Supervised Self Organizing Maps for Classification and Determination of Potentially Discriminatory Variables: Illustrated by Application to Nuclear Magnetic Resonance Metabolomic P...

ARTICLE *in* ANALYTICAL CHEMISTRY · JANUARY 2010

Impact Factor: 5.64 · DOI: 10.1021/ac9020566 · Source: PubMed

Anal. Chem. **2010,** *82,* 628–638

Supervised Self Organizing Maps for Classification and Determination of Potentially Discriminatory Variables: Illustrated by Application to Nuclear Magnetic Resonance Metabolomic Profiling

Kanet Wongravee,[†] Gavin R. Lloyd,[†] Christopher J. Silwood,[‡] Martin Grootveld,[§] and Richard G. Brereton*[,†]

Centre of Chemometrics, School of Chemistry, University of Bristol, Cantocks Close, Bristol, BS8 1TS, U.K., London South Bank University, Department of Applied Science, Faculty of Engineering Science and The Built Environment, London SE1 0AA, U.K., and Centre for Materials Research and Innovation, University of Bolton, Deane Road, Bolton BL3 5AB, U.K.

The article describes the extension of the self organizing maps discrimination index (SOMDI) for cases where there are more than two classes and more than one factor that may influence the group of samples by using supervised SOMs to determine which variables and how many are responsible for the different types of separation. The methods are illustrated by an application in the area of metabolic profiling, consisting of a nuclear magnetic resonance (NMR) data set of 96 samples of human saliva, which is characterized by three factors, namely, whether the sample has been treated or not, 16 donors, and 3 sampling days, differing for each donor. The sampling days can be considered a null factor as they should have no significant influence on the metabolic profile. Methods for supervised SOMs involve including a classifier for organizing the map, and we report a method for optimizing this by using an additional weight that determines the relative importance of the classifier relative to the overall experimental data set in order to avoid overfitting. Supervised SOMs can be obtained for each of the three factors, and we develop a multiclass SOM discrimination index (SOMDI) to determine which variables (or regions of the NMR spectra) are considered significant for each of the three potential factors. By dividing the data iteratively into training and test sets 100 times, we define variables as significant for a given factor if they have a positive SOMDI in the training set for the factor and class of interest over all iterations.

SOMs (self organizing maps)[1−3] were first reported by Kohonen 20 years ago and have been widely employed for visualization of relationships between samples. They are a powerful alternative to principal components analysis (PCA)[4−6] for several reasons. First, PCA models are linear when we may expect nonlinearities in the data; second, there are many facile ways of graphical display using SOMs, and third, PC models can be strongly influenced by outliers, common problems especially with complex biomedical data sets. However, currently, unlike traditional approaches, they are not used commonly in areas such as analytical chemistry or metabolomic profiling, probably because they are computationally intensive and there is limited packaged software available. However, with the rapid growth of desktop computing power, SOMs are now much more feasible for real world problems in analytical chemistry data analysis.[7−11]

Unsupervised SOMs as traditionally employed are used primarily for exploratory data analysis[12−14] to reveal relationships between samples in data. They give an advantage to visualize a large number of samples especially as is common in metabolomic studies in limited space. However, it is also possible to employ these in a supervised mode as well. In most chemometrics, there are both exploratory approaches (for example, to see whether there are groupings in samples without any preconceptions) and supervised approaches (used for modeling and prediction, e.g.,

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

[†] University of Bristol.

[‡] London South Bank University.

[§] University of Bolton.

(1) Kohonen, T. *Construction of Similarity Diagrams for Phonemes by a Self-Organising Algorithm*; Helsinki University of Technology: Espoo, Finland, 1981.

(2) Kohonen, T. *Biol. Cybern.* **1982**, *43*, 59–69.

(3) Kohonen, T. *Self-Organizing Maps*; Springer-Verlag: New York, 2001.

(4) Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.

(5) Jackson, J. E. *A User's Guide to Principal Components*; Wiley: New York, 1991.

(6) Brereton, R. G. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*; Wiley: Chichester, U.K., 2003.

(7) Suna, T.; Salminen, A.; Soininen, P.; Laatikainen, R.; Ingman, P.; Mäkelä, S.; Savolainen, M. J.; Hannuksela, M. L.; Jauhiainen, M.; Taskinen, M. R.; Kaski, K.; Ala-Korpela, M. *NMR Biomed.* **2007**, *20*, 658–672.

(8) Mäkinen, V. P.; Soininen, P.; Forsblom, C.; Parkkonen, M.; Ingman, P.; Kaski, K.; Groop, P. H.; Ala-Korpela, M. *Mol. Syst. Biol.* **2008**, *4*, 167.

(9) Tukiainen, T.; Tynkkynen, T.; Mäkinen, V. P.; Jylänki, P.; Kangas, A.; Hokkanen, J.; Vehtari, A.; Gröhn, O.; Hallikainen, M.; Soininen, H.; Kivipelto, M.; Groop, P. H.; Kaski, K.; Laatikainen, R.; Soininen, P.; Pirttilä, T.; Ala-Korpela, M. *Biochem. Biophys. Res. Commun.* **2008**, *375*, 356–361.

(10) Marini, F.; Bucci, R.; Magri, A. L.; Magri, A. D. *Microchem. J.* **2008**, *88*, 178–185.

(11) Marini, F.; Zupan, J.; Magri, A. L. *Anal. Chim. Acta* **2005**, *544*, 306–314.

(12) Lloyd, G. R.; Brereton, R. G.; Duncan, J. C. *Analyst* **2008**, *133*, 1046–1059.

(13) Xiao, Y. D.; Clauset, A.; Harris, R.; Bayram, E.; Santago, P., II; Schmitt, J. D. *J. Chem. Inf. Model.* **2005**, *45*, 1749–1758.

(14) Melssen, W. J.; Smits, J. R. M.; Rolf, G. H.; Kateman, G. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 195–204.

10.1021/ac9020566 © 2010 American Chemical Society
Published on Web 12/28/2009

of unknowns), supervised SOMs belonging to the second category which results in analogies to more traditional statistical methods of prediction allowing numerical values to be attached, e.g., to determine whether a sample is a member of a group or which variables are more important as diagnostic markers.

Supervised SOMs[15−18] have been proposed for classification purposes whereby an additional vector of class information is included in the training. This then introduces an additional factor that organizes the map. The degree to which the class information influences that map can be controlled, and in this article we introduce a class weight that can be adjusted according to how far the class membership information is used to train the map: a low value results in a map that is close to unsupervised, whereas a high value may overfit the data. This parameter is described in the section Optimum Scaling Value. According to the need of the analysis, the classifier can assume high or low importance: in approaches such as partial least squares-discriminant analysis (PLS-DA),[19−24] the classifier and the experimental (analytical) variables are assumed to have equal significance, whereas supervised SOMs allow this relative significance to be varied.

An important use of SOMs is in variable selection,[18,25] which is used to find which compounds or regions of a spectrum or chromatogram are potential markers for a group of samples. We have reported a SOMDI (SOM discrimination index) previously,[18] but only for two class unsupervised SOMs. In this article, we show that this SOMDI can be generalized and combined with supervised SOMs. In many practical situations, there are more than two groups in the data, and so we show an extension that copes with this situation. However, another important aspect, especially of designed experiments, and increasingly common in metabolomics studies, is that we may wish to study several factors that could influence the metabolic profiles in biological or clinical investigations. Using supervised SOMs, we can independently train each map for a different factor and therefore find markers for each of the factors independently. Without using supervised SOMs, we can only find markers for the most dominant factors that influence the appearance of an unsupervised map. We propose an approach in this paper, which also determines which of the variables are significant, according to how many times they are selected as being a potential marker on 100 iterative reformulations of the SOM map.

