12) Dixon S. J.; Brereton, R. G.; Soini, H. A.; Novotny, M. V.; Penn, D. J. *J. Chemom.* **2006**, *20*, 325−340.
(13) Gower, J. G.; Legendre, P. *J. Classi*. **1986**, *3*, 5−48.
(14) Kleivi, K. M.; Teixeira, R.; Eknæs, M.; Diep, C. B.; Jakobsen, K. S.; Hamelin, R.; Lothe, R. A. *Cancer Genet. Cytogenet.* **2004**, 155, 119−131.
(15) Borg, I.; Groenen, P., *Morden Multidimensional Scaling, Theory and Applications;* Springer-Verlag New York Inc.: New York, 1997.
(16) Izrailev, S.; Agrafiotis, D. K. *J. Mol. Graphics Modell.* **2004**, *22*, 275−284.
(17) Rassokhin, D. M.; Agrafiotis, D. K. *J. Mol. Graphics Modell.* **2003**, *22*, 133−140.
(18) da Silva, J. C. G. E; Ferreira, M. A.; Machado, A. A. S. C.; Rey, F. *Anal. Chim. Acta* **1996**, *333*, 71−82.
(19) Jagadish, V.; Robertson, J. Gibbs, A. *Forensic Sci. Int.* **1996**, *79*, 113−121.
(20) Gower, J. C. *Biometrika* **1966**, 53, 325−338.
(21) Jaccard, P. *Bull. Soc. Vaud. Sci. Nat.* **1908**, *44*, 223−270.
(22) Pinheiro, H. P.; Pinheiro, A. de S.; Sen, P. K. *J. Stat. Plan. Infer.* **2005**, *130*, 325−339.
(23) He, M. X.; Petoukhov, S. V.; Ricci, P. E. B. *Math. Biol.* **2004**, *66*, 1405−1421.
(24) Martín, A. P. M. S.; Adamec, L.; Suda, J.; Mes, T. H. M.; Štorchová, H. *Aquat. Bot.* **2003**, *75*, 159−172.
(25) Jäggi, C.; Wirth, T.; Baur, B. *Biol. Conserv.* **2000**, *94*, 69−77.
(26) Johnson, R.; Wichern D. *Applied Multivariate Statistical Analysis*; Prentice Hall Inc.: Englewood Cliffs, NJ, 1982.
(27) Petrich, W.; Staib, A.; Otto, M.; Somorjai, R. L. *Vib. Spectrosc.* **2002**, *28*, 117−129.
(28) Vigneau, E.; Loisel, C.; Devaux, M. F. Cantoni, P. *Powder Technol.* **2000**, *107*, 243−250.
(29) Gutheil, W. G. *Biophys. Chem.* **1998**, *70*, 185−201.

independent samples. However, χ^2 tests make an assumption that the similarities between samples follow a normal distribution, while as a nonparametric test, K−S tests make no such assumption. In a two-sample K−S test, the null hypothesis is that two independent samples come from the same population. If there is insufficient evidence to reject the null hypothesis, the two-sample, number-based cumulative distribution functions are fairly close and their maximal difference would approach zero. However, if the null hypothesis does not hold, the maximal difference would significantly deviate from zero.

In this paper, we illustrate the application of the methods to sweat, which is a complex human excretion. In general, GC/MS is the analytical technique of choice employed to qualitatively and quantitatively determine chemical components in human sweat.[30−33] However, most studies involve fairly crude analyses, for example, finding specific marker compounds in high abundance and do not involve looking at the patterns among the large number of peaks present in low quantities. Sweat is a complex matrix, whose GC/MS appears more complex than common emanations such as urine and saliva, yet it is potentially rich in information. In this paper, a large-scale investigation on human sweat has been conducted by using GC/MS. An objective of this study is attempting to determine whether there are individual chemical fingerprints from human sweat samples. Nearly 200 subjects, from an isolated population in Carinthia, Southern Austria, all of whose family history was known, participated in the study: the sweat samples were collected 5 times from each subject over a 10-week period, once each fortnight. The sweat samples were then analyzed by GC/MS, and the information in the chromatograms was used for subsequent data analysis.

EXPERIMENTAL SECTION

Reagents and Materials. Standard compounds were purchased from Aldrich (Milwaukee, WI). Stir bars (Twister, 10 mm, 0.5-mm film thickness, 24-μL poly(dimethylsiloxane) volume) used for sample collection were purchased from Gerstel GmbH (Mülheim an der Ruhr, Germany). Stir bars were conditioned prior to and between each used in the TC 2 tube conditioner (Gerstel GmbH) at 300 °C under helium flow. Volatile and semivolatile compounds were collected from skin using methodology described elsewhere.[34] 7-Tridecanone was added to the stir bars as an internal standard.

