Metabonomic Assessment of Physiological Disruptions Using $^{1}H-^{13}C$ HMBC-NMR Spectroscopy Combined with Pattern Recognition Procedures Performed on Filtered Variables

Marc-Emmanuel Dumas,[†] Cécile Canlet,[†] François André,[‡] Joseph Vercauteren,[§] and Alain Paris*,[†]

Laboratoire des Xénobiotiques, UMR 1089, INRA, BP 3, 180, chemin de Tournefeuille, 31931 Toulouse Cedex 9, France, Laboratoire de Pharmacognosie, EA 491, Université Victor Segalen, 146, rue Léo Saignat, 33076 Bordeaux Cedex, France, and LABERCA, Ecole Nationale Vétérinaire, route de Gachet, BP 50707, 44307 Nantes Cedex 3, France

Metabonomic characterization of long-lasting although weak physiological events such as anabolic disruptions remains poorly investigated. We have validated $^{1}H-^{13}C$ HMBC-NMR as a suitable generator of instrumental variables that are strongly linked to the concentration of endogenous metabolites in biological fluids. This method is interfaced to multivariate pattern recognition procedures. Fingerprints established from urine sample collected on cattle treated with anabolic steroids were used to validate this method. Four main results arise from this study. (i) 2D NMR is as informative as 1D NMR. (ii) 2D NMR variable clustering highlights successfully a contingent redundancy of variables, although a relevant hierarchical model of statistical correlations covering from structural relationships to physiologic links can also be evidenced. (iii) To enhance pattern recognition performances, we have validated a variable selection algorithm for accurate prediction of unknown individuals belonging to predetermined groups achieved by linear discriminant analysis (LDA). This algorithm synthesizes the whole information contained in the data set by selecting preferentially nonredundant variables. Parameters generating variable subsets are validated by predicted variance efficiency obtained when minimizing error rates calculated by cross-validation methods. (iv) Provided variables are correctly filtered, LDA fairly competes with partial least-squares methods for both classification of individuals and statistical interpretation of metabolic responses obtained in such a physiological disruption context.

Finding comprehensive relationships between molecular mechanisms studied at a cellular level and specific biological functions, evidenced in organs or organisms, has become a promising domain in biology. As alternatives to postgenomics and proteomics, metabonomics[1,2] deals with the study of general metabolism thanks to recent buildup of tools able to detect significant functional variations in a metabolic network when integrated at the whole-organism scale by analyzing biofluids and tissues, and metabolomics[1] investigates at a more tenuous scale, i.e., in cells or cell types, metabolic regulations, and adaptations of fluxes of biomolecules. In fact, these two methodologies are based on spectroscopic methods, especially infrared spectroscopy (IR), mass spectrometry (MS), and nuclear magnetic resonance spectroscopy (NMR).[3−7] All metabonomic methods are interfaced to multivariate statistical analyses to underline any subset of relevant biomarkers of the biological situation under study.[8] Since spectrometric methods such as NMR or MS allow observations of metabolism without any a priori hypotheses on the disturbances that are awaited, the serendipity principle is warranted by a relevant experimental design and an exhaustive exploration of the multivariate space by statistics.

Until now, multidimensional analysis of NMR spectra is mainly involved in diagnosis[9−11] and drug screening applications;[12] ^{1}H NMR has also shown its ability to give reproducible fingerprints from biological fluids and therefore provide a suitable investigation tool in toxicology.[13] However, if 2D homonuclear and hetero-

[*] Corresponding author. Phone: +33 561 285 394. Fax: +33 561 285 244. E-mail: aparis@toulouse.inra.fr.

[†] Laboratoire des Xénobiotiques.

[‡] LABERCA.

[§] Université Victor Segalen.

(1) Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E. *Nat. Rev. Drug Discovery* **2002**, *1*, 153−162.
(2) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, *29*, 1181−1189.
(3) Goodacre, R.; Kell, D. B. *Anal. Chem.* **1996**, *68*, 271−280.
(4) Goodacre, R.; Kell, D. B. *Curr. Opin. Biotechnol.* **1996**, *7*, 20−28.
(5) Goodacre, R.; Timmins, E. M.; Rooney, P. J.; Rowland, J. J.; Kell, D. B. *FEMS Microbiol. Lett.* **1996**, *140*, 233−239.
(6) Jellum, E.; Bjornson, I.; Nesbakken, R.; Johansson, E.; Wold, S. *J. Chromatogr.* **1981**, *217*, 231−237.
(7) Holmes, E.; Nicholls, A. W.; Lindon, J. C.; Connor, S. C.; Connelly, J. C.; Haselden, J. N.; Damment, S. J.; Spraul, M.; Neidig, P.; Nicholson, J. K. *Chem. Res. Toxicol.* **2000**, *13*, 471−478.
(8) Beckwith-Hall, B. M.; Nicholson, J. K.; Nicholls, A. W.; Foxall, P. J.; Lindon, J. C.; Connor, S. C.; Abdi, M.; Connelly, J.; Holmes, E. *Chem. Res. Toxicol.* **1998**, *11*, 260−272.
(9) Hagberg, G. *NMR Biomed.* **1998**, *11*, 148−156.
(10) Howells, S. L.; Maxwell, R. J.; Peet, A. C.; Griffiths, J. R. *Magn. Reson. Med.* **1992**, *28*, 214−236.
(11) Tate, A. R.; Griffiths, J. R.; Martinez-Perez, I.; Moreno, A.; Barba, I.; Cabanas, M. E.; Watson, D.; Alonso, J.; Bartumeus, F.; Isamat, F.; Ferrer, I.; Vila, F.; Ferrer, E.; Capdevila, A.; Arus, C. *NMR Biomed.* **1998**, *11*, 177−191.
(12) Robertson, D. G.; Reily, M. D.; Sigler, R. E.; Wells, D. F.; Paterson, D. A.; Braden, T. K. *Toxicol. Sci.* **2000**, *57*, 326−337.
(13) Gartland, K. P.; Sanins, S. M.; Nicholson, J. K.; Sweatman, B. C.; Beddell, C. R.; Lindon, J. C. *NMR Biomed.* **1990**, *3*, 166−172.