Whereas there are a number of traditional approaches for determining markers as described below, common especially in metabolomics, these have a number of drawbacks. Statistically based methods such as ANOVA and the F-statistic often assume normal distributions for the variables in order to yield significance values, but most metabolomic data sets fail traditional normality tests. Furthermore, these approaches do not take into account the interactions between the variables, and multivariate methods are usually preferred. PLS-DA[19−24] is often employed in chemometrics, and when the problem is fairly simple, e.g., involving only two groups, it is valuable, and PLS weights and regression coefficients can be used as indicators of significance.[26,27] However, this has a drawback that the classification information and the variable information are weighted as being of equal significance: such a weighting is often undesirable, and the implementation of supervised SOMs outlined in this article allows us to attach any relative significance we like to the classifier and the experimental data, for example, we may want to weight the classifier as completely unimportant so it will have no influence on the map, or very important in that the appearance of the map will primarily be influenced by class membership of samples. Furthermore, although PLS can indeed be extended to multiclass situations, implementations can be complicated because it is typically implemented as a series of one versus all binary decisions (one for each class), and there can be a problem in determining a consensus decision criterion.[24]

In this article we report the use of supervised SOMs for variable selection and illustrate its application to an NMR-based metabolomic data set, involving saliva samples analyzed before and after treatment with an oral rinse formulation, 16 donors, and 3 sampling days, so there are three different types of factors, each in turn containing a different number of groups. Codes for SOMs as referenced to in ref 24 will be made available on the www.spectroscopynow.com Web site.

EXPERIMENTAL SECTION

More detail is discussed elsewhere,[18] and hence only a summary is presented below.

^{1}H NMR Analysis of Human Saliva Samples. Saliva samples were collected from 16 healthy volunteers every day on 3 days within 5 min of waking in the morning. Prior to collection of the samples, the volunteers were asked to avoid any oral activities (eating, drinking, smoking, tooth brushing, etc.) and were asked to collect their saliva into a plastic tube containing sufficient sodium fluoride (15 μmol) to ensure that compounds were not generated or metabolized by micro-organisms during storage. The samples were stored on ice during transportation and immediately centrifuged at 3500 rpm for 15 min on arrival. The resulting supernatants were stored at −70 °C prior to analysis. All samples were divided into two equal parts, each 0.60 mL in volume. The first was treated with the oral rinse product (3.0 mL) and the second retained as an untreated control with HPLC-grade H_2O (3.0 mL) added. A total of 96 samples (16 volunteers × 3 sampling times × 2 groups) were collected in total. Note that the sampling days although sequential for each donor are not on the same days, and this factor can be considered a null or dummy factor as a control for our methods, although there

[15] Melssen, W.; Wehrens, R.; Buydens, L. *Chemom. Intell. Lab. Syst.* **2006**, *83*, 99–113.

[16] Xiao, Y. D.; Clauset, A.; Harris, R.; Bayram, E.; Santago, P.; Schmitt, J. D. *J. Chem. Inf. Model* **2005**, *45*, 1749–1758.

[17] Melssen, W.; Ustun, B.; Buydens, L. *Chemom. Intell. Lab. Syst.* **2007**, *86*, 102–120.

[18] Lloyd, G. R.; Wongravee, K.; Silwood, C. J. L.; Grootveld, M.; Brereton, R. G. *Chemom. Intell. Lab. Syst.* **2009**, *98*, 149–161.

[19] Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1–17.

[20] Baker, M.; Rayens, W. *J. Chemom.* **2003**, *17*, 166–173.

[21] Ståhle, L.; Wold, S. *J. Chemom.* **1987**, *1*, 185–196.

[22] 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.

[23] Martens, H.; Naes, T. *Multivariate Calibration*; Wiley: Chichester, U.K., 1989.

[24] Brereton, R. G. *Chemometrics for Pattern Recognition*; Wiley: Chichester, U.K., 2009.

[25] Corona, F.; Reinikainen, S. P.; Aaljoki, K.; Perkkiö, A.; Liitiainen, E.; Baratti, R.; Lendasse, A.; Simula, O. *J. Chemom.* **2008**, *22*, 610–620.

[26] Sanchez, E.; Kowalski, B. R. *J. Chemom.* **1988**, *2*, 247–263.

[27] Cloarec, O.; Dumas, M. E.; Trygg, J.; Craig, A.; Barton, R. H.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *Anal. Chem.* **2005**, *77*, 517–526.

may be a very small variation due to the sequential nature of the days (day 1 always being earlier than day 2, etc.). The samples were equilibrated at 37 °C for 30 s. Samples were prepared for analysis by adding 0.05 mL of D_2O and 0.05 mL of a 5.0 × 10^{-3} mol dm^{-3} solution of sodium 3-trimethylsilyl [2,2,3,3-$^2H_4$] propionate in D_2O. Each sample was then analyzed by ^1H NMR (Bruker Avance AX-600 spectrometer operating at 600.13 MHz). Spectral regions 1.03−1.35, 1.88−1.94, 2.42−2.79, 3.35−3.38, and >7.92 ppm were removed since they contained signals from the added oral rinse. Furthermore, the region 4.62−4.94 ppm was removed due to the intense residual H_2O/HOD signal. Data was subjected to the "intelligent bucketing" procedure developed by ACD/Laboratories,[28] resulting in 146 buckets. In the final step, any bucket containing less than 1% of the maximum summed intensity from all 96 spectra was rejected since it may correspond to noise and is regarded as uninformative, resulting in 49 buckets. Hence a data set of dimensions 96 samples × 49 buckets is employed for the calculations in this article. A typical example of an original NMR spectrum and the corresponding bucketed NMR data after the removal of specific regions corresponding to compounds which are present in the oral rinse and the broad residual H_2O/HOD signal are shown in Figure S-1 in the Supporting Information.

Preprocessing. The next stage after bucketing the NMR spectra is to prepare the data for pattern recognition. In this article, we report results using only one approach for simplicity. First, the bucketed data are square-rooted in order to reduce the influence of large resonances. Next, centering is applied to the square rooted data for variable selection. In classification where data is split into training sets and test sets,[24] centering is performed on only the training set samples as appropriate; the test set is centered according to parameters (mean) obtained from the training set to ensure that the test set samples do not influence the model. More considerations about the optimal strategy are discussed elsewhere.[22] This provides us with a datamatrix **X** of dimensions $I \times J$ (or in this paper 96 × 49) that is used as input to the SOMs and all subsequent computations. Notations and definitions are presented in Tables 1 and 2, respectively.

Software. The intelligent bucketing procedure was applied to the NMR spectra acquired by using the ACD/Laboratories 1D NMR Manager Software package. The data set after bucketing was exported into Microsoft Excel. All data preprocessing, visualization, and analysis programs were written in-house using MATLAB version 7.0.4.365, release 14, service pack 2.

Table 1. Notations

symbol	description
$\mathbf{d}$	supervised sample vector (SSV)
$\mathbf{d}_r$	variable sample vector (VSV)
$\mathbf{d}_s$	class sample vector (CSV)
$\mathbf{d}_{train}$	VSV of training set samples
$\mathbf{d}_{test}$	VSV of test set samples
I	number of samples
J	number of variables
K	number of classes
M	number of row units on SOM map (= 15)
N	number of column units on SOM map (= 20)
Q	Euclidean distance between v_{BMU} and d_{test}
s_{in}	SOMDI score of "in-group" group
s_{out}	SOMDI score of "out-group" group
s_k	score index of class k
s_Δ	difference between s_{in} and s_{out} ($s_\Delta = s_{in} - s_{out}$)
u	unit on the SOMs map
U	total number of units on the SOM map
$\mathbf{v}_r$	variable weight vector (VWV)
$\mathbf{v}_s$	class weight vector (CWV)
$\mathbf{v}_{BMU}$	weight vector of best map unit (BMU)
$\mathbf{v}$	supervised weight vector (SWV)
$\mathbf{V}_s$	variable weight matrix (VWM)
$\mathbf{V}_{in}$	in-group variable weight matrix
$\mathbf{V}_{out}$	out-group variable weight matrix
w	weight values
$\mathbf{X}$	data matrix
ω	scaling value