Instrumentation. GC equipment for quantitative analysis consisted of an Agilent 6890N gas chromatograph connected to a 5973i MSD mass spectrometer (Agilent Technologies) with a thermal desorption autosampler (TDSA, Gerstel). Positive electron ionization mode at 70 eV was used with a scanning rate of 4.51 scans/s over the mass range of 35−350 amu. The ion source and quadrupole temperatures were set at 230 and 150 °C, respectively. The separation capillary was DB-5MS (20 m, 0.18-mm i.d., 0.18-μm film thickness) from Agilent. Samples were thermally desorbed in a TDSA automated system, followed by injection into the column with a cooled injection system, CIS-4. The TDSA operated in a splitless mode. The temperature program for desorption was 20 °C (0.5 min) and then 60 °C/min to 250 °C (3 min). The temperature of the transfer line was set at 280 °C. The CIS was cooled with liquid nitrogen to −80 °C. After desorption and cryotrapping, the CIS was heated at 12 °C/s to 280 °C with the hold time of 10 min. The CIS inlet was operated in the solvent vent mode, a vent pressure of 14 psi, a vent flow of 50 mL/min, and a purge flow of 50 mL/min. The temperature program in the GC operation was 50 °C for 1 min and then increasing to 160 °C at the rate of 5 °C/min, followed by the second ramp at the rate of 3 °C/min to 200 °C (hold time 10 min). The carrier gas head pressure was 14 psi (flow rate, 0.7 mL/min at constant flow mode). The GC temperature program lasted for 52.33 min, with mass spectrometric detection commencing after a deadtime of 1.88 min. To increase throughput, two instruments of identical specifications were used to analyze the samples. The configuration of both instruments was the same, and tests had been done to ensure reference samples analyzed on each instrument were of acceptable similarity. The long-term RSD of the internal standard was 14.3%.[34]

Data Set. The data set studied consisted of human sweat samples primarily of five repeats from individuals, sampled once a fortnight for five fortnights between June and August 2005 obtained from an isolated population in Carinthia, Southern Austria, whose family histories were all known. A total of 197 individuals took part in the survey, and each individual was grouped into a family. However, a few individuals were not able to participate each fortnight and so this reduced the data set size from a possible 985 samples down to 965. For one individual, only one sample was recorded. Further details have been reported elsewhere.[12] The repeats from one individual were distributed as evenly as possible between two analytical instruments; in most cases three samples were analyzed on instrument 1, and the remaining two on instrument 2. For the purpose of this paper, the split of samples between instruments is not particularly important. A typical GC/MS of sweat from individual A2 is presented in Figure 1. More details of the compounds detected are presented elsewhere.[35]

Software. The GC/MS instrument was controlled using Chemstation software. GC/MS data were exported to AIA/netCDF (network Common Data Format) format and then imported into MATLAB (The Mathworks, Inc., Natick, MA) using a freely available conversion tool.[36] All data processing was performed using MATLAB version 7.0.4.365, Release 14, Service Pack 2.

3. METHODS

Construction of the Peak Table. The first step is to construct a peak table, which is a matrix whose rows correspond to samples and whose columns correspond to peak areas of compounds using the summed intensity of the mass fragments corresponding to that peak. The method for peak identification and peak alignment has been described in detail elsewhere[12] and is not repeated

(30) Follador, M. J. D.; Yonamine, M.; Moreau, R. L.d. M.; Silva, O. A. *J. Chromatogr., B* **2004**, *811*, 37−40.
(31) Kidwell, D. A.; Kidwell, J. D.; Kidwell, F.; Harper, C.; Harper, K.; Bernadt, K.; McCaulley, R. A.; Smith, F. P. *Forensic Sci. Int.* **2003**, *133*, 63−78.
(32) Steinmeyer, S.; Ohr, H.; Maurer, H. J.; Moeller, M. R. *Forensic Sci. Int.* **2001**, *121*, 33−36.
(33) Thieme, D.; Anielski, P.; Grosse, J.; Sachs, H.; Mueller, R. K. *Anal. Chim. Acta* **2003**, *483*, 299−306.
(34) Soini, H. A.; Bruce, K. E.; Klouckova, I.; Brereton, R. G.; Penn, D. J.; Novotny, M. V. *Anal. Chem.* **2006**, *78*, 7161−7168.
(35) Penn, D. J.; Oberzaucher, E.; Grammer, K.; Fischer, G.; Soini, H. A.; Wiesler, D.; Novotny, M. V.; Dixon, S. J.; Xu, Y; Brereton, R. G. *J. R. Soc. Interface* **2007**, *4*, 331−340.
(36) The toolbox is available at http://mexcdf.sourceforge.net/index.html.

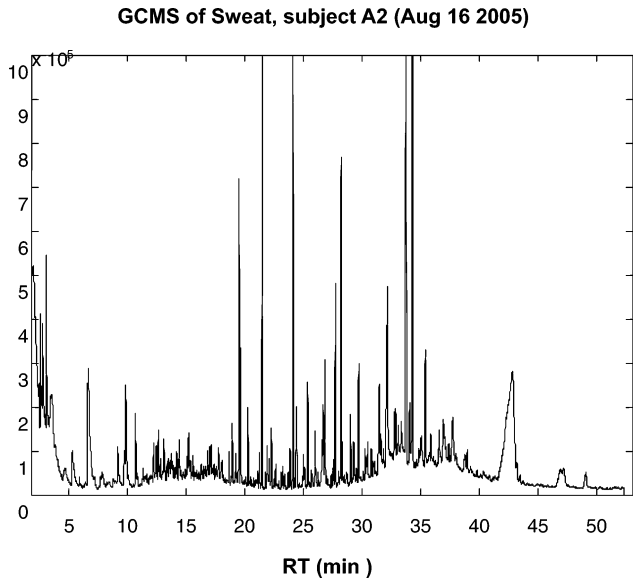

Figure 1. GC/MS of sweat sample of subject A2.

for brevity. Several peaks were present, which were known to be artifacts from the analytical process (such as siloxanes from the sample vial septum). These peaks had characteristic mass spectra and so could be removed from the peak table. Biologically interesting compounds are likely to be present over repeat samples from an individual, and so the final step is to retain only peaks that are present in at least one individual over the five repeats 4 out of 5 times (or 3 out of 4 from the few subjects that were sampled on only 4 fortnights). These steps reduced the peak table from 5080 compounds to 373 unique peaks, yielding a matrix **X** of dimensions 965 × 373.