10.1021/ac0156870 CCC: $22.00 © 2002 American Chemical Society
Published on Web 04/17/2002

nuclear NMR spectroscopies are widely used in the chemical structure elucidation of pure compounds, they are not currently used to perform fingerprinting from biofluids. Nevertheless, modulating the signal along a second dimension supplies more numerous resolved signals leading to seemingly deconvoluated although redundant variables. Recently, it has been used for structure elucidation of mixtures of analytes[14] and metabolites in biological fluids[15−17] and exceptionally for discriminating polyphenolic fingerprints from wine.[18] From the different heteronuclear techniques available, $^{1}H-^{13}C$ heteronuclear multiple bonding correlation (HMBC)[19−21] can be considered as an investigation tool of metabolic variations, with a particular focus on the intensity of carbohydrate backbone structural correlations that can be used as quantitative variables. From a statistical point of view, each 2D NMR spectrum becomes an individual and the different heteronuclear correlation intensities correspond to variables.

Taking into account the correlation between those responsive variables, one can perform multivariate analyses. Contrary to exploratory methods such as principal component analysis (PCA) and hierarchical clustering trees (HCT), pattern recognition methods take advantage of the analysis of the class distribution. Linear discriminant analysis (LDA)[22−24] takes into account the information contained in the data set partition. Robustness evaluation of classification algorithms is given by the error rate calculated by cross validation (CV). An unbiased error rate is estimated by predicting test samples, randomly removed from the whole set during the calibration step.

Contrary to 2D electrophoretic methods that can generate nonstructurally redundant variables due to the almost unequivocal assignment of one spot to a unique protein,[25,26] spectrometric methods are able in most cases to give different signals for a single analyte. Conversely, a fragment in MS or a 2D correlation spot in NMR can result from few overlapped or superimposed signals that are characteristic of different metabolites. Consequently, conjugation of both aspects leads to an intrinsic redundancy of information or a noisy one. The spectroscopic information redundancy coupled to the rank deficiency problem—i.e., the number of individuals is lower than the number of variables—is statistically equivalent to the problem of multicollinear variables. Therefore, the noisy information can explain a lower performance of classification algorithms of unknown individuals. Eigenvalue decomposition is solved analytically in PCA and LDA and then is very sensitive to the pseudosingularity caused by multicollinearity and rank deficiency and, therefore, needs to be performed on nonsingular variance−covariance or correlation matrices. To face this problem, a conventional strategy used in statistics is to reduce the dimensionality without a detrimental loss of information. Different ways for selecting variables based upon their ability to explain a model have been reviewed by Tate.[27] Most of these methods perform data reduction by ANOVA, by correlation analysis (Spearman or Pearson correlation coefficients), or by taking the first principal components of a PCA model before using a pattern recognition algorithm. Another strategy is based on partial least-squares (PLS) methods, essentially developed by Wold and Wold,[28,29] performing data reduction in multicollinear and rank deficiency contexts, thanks to a numeric estimation of eigenvalues validated by iterations and CV. This explains why the most recent works on metabonomics involve PLS, mainly PLS discriminant analysis (PLS-DA),[30] and soft independent modeling of class analogy (SIMCA).[7,31]

Anabolic steroids are characterized physiologically as potent hormonal disrupters that alter lipid, glucide, and amino acid metabolism.[32] Both muscle accretion and decreasing fat deposition are corresponding to morphological changes awaited in high-performance sports or in intensive cattle breeding that mainly justify their use.[33] More generally, steroids are also involved in some developmental disorders evidenced clinically in children[34] or experimentally in the rat.[35] Even sex reversal phenomena were evidenced after steroidal treatments in lower vertebrates[36−39] without a threshold dose effect in the case of 17β-estradiol.[40] Due to the pleiotropic effect of such hormones on metabolism, a global analytical approach could be of considerable help for detecting different kinds of physiological disruptions induced by anabolics and revealing at the same time a clear insight into the different physiological targets involved in endocrine disruption processes.

In the current study, we perform 2D NMR for metabonomic assessment of long-term and weak physiological variations. We generate quantitative variables by $^{1}H-^{13}C$ HMBC-NMR spectroscopy, tracing out the chemical structures of analytes quantitatively affected by such endocrine disruptions. Meanwhile, a filtering procedure of those variables is evaluated for pattern recognition

[14] Lin, M.; Shapiro, M. J. *Anal. Chem.* **1997**, *69*, 4731−4733.

[15] Holmes, E.; Foxall, P. J.; Spraul, M.; Farrant, R. D.; Nicholson, J. K.; Lindon, J. C. *J. Pharm. Biomed. Anal.* **1997**, *15*, 1647−1659.

[16] Nicholson, J. K.; Foxall, P. J.; Spraul, M.; Farrant, R. D.; Lindon, J. C. *Anal. Chem.* **1995**, *67*, 793−811.

[17] Willker, W.; Leibfritz, D. *Magn. Reson. Chem.* **1998**, *36*, 79−84.

[18] Forveille, L.; Vercauteren, J.; Rutledge, D. N. *Food Chem.* **1996**, *57*, 441−450.

[19] Bax, A.; Summers, M. F. *J. Am. Chem. Soc.* **1986**, *108*, 2093−2094.

[20] Hurd, R. A.; John, B. K. *J. Magn. Reson.* **1991**, *91*, 648−653.

[21] Hurd, R. E. *J. Magn. Reson.* **1990**, *87*, 422−428.

[22] Fisher, R. A. *Ann. Eugenics* **1936**, *7*, 179−188.

[23] Rao, C. R. *J. R. Stat. Soc. Ser. B* **1948**, *10*, 159−203.

[24] Mahalanobis, P. C. *Proc. Natl. Inst. Sci. (India)* **1936**, *12*, 49−55.

[25] Anderson, N. L.; Esquer-Blasco, R.; Hofmann, J. P.; Meheus, L.; Raymackers, J.; Steiner, S.; Witzmann, F.; Anderson, N. G. *Electrophoresis* **1995**, *16*, 1977−1981.

[26] Plomion, C.; C., P.; Brach, J.; Costa, P.; H., B. *Plant Physiol.* **2000**, *123*, 959−969.

[27] Tate, A. R. *J. Magn. Reson. Anal.* **1997**, *3*, 63−78.

[28] Wold, H. *Research papers in statistics. Festshrift for Jerzy Neuman*; John Wiley: New York, 1966; pp 411−444.

[29] Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735−743.

[30] Gavaghan, C. L.; Holmes, E.; Lenz, E.; Wilson, I. D.; Nicholson, J. K. *FEBS Lett.* **2000**, *484*, 169−174.