Table 2. Definitions

abbreviation	full name	dimension
BMU	best map unit	$1 \times (J + K)$
CSM	class sample matrix	$I \times K$
CSV	class sample vector	$1 \times K$
CWM	class weight matrix	$U \times K$
CWV	class weight vector	$1 \times K$
SSM	supervised sample matrix	$I \times (J + K)$
SSV	supervised sample vector	$1 \times (J + K)$
SWM	supervised weight matrix	$U \times (J + K)$
SWV	supervised weight vector	$1 \times (J + K)$
VSM	variable sample matrix	$I \times J$
VSV	variable sample vector	$1 \times J$
VWM	variable weight matrix	$U \times J$
VWV	variable weight vector	$1 \times J$

METHODS

There are several supervised methods available in the literature such as linear discriminant analysis (LDA),[6,24,29,30] partial least squares-discriminant analysis,[19−24] and support vector machines (SVM).[24,31] These methods are used to classify unknown samples to a group. PLS-DA is often used for other applications such as selecting the significant variables using PLS weights[24,32] and PLS regression coefficients.[24,27,32,33] All these classical methods involve performing a single reproducible calculation. More intense approaches such as those based on neural networks are less commonly applied in metabolomic profiling, possibly because they are computationally intense and there is less readily available packaged software. However the SOM based methods reported here provide valuable alternatives often with greater power for visualization, particularly in the contexts of classification and variable selection.

Self Organizing Maps. Self-organizing maps (SOMs) are an unsupervised learning method using artificial neural networks[12,34]

[28] Lefebvre, B. Intelligent Bucketing for Metabonomics, ACD/Labs Technical Note, 2004, http://www.acdlabs.com/publish/publ04/enc04_intelli_bucket.html (accessed August 13, 2009).

[29] Fisher, R. A. *Ann. Eugenics* **1936**, *7*, 179–188.

[30] Brereton, R. G. *Applied Chemometrics for Scientists*; Wiley: Chichester, U.K., 2007.

[31] Xu, Y.; Zomer, S.; Brereton, R. G. *Crit. Rev. Anal. Chem.* **2006**, *36*, 177–188.

[32] Wongravee, K.; Lloyd, G. R.; Hall, J.; Holmboe, M. E.; Schaefer, M. L.; Reed, R. R.; Trevejo, J.; Brereton, R. G. *Metabolomics* DOI: 10.1007/s11306-009-0164-4.

[33] Wongravee, K.; Heinrich, N.; Holmboe, M.; Schaefer, M. L.; Reed, R. R.; Trevejo, J.; Brereton, R. G. *Anal. Chem.* **2009**, *81*, 5204–5217.

[34] Kohonen, T.; Kaski, S.; Lappalainen, H. *Neural Comput.* **1997**, *9*, 1321–1344.

to visualize different patterns in data and to determine the relationship between experimental measurements and samples. However, there are a small number of articles[15−18] that describe a supervised modification to SOMs for classification purposes, although this approach is not widespread. Training maps generated using unsupervised SOMs mainly represent the major factors that influence the similarities between samples. In some cases, minor variation is much more interesting (e.g., variation due to donors, instruments, sampling days, etc.) and as such are often not well represented using unsupervised SOMs. Supervised SOMs offer the opportunity to study each of the known factors influencing variation by increasing their importance on the organization of the maps. The degree to which each of these factors is allowed to influence the appearance of the map can be controlled, as described in this article both to improve the representation of the factor on the maps while preventing overfitting. In addition, maps can be employed for selecting which variables are the most significant and by using different factors to weight the map, one can determine which variables are most significant for each different source of variation (in this article, there are three sources, specifically treatment, sampling day, and donor).

In general, an initial SOM map represented by a two-dimensional map space containing $M \times N = U$ units is generated. The units are can be square or hexagonal, but in this article only hexagonal units are used as this is a commonly accepted method of presentation. The number of units can be chosen according to the complexity of the problem, more units being required if several complex trends or groups are to be represented. In this article, we set $M \times N = 15 \times 20$ units on the map (300 in total) which is approximately 3 times the number of samples in the data (96 samples) allowing samples to spread out over the map but ensuring that the map is not too sparse. For unsupervised SOMs, each unit is characterized by a weight vector $\mathbf{v}_r = [w_{u1}, w_{u2}, \ldots, w_{uJ}]$, which we call for the purpose of this article a "variable weight vector" (VWV) with dimensions $1 \times J$, where J equals the number of variables in the data and u is the map unit.[12] To initialize the map, w_{uj} is randomly generated by a uniform distribution between the maximum and minimum values of variable j in the data. In contrast, for supervised SOMs, the weight vectors have dimensions $1 \times (J + K)$, where K is the number of classes in the data. The weight vector of each unit for a supervised SOM is $\mathbf{v} = [\mathbf{v}_r\mathbf{v}_s]$ which we will call a "supervised weight vector" (SWV), where $\mathbf{v}_s = [w_{u1}, w_{u2}, \ldots w_{uK}]$ which we call a "class weight vector" (CWV) and includes information about class membership, and w_{uk} is a positive number; the higher it is, the more likely it is to be a member of class k. The maximum possible value of w_{uk} is chosen as described in the section Optimal Scaling Value (ω) and Performance of Classifier, and relates to how influential the class membership information is on the formation of the map. It is initialized as a uniform number between 0 and its maximum, in the same manner as the values are of $\mathbf{v}_r$.

The input vector of each sample to train the supervised SOM map contains two parts $\mathbf{d}_r$ which is the preprocessed "variable sample vector" (VSV) in the data ($1 \times J$) and $\mathbf{d}_s$ that is a vector containing class information called a "class sample vector" (CSV) of dimensions ($1 \times K$) as described below. Combination of the sample vector and the scaled class vector gives $\mathbf{d} = [\mathbf{d}_r\mathbf{d}_s]$ which we call a "supervised sample vector" (SSV) in this article. The dimensions of the CSV are dependent on the number of classes in the data ($= K$). For example, if the data contains three classes ($K = 3$), A, B, and C, then the CSV $\mathbf{d}_s = [\omega\ 0\ 0]$, $[0\ \omega\ 0]$, and $[0\ 0\ \omega]$ for samples that are members of classes A, B, and C, respectively, where ω is a scaling value. Therefore, the training procedure of the supervised SOM map is performed using a SSV $[\mathbf{d}_r\mathbf{d}_s]$ in contrast to unsupervised SOMs which use only the VSV, $\mathbf{d}_r$. The definitions of these vectors are presented in Figure S-2 in the Supporting Information.

When all samples or units in the SOM are collected together, the CSVs become matrices, so that, for example, the "class sample matrix" (CSM) has dimensions $I \times K$ containing ω, where a sample is a member of the specific class and 0 otherwise, and the "variable weight matrix" (VWM) has dimensions $U \times J$.

The SSV $\mathbf{d}$ is compared to each unit in the SOM map, and the unit whose SWV is most similar to it is assigned as the best matching unit (BMU). The SWV of the BMU and also neighboring units are updated to become similar to the input sample as discussed previously.[12] The initial neighborhood width σ_0 was chosen to be 7 units, and only stage 1 of the learning was employed with an exponential reduction in the learning rate and an initial learning rate $\alpha_0 = 0.1$. A total of 5000 iterations are used each time updating the map. After the learning process, samples that have similar characteristics based on a consensus of class membership and similarity in measured variables should be assigned to similar regions of the map. The separation between different groups of samples using the supervised method described are influenced by the scaling value ω. The larger ω is relative to the measured variables, the more important the class information becomes and so more samples are forced into predefined groups on the map. For low values of ω, class membership has little influence on learning, and classes may not always be clearly separated, particularly if the factor of interest is not the primary cause of variation in the data set. The algorithms to determine the BMU, adjusted learning rate, neighborhood widths, etc. during the training of a SOM map have been described in more detail elsewhere.[12,35,36]

Classification. In supervised SOMs, the map is updated using information both from $\mathbf{d}_r$ and $\mathbf{d}_s$, which means the BMU is defined using both variable and class information. The supervised SOM map can therefore be used as a classifier to determine the class of an unknown sample by locating the BMU of the unknown sample using only the VWV $\mathbf{v}_r$ weights for each unit, and assigning the sample to the class in the CWV $\mathbf{v}_s$ of the BMU that has the largest value. The appearance of the resultant SOM is influenced by ω. The larger ω, the higher the risk of overfitting; i.e., the SOM may not be able to successfully classify test set samples but will force the training set into predefined groups. Optimization of ω and validation of the classifier are therefore required to ensure the best performance. Methods for the validation and optimization of ω are described below.

