Available online at www.sciencedirect.com
SCIENCE DIRECT®

Talanta 68 (2006) 1683–1691

Talanta

www.elsevier.com/locate/talanta

Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms

Z. Ramadan*, D. Jacobs, M. Grigorov, S. Kochhar

Nestlé Research Center, Vers-chez-les-Blanc, CH-1000 Lausanne 26, Switzerland

Received 11 April 2005; received in revised form 27 July 2005; accepted 10 August 2005
Available online 19 September 2005

Abstract

The aim of this study was to evaluate evolutionary variable selection methods in improving the classification of ^{1}H nuclear magnetic resonance (NMR) metabonomic profiles, and to identify the metabolites that are responsible for the classification. Human plasma, urine, and saliva from a group of 150 healthy male and female subjects were subjected to ^{1}H NMR-based metabonomic analysis. The ^{1}H NMR spectra were analyzed using two pattern recognition methods, principal component analysis (PCA) and partial least square discriminant analysis (PLS-DA), to identify metabolites responsible for gender differences. The use of genetic algorithms (GA) for variable selection methods was found to enhance the classification performance of the PLS-DA models. The loading plots obtained by PCA and PLS-DA were compared and various metabolites were identified that are responsible for the observed separations. These results demonstrated that our approach is capable of identifying the metabolites that are important for the discrimination of classes of individuals of similar physiological conditions.

Keywords: PCA; PLS-DA; Metabonomics; NMR; Genetic algorithms

1. Introduction

The metabonomics approach evolved from the pioneering work of Nicholson and co-workers to become a novel analytical technique for rapid discovery of biological dysfunctions in pharmaceutical and clinical applications. It consists in using high-resolution ^{1}H nuclear magnetic resonance (NMR) spectroscopic profiling of biological fluids combined with multivariate analysis, to identify the metabolites that correlate with changes of physiological conditions [1,2]. It is defined as "the quantitative measurement of the dynamic multi-parametric response of living systems to pathophysiological stimuli or genetic modification" [3]. The technique provides a global quantitative description of hundreds of low-molecular endogenous metabolites present in a biological sample, such as urine, plasma, or tissue [4,5]. The high-resolution ^{1}H NMR spectroscopy datasets are represented as complex matrices with several hundreds of proton signals originating from the various metabolites. This complexity can be untangled by the application of chemometrics methods that are used to reduce the dimension of the ^{1}H NMR data for visualization purposes and to identify inherent patterns among sets of spectral measurements. It can be used to generate models for classification of disease states [6], toxic shocks [7,8], genetic modifications [9], or even of change in diet [10]. The typical examples of chemometric methods applied to analyze metabonomics data are principal component analysis (PCA) [11], soft independent modeling of class analogy (SIMCA) [4], partial least square discriminant analysis [12] (PLS-DA), and neural networks [9,13].

In this article, we are presenting the results of a recent investigation, where we scrutinized the inherent metabolic variability in a control human population in an attempt to derive multivariate boundaries of physiological normality. Three biological fluids, namely plasma, urine, and saliva samples, were collected and analyzed employing high-resolution NMR and multivariate statistical data analysis. The identification of the metabolite profiles in the respective biological fluids unravelled patterns related to key parameters such as gender, age, and life style. The metabolic profiles, annotated with these parameters, were compiled in a lifestyle database of healthy human beings necessary to identify disease- and nutrient-related metabolic fingerprints in follow-up studies. In this paper, only the gender parameter will

* Corresponding author. Tel.: +41 21 785 8020; fax: +41 21 785 9486.
E-mail address: ziad.ramadan@rdls.nestle.com (Z. Ramadan).

0039-9140/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2005.08.042

be considered. The chemometrics methods used for the pattern recognition model were PCA and PLS-DA, while genetic algorithms (GA) enabled the selection of the relevant variables in the ^1H NMR dataset that provided better classification models.

2. Material and methods

2.1. Study set up and sample preparation

The study was conducted as a screening trial with one-time sampling of blood, saliva, and urine samples provided by 150 healthy adult volunteers (48% females and 52% males). The volunteers were recruited at Nestlé Research Center (NRC). The samples were collected in May–June 2003 and immediately stored at −20 °C. All subjects were required to provide an informed consent and to comply with all of the following criteria: completion of a confidential questionnaire on health status, completion of a detailed life style questionnaire, and not being pregnant. Acutely ill subjects (cold, flu, fever, etc.) under medication (antibiotic therapy, anti-inflammatory drugs, etc.) were excluded from the study. For women, the sampling was restricted to the first 10–15 days (inclusive) of the menstruation cycle. The subjects were classified according to their gender, age, and answers in the life style questionnaire (for example, sport and exercise activities, alcohol consumption, coffee consumption, specific dietary regimes, etc.). The personnel of the Metabolic Unit at NRC entered Demographic and life style data in a case report form (CRF). Finally, all of the information has been encoded to protect confidentiality. The trial was conducted according to the relevant legal requirements and approved by the local ethical committee.

2.1.1. Urine samples

Second morning spot urine (minimum 20 ml) were collected by the subject and brought to the Metabolic Unit at the time of blood sampling. They were immediately frozen and stored at −20 °C. Urine samples were prepared by diluting (2:1) urine with phosphate buffer (0.2 M Na_2HPO_4/0.2 M NaH_2PO_4, pH 7.4; 80% H_2O/20% D_2O). The samples were subsequently centrifuged and filled in NMR tubes.