[31] Holmes, E.; Nicholson, J. K.; Tranter, G. *Chem. Res. Toxicol.* **2001**, *14*, 182−191.

[32] Meyer, H. H. *Apmis* **2001**, *109*, 1−8.

[33] Herschler, R. C.; Olmsted, A. W.; Edwards, A. J.; Hale, R. L.; Montgomery, T.; Preston, R. L.; Bartle, S. J.; Sheldon, J. J. *J. Anim. Sci.* **1995**, *73*, 2873−2881.

[34] Sultan, C.; Paris, F.; Terouanne, B.; Balaguer, P.; Georget, V.; Poujol, N.; Jeandel, C.; Lumbroso, S.; Nicolas, J. C. *Hum. Reprod. Update* **2001**, *7*, 314−322.

[35] Thayer, K. A.; Ruhlen, R. L.; Howdeshell, K. L.; Buchanan, D. L.; Cooke, P. S.; Preziosi, D.; Welshons, W. V.; Haseman, J.; vom Saal, F. S. *Hum. Reprod.* **2001**, *16*, 988−996.

[36] Cheek, A. O.; Vonier, P. M.; Oberdorster, E.; Burow, B. C.; McLachlan, J. A. *Environ. Health Perspect.* **1998**, *106* (Suppl. 1*)*, 5−10.

[37] Crews, D.; Bull, J. J.; Wibbels, T. *Gen. Comput. Endocrinol.* **1991**, *81*, 357−364.

[38] Bogart, M. H. *J. Theor. Biol.* **1987**, *128*, 349−357.

[39] Xu, J.; Liao, L.; Ning, G.; Yoshida-Komiya, H.; Deng, C.; O'Malley, B. W. *Proc. Natl Acad. Sci. U.S.A* **2000**, *97*, 6379−6384.

[40] Sheehan, D. M.; Willingham, E.; Gaylor, D.; Bergeron, J. M.; Crews, D. *Environ. Health Perspect.* **1999**, *107*, 155−159.

of 2D $^1H-^{13}C$ HMBC-NMR metabonomic data. The metabonomic procedure is performed on steers and cows submitted to hormonal manipulations with anabolic steroids and on their respective control.

EXPERIMENTAL SECTION

Animals and Hormonal Treatments. Animals ($n = 91$) were submitted to different conditions of physiological treatment as follows: control steers (MC, $n = 16$), treated steers (MR, $n = 37$) with Revalor-S implants (Hoechst-Roussel Vet, Sommerville, NJ) containing 17β-estradiol (24 mg) and trenbolone acetate (140 mg), control cows (FC, $n = 22$), and treated cows with testosterone enanthate (FE, $n = 16$). Treated steers were implanted with one, two, or four implants as described elsewhere.[41] Females were subjected to a unique intramuscular injection of Androtardyl (testosterone enanthate; 250 mg; Schering, Lys-lez-Lannoy, France).[42,43] Moreover, physiological variability took into account parameters such as age, breeding, feeding, and, for cows, the reproductive stage in order to increase the discrimination robustness, and therefore, those parameters were not entered in statistical models. Urine was collected on the 10th and 23rd days after treatment on steers and cows and on the 90th day after implantation on steers only. Each sample was aliquoted twice and then freeze-dried in order to obtain 500 mg of dried materials, which was dissolved in 1 mL of H_2O. Before NMR analysis, aliquots were centrifuged for 5 min at 9200*g* to remove insoluble materials. A 10% (v/v) of phosphate-buffered saline prepared in D_2O (pH 7.4) containing 10 mM deuterated trimethylsilylpropionate was added to the supernatant in order to provide an internal reference (1 mM) for ^2H lock frequency calibration.

NMR Spectroscopy. Spectra were recorded on Bruker AMX-500 and Avance-500 spectrometers operating at 500.13-MHz ^1H resonance frequency, with a *z*-field gradient facility, a reverse probe, and a temperature control held at 303 ± 0.1 K. ^1H experiments were recorded using WATER GrAdient Tailored Excitation (Watergate)[44] with three 1.2-ms *z*-gradient pulses, and 0.8-ms recovery delay, accumulating 128 FIDs into 32K points on a 10.098 ppm spectral width. Fourier transform was computed on 64K points. $^1H-^{13}C$ HMBC[19-21] spectra were recorded with a 70-ms delay for long-range $^1H-^{13}C$ coupling (^2J, ^3J, ^4J) selection, with 10.098 (^1H) and 220 ppm (^{13}C) spectral widths, for 2 h. Prior to Fourier transform (FT) in magnitude mode, the FIDs were zero-filled in the ^{13}C dimension, and a shifted sine bell function was applied to both dimensions. Further metabolite assignments were achieved thanks to $^1H-^1H$ total correlation spectroscopy (TOCSY)[45,46] and $^1H-^{13}C$ heteronuclear single quantum coherence (HSQC)[47] experiments performed on representative urinary samples.[48] To confirm the ^1H assignments, TOCSY was performed with water presaturation with a mixing time of 300 ms. Spectra were collected in the phase-sensitive mode using a time-proportional phase incrementation (TPPI).[49] The spectral width was 12 ppm, with data collected in 1024 time domain points. A total of 360 increments was measured with 80 transients per increment, the data set being zero-filled to 1024 in t_1, and a shifted sine-bell apodization function was applied prior to FT. HSQC spectra were collected with ^1H detection. A relaxation delay of 1 s and a refocusing delay of 4.2 ms were employed. A total of 1024 data points with 224 scans per increment and 280 experiments were acquired with spectral widths of 12 ppm in F2 and 140 ppm in F1. The data set was zero-filled to 1024 in t_1 and sine-bell-shifted.

Integration was performed with Aurelia/Amix (v2.8.11, Bruker SA, Wissembourg, France). 1D spectra were bucketed at 0.04, 0.02, and 0.01 ppm. HMBC spectra were integrated by Aurelia by summing all intensity values within referenced intervals surrounding correlation spots in both dimensions as described previously[18] and exported to statistical software by a C++ routine specifically developed for this purpose. Because of a high dosage of testosterone enanthate administered to cows, variables potentially overlapping spots corresponding to α-testosterone, the major metabolite from testosterone in cattle, were checked. Since statistical interpretations of discriminations between the different groups are not substantially affected when they are discarded, since their contribution in discriminant axes is weak, and since the spectra do not reveal directly any presence of α-testosterone in urine, those variables were kept for further statistical analyses. The matrices that were analyzed contained 182 rows corresponding to 91 individuals for which samples were duplicated before being freeze-dried and analyzed by NMR and 375 (2D), 870 (0.01 ppm), 435 (0.02 ppm), or 218 (0.04 ppm) columns as NMR variables.