[35] Vesanto, J. *Intell. Data Anal.* **1999**, *3*, 111–126.

[36] Ultsch, A.; Siemon, H. P. *Proceedings of the INNC'90 International Neural Network Conference*, Dordrecht, The Netherlands, 1990.

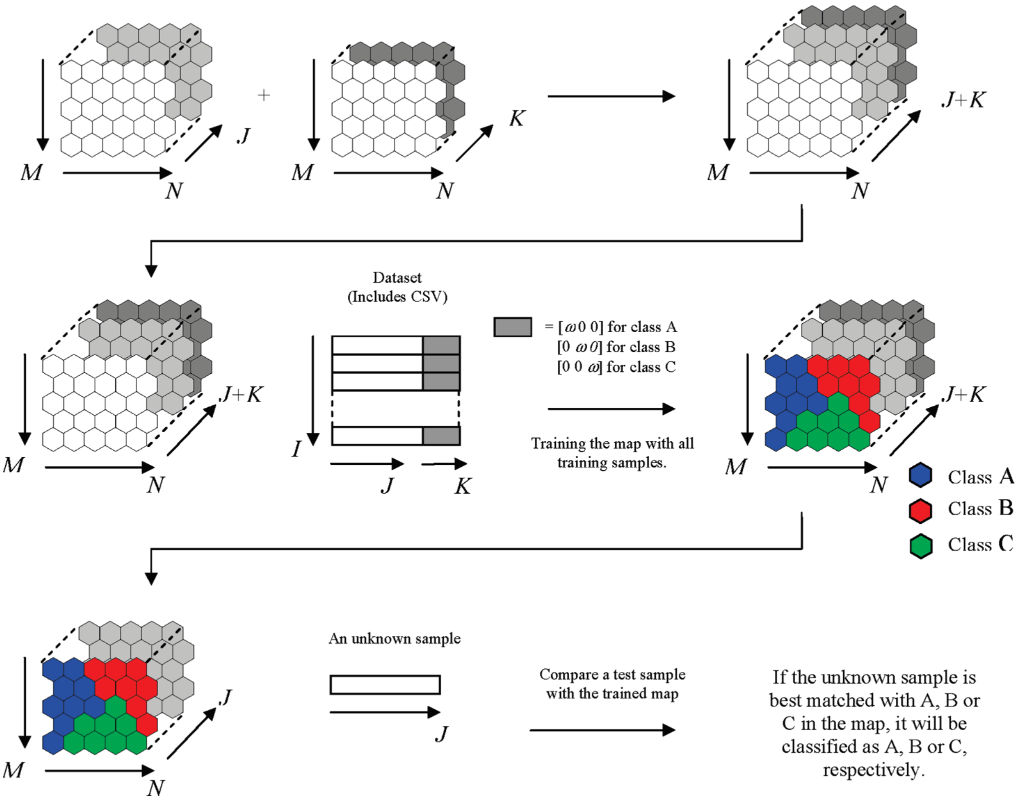


Figure 1. Scheme for classification of an unknown sample using Supervised SOMs for *K* classes and *J* variables.

Validation. Proper validation[24,37,38] is necessary to estimate the performance of a classifier and is usually performed by dividing the data set into 2 subsets, called the training set and test set. The training set is used to train the model for the classifier, which in our case, is the SOM map, and the test set is used to determine the performance of the classifier. In this article, we randomly select two-thirds of the samples in each class for the training set and the remaining are assigned to the test set. The supervised SOM map is trained using the SSVs of the training set samples to provide a map of VWVs that is used to classify the test set samples. In this article, the procedure is repeated 100 times using different randomly selected training and test sets, in order to ensure that the map is not unduly influenced by outliers or typical samples from the training set.

The overall scheme for validating the supervised SOM map and classifying the unknown sample is shown in Figure 1. The percent correctly classified (%CC)[24,33] was obtained by a majority vote,[24] which involves assigning a sample into a class in which it is classified into the maximum number of times, to assess the performance of the classifier. %CC can be calculated for both the training and test sets. High %CC could ideally be obtained for both the training and test sets. However, it is possible to obtain a high %CC for the training set and a low %CC for the test set which indicates that the SOM is overfitted.

Optimum Scaling Value. As discussed in Self Organizing Maps, the choice of the scaling value ω is a key step in supervised SOMs. During training of the supervised SOM maps, supervised sample vectors (SSVs) will be used to determine the BMU on the map. The CSV will have a large influence in comparison to the VSV during the training of the map if the scaling value ω is too large, resulting in overfitting. Since we need to optimize only one parameter (ω) and its behavior is unimodal, it is necessary to establish a parameter that assesses the performance of ω to find the optimum. For each test set sample i, we calculate the variable sample vector (VSV) and the corresponding variable weight vector (VWV) of the BMU ($\mathbf{v}_{\mathrm{BMU}}$) and then compute the Euclidean distance[6,24] between them

$$Q_i = \sqrt{(\mathbf{d}_{\mathrm{r},i} - \mathbf{v}_{\mathrm{BMU}})(\mathbf{d}_{\mathrm{r},i} - \mathbf{v}_{\mathrm{BMU}})'}$$

The more similar, the lower this value. We call Q_i the "Q distance" in this article. We use the average value over the test set $\bar{Q}_{\mathrm{test}}$ as a measure of how different the VSV and VWV are after training. Note that this is computed for each test set separately, so there may be different optimal values of ω. If this value is very high, it will have a large influence on the final map and Q will be high because the map is overfitted. Although the map appears to show a good separation between classes, the BMU is actually a poor representation of samples in the data set. The most appropriate value of ω yields a minimum value of $\bar{Q}_{\mathrm{test}}$ since this will be the map with weights that are the most representative of the data set. We determine this value of ω using the golden search method[39,40] for values of ω between the minimum and maximum of the squared data matrix **X**.

Variable Selection. Supervised SOMs can be used for both visualizing groupings in samples and the classification of unknown

[37] Brereton, R. G. *TrAC, Trends Anal. Chem.* **2006**, *25*, 1103–1111.

[38] Dixon, S. J.; Brereton, R. G. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 1–17.

[39] Forsythe, G. E.; Malcolm, M. A.; Moler, C. B. *Computer Methods for Mathematical Computations*; Prentice-Hall: Englewood Cliffs, NJ, 1976.

[40] Brent, R. P. *Algorithms for Minimization without Derivatives*; Prentice-Hall: Englewood Cliffs, NJ, 1973.

samples as described in Self Organizing Maps and Classification. These two applications are mostly related to the behavior of samples in different groups. After the supervised SOMs are trained, the influence of variables can be visualized by the layers of the map, called "Component Planes".[18] As described in Self Organizing Maps, the advantage of using supervised SOMs is that each source of variation can be studied. In the context of variable selection, this means that the significant variables corresponding to each source of variation can be determined. In this article, the major source of variation is the oral rinse treatment status of the samples and the minor source arises from differences between donors; in addition, spectra may also be influenced by sampling day, but we would not expect this factor (as it differs for each donor) to have a major influence although it can be considered a null factor.