Dissimilarity Measures. The next step involves determining a measure of dissimilarity between samples. It is important to recognize that the analytical procedure is semiquantitative. There are many reasons for this; the first is that it is not easy to obtain quantitative samples of sweat. Unlike urine, for example, it is not easy to introduce internal standards into sweat because the volume is hard to control. In addition, the amount a person sweats and the absolute amount of compounds may depend on a variety of biologically uninteresting factors, and as such, the analysis of sweat differs substantially from urine where there are often direct correlations between the concentrations of metabolites and, for example, a patient's condition. Although an internal standard is available, this is to determine the quality of the analytical method rather than to determine the ratio of concentrations in the sweat. However, sweat is very rich in compounds and contains substantial information about a person's biology and their environment; this information though is more likely to be reflected in the presence or absence of metabolites or in their relative, rather than absolute, concentrations.

In this paper, we compute two types of dissimilarity measure.

(1) Qualitative Distance Metric or Dissimilarity Measure. There are a large variety of methods available for constructing similarity measures for categorical or qualitative data.[13] The features in two samples are compared, and the number of matches is counted. In our application, we are primarily interested in whether a compound is detected or not, so the more compounds common to two samples, the more similar they are. Many of the criteria in the literature differ according to the significance attached to double negatives, that is to whether it is significant if a specific compound is absent in two samples. In our application, we do not attach significance to double negatives because the average number of peaks detected per sample is 67 out of the 373 peaks in the original data set; so for each pair of samples, the majority of peaks are not detected. We do, however, consider it significant if a specific compound is detected in both or one of a pair of samples.

In this paper, we use the square-rooted Jaccard distance.[21] Suppose for the *i*th and *j*th samples, there are α peaks common to both samples and β and γ peaks unique to the *i*th and *j*th sample, respectively; the dissimilarity measure is calculated as

$$^{1}D_{ij} = \sqrt{1 - \frac{\alpha}{\alpha + \beta + \gamma}} \qquad (1)$$

The dissimilarity measure, defined above, is based on presence/absence criteria as the information on the number of the common peaks between a pair of samples is used. It characterizes differences between samples due to the presence and absence of peaks detected in GC/MS.

(2) Quantitative Dissimilarity. As an alternative, the Euclidean distance between peak table area vectors is used to represent another type of dissimilarity measure that takes into account quantitative differences between peak areas. Each peak table vector is preprocessed as follows.

1. The first step involves taking the square root of the elements of **X.** We use square roots to reduce the influence of large peaks,[11,37,38] which is an alternative to log scaling,[39−41] which latter approaches poses problems if there are a significant number of peaks with zero intensity.

2. The next step involves normalizing each scaled chromatogram to sum to a constant total of 1.

3. The final step involves standardizing the elements in the matrix formed in step 2 to give a matrix with elements z_{ni}, where *n* is the sample number and *i*. The Euclidean distance is obtained as

$$^{2}D_{ij} = \sqrt{(\mathbf{z_i} - \mathbf{z_j}) \cdot (\mathbf{z_i} - \mathbf{z_j})'} \qquad (2)$$

It can be shown that the PCO scores using this distance metric are equivalent to those by PCA.

Principal Coordinates Analysis. The next step involves visualizing the dissimilarity matrix, using PCO. The following steps are performed.

1. The matrix $\mathbf{D}^{(2)}$ is calculated, by squaring the elements of the original dissimilarity matrix.

2. A matrix **A** is computed, whose elements are given by

(37) Kubicka, D.; Ronnholm, M.; Reinikainen, S. P.; Salmi, T.; Yu, D. *Anal. Chim. Acta* **2005**, *537*, 339−348.
(38) Osán, J.; de Hoog, J.; Worobiec, A.; Ro, C. U.; Oh, K. Y.; Szalóki, I.; Van Grieken, R. *Anal. Chim. Acta* **2001**, *446*, 209−220.
(39) Dixon, S. J.; Brereton, R. G.; Carter, J. F.; Sleeman, R. *Anal. Chim. Acta* **2006**, *559*, 54−63.
(40) Reichenbach, S. E.; Kottapalli, V.; Ni, M. *J. Chromatogr., A* **2005**, *1071*, 263−269.
(41) McCalley, D. V.; Brereton, R. G. *J. Chromatogr., A* **1998**, *828*, 407−420.

$$a_{in} = -1/2 d_{in}^{(2)} \quad (3)$$

3. Finally, a column and row centered matrix **G** is computed so that

$$g_{in} = a_{in} - \bar{a}_i - \bar{a}_n + \bar{a} \quad (4)$$

where $\bar{a}_i$ and $\bar{a}_n$ represent the row and column means of matrix **A** and $\bar{a}$ the overall mean.