Statistical Analyses. Multidimensional statistical analyses were performed on log-transformed variables using Splus 2000 (v2.0, Mathsoft Inc., Seattle, WA) with MASS (http://lib.stat.cmu.edu/DOS/S) and Multidim (http://www.lsp.ups-tlse.fr/Carlier/Logiciel.html) libraries and SIMCA-P (v8.0, Umetrics AB, Umea, Sweden).

Multivariate Redundancy Analysis and Variable Selection Algorithm in $^1H-^{13}C$ HMBC Spectra. Redundancy of information in 2D NMR has been explored by hierarchical ascending clustering of variables. Clusters of variables are made by a measure of distance *d*, or a dissimilarity coefficient, which is symmetric, positive with $d(A, A)$ equal to zero. Several criteria can be used to aggregate close clusters. Among these, using the largest dissimilarity distance between two variables issued from two clusters is known as the complete linkage algorithm. Pearson's correlation coefficient is symmetric, but negative correlations are allowed and $r(A, A) = r_{AA}$ is equaled to 1. To solve this, two metrics, $1 - r_{ij}$ and $1 - |r_{ij}|$ were used as dissimilarity coefficients. In brief, two variables *i* and *j* are aggregated if the correlation r_{ij} between them is high, and then the corresponding distance $1 - r_{ij}$ is low. The significance of Pearson's correlation (null hypothesis H_0: $|r_{ij}| = 0$) was checked[50] and gave therefore the significance of aggregation indexes (null hypothesis H_0: metric = 1).

[41] Maume, D.; Deceuninck, Y.; Pouponneau, K.; Paris, A.; Le Bizec, B.; Andre, F. *Apmis* **2001**, *109*, 32−38.

[42] Ferchaud, V.; Le Bizec, B.; Monteau, F.; Andre, F. *Analyst* **1998**, *123*, 2617−2620.

[43] Ferchaud, V.; Le Bizec, B.; Monteau, F.; Andre, F. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 652−656.

[44] Piotto, M.; Saudek, V.; Sklenar, V. *J. Biomol. NMR* **1992**, *2*, 661−665.

[45] Bax, A.; Davis, D. G. *J. Magn. Reson.* **1985**, *65*, 355−360.

[46] Edwards, M. W.; Bax, A. *J. Am. Chem. Soc.* **1986**, *108*, 918−923.

[47] Bodenhausen, G.; Ruben, D. J. *Chem. Phys. Lett.* **1980**, *69*, 185−188.

[48] Braun, S.; Kalinowski, H. O.; Berger, S. *150 and more basic NMR experiments. A practical course*; Wiley-VCH: Weinheim, 1998.

[49] Marion, D.; Wuthrich, K. *Biochem. Biophys. Res. Commun.* **1983**, *113*, 967−974.

[50] Dagnélie, P. *Statistique Théorique et Appliquée*; De Boeck-Université: Paris, 1998; Vol. 2.

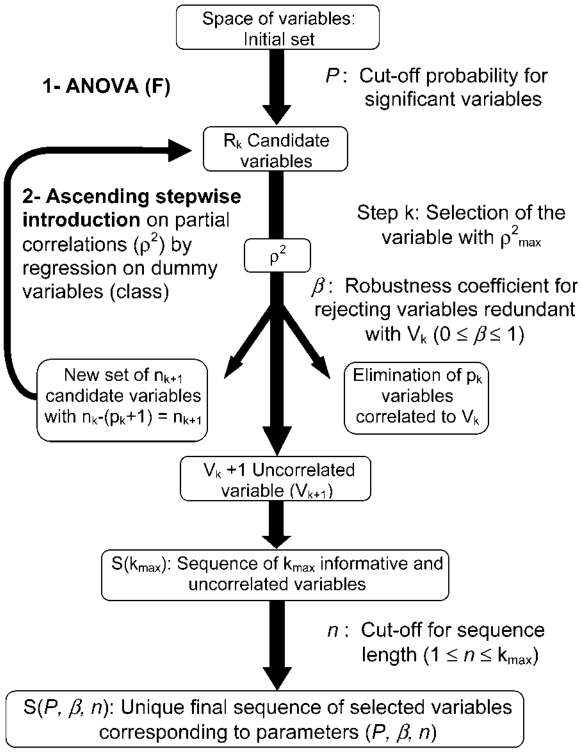
Space of variables: Initial set
1- ANOVA (F)
P: Cut-off probability for significant variables
R$_k$ Candidate variables
2- Ascending stepwise introduction on partial correlations (ρ^2) by regression on dummy variables (class)
Step k: Selection of the variable with ρ$^2_{max}$
ρ^2
β: Robustness coefficient for rejecting variables redundant with V$_k$ (0 ≤ *β* ≤ 1)
New set of n$_{k+1}$ candidate variables with n$_k$-(p$_k$+1) = n$_{k+1}$
Elimination of p$_k$ variables correlated to V$_k$
V$_k$ +1 Uncorrelated variable (V$_{k+1}$)
S(k$_{max}$): Sequence of k$_{max}$ informative and uncorrelated variables
n: Cut-off for sequence length (1 ≤ *n* ≤ k$_{max}$)
S(*P*, *β*, *n*): Unique final sequence of selected variables corresponding to parameters (*P*, *β*, *n*)

Figure 1. Algorithm used for selecting variables, adapted from Carlier.[51]

The algorithm for selecting 2D NMR variables is based upon a model explanation of the variance generated by the experimental design used. Variables were selected with a two-step algorithm (Figure 1):

(i) An ANOVA revealing candidate variables, which are significantly affected by the experimental design. The effect of the anabolic treatment with four modalities is checked for each variable. The cutoff parameter corresponds to a given probability value, P_0. Variables, which are revealed by ANOVA as being more significant than a threshold value P_0 ($P < P_0$), are considered as candidate variables and selected for the next phase.