The number of component planes in a supervised SOM, excluding those layers corresponding to the class membership matrix, is equal to the number of variables. This means that the significant variables can be determined directly from the corresponding component planes of a supervised SOM. A visual method is usually employed by shading the plane map according to the intensity of the weights and comparing this to a map with the labeled sample BMUs. Appropriate shading of the component planes therefore allows variables to be identified visually that relate to a specific group of samples. However, when there are a large number of variables, it can become impracticable to visualize all the possible component planes and hence an automated algorithm is required to determine which variables serve as diagnostic indices of particular groups. Previously we introduced a SOMDI[18] for a two class data set which describes numerically how well the component plane of a variable corresponds to the component plane calculated for each class. The SOMDI is calculated by scaling the component plane weights and multiplying it by the class weights for a map. Regions of the map that are representative of both the variable and the class have large values, and taking the summation over all the map units relative to the total number of units indicates how well the two layers correspond. If a component plane gives a significantly higher SOMDI value for one class than it does for the other, then the variable corresponding to this component plane is determined to be a significant variable.

In this article, the SOMDI algorithm is extended to determine significant variables in data sets that may consist of more than two classes and is combined with supervised SOMs. To identify the markers for any specific group, we use "in-group" vs "out-group" comparisons. For example, if we are interested in finding markers for class A, we will define class A as the "in-group" and all other classes as the "out-group" using a two class one vs all comparison.[24] Therefore, if there are K groups in the data, K such comparisons will be made. This process is analogous to the one vs all procedures often employed for other classifiers, e.g., PLS-DA, SVM.[24,31]

In each of K comparisons (where $K = 16$ donors, 2 treatments, and 3 sampling days), we can calculate a SOMDI for both the "in-group" and "out-group". Usually the "in-group" corresponds to a small proportion of the samples, e.g., it may correspond to an individual donor, and we are asking primarily whether a specific variable is more often found (or in greater intensity) in this "in-group" than the rest of the samples. Markers will only be useful if the "in-group" SOMDI is greater than the "out-group" one. (1) For each group k the following procedure is performed. (2) The CWM containing K layers (each layer consisting of U units), denoted by $\mathbf{V}_s$, is combined into two layers: the first layer corresponds to membership of a single group k ($\mathbf{V}_{in}$) and the second layer corresponds to membership of any of the other groups ($\mathbf{V}_{out}$), which is equal to the summation of the values in the remaining $K - 1$ layers. (3) The SOMDI scores for all variables are calculated for both the "in-group" k ($\mathbf{s}_{in}$) and "out-group" ($\mathbf{s}_{out}$), as described previously[18] to give two vectors of dimensions $1 \times J$. The difference $\mathbf{s}_\Delta = \mathbf{s}_{in} - \mathbf{s}_{out}$ is computed. (4) Any variable with a negative values of s_Δ is considered as an unranked variable for group k. All variables with positive s_Δ values are assigned a rank, according to their magnitude (the higher the value the higher the rank) and is considered a candidate marker for group k. Rank 1 means the most significant. (5) Steps 1−4 are repeated for all K groups.

From steps 1 to 5, the candidate markers are determined for each group k. In order to determine the stability of the map, the method above is repeated for 100 iterations. A different subset of around $^2/_3$ (= 64) of the samples is randomly chosen for each iteration to produce a training set, each time producing a different rank list for each group: this tests both for the reproducibility of the map but also reduces the influence of potential outliers since they will not always be part of the training set. For each group k, any variable that is unranked (i.e., s_Δ is negative), in at least one of the iterations is assumed not to be a significant variable. All other variables are retained, and the average rank is calculated over all 100 iterations to provide an overall rank for these retained variables. It is anticipated that variables that are not significant markers for a specific group will be consistently unranked or will occasionally be unranked, whereas strong markers should always be ranked in all training sets.[33] This mechanism allows us to determine how many variables are significant for each factor and group in addition to determining their relative significance. For the most important factor (oral rinse treatment in this case), we anticipate several variables will be significant, whereas for sampling day (which can be considered a null factor), we would predict few or none to be significant. It also allows variables to be found that are significant for more than one group (e.g., donors).

RESULTS AND DISCUSSION

There are three factors of interest for the data set described in this article. The main factor (source of variation) is clearly attributable to the effect of the added oral health care product. The factors due to donor is likely to be a minor source of variation and those due to sampling day are a null factor, which we could not expect to result in many or any significant variables.

Optimal Scaling Value (ω) and Performance of Classifier. Supervised SOMs are frequently used to classify an unknown sample into a group by using the trained map as a classifier. Like other classification methods, the method validation described in the Validation section is required to determine whether the classifier is appropriate for the study in hand and to avoid overfitting. The percentage correctly classified (%CC) by the

Table 3. Overall Percent Correctly Classified (%CC) over 100 Iterations of Training and Test Sets Using the Majority Vote Criterion Using the Optimal Scaling Values (ω)[a]

	% correctly classified		
factor	training set	test set	random
treatment	94.72	70.79	50
sampling day	92.36	38.19	33.33
donor	89.26	19.53	6.25

[a] The column to the right represents the classification that would be achieved if data were randomly assigned to each class.

majority vote[24,41] was used to determine the performance of the classifier. For the method described in this article to train a supervised SOM, the %CC results are dependent on the chosen scaling value (ω). If ω is too small, it will not have influence on the classifier and the results obtained will be comparable to an unsupervised SOM. In contrast, if too a large value of ω is employed, it will dominate the analysis and the classifier will overfit the data, resulting in poor test set performance. It is therefore necessary to optimize the scaling value as it will influence the performance of the classifier. The procedure is described in Optimum Scaling Value. The minimum $\bar{Q}_{\text{test}}$ values for treatment, sampling day, and donor were found to be ω = 0.0311, 0.0340, and 0.0076, respectively. Therefore, these values are used to produce the supervised SOM maps for the purposes of classification and variable selection.

The %CC of the training set and test sets for all studies using a majority vote are shown in Table 3. In addition, the %CC that would have been achieved if samples were randomly assigned to each group is listed: this value depends only on the number of classes, so if there are 16 classes (e.g., donor), this equals 100 × (1/16) = 6.25%. The training set %CC has little meaning as to how well the classifier can predict data but does serve as an indication of how well the model is optimized. From Table 3, it is clear that the training set %CC for all cases is around 90% and higher than the test set %CC for all cases. This suggests that the trained maps are well organized and successfully classified training set samples for the optimal values of ω chosen. The test set %CC for the treatment factor is 70.79% suggesting that the oral rinse treatment is a major source of variation. However, the test set %CC for the donor (= 19.53%) is also high when compared to a random model (6.25%), while the test set %CC for sampling day (38.19%) is very close compared to a random model (33.33%). This suggests the donor has a small effect on the data set, and the sampling day has a limited effect on the data, which is to be expected for a null factor: there may be small temporal variations (see below), which while data are not aligned exactly in time, a weak influence on the signal is exhibited.

Visualization. Unsupervised SOMs were trained using the preprocessed data matrix **X** for visualization in order to investigate the variation arising from treatment (oral rinse vs H_2O addition), donor, and sampling day. The maps are shown in Figure 2a which presents the three different maps shaded according to the sources of variation involved in this study. It can be seen that there are very distinct regions for samples treated with the oral rinse and the H_2O controls: this suggests that addition of the oral rinse has had a significant impact on the chemicals detected in human saliva. On the other hand, there is not such a good separation between sampling days and donors on the unsupervised SOM maps, suggesting that variability from these factors has a relatively small influence on the overall map. However, we might expect some small separation between donors as they have different habits (e.g., foods, exercise, drink) and characteristics (e.g., age, gender, body mass index), but because these are minor compared to treatments, we cannot easily observe these effects when using unsupervised methods.