4. Eigen decomposition is performed so that

$$\mathbf{G} = \mathbf{V}\cdot\Lambda\cdot\mathbf{V}' + \mathbf{E} \quad (5)$$

A scores matrix $\mathbf{T} = \mathbf{V}.\ \Lambda^{1/2}$ can be computed in analogy to PCA and the first l components are retained.

The visualization can be done by plotting one column of **T** against another.

It can be seen that, for the PCO algorithm, Λ needs to be nonnegative; i.e., the matrix **G** has to be positive semidefinite (psd). Otherwise, some of the PCs will exist in imaginary space. Under such circumstances, the original configuration of data points represented by **D** cannot be perfectly reconstructed by PCO. It has been shown by Gower and Legendre[13] that although the **G** derived from Jaccard distance matrix cannot be guaranteed to be psd, the **G** derived from the square-root Jaccard distance matrix is always psd. Therefore, we used the square-root Jaccard distance in this study. In addition, for the Euclidean distance, the matrix will always be psd.

Once a PCO scores matrix has been obtained, the multivariate separation between two groups of samples (each belonging to different individuals) can be numerically assessed using the two-sample t^2 statistic[42] via the following equation:

$$T^2 = \frac{n_a\cdot n_b\cdot(\overline{\mathbf{x}_a} - \overline{\mathbf{x}_b})\cdot\mathbf{S}^{-1}\cdot(\overline{\mathbf{x}_a} - \overline{\mathbf{x}_b})'}{n_a + n_b} \quad (6)$$

where n_a and n_b are the number of repeat samples from two individuals a and b, $\overline{\mathbf{x}_a}$ and $\overline{\mathbf{x}_b}$ are the mean measurements of these two individuals (in our case the scores from PCO) and **S** is the pooled variance−covariance matrix. Although the statistic could be converted into a probability, if required, we simply report t^2 values, the higher the greater the separation between two groups.

Kolmogorov−Smirnov Rank Test. Dissimilarity measures as described in Dissimilarity Measures can be employed for PCO but can also be used for rank-based approaches. These approaches look at which samples are most similar and rank all pairs of samples according to their mutual similarity. A hypothesis can then be set up that certain groups of samples are more similar to each other than other groups. For example, if there is a consistent individual signal, we would expect the similarities between repeat samples taken from individuals on average to be greater than the similarities between samples from different individuals. The K−S test can be performed to determine whether the within and between individual ranks are significantly different. In this data set, there are 196 within individual (one individual is repeated only one time and thus there is no within individual comparison) and 19 306 ($197 \times 196/2 = 19\,306$) between individual comparisons, so 19 502 comparisons in total. Note occasionally for reasons of sampling and analysis there will be cases where there are outlying samples, which means that a few of the repeats from the same individual will have a low similarity. In this paper, there is no attempt at outlier rejection.

(1) Ranking. The first step is to obtain an average dissimilarity $\bar{D}_{aa}$ for samples originating from the same individual. If n_a samples are collected from the same individual a (normally $n_a = 5$ but a few individuals were sampled less than 5 times)

$$\bar{D}_{aa} = \frac{\sum_{i=2}^{n_a}\sum_{j=1}^{i} D_{ij}}{\sum_{k=1}^{n_a - 1} k} \quad (7)$$

where D_{ij} represents the dissimilarity between samples i and j. If an individual is sampled only once, this statistic is not calculated.

Next we determine the similarity $\bar{D}_{ab}$ between samples collected from different individuals a and b, then

$$\bar{D}_{ab} = \frac{\sum_{i=1}^{n_a}\sum_{j=1}^{n_b} D_{ij}}{n_a n_b} \quad (8)$$

where D_{ij} represents the dissimilarity between sample i of individual a and sample j of individual b, where individuals a and b are sampled n_a and n_b times, respectively.

If there are P individuals in a population, there are $P \times (P - 1)/2$ average dissimilarities between samples obtained from different individuals and Q dissimilarities between the samples originating from the same individual, where $Q \leq P$, (if all individuals are sampled more than once Q will equal P). Thus, $P \times (P - 1)/2 + Q$ ($=\vartheta$) dissimilarities can be obtained overall, corresponding to $197 \times 196/2 = 19\,306$ possible pairwise between individual and 196 within individual similarity measures in our population. The dissimilarities can be then ranked from $\theta = 1$ (the least dissimilar) to ϑ (most dissimilar).

(2) Kolmogorov−Smirnov Test. After ranking the overall similarities between samples as above, two lists can be constructed. A rank list denoted by AA consists of the average dissimilarities between repeat samples for the same individual while the information on the average dissimilarity between samples originating from different individuals is contained in another rank list AB. If there is strong evidence for an individual fingerprint, we would expect the values of the ranks in the "AA" list to be significantly lower than the "AB" list; if not, we expect both lists to arise from the same distribution. The proportion of dissimilarity measures exceeding any rank can then be calculated for each distribution. For the entire population of dissimilarity measures, the proportion must be linearly related to rank, but if a population is split into two, then if one subpopulation is more similar than the other, the values of the ranks will be lower on average.

(42) Conover, W. J., *Practical Nonparametric Statistics,* 3rd ed.; John Wiley & Sons, Inc.: New York, 1999.