(ii) An ascending stepwise introduction of variables. For each candidate variable, a regression is made on the dummy variable coding for the distribution describing the anabolic treatment with four levels.[51] If V_k is the explicative subspace at iteration k, X_k corresponds to the set of variables selected at step k generating V_k and R_k to the normalized variables still to be selected. The algorithm projects the remaining variables R_k on the V_k orthogonal subspace. The correlation coefficient of these orthogonal projections of variables on the V_k orthogonal subspace corresponds to the partial correlation coefficient ρ_k.[51] The variable with the strongest correlation coefficient is selected, and the ρ_k variables linearly correlated to V_k are rejected. The corresponding parameter for throwing out the correlated variables is called robustness coefficient, noted β ($0 \leq \beta \leq 1$). When β equals 1, the selection is classical, and when β is null, the throwing out is drastic.

This algorithm ends when all variables are selected or when the correlation coefficients ρ_k are smaller than a final threshold value, chosen here to 10^{-4}, considering that variables already belong to the selected subspace. In the end, each set of parameters P and β gives a sequence of $k_{\max}$ independent variables, corresponding to the iterative selection of variables. The parameter n determines the final number of explicative variables, which are effectively selected. Each set of parameters (P, β, n) leads to a unique combination of independent and informative variables.

Multivariate Pattern Recognition of $^1H-^{13}C$ HMBC-NMR Spectra. Variables were log-transformed. Instead of giving unsupervised classification of individuals that can be obtained by PCA, LDA calculates orthogonal linear combinations of variables (linear discriminants) in order to discriminate group centroids.[22,23] Those linear discriminants are computed through single value decomposition. LDA was performed using classical criteria for homoscedastic variances and *plug-in* classification rules.[52,53] Multivariate analyses were validated by 10-fold CV giving meta-statistics. Ten partitions between a training set and its complementary test set were randomly generated from the whole set of individuals. The pattern recognition calibration was run on the training set and an error rate was calculated from the test set prediction. Error rates provided for the different (P, β, n) combinations were plotted for an overall comparison by a quadratic local fitting for each β value, along the two dimensions, P and n, with a smoothing parameter equaled to 0.1.

PLS methods (PLS-DA, SIMCA) share the ability to iteratively estimate eigenvalues instead of analytically solving decomposition into singular values. PLS-DA is a multiregression of spectroscopic variables on dummy variables. SIMCA is a composite model made of class-specific PCA models. SIMCA tests individuals for belonging to each of the independent PCA models. Every significant component was included in models following Wold's methodology.[28,29] Model memberships and Hotelling's T^2 were computed with a 5% α risk.

RESULTS AND DISCUSSION

Generation of Metabolic Descriptors by NMR. Efficient 1D water signal suppression was obtained with the ^{1}H Watergate method as shown earlier,[44] and a short acquisition time (17 min) was sufficient to obtain 1D fingerprints from urinary spectra that can be then submitted to statistical analyses (Figure 2). Whatever the chemical shift interval used (0.04, 0.02 or 0.01 ppm) to bucket 1D spectra, less than 20 raw variables are potentially informative to perform subsequent classification tests (Figure 3a). In fact, ^{1}H NMR spectra of urine appears as highly complex and contain hundreds of metabolite resonances, many of which are overlapped, resulting in indiscernible multiplets (Figure 2). The main peaks arise from endogenous metabolites with a low molecular weight that are filtered by kidney glomerules before being excreted in urine. Aiming at a low signal-overlapping technique, 2D signal modulation by the ^{13}C dimension appears as a quite interesting way to give signals that are better resolved and to generate more numerous independent variables after integration. This is obtained by ^{1}H$-^{13}$C HMBC, which provides a convenient filtering procedure to observe the carbohydrate backbone of the urinary metabolites (Figure 2). Because of the ^{13}C natural-abundance rate (1.1%), inverse detection by the ^{1}H channel is chosen. Enhance-

[51] Carlier, A. In *Analyse discriminante sur variables qualitatives*; Celeux, G., Nakache, J. P., Eds.; Polytechnica: Paris, 1994; Chapter 5.

[52] Venables, W. N.; Ripley, B. D. *Modern applied statistics with S−PLUS*, 3rd ed.; Springer-Verlag: New York, 1999.

[53] Ripley, B. D. *Pattern recognition and neural networks*; Cambridge University Press: Cambridge, U.K., 1996.