In the next step, supervised SOMs using optimal scaling values obtained as discussed above are applied to the data set. These are shown in Figure 2b. It can be seen that there is improved separation between groups on these maps for all cases, especially for the minor factors. Note that for many donors, the samples now fall into one region of the map rather than two regions. However, the main application of the supervised SOM technique is not for visualization because the SOM maps using this technique will tend to overfit the data according to the class variable, although they could be regarded as analogous to PLS-DA scores where one can get separation according to the covariance between the classifier and the measured experimental variables.

Markers. A more useful application of supervised SOMs is to find the discriminatory variables corresponding to different sources of variation. As described in Variable Selection, variables with positive values of s_{Δ} for all 100 different training sets are considered as potential markers. The number of significant variables found in each group is illustrated in Figure 3.

In case of the treatment, there are 23 variables (buckets) that pass this criterion for either the H_2O control or oral rinse groups. These are listed in Table 4, which includes both known and tentative assignments for each bucket ordered by the average rank of the variables. In the selected buckets, there are only 6 markers for the "treated" (oral rinse) group but there are 17 makers corresponding to the "control" (H_2O control) group; this is not unexpected because chloride anion (ClO_2^-) present in the oral rinse utilized is known to react with and/or oxidatively consume many salivary biomolecules.[42] The scaled component planes and the intensity distributions of variables with the highest ranks for the "treated" and "control" groups are presented in Figure 4.

In the next step, the method described under Variable Selection is applied to determine the markers for the other sources of variation, specifically donor and sampling day, which involve the use of multiclass models. The buckets that were selected for each sampling day are listed in Table S-1 in the Supporting Information. We can see that unlike the treatment/control factor, where 23 variables are selected, only 3 are found, of which none characterize day 2. The component planes including the intensity distribution plots of the top ranked markers are shown in Figure S-3 in the Supporting Information, where it can be seen that the variables are only weakly discriminatory. These observations are as anticipated for a null factor and suggest that sampling day has only a small influence on the observed NMR spectra. There are

[41] Lloyd, G. R.; Ahmad, S.; Wasim, M.; Brereton, R. G. *Anal. Chim. Acta* **2009**, *649*, 33–42.

[42] Lynch, E.; Sheerin, A.; Claxson, A. W. D.; Atherton, M. D.; Rhodes, C. J.; Silwood, C. J. L.; Naughton, D. P.; Grootveld, M. *Free Radical Res.* **1997**, *26*, 209–234.

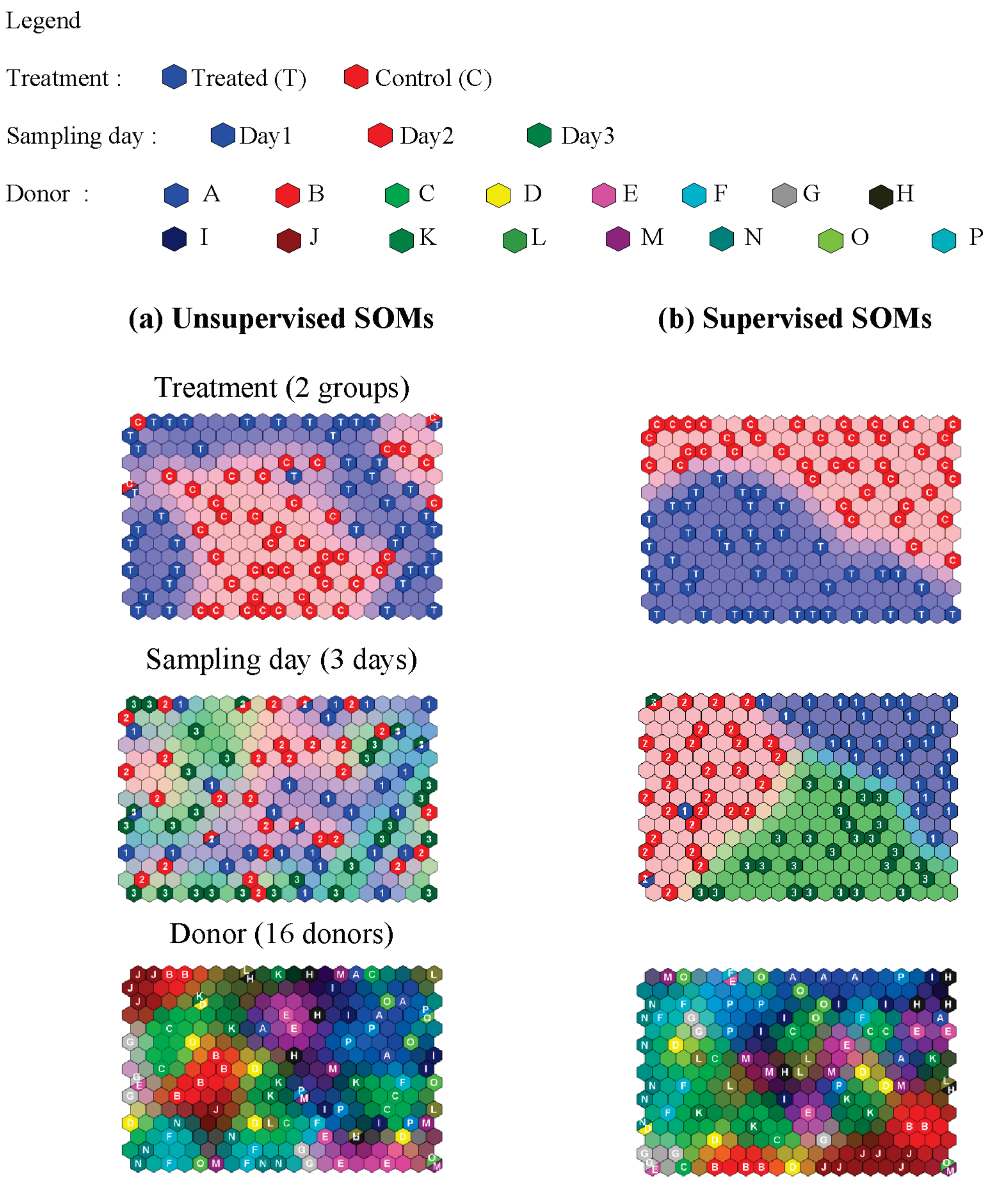


Figure 2. Unsupervised SOMs (left) and supervised SOMs (right) of the three factors, which are treatment, sampling day, and donor, for the intelligently bucketed data set. The optimal scaling values (ω) for each factor were employed to obtain the supervised SOMs.

various reasons why we may expect a small number of variables to be selected. The first is that even if the underlying distribution of a variable is random, we expect there to be an uneven distribution occasionally, rather like tossing a coin 10 times, it will occasionally come up 9 Heads. One solution would be either to take more samples or to perform more iterations of the SOM (in this article we use 100). However even if we were to select variables for example using a 99% significance threshold, this would imply that 1 in a 100 variables that is not significant is above this threshold, so it is encouraging that only very few such variables are found.

In the case of the donor, it can be seen from a supervised SOM map (Figure 2b) that for donors J and N all six samples are grouped together: note that of course the value of ω can be increased to provide better clustering within samples belonging to individual donors but this can result in overfitting. However, both these donors can be described by several characteristic variables according to our criteria. The component planes for the most significant variables characteristic of each of these donors are illustrated in Figure 5 together with their distributions and suggest that there are indeed some regions of the spectra characteristic of individuals. A total of 29 characteristic variables are suggested. Indeed, the 29 variables are found to be characteristic of at least one individual (Table S-2 in the Supporting Information). Note that a marker may be characteristic of several donors rather than just one; for example, these donors may have similar habits or genetic backgrounds.