If $G_{AA}(\theta)$ and $G_{AB}(\theta)$ are the two-sample, number-based cumulative distribution functions from the rank lists AA and AB, respectively (the proportion of the rank values less than or equal to θ, where θ varies from 0 to ϑ), the two-sample K−S statistic (S_{K-S}) is as follows.

$$S_{K-S} = \max(|G_{AA}(\theta) - G_{AB}(\theta)|) \quad (9)$$

In this paper, the hypothesis is that samples from the same individual should be more similar to each other than samples from different individuals whereas the null hypothesis is that both two lists arise from the same underlying distribution. If the null hypothesis cannot be rejected, $G_{AA}(\theta)$ and $G_{AB}(\theta)$ would be fairly close for all values of θ and S_{K-S} is almost zero. Thus, the two-sample K−S statistic is used to test the hypothesis that the ranks in the AA list are significantly lower than those in AB in this study. The p value for the one-sided test can be approximated by using the following equations:

$$\lambda = \left(\sqrt{K} + 0.12 + \frac{0.11}{\sqrt{K}}\right)S_{K-S} \quad \text{and} \quad p = e^{-2\lambda^2} \quad (10)$$

where $K = K_1K_2/(K_1 + K_2)$ and K_1 and K_2 are the number of objects in each group.[42]

The data may also be presented graphically, where the value of θ is along the horizontal axis, and the vertical axis represents the proportion exceeded by a certain value. The principles of this type of graph are illustrated in Figure 2. In Figure 2a, the entire population is plotted, whereas in Figure 2b and c, the population is split into two portions. Because the size of the AA rank list is far smaller than the AB rank list, we would expect to see the pattern of Figure 2c if our hypothesis is correct.

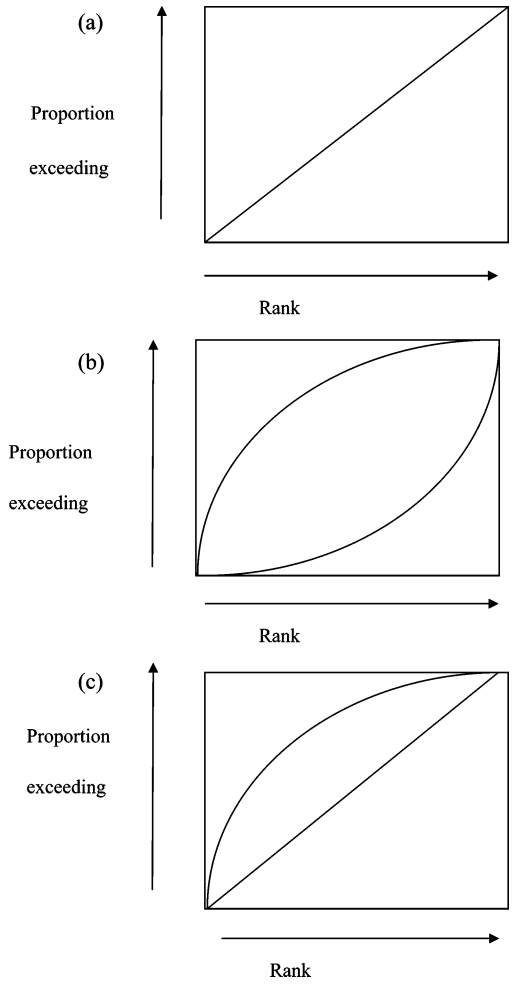

Figure 2. Illustration of rank charts. (a) A graph of proportion of dissimilarity measures exceeding a given rank against rank for an entire population. (b) Separating the population of dissimilarity measures into two groups of equal size, one that is more similar to each other, and the other that is less similar. (c) Separating the population of dissimilarity measures into two groups of very unequal size.

RESULTS AND DISCUSSION

Principal Coordinates Analysis. In this paper, two different kinds of dissimilarity data matrices denoted by $^1\mathbf{D}$ (square-rooted Jaccard distance, relating to presence and absence of detectable peaks) and $^2\mathbf{D}$ (Euclidean distance, relating to quantitative similarity) are used to construct matrices for PCO.

In order to simplify the PCO plots, individuals were grouped into families, as it is hard to see the main trends with 965 objects in a graph. The rank tests below are used to look at all the individuals together. Two families are chosen to illustrate the methods, family A consisting of 9 individuals (44 samples) and family B of 10 individuals (50 samples). The family trees are presented in Figure 3. Individuals were numbered with the first letter representing their family and the second a unique numerical individual identifier, so that individual A1, for example, was individual 1, who was from family A. PCO and similarity matrices were computed separately on each family. Figure 4 shows the PCO plots using both distance measures for both families. As discussed in the introduction, when the Euclidean distance is used to construct the distance matrix, the result obtained by PCO is equivalent to the result of PCA on the original data matrix. Therefore, the scores plots of the quantitative distance ($^2\mathbf{D}$) measure are actually the same as PCA scores plots, and PCA can be considered as a special case of PCO.