2.1.2. Blood samples

Two milliliters of blood was drawn from the antecubital vein by single puncture, using Sarstedt syringes with heparin as anticoagulant. Plasma were immediately separated by centrifugation and stored at −20 °C. For NMR measurements, the samples were diluted twice with 90% H_2O/10% D_2O phosphate buffer (0.2 M) adjusted to pH 6.0. After centrifugation, the samples were placed in 5 mm NMR tubes for data acquisition.

2.1.3. Saliva samples

A maximum of 2 ml was collected with “Salivettes” from Sarstedt at least 1 h after brushing the teeth in the morning. The NMR spectra were recorded after lyophilizing 500 µl of saliva and subsequent reconstitution in 550 µl of deuterated phosphate buffer (0.2 M) adjusted to pH 7.4.

All NMR samples contain sodium azide to prevent bacterial contamination, DSS (3-(trimethylsilyl)-1-propanesulfonic acid, sodium salt) as internal reference substance (except the plasma samples) and imidazole to check the pH.

2.2. NMR data collection

All one-dimensional ^1H NMR spectra were acquired on a Bruker DRX-600 NMR spectrometer operating at 600.13 MHz. Samples were measured in 5-mm-o.d. NMR tubes and at 300 K. For all samples, a standard 1D ^1H pulse sequence with water pre-saturation was applied. In addition to the standard 1D spectrum, Carr–Pucell–Meiboom–Gill (CPMG) spectra with spin echo sequence $\pi/2$-t_D-π-t_D, were acquired for plasma and saliva samples to attenuate broad signals arising from protein and lipoproteins [14]. The spin echo loop time was adjusted to 64 ms. A total of 256, 128, and 256 transients were collected for urine, plasma, and saliva, respectively. Typical acquisition parameters included 32 k data points, a spectral width of 8389 Hz, an acquisition time of 1.95 s, and a relaxation delay of 2 s. As with the standard 1D spectra, an exponential line-broadening function of 0.3 Hz was applied to the free induction decay (FID) prior to Fourier transformation. All spectra were processed for phase and baseline correction. The urine and saliva spectra were referenced to DSS (δ0 ppm) and the plasma spectra to lactate (CH_3, δ1.33 ppm).

2.3. Data reduction and pattern recognition

Each NMR spectrum was reduced to smaller number of variables, calculated by integrating regions of equal bucket size of 0.02 ppm and variable bucket size where large variations in chemical shift were expected using an in-house routine written in MATLAB (The MathWorks, Natick, MA). Several spectral regions were excluded as shown in Table 1, mainly to eliminate variation in water suppression efficiency and imidazole peaks.

The datasets were rearranged in such a way that the rows of each data matrix represent the subjects and the columns represent chemical shift (variable). The size of the dataset for the plasma and saliva samples was 150 × 493. The size of the dataset for urine samples was 150 × 409. Each spectral dataset was normalized to the total sum of the integrals to partially compensate for differences in concentrations. The gender affiliation was used as a dependent variable, i.e., the variable to be predicted by pattern recognition methods. The entire dataset was divided into two parts: a training set that was used to build a model and a test set that was used to test the model’s predictive ability. The

Table 1
Spectral regions excluded from ^1H NMR spectra

Sample type	Spectral regions excluded (δ)
Plasma and saliva	<0.5, >9.5, 4.5–5.1 (water peak), 7.341–7.397, 8.449–8.501 (imidazole peak)
Urine	<0.4, >9.0, 4.5–5.1 (water peak), 7.378–7.480, 8.467–8.679 (imidazole peaks)

training set was prepared by taking every first pair of samples, while every third sample was included in the test set. Therefore, the training set was composed of 100 ^{1}H NMR spectra and the test set was composed of 50 spectra. The test set was separated from the data and was not used to monitor the training process. This procedure prevented any possibility that the best regression models selected had a chance correlation to peculiarities in the measurements of the test set and reduced the risk of over-fitting. Two scaling methods were applied, auto-scaling and Pareto scaling. In auto-scaling, the variable mean was subtracted from each variable (column of the data) and then each variable was divided by its standard deviation. The same process was repeated in Pareto scaling except that the square root of the standard deviation was used. Pareto scaling falls in between no scaling at all and auto-scaling and gives the variable a variance equal to its standard deviation instead of unit variance. The Pareto scaling was applied when PCA or PLS-DA were used, while auto-scaling was applied during the training process of the GA. This gave equal weights to all of the chemical shifts during the training process of variable selection with GAs. Three latent variables were used in all PLS-DA classification models to avoid over-fitting and for comparison purposes.

Multivariate analyses such as principal component analysis, partial least square discriminant analysis, and genetic algorithms were performed using PLS_Toolbox 3.0 (Eigenvector Research, Inc., Manson, WA 98831) for MATLAB and the software package SIMCA (Version 8, Umetrics AB, Umeå, Sweden).