In Table S-3 in the Supporting Information, we list all the buckets and whether they are markers for treatment, donor, and sampling day and if so, which for which group. We find that 7 out of the 49 buckets are never characterized as being markers for any of the factors, whereas 29 for only one of the factors and 13 for more than one factor. It is important to realize that the signals in each bucket do not necessarily originate from a single compound, so this is not unexpected, although there still appears a lot of selectivity, especially for dominant factors: therefore, it is possible for a region of an NMR spectrum to be influenced by more than two unrelated factors. Interestingly, the 6 bucket

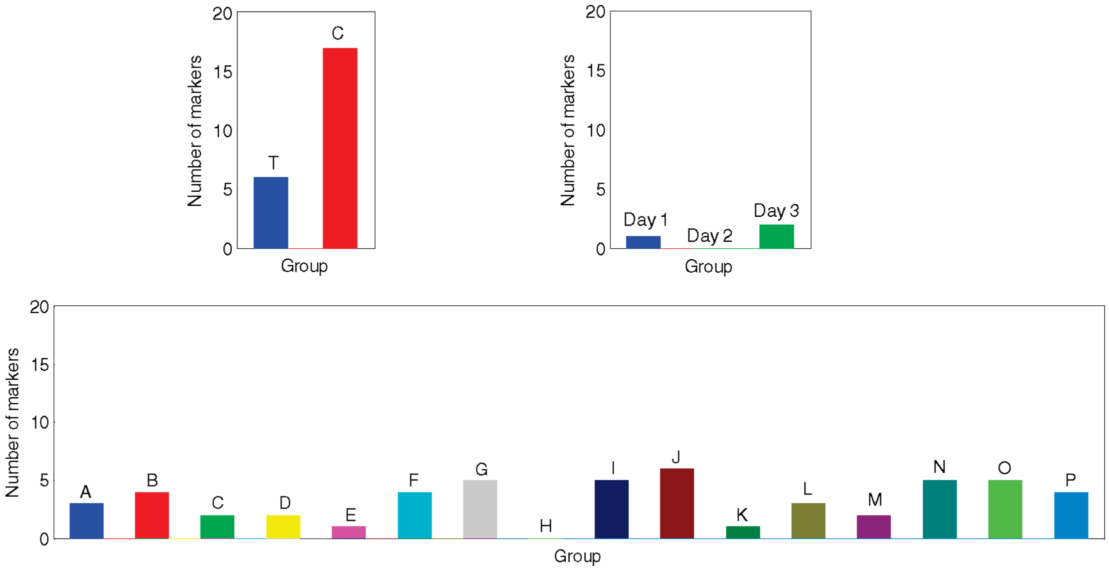


Figure 3. Number of significant variables found for each group and source of variation.

Table 4. 23 ^{1}H-NMR Chemical Shift Buckets, Their Tentative Assignments, and Rankings for Their Ability to Distinguish between Oral-Rinse-Treated and H_2O-Treated Control Salivary Supernatant Specimens[a]

rank	variable	chemical shift	group	tentative assignment
1	45	4.24–4.29	treated	part cysteine-sulphinate-α-CH
2	48	7.20–7.22	control	part protein tyrosine residue-Ar-H2, H6; tryptophan ring-H6
3	26	2.35–2.38	control	pyruvate-CH_3; glutamate-γ-CH_2; proline-β-CH_2
4	5	0.99–1.03	control	isoleucine-β-CH_3; valine-CH_3s; propionate-CH_3
5	4	0.96–0.99	control	leucine-γ-CH_3s; valine-CH_3s
6	13	1.72–1.78	control	lysine-δ-CH_2
7	33	3.24–3.28	control	taurine-$CH_2NH_3^+$; betaine-$^+N(CH_3)_3$; carnitine-$^+N(CH_3)_3$; arginine-δ-CH_2; β-glucose-H2; phenylalanine-β-CH_2; trimethylamine-*N*-oxide-$(CH_3)_3NO$; histidine-β-CH_2; *myo*-inositol-H2.
8	24	2.28–2.31	treated	γ-aminobutyrate-α-CH_2; acetoacetate-γ-CH_3
9	7	1.45–1.51	control	alanine-CH_3; isoleucine-γ-CH_2; pyruvate hydrate-CH_3
10	42	4.02–4.07	treated	phosphorylethanolamine-O-CH_2; choline-CH_2OH; creatinine-$COCH_2N$
11	17	1.99–2.01	control	isoleucine-β-CH
12	35	3.32–3.35	control	tryptophan-β-CH; caffeine-NCH_3 (C3)
13	12	1.70–1.72	control	leucine-β- and γ-CH_2s; arginine-γ-CH_2
14	20	2.09–2.15	control	methionine-S-CH_3 and -β-CH_2, glutamate-β-CH_2 glutamine-β-CH_2
15	16	1.86–1.88	control	γ-aminobutyrate-β-CH_2; citrulline-β-CH_2
16	1	0.84–0.86	treated	*n*-valerate-CH_3; fatty acid-CH_3
17	30	3.02–3.08	control	lysine-ε-CH_2; γ-aminobutyrate-γ-CH_2; creatine-N-CH_3; creatinine-N-CH_3; cysteine-CH_2; ornithine-δ-CH_2; phenylalanine-β-CH_2; tyrosine-β-CH_2
18	49	7.38–7.43	control	phenylalamine-Ar-H4; phenylalanine-Ar-H3,H5; aspartame-Ar-H4; aspartame-Ar-H3,H5
19	37	3.40–3.45	control	taurine-$^-O_3SCH_2$; proline-δ-CH_2NH-; carnitine-γ-CH_2,-α-CH_2
20	36	3.38–3.40	control	Proline-δ-CH_2NH-; β-glucose-H4; methanol-CH_3
21	18	2.01–2.03	control	N-Ac-CH_3[b]
22	2	0.86–0.91	treated	fatty acid-CH_3; *n*-butyrate-CH_3; iso-caproate-δ-CH_3s
23	47	5.39–5.44	treated	unidentified

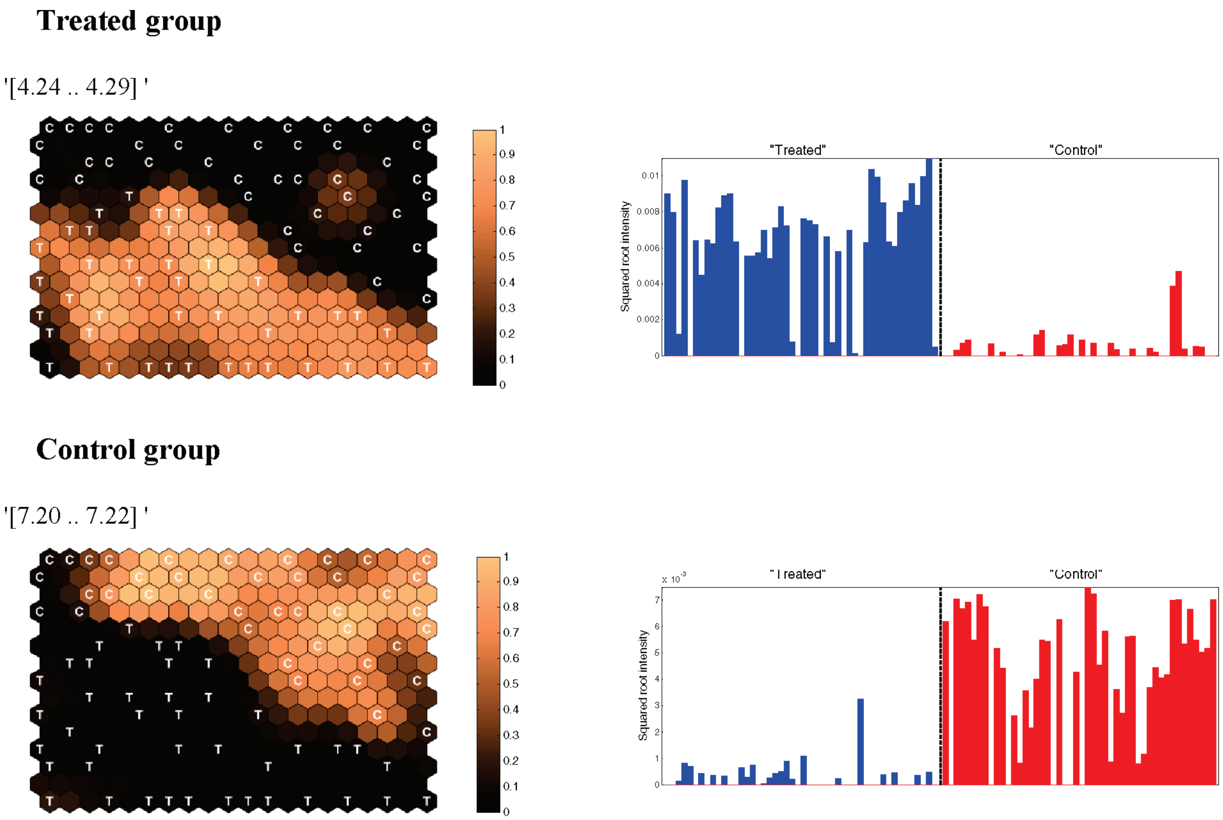
[a] These buckets were selected over all 100 different sets of samples including group. [b] N-Ac-CH_3 represents acetamido group protons of N-acetylsugars present in the molecularly mobile carbohydrate side-chains of "acute-phase" glycoproteins, hyaluronate (HA) and/or further glycosaminoglycans (broad resonances), together with "free" N-acetylsugars and those present in oligosaccharides (sharp signals) derived from the hyaluronidase-mediated depolymerization of or reactive oxygen radical species/hypochlorite attack on HA (such sharp resonances can also be generated from the interaction of oral rinse ClO_2^- and/or $ClO_2^\bullet$ with this glycosaminoglycan); the sharper signals are also ascribable to N-acetyl amino acids such as *N*-acetylaspartate.