Using the qualitative distance measure ($^1\mathbf{D}$) we can see that repeats for most individuals appear to be clustered quite closely

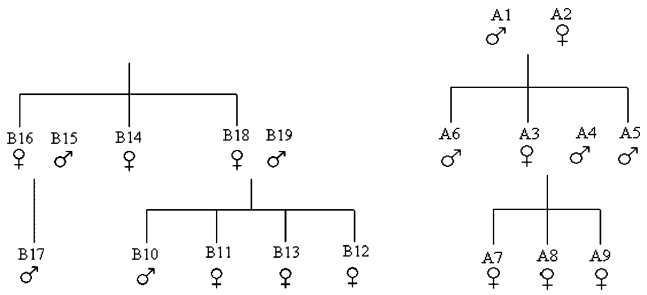

Figure 3. Family trees of families A and B: only people sampled in the study are indicated, All subjects have 5 repeat GC/MSs except A4, who was sampled only 4 times.

to each other. Some individuals stand out as being very characteristic; for example. individuals A1, A2, and A9 are on the right-hand side of Figure 4a, and B11 is well characterized on the right-hand side of Figure 4c. This suggests that most individuals can be well characterized by qualitative (presence/absence) fingerprints and that the repeats in the majority of cases for any one individual cluster closely together. There is inevitably some overlap between individuals, but this is because the entire

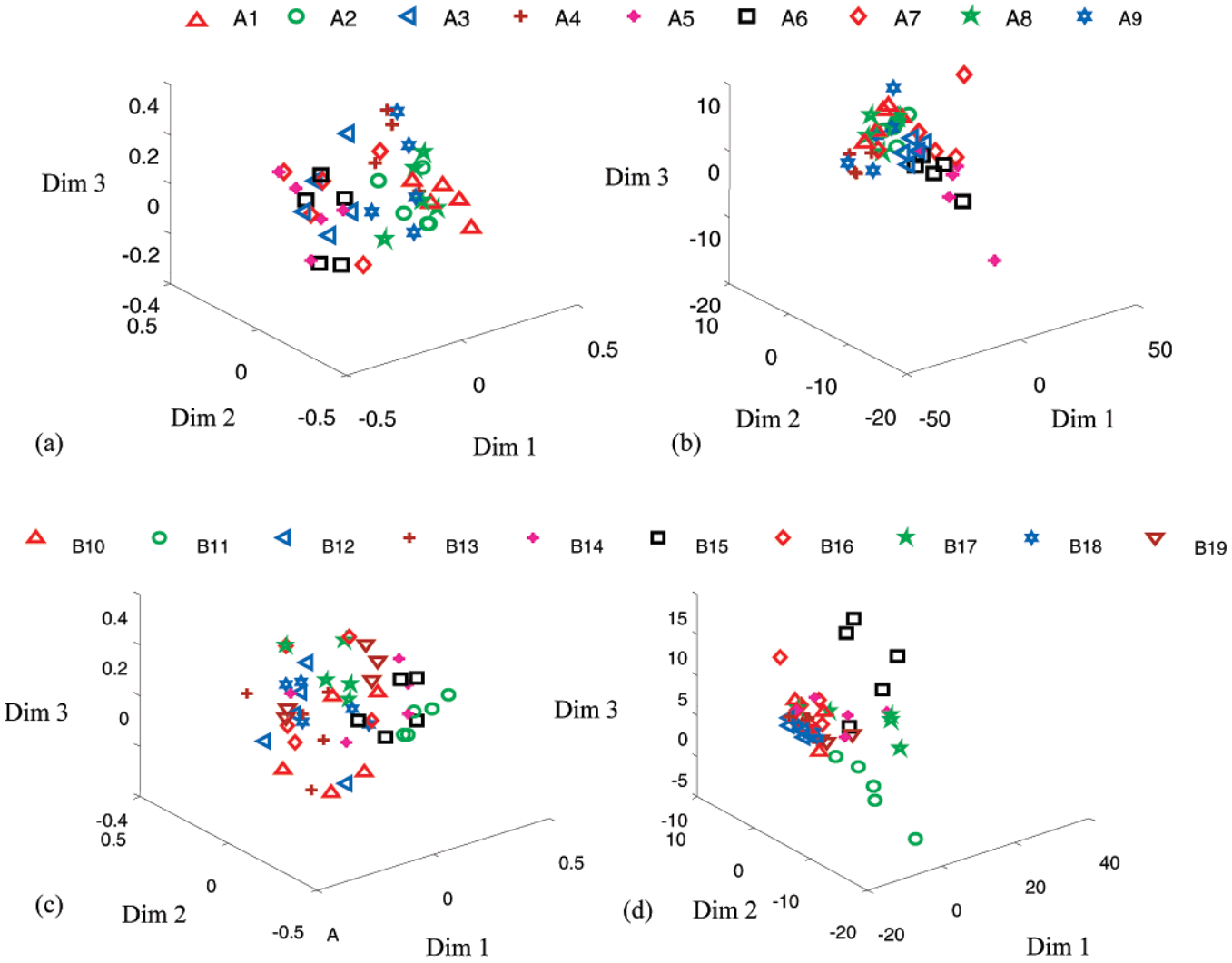

Figure 4. Principal coordinates scores plots for family A using (a) ^{1}**D** (b) ^{2}**D** and family B using (c) ^{1}**D** (d) ^{2}**D**. Note that ^{2}*D* plots are the same as PCA score plots.

chromatographic profile is used, and there will be other factors such as environment, personal habits, and even emotional state that could influence the signatures in sweat. However, the reproducibility does suggest that there is strong evidence for an individual signature, whether due to genetics or to specific repeatable personal habits is uncertain. The similarity (or PCA) plots based on the quantitative measure (^{2}**D**) are presented in Figure 4b and Figure 4d. Using these criteria, there are some very characteristic individuals; for example, B11 and B15 can be easily distinguished from the others in family B; in family A, A7 shows quite a characteristic outlier on the top right of the graph. However, the quantitative similarity appears to be more useful for distinguishing a few very distinctive people very well, rather than the majority of people. Possibly this is because some people sweat more easily and may have very characteristic compounds in high relative concentrations in sweat. It is important to recognize that a fully quantitative analysis of sweat is very hard because of the sampling problems.