2.4. Variable selection using genetic algorithms

GAs and evolutionary programming are optimization techniques [15,16] based on the concepts of natural selection and evolution and they have been used to efficiently solve variable selection problems [17,18]. In this approach, the variables are represented as genes on a chromosome, and they are generally coded as binary strings. Through a simulated natural selection and the action of the genetic operators mutation and recombination, chromosomes that satisfy at best to a predefined fitness function are found. The fitness function is deduced from the gene composition of a chromosome. In our case, we used the percent of correct prediction of gender affiliation as the relevant fitness function to drive the optimization process. Using the recombination operator, the GA combines genes from two parent chromosomes to form two new chromosomes (children) that have a high probability of having better fitness than their

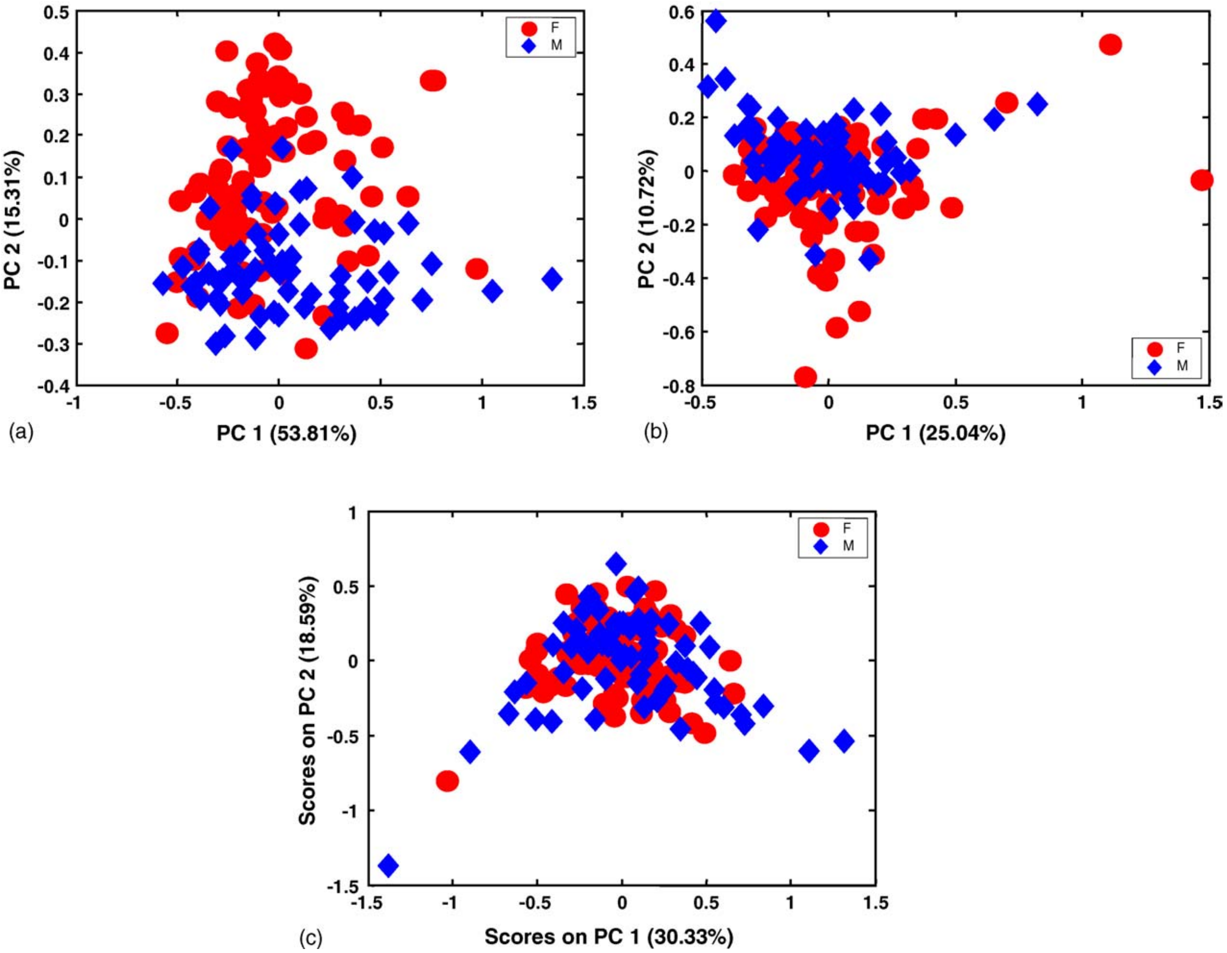

Fig. 1. Scores plot of a two-component PCA model of ^{1}H NMR spectra showing sample clustering according to Gender with percentage of variance captures by each PC for: (a) plasma, (b) urine, and (c) saliva dataset.

parents. GA offers a generational improvement in the fitness of the chromosomes and after many generations will create chromosomes containing the optimized variable settings. GA has several advantages when compared to other optimization algorithms. They have the ability to move from local optima present on the response surface. They require no knowledge or gradient information about the response surface and can be employed for a wide variety of optimization problems [19]. The major drawback of GA is that there can be difficulties in finding the exact global optimum, which requires a large number of response (fitness) function evaluations and prohibitively long computation time [20].

In GA variable selection method, a population of strings is randomly created where each string is a row vector containing as many elements as there are variables. Each element is coded as 1, if the corresponding variable was selected, and 0 if it was not selected. The fitness of the string is equal to the evaluation response that is based on the predictive ability with a given subset of selected variables. The method used in the GA variable selection is designed to select variables with lowest prediction error. Thus, at each step, half of the PLS-DA models formed with the lowest prediction error are allowed to live and breed. The prediction errors were determined using random crossvalidation procedure. Pairs of these model forms are randomly selected. The sets of strings belonging to these two models are used for breeding using a crossover technique. Then, the mutation operator is introduced to prevent premature convergence to local optima by randomly sampling new points in the search space. It sets the fraction of bits in the binary strings, which are randomly flipped in each generation. The procedure is repeated several times until it converges. In all the GA runs, the maximum number of generations was set to 500, the population size was set 128, the breeding crossover rule was set to double crossover, and the default mutation rate was used (**0.005**). Finally, the GA training process was repeated 10 times with different pseudo-random starting points to obtain the global optimum and check the stability of the GA model. Most of the GA models gave similar results; thus, only one GA model for each dataset will be presented in this paper. Interestingly, genetic algorithms were recently applied as a supervised learning procedure to discriminate urine proteomics and metabonomics patterns in the diagnosis of interstitial and bacterial cystitis. The technique was applied jointly to NMR and MS metabolic fingerprints of healthy individuals and disease patients, and was able to predict unaffected subjects with a success rate of nearly 84% [21].