regions that are significant markers for oral rinse treatment do not appear to be characteristic of either of the other factors. Also several donors appear to follow common trends, for example, donors I and P and also O and N.

CONCLUSIONS

Unlike traditional methods such as PLS or PCA, SOMs are more rarely employed in metabolomic profiling although there are a few papers in the literature.[7−9] This is probably because the software is not so widely available in packaged form, and is more computationally intensive than traditional algorithms. SOMs have mainly been employed in an unsupervised form.

In this article, we show the power of supervised SOMs and propose an approach for optimizing the SOM, to include the classifier but to avoid overfitting. In many methods for classification such as PLS-DA, the classifier and experimental data matrix are weighted equally, but in this article we propose a new method for optimizing the contribution of the classifier. We show that the

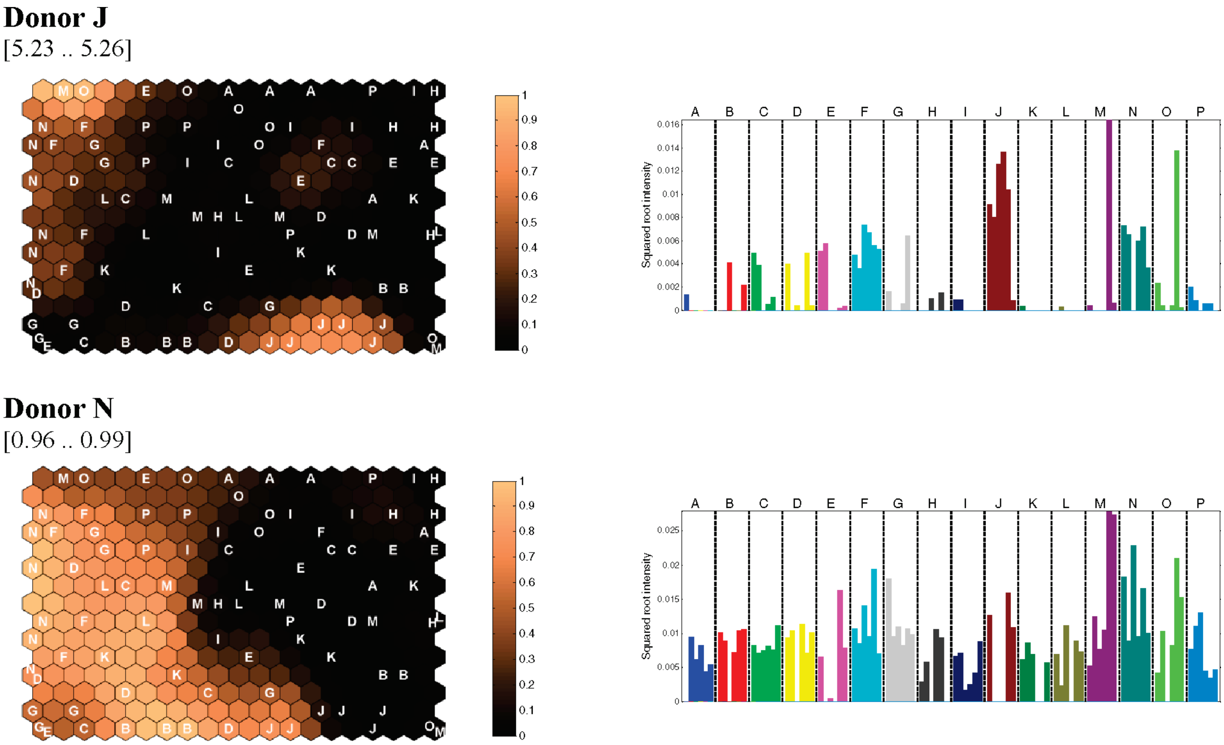

Treated group

'[4.24 .. 4.29] '

Control group

'[7.20 .. 7.22] '

Figure 4. Class component planes (left) and the square root intensity distribution (right) of the two highest discriminatory variables for each of the control and treatment groups. The map units are shaded from highly representative (light) and nonrepresentative (dark) of the classes. Hexagons are numbered according to the BMU representative of each class in the data set (T = treated, C = control).


Donor J

[5.23 .. 5.26]

Donor N

[0.96 .. 0.99]

Figure 5. Class component planes (left) and the square root intensity distributions (right) of the top discriminatory variable for donors J and N. The map units are shaded from highly representative (light) and nonrepresentative (dark) of the classes. Note that hexagons are numbered according to the BMU representative of the donor in the data set.

SOMDI index that has previously been reported can be extended to deal with cases where there are several factors and each factor consists of several groups, to provide a versatile approach for variable selection. Since biomarker discovery is one of the most important aspects of metabolomics, we feel that the methods discussed in this article are of general value to analytical chemists involved in such studies. PLS-DA, while traditional in metabolomics, has many disadvantages. The first is that it is a less powerful visualization tool, the second is that it assumes linear models, and the third is that it can be overly influenced by outliers as it is a least-squares method. For straightforward linear problems, PLS-DA, being computationally less intensive than SOMs, is probably sufficient, but when extended to more complex and nonlinear application areas that are of increasing importance in modern analytical science, it is not always the best method of choice.

It is important to recognize that the implementation of supervised SOMs in this article is aimed primarily at determining which variables are most important for discriminating between groups rather than for primarily visualizing class structure. Whereas it may appear that there is classification between groups if the SOM is forced (for example, for the null factor, sampling day), we find that there are in fact very few significant variables suggesting that this factor is not important for group separation. The protection of reformulating training sets 100 times and seeing whether there are stable models guards against overfitting.

Finally, the methods in this article could be extended to provide probabilities, e.g., that each sample belongs to its own (or another class) or for an unknown, for example, if a sample is chosen 30 times to be a member of a test set and predicted to be a member of class A 20 times, then the (empirical) probability to being a member of class A is 67%. Similarly samples of unknown origin could also be predicted this way. Hence, additional information could be obtained about an individual test set or unknown samples.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review September 13, 2009. Accepted December 4, 2009.

AC9020566