Such visual observations are confirmed by using Hotelling's t^2 statistic based on the first three PCs. The results are shown in Table 1 for both similarity measures. The statistic T^2 is calculated between each pair of individuals. For family A, T^2 is higher for 26 of the comparisons using the qualitative measure, and lower on 10 occasions, using the quantitative measure (PCA); of these 10 comparisons, 8 involve individual A7, who has a very characteristic quantitative signal. For family B, the difference is not so dramatic, the qualitative measure being better in the case of 26 as opposed to 19 comparisons where PCA is preferred; however, 17 of the cases where PCA perform better involve individuals B11 and B15, which are clear outliers in the scores plot.

Rank Analysis Based on the Dissimilarity for 197 Individuals. PCO plots for the entire population would be crowded with 965 symbols and of very little value. However, the closeness of the individual signatures in the PCO plots can be quantified and the population as a whole can be visualized using the methods of the K−S Rank Test. The results from the qualitative distance (square-rooted Jaccard distance) and the quantitative distance (Euclidean distance—equivalent to PCA) are compared.

The cumulative distribution for the entire population is presented in Figure 5, showing a very distinct difference between the two sets of comparisons. Note that the AB rank list is approximately linear because only a small portion of the measurements are from the AA rank list. The curves from the qualitative distance are quite distinct and clearly different and have a K−S

Table 1. Hotelling's Multivariate *t*-Statistic Ion Families A and B Using the First Three Components

(a) Qualitative Metric (Jaccard Distance)

family A	A1	A2	A3	A4	A5	A6	A7	A8	A9
A1									
A2	19.8								
A3	70.5	40.1							
A4	35.9	41	56.2						
A5	49	47.9	14	30.4					
A6	53.4	46	14.7	36.1	21.2				
A7	40.9	22	31.9	19.6	34.6	22.5			
A8	13.7	15.3	60.4	20.6	69.3	65.5	21.3		
A9	20.6	14.7	41.9	26.4	90.4	61.9	37.3	22.1	

family B	B10	B11	B12	B13	B14	B15	B16	B17	B18
B10									
B11	39.9								
B12	14.4	62.8							
B13	13.9	61	14						
B14	16	44.5	36.1	6.7					
B15	28.9	37.2	45.2	43	34.7				
B16	49	41.1	41.6	30.8	23.5	22.8			
B17	24.7	19.2	24	39.6	35.7	42.2	24.7		
B18	14.2	48	32.48	13.9	18.6	32.5	17.2	53.9	
B19	22.5	37.8	22.6	24.8	35	46	19	24.1	18.5

(b) Quantitative Metric (Euclidean Distance Equivalent to Principal Components Analysis)

family A	A1	A2	A3	A4	A5	A6	A7	A8	A9
A1									
A2	11.8								
A3	10.2	7.7							
A4	12.4	21.6	12.3						
A5	40.7	30.3	8.9	31.8					
A6	33	25.8	6.8	34.3	22.9				
A7	50.9	32.4	33.8	42.2	46.5	30.1			
A8	3.5	3.1	12.2	20.3	47.9	37.8	43.7		
A9	5	6	15.1	13.9	43	32.9	63.1	4.7	

family B	B10	B11	B12	B13	B14	B15	B16	B17	B18
B10									
B11	118.5								
B12	10.1	76.4							
B13	3.3	67.4	1.8						
B14	6.1	72.2	7.7	10.4					
B15	117.2	131.7	116.5	99	86.9				
B16	6.3	87	12.4	10.1	18.6	126.9			
B17	7.4	81.6	23.1	40.1	25.5	49.2	14.3		
B18	10.4	79	13.8	12.9	17.2	149.2)	5	29.7	
B19	14	114.8	13.3	5.8	13.4	160.8	2	20.5	5.8