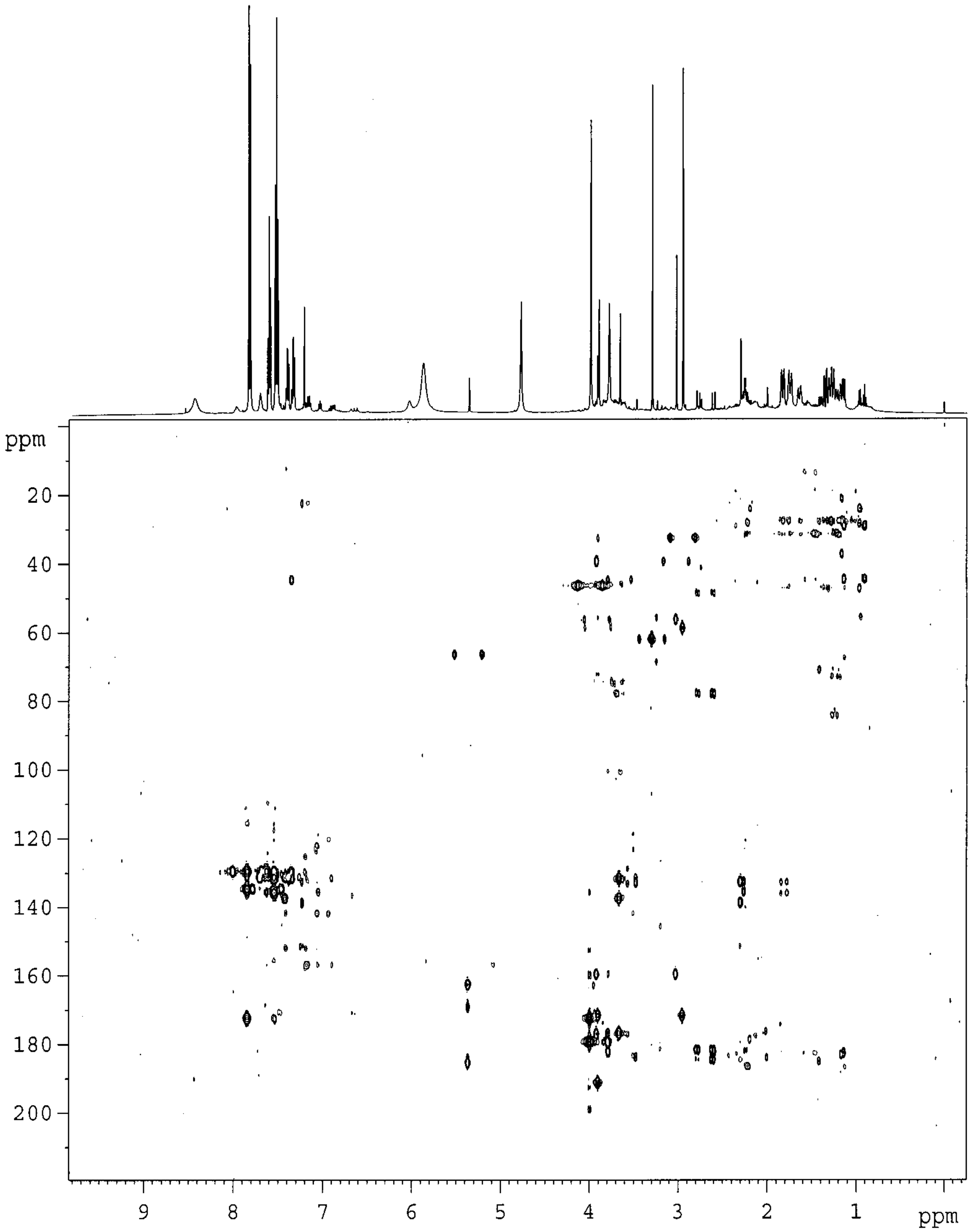

Figure 2. ^1H Watergate and $^1H-^{13}C$ HMBC-NMR spectra of urine collected from a treated steer.

ment of multiple quantum filtering and water suppression is obtained by means of a *z*-field gradient.[21] Nevertheless, to reduce the acquisition time, sample freeze-drying is necessary to concentrate urinary metabolites up to 500 mg/mL and to achieve a 2D NMR analysis in almost 2 h. Probably, for some urinary metabolites, the concentration step of samples leads to some problem of solubility when samples are reconstituted to such a concentration, and some metabolites can be lost. Furthermore, D_2O was added to samples to lock the ^2H frequency just before performing 2D NMR analysis, and some partial deuteration of acidic protons can occur during analysis. All these difficulties concerning the preparation of samples should increase the residual variance of some variables describing specific chemical functions of those analytes. Nevertheless, we can hypothesize that, in the near future, use of a cryoprobe should avoid the sample concentration step and should improve at the same time generation of more numerous relevant variables, as far as the factorial design is concerned.

Effects of anabolic treatments on general metabolism were compared by an ANOVA performed on variables generated by HMBC or by 1D NMR bucketed with 0.04, 0.02, and 0.01 ppm intervals (Figure 3).[8,15,31,54,55] Each variable is checked for the effect of treatment with four modalities and the *F*-statistic is computed. Over the first 100 raw variables with highest *F*-statistics, *F* values obtained from 2D NMR variables are greater than any of those calculated for 1D NMR (Figure 3a). Information handled by raw 2D NMR variables is clearly more significant than 1D information,

[54] Gray, H. F.; Maxwell, R. J.; Martinez-Perez, I.; Arus, C.; Cerdan, S. *NMR Biomed.* **1998**, *11*, 217−224.

[55] Maxwell, R. J.; Martinez-Perez, I.; Cerdan, S.; Cabanas, M. E.; Arus, C.; Moreno, A.; Capdevila, A.; Ferrer, E.; Bartomeus, F.; Aparicio, A.; Conesa, G.; Roda, J. M.; Carceller, F.; Pascual, J. M.; Howells, S. L.; Mazucco, R.; Griffiths, J. R. *Magn. Reson. Med.* **1998**, *39*, 869−877.

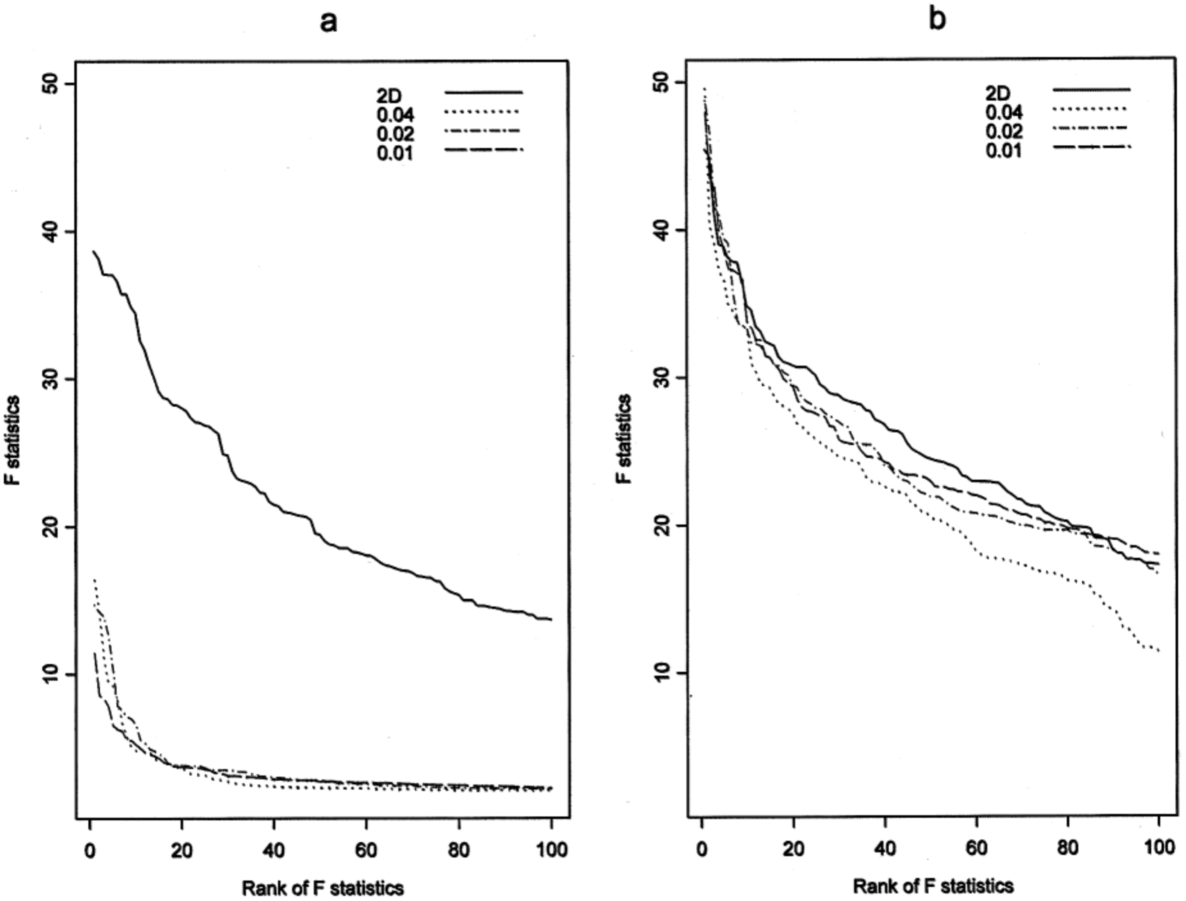

Figure 3. Reliability of 1D and 2D NMR analyses by ANOVA: raw variables (a) and Log-transformed variables (b). The effect of treatment on variables was tested by ANOVA. Fisher's *F*-statistics were performed to assess the ANOVA significance. *F*-Statistics for the different variables are ranked by decreasing values. 2D NMR as integrated by Aurelia/Amix software is compared to 1D NMR bucketed with spectral width of 0.01, 0.02, and 0.04 ppm.

which is obviously inherent to spectroscopic principles involved in such analyses. Translated into statistical language, both weak peak resolution and baseline deformation result in an increase of residual variance which corresponds to the unexplained variance calculated by ANOVA. When raw variables are log-transformed to obtain a better variance homogeneity (Figure 3b), *F*-statistics for 1D NMR variables gain almost 1 order of magnitude and become as informative as 2D NMR variables, which remains nearly at the same magnitude (+25%) as the nontransformed 2D NMR variables. These results corroborate the fact that $^1H-^{13}C$ HMBC allows a convenient spectrochemical modulation of the 1D NMR information obtained on analytes present in biofluids by giving more useful raw variables. But inherently, HMBC, as with other 2D NMR methods, generates for a same compound several signals that are statistically highly correlated.

Representation of Data Set Information Redundancy and Hierarchical Clustering of Variables. We have applied clustering methods in order to investigate the information redundancy due to the presence of highly correlated variables generated in such a physiological context of hormonal manipulation. At each step, variables or clusters of variables are aggregated to other variables or clusters on the basis of a distance or similarity metric. One suitable way to account for redundancy between variables is to use metrics drifting from Pearson's correlation coefficient r[56] or metrics such as $1 - r_{ij}$ or $1 - |r_{ij}|$ and to build complete linkage hierarchical clustering dendrograms (Figure 4a). All the variables from a cluster are correlated and can be considered as redundant. As functions of Pearson's correlation coefficient, the metrics $1 - r_{ij}$ and $1 - |r_{ij}|$ enable us to test the aggregation level significance through the significance of correlation coefficients.[50]

When the variables are clustered using the metric $1 - r_{ij}$ (Figure 4a), similar variables or clusters of variables are aggregated along a metric index in correspondence with the correlation coefficient, from the bottom to the top. At the first step, the best correlated variables, i.e., var206 and var213, have a correlation coefficient close to 1, and the significance test assumes a null probability ($P < 10^{-16}$) for this correlation being null with 181 degrees of freedom (DF). These are the first variables to be aggregated at the bottom of the tree with an aggregation index value close to zero (2×10^{-16}). This first cluster (var206−var213, $r = 1$, index = 0) is aggregated later with the cluster composed of var13−var9 ($r = 0.718$, index = 0.282) and variables var3−var16 ($r = 0.698$, index = 0.302). Index values over 1 mean negative correlations. The two clusters aggregated in the last step have an aggregation index of 1.682, which corresponds to the minimal correlation value of −0.682 between var291 and var281 ($P = 0$). The clusters aggregated in the index interval comprised between 0.876 and 1.124 are not significantly correlated. Because

[56] Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. *Proc. Natl Acad. Sci. U.S.A* **1998**, *95*, 14863−14868.

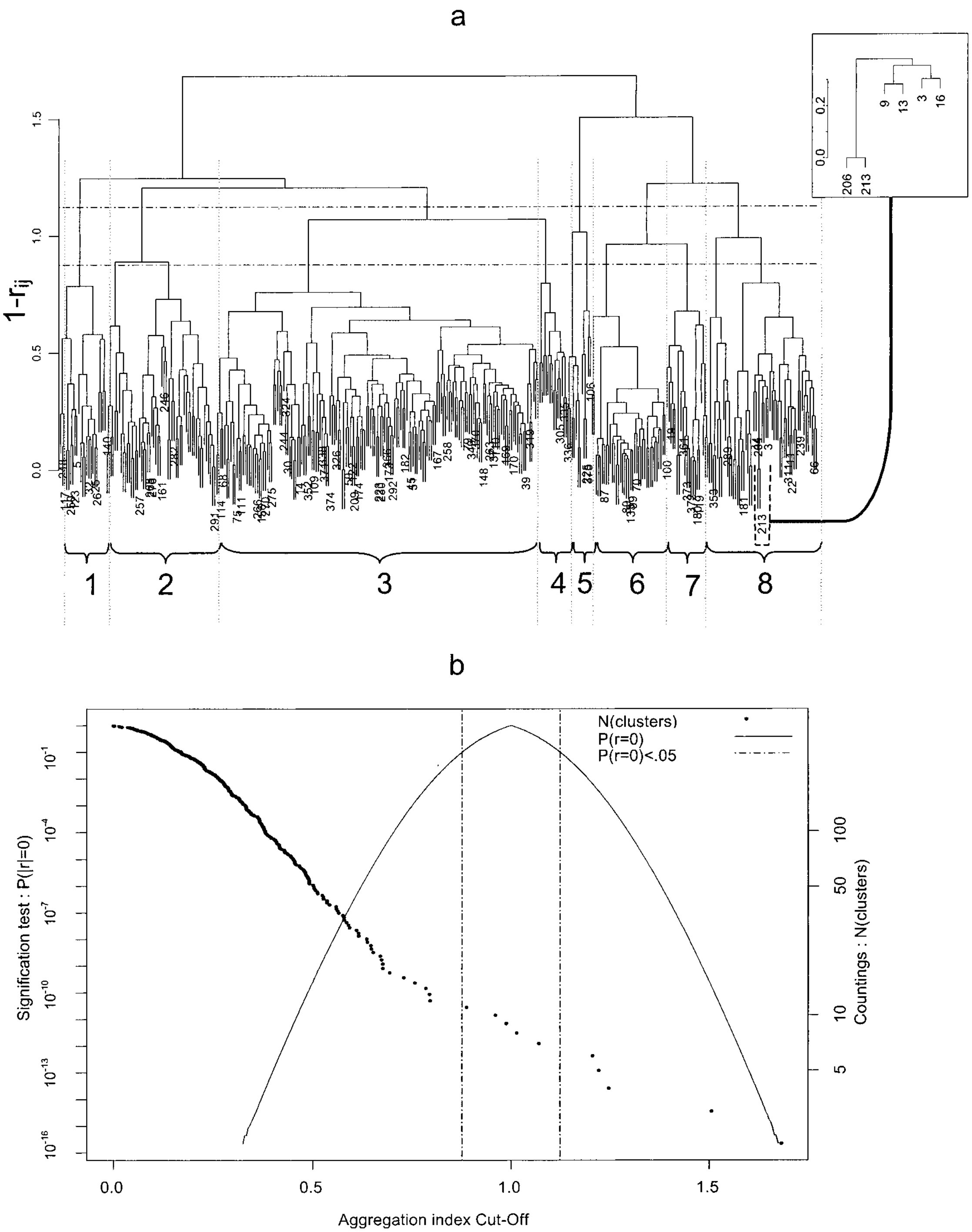

Figure 4. Clustering of variables in a physiological disruption context (375 variables). (a) Hierarchical classification trees (HCT) using the metric $1 - r_{ij}$. Note the nonsignificant region between the horizontal dashed lines. Numbered clusters correspond to fully homologous sequences of variables whatever the metric used ($1 - r_{ij}$ or $1 - |r_{ij}|$). When $1 - |r_{ij}|$ is used, clusters are sorted in the following sequence: 1 4 5 2 8 7 6 3. Labels correspond to selected variables after the parametrization step of the selection algorithm. (b) Signification of clusters and correlations. Clusters and their corresponding correlations in function of the aggregation index provide confidence intervals to the number of clusters. Probabilities with $P_{|rij|=0} < 10^{-16}$ are rounded to zero and are not shown in the logarithmic plot.

of this nonsignificant interval, incertitude lies on the negative correlation accuracy. To disregard this region and prove the accuracy of negative correlation clusters, we changed the metric $1 - r_{ij}$ to $1 - |r_{ij}|$ (not shown). This leads to a modification of the aggregation sequence of variables. Some clusters previously identified with the $1 - r_{ij}$ metric are now inserted between other contiguous clusters, with a 100% sequence homology inside each cluster. Using this sequence homology property whatever the metric used to build clusters, we isolated eight meta-clusters numbered from 1 to 8 (Figure 4a). In fact, the sequence homology inside the meta-clusters obtained with the metric $1 - |r_{ij}|$ validates the existence of negatively correlated clusters observed above the uncorrelated region with the metric $1 - r_{ij}$. Moreover, the curve of probabilities associated with the significance test of the

Table 1. Summary of Major Urinary Metabolites Identified by ^{1}H, TOCSY, HSQC, and HMBC

		chemical shift			variables,	meta-		
compound	multiplicity	^{1}H	^{13}C	assignmt	**no.**{centroid δ^{1}H, centroid δ^{13}C}	cluster	SCR_{181}	SCR_{90}
citrate	doublet	2.62	49.2	CH2	**117**{2.62, 78.7}; **250**{2.62, 184.9}; **7**{2.62, 49.2}; **249**{2.62, 181.9}	1		
	doublet	2.79	49.2	CH2	**248**{2.79, 181.9}; **123**{2.79, 49.2}; **118**{2.79, 78.7}	1	7/22	7/98
			78.7	Cq[a]				
			181.9; 184.9	CO				
allantoin	singlet	6.02		NH				
	singlet	5.36	66.5	CH	**360**{5.38, 169.3}; **362**{5.38, 186.1}; **359**{5.38, 162.3}	8	3/3	3/3
hippurate			179.7; 173.1	CO				
			136.3	Cq[a]				
	triplet	7.62	135.2	CH para	**288**{7.62, 130.2}	2	5/7	5/10
	triplet	7.53	131.8	CH meta	**283**{7.53, 136.3}; **294**{7.53, 173.1} **287**{7.53, 131.8}; **290**{7.53, 130.2}	2		
	singlet	8.02		NH	**278**{8.02, 130.2}	5		
	doublet	7.83	130.2	CH ortho	**279**{7.83, 130.2}; **280**{7.83, 173.1}; **281**{7.83, 135.2}	5	5/6	5/28
					299{7.62, 135.2}	5		
	singlet	3.98	46.5	CH2	**189**{3.98, 179.7}; **188**{3.98, 173.1}	8	2/2	2/2
creatine	singlet	3.90	57.0	CH2	**184**{3.91, 160.5}; **78**{3.91, 40.0}	7		
	singlet	3.01	40.0	CH3	**180**{3.01, 160.5}	7	3/3	3/3
					111{3.01, 57.0}	3		
					186{3.91, 177.5}	2		
			160.5; 177.5	Cq[a]; CO				
creatinine	singlet	2.94	33.3	CH3	**112**{2.94, 59.5}	2		
	singlet	3.89	59.5	CH2	**257**{3,89, 171.9}	2	2/2	2/2
					108{3.89, 33.3}	3		
					185{3.89, 191.5}	8	2/2	2/2
					181{2.94, 171.9}	8		
			171.9; 191.5	CO				
TMAO	singlet	3.28	62.4	CH3	**105**{3.28, 62.4}	2		
dimethylamine	singlet	2.72	37.8	CH3	**122**{2.72, 37.8}	3		

[a] Cq, "quaternary carbon".

aggregation index (Figure 4b) shows that both positive and negative correlations are highly significant, except for a metric index comprised between 0.876 and 1.124 where the clusters are noncorrelated. Five clusters are inside this interval and correspond to nonsignificant regrouping of some meta-clusters. Above 1.124, five other aggregations between meta-clusters are significant and correspond in fact to highly anticorrelated variables