3. Results and discussion

3.1. PCA

PCA of the ^{1}H NMR spectra of the three biofluids (plasma, urine, and saliva) for the gender differences is presented in Fig. 1. In Fig. 1a, the representative points of the plasma samples are mapped in the space spanned by the first two principal components PC1 versus PC2. This scores plot is illustrating a reasonable clustering appearing according to gender membership. The same trends are shown in Fig. 1b that represents the urine samples, and Fig. 1c illustrating the distribution of the saliva samples. PCA unravelled the existence of differences in plasma composition (54% of variance was captured by first PC) in the male (blue) and female (red) subjects, which were missing in urine or saliva.

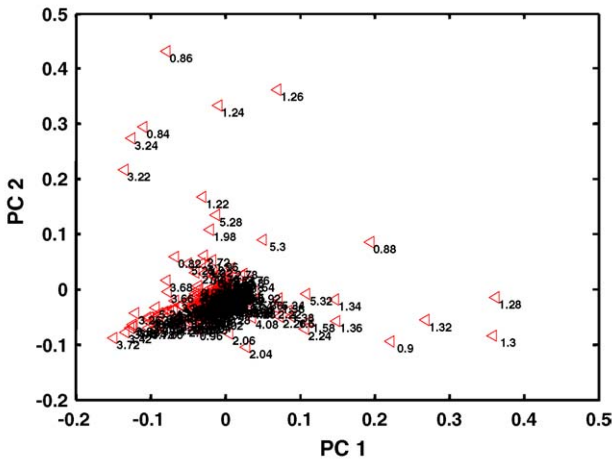

Fig. 2. PCA loadings plot of Life Style Feature "gender" in plasma.

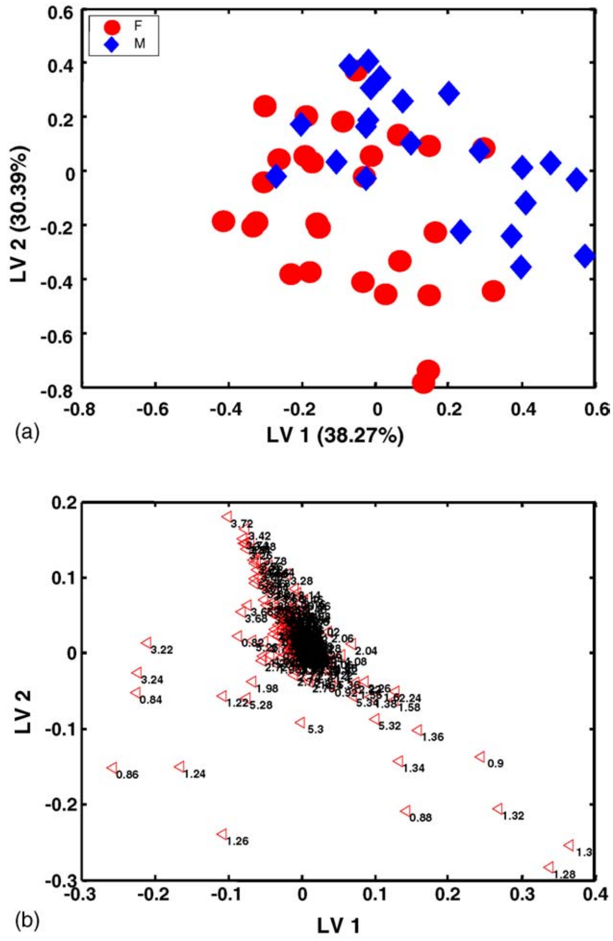

Fig. 3. (a) PLS-DA scores plot of Life Style Feature "gender" in plasma for the test set along with the percentage of variance capture by each latent variable and (b) corresponding loadings plot for the PLS-DA model.

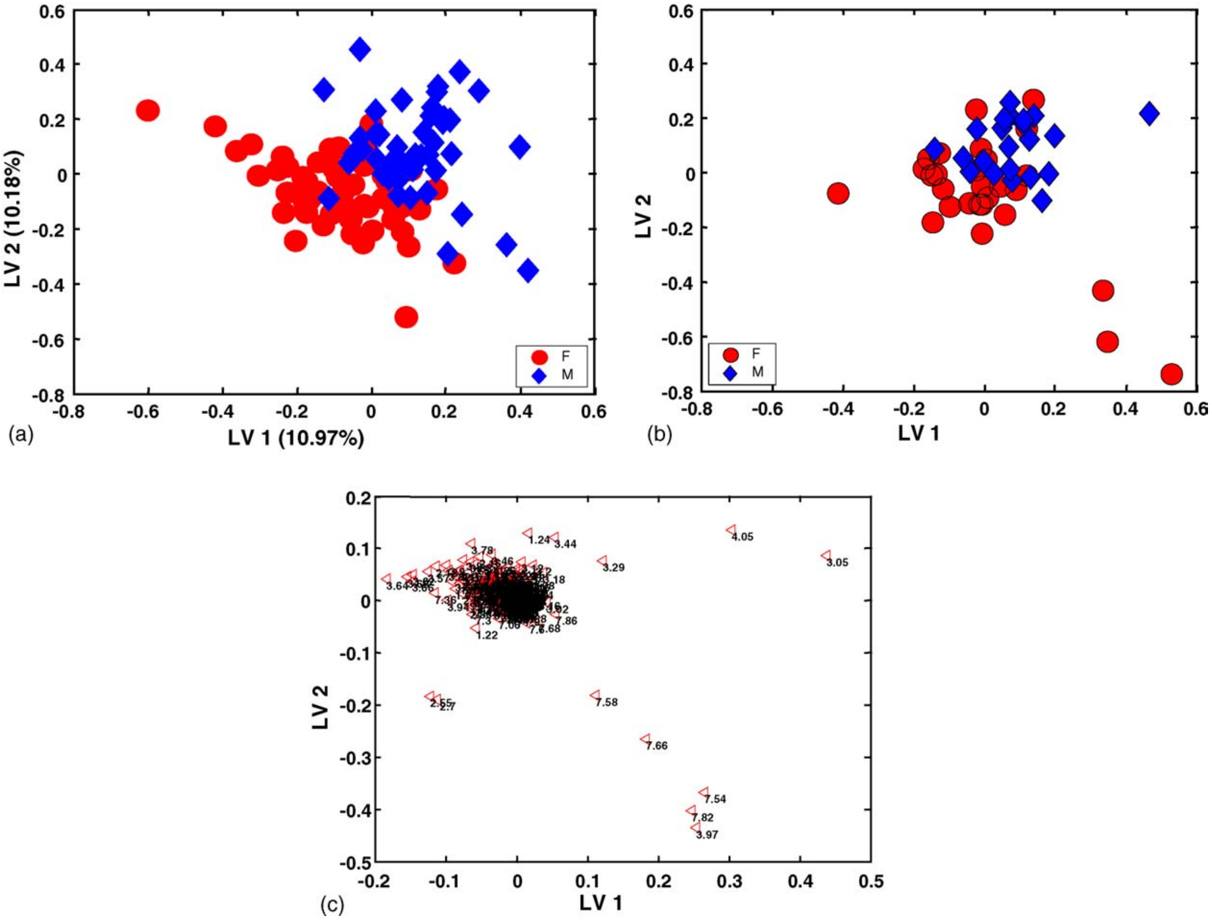

Fig. 4. PLS-DA sores plot of urine dataset: (a) training set, (b) test set, and (c) the corresponding loadings plot.

To find out precisely which regions of the NMR spectra have caused the separation between the male and female populations when scrutinizing the plasma samples, the loadings plot of the related PCA model is shown in Fig. 2.

This loadings plot shows the regions of the NMR spectra, which are responsible for the clustering appearing in the scores plot of the plasma samples. Peaks with different levels between the two genders appeared in the same region of the scores and the loadings plot. In Fig. 1, it is clearly shown that PC2 is capturing most of the variations between the genders. The loadings of PC2, those were most influential for the gender separation in plasma, are summarized in Table 2.

The inspection of Table 2 indicated that the main difference between the two genders occurred in their plasma lipid profiles. The total cholesterol level was found somewhat higher in females than males. Since these lipids are the most abundant metabolites in the blood, most of the variations in PCA were due to these metabolites. The scrutiny of the higher order PCs (four and five PCs) indicated no additional separation between genders.

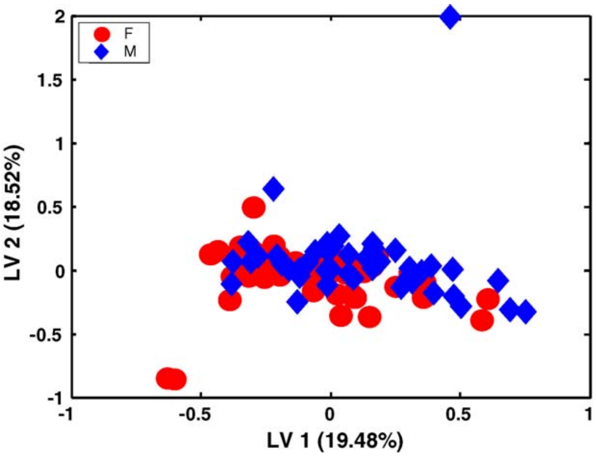

Fig. 5. PLS-DA scores plot of saliva dataset.

Table 2
PCA-detected ^{1}H NMR spectral regions that cause separation between genders (plasma samples)

Chemical shift (ppm)	Metabolites
δ0.84–δ0.86	Mainly due to CH_3 groups from fatty acid side chains in lipids of HDL particles
δ1.24–δ1.26	Mainly due to $(CH_2)_n$ groups from fatty acid side chains in lipids of HDL particles
δ3.22–δ3.24	Mainly due to choline $–N(CH_3)_3^+$ principally phosphatidylcholine from lipoproteins, mainly HDL
δ5.28	Mainly due to CH of lipids
δ1.98	Mainly due to $CH_2C{=}C$ of lipids

The supervised clustering method PLS-DA was carried out to enhance the poor separation obtained with the PCA model for the urine and the saliva NMR spectra. The model was validated with an independent test set. Again, a good gender class separation was attained after PLS-DA for the integrated ^{1}H NMR spectra of plasma. The scores plot of the first and the second latent variable is shown in Fig. 3a for the plasma (test set) and the loading plot of the latent variables is shown in Fig. 3b.

In general, loadings plot (Fig. 3b) indicated that the same regions of the spectra that contributed to the clustering in the PCA analysis also contributed to the clustering seen after application of PLS-DA. The rate of correct classification was 84% within the training set and 84% within the test set.

In what concerns the urine dataset, a better separation was attained after applying PLS-DA. The scores plot for the training and the test set is shown in Fig. 4a and b, respectively.

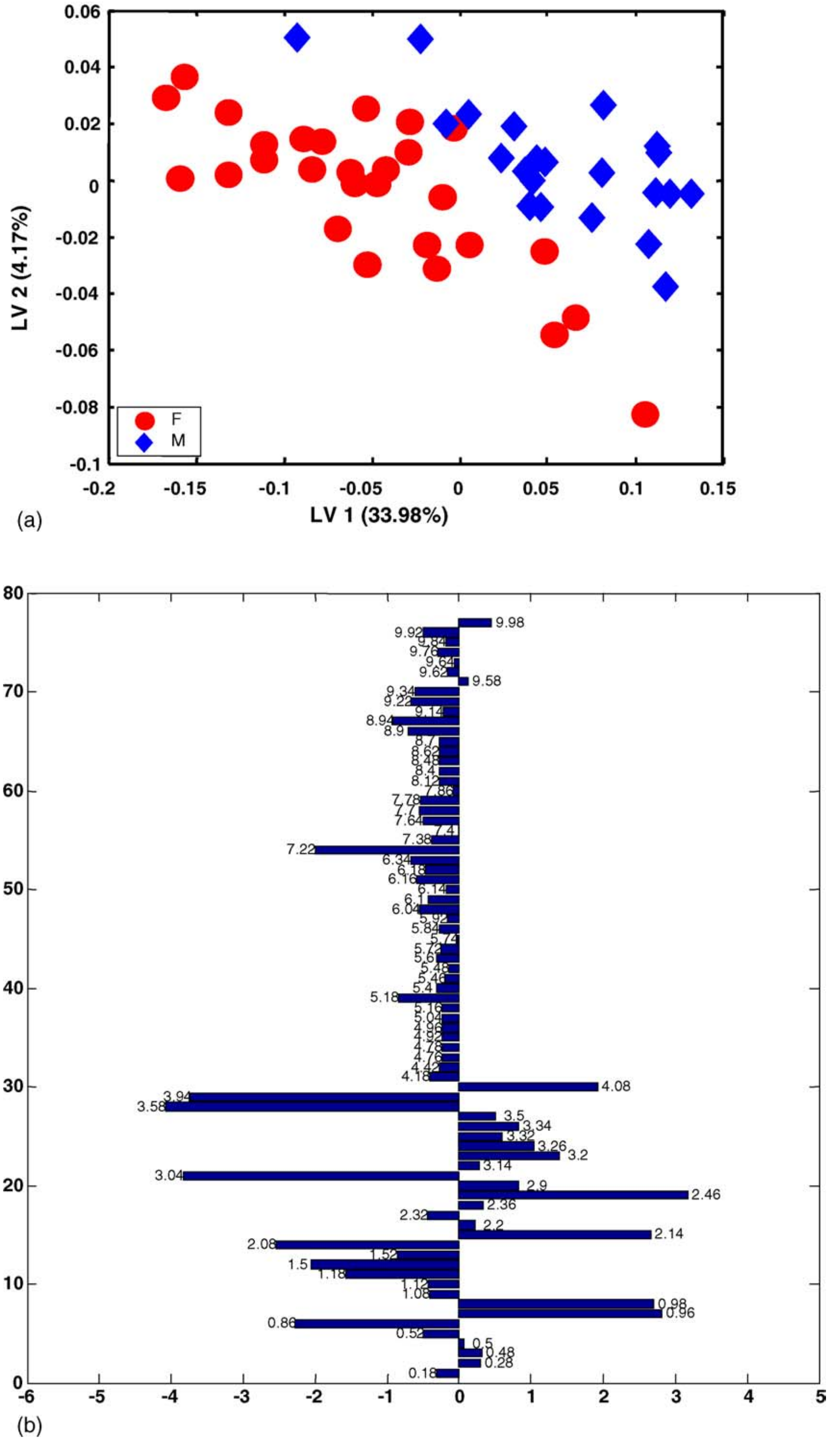

Fig. 6. (a) PLS-DA sores plot in plasma after GA and (b) the regression coefficients of the PLS-DA model shown in (a).

The classification percentage of the training set was 93% and 76% for the test. The corresponding loadings plot of the PLS-DA model is shown in Fig. 4c. The loadings that were the most influential in the separation of genders were situated around δ2.55 and δ2.7 (due to citrate), δ3.05 and δ4.05 (due to creatinine), δ7.58, δ7.66, δ7.54, and δ7.52 (due to hippurate), δ3.44 (due to taurine), and δ3.29 (due to trimethylamineoxide).

In the case of saliva samples, a poor separation between male and female classes was attained after applying PLS-DA. The scores plot for the training set is shown in Fig. 5. The classification percentage of the training set was 70% and 52% for the test.

In order to optimize the separation between the classes under investigation and to improve the performance of the subsequent

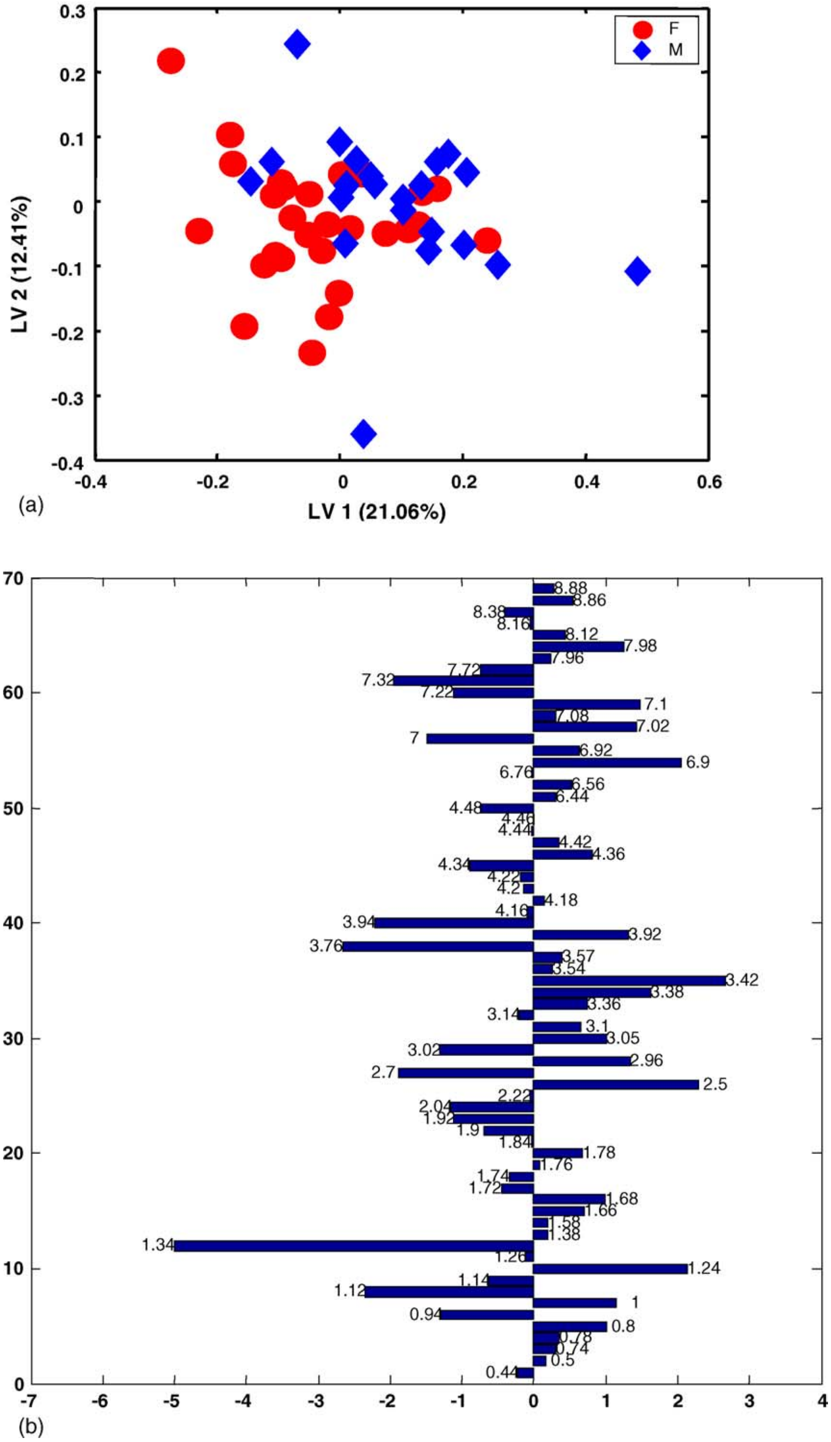

Fig. 7. (a) PLS-DA scores plot in urine after GA and (b) the regression coefficients of the PLS-DA model shown in (a).

multivariate pattern recognition analysis, a variable selection method was applied. GA was applied to achieve a better separation of the classes and remove variables that were not related to gender differences. After GA, the number of variables was reduced to 78 for the plasma samples, 69 for the urine samples, and 68 for the saliva samples. The classification percentage was improved in both the training and the test set for both plasma and urine dataset, but not in the saliva dataset. For the plasma dataset, the scores plot (test set) after GA is shown in Fig. 6a, along with the regression coefficients plot for the variables that survived the GA variable selection method shown in Fig. 6b.

After application of GA, the male and female groups were well separated in the PLS-DA scores plot of LV1 and LV2 (Fig. 6a). A negative value in the plot of the regression coefficients (Fig. 6b) indicated a relatively high concentration of the metabolites present in the female plasma samples, while a positive value indicated a relatively lower concentration. The inverse was true in what concerns the male plasma samples. In general, the regression coefficients plot indicated that similar regions of the spectra that contributed to the clustering in the original dataset also contributed to the clustering seen after application of the GA variable selection procedure. Also, GA selected new chemical shifts that are important in discrimination between the two genders. These chemical shifts are: δ3.04 and δ3.93 (due to creatine), δ3.58 (due to glucine), δ4.08 (due to choline), δ2.46 (due to glutamine), and δ2.14 (unknown). The classification percentage was 97% within the training set and 96% within the test set.

The classification of the two genders when using the urine dataset was improved after GA. The scores plot for the first and the second PC for the test set is shown in Fig. 7a, along with the coefficients plot for the variables that survived the GA variable selection methods shown in Fig. 7b.

The regression plot (cf. Fig. 7) indicated the regions of the spectra that contributed to the clustering in the clustering seen after application of GA. The regression coefficients plot for the training set indicated that the spectral regions that contributed the most to the discrimination of the classes were: δ3.05 (creatinine) and δ2.5–δ2.7 (citrate). The classification percentage was 98% for the training set and 84% for the test set.

For the saliva samples, a good PLS-DA predictive model was not obtained even after the application of GAs. The best classification rate was 74% for the training set and 52% for the test set. A summary of the classification percentage before and after GA for the training set and test set is shown in Tables 3 and 4.

Table 3
Summary of classification percentages for the training set

Sample type	PLS-DA	PLS-DA (after GA)
Plasma	84	97
Urine	93	98
Saliva	70	74

Table 4
Summary of classification percentages for the test set

Sample type	PLS-DA	PLS-DA (after GA)
Plasma	84	96
Urine	76	84
Saliva	52	52

4. Conclusion

The ^{1}H NMR metabonomic investigation presented in this work, combined with variable filtration algorithms and pattern recognition procedures, gave an evidence for the existence of clear metabolic differentiation of individuals, according to their gender or life style. In the first phase of this work, unsupervised (PCA) and supervised (PLS-DA) data mining methods, applied in combination with correct and rigorous pre-processing of the data, demonstrated that it is possible to obtain models that accurately classify metabonomic data samples. In the second part of the work, genetic algorithm procedures were incorporated to optimize the resulting PLS regression equations by removing irrelevant variables. GAs variable selection method dramatically improved the separation between the two genders and different metabolites were identified that were involved in this separation.

References

[1] A.W. Nicholls, J.K. Nicholson, J.N. Haselden, C.J. Waterfield, Biomarkers 5 (2000) 410.
[2] J.K. Nicholson, I.D. Wilson, Nat. Rev. Drug Discov. 2 (2003) 668.
[3] J.K. Nicholson, J.C. Lindon, E. Holmes, Xenobiotica 29 (1999) 1181.
[4] E. Holmes, A.W. Nicholls, J.C. Lindon, S.C. Connor, J.C. Connelly, J.N. Haselden, S.J. Damment, M. Spraul, P. Neidig, J.K. Nicholson, Chem. Res. Toxicol. 13 (2000) 471.
[5] J.C. Lindon, J.K. Nicholson, E. Holmes, J.R. Everett, Concepts Magn. Reson. 12 (2000) 289.
[6] J.T. Brindle, H. Antti, E. Holmes, G. Tranter, J.K. Nicholson, H.W.L. Bethell, S. Clarke, P.M. Schofield, E. McKilligin, D.E. Mosedale, D.J. Grainger, Nat. Med. 9 (2003) 477.
[7] H.C. Keun, T.M.D. Ebbels, M.E. Bollard, O. Beckonert, H. Antti, E. Holmes, J.C. Lindon, J.K. Nicholson, Chem. Res. Toxicol. 17 (2004) 579.
[8] J.C. Lindon, J.K. Nicholson, E. Holmes, H. Antti, M.E. Bollard, H. Keun, O. Beckonert, T.M. Ebbels, M.D. Reily, D. Robertson, G.J. Stevens, P. Luke, A.P. Breau, G.H. Cantor, R.H. Bible, U. Niederhauser, H. Senn, G. Schlotterbeck, U.G. Sidelmann, S.M. Laursen, A. Tymiak, B.D. Car, L. Lehman-McKeeman, J.M. Colet, A. Loukaci, C. Thomas, Toxicol. Appl. Pharmacol. 187 (2003) 137.
[9] E. Holmes, J.K. Nicholson, G. Tranter, Chem. Res. Toxicol. 14 (2001) 182.
[10] K.S. Solanky, N.J.C. Bailey, B.M. Beckwith-Hall, A. Davis, S. Bingham, E. Holmes, J.K. Nicholson, A. Cassidy, Anal. Biochem. 323 (2003) 197.
[11] E. Holmes, H. Antti, Analyst 127 (2002) 1549.
[12] C.L. Gavaghan, I.D. Wilson, J.K. Nicholson, FEBS Lett. 530 (2002) 191.
[13] T. Ebbels, H. Keun, O. Beckonert, H. Antti, M. Bollard, E. Holmes, J. Lindon, J. Nicholson, Anal. Chim. Acta 490 (2003) 109.
[14] T.D.W. Claridge, High-Resolution NMR Techniques in Organic Chemistry, Elsevier Sciences Ltd., 1999.
[15] C.B. Lucasius, G. Kateman, TrAC, Trends. Anal. Chem. (Pers. Ed.) 10 (1991) 254.
[16] R. Leardi, J. Chemom. 15 (2001) 559.

[17] D. Broadhurst, R. Goodacre, A. Jones, J.J. Rowland, D.B. Kell, Anal. Chim. Acta 348 (1997) 71.

[18] Z. Ramadan, X.H. Song, P.K. Hopke, M.J. Johnson, K.M. Scow, Anal. Chim. Acta 446 (2001) 233.

[19] R. Leardi, A.L. Gonzalez, Chemom. Intell. Lab. Syst. 41 (1998) 195.

[20] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 25 (1994) 99.

[21] Q.N. Van, J.R. Klose, D.A. Lucas, D.A. Prieto, B. Luke, J. Collins, S.K. Burt, G.N. Chmurny, H.J. Issaq, T.P. Conrads, T.D. Veenstra, S.K. Keay, Dis. Markers 19 (2004) 169.