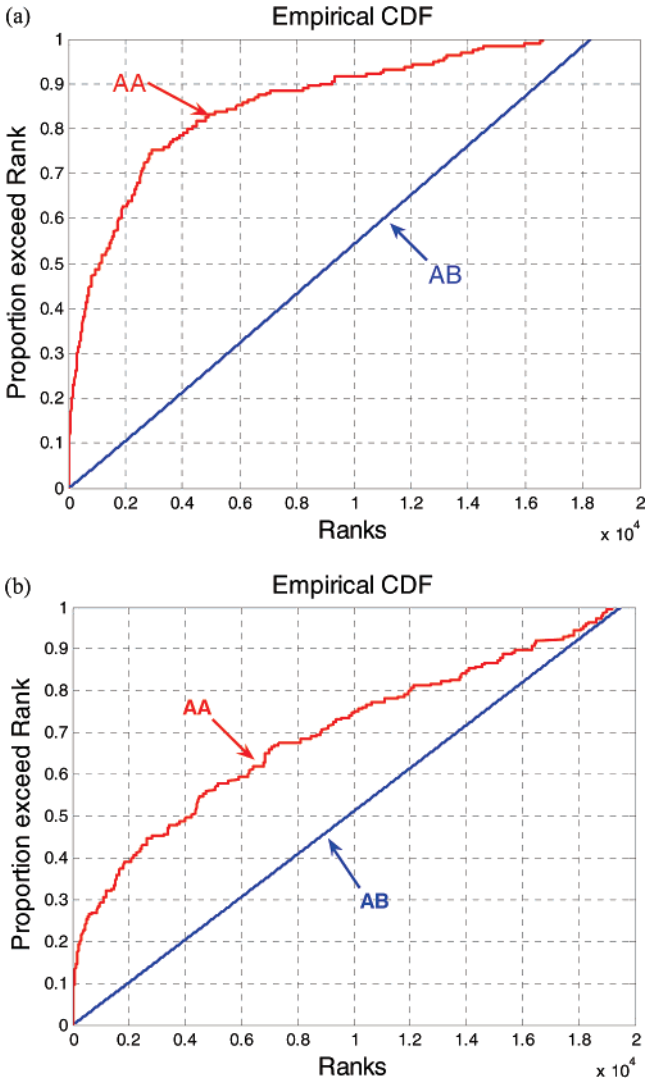

Figure 5. Rank plot for similarities for all subjects in the study. (a) The result obtained by using the qualitative distance measure; (b) the result obtained by using the quantitative distance measure.

statistic of 0.61 and a probability of 1.50×10^{-64}. The curves from the quantitative distance (PCA) are less distinct, but still show the same trend, and yield a K−S statistic of 0.29 and a probability of 2.18×10^{-15}. This means that, by using both qualitative and quantitative distance measures, there are always statistically striking differences between the AA and AB rank lists. The evidence in this paper demonstrates that it is possible to track individuals over a period of several months within a controlled population. The underlying reason for this cannot be determined for certain from this paper; it may be due to consistent personal habits, genetics, or where people live and work. It may change over time as people age. It would only be possible to study the long-term trend over several years, which would require careful attention to maintaining the same sampling regime and studying problems of long-term storage and instrumental stability. Model studies can be better performed on animals with shorter lifespans such as mice, which allow tracking an individual fingerprint over a lifetime within a more controlled environment.

In addition, it is also easier to characterize individuals according to the presence or absence of compounds rather than their relative amounts. This observation will not be universally true, but for the particular study where absolute quantitation is hard and certain individuals may have very characteristic metabolism for which results in certain specific compounds being produced, the qualitative measure is superior. A major advantage of PCO is that several different similarity measures can be investigated and compared.

It is interesting that not all compounds that appear characteristic of specific individuals are universally present in every sample from that individual. This does not necessarily militate against a unique genetic origin. For example, we may wish to distinguish males from females. The presence of certain compounds characteristic of females may be a function of their menstrual cycle, so not universally found in every female but still will be a characteristic genetically derived marker. Other examples can be the ability to metabolize certain types of food, which may relate to

the presence of specific enzymes, which has a genetic basis: specific breakdown products might be highly characteristic of individuals with certain enzymes, and so good individual marker compounds, but only detected if the individual has eaten a particular type of food recently. We expect the individual chemical signature to be multivariate in nature, even if in part or in whole of genetic origins, and so no one compound is likely to be uniquely and universally characteristic of a specific individual.

CONCLUSIONS

In most studies in human metabolomics, methods based on PCA or PLS are employed. In this paper, we present a complementary method that allows the use of dissimilarities between samples. This has an advantage in that there are a wide variety of potential similarity measures, especially those that can take into account presence and absence of signals, which is quite common in many biological situations. Whereas PCA can cope with missing data or undetected signals, there can be problems if matrices are sparse and if the differences between signals are based primarily on the presence or absence of features rather than quantitative criteria such as ratios. Sweat is a matrix that is very rich in information, some probably related to genetics, others to the environment, and others to personal habits, and so has the potential to provide a highly characteristic fingerprint, but is inherently difficult to study first because of the problem of sampling, especially quantitatively, and second because it is influenced by so many factors.

This paper presents the application of similarity-based approaches (PCO and the Kolmorogov−Smirnov rank test) to the GC/MS of sweat of a substantial human population and demonstrates that there appears to be a characteristic individual signature over the time period studied. To determine the origin of this fingerprint (whether due to genetics or personal habits) and whether it is stable over an extensive period of time or whether it varies with age (such as, for example, a person's appearance) would require a much larger study. Furthermore, such individual signature appeared much clearer when the qualitative distance measure was used, suggesting that using such dissimilarity measure can provide complementary information to more commonly used quantitative methods such as PCA.

The methods in this paper can be extended to any study of similarities, for example, taking samples from patients with and without disease, and from people of different groups, and provide an important novel method for visualizing metabolomic analytical chemical data. They also allow individual fingerprinting from the chemosignal.

The approaches in this paper are not in common use in metabolomics, yet one of the hardest problems is undoubtedly studying individual repeatability and determining what elements of an individual signature are constant. In order to do this, it is hard to obtain a large number of genuine repeats for each individual, and new approaches are required to tackle this important and rapidly developing application area.

ACKNOWLEDGMENT

We thank Hejun Duan and Thomas Parfrey of the Centre for Chemometrics for help in data organisation. Alexandra Katzer is thanked for her excellent organisational skills. This work was sponsored by ARO Contract DAAD19-03-1-0215. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Received for review January 23, 2007. Accepted May 25, 2007.

AC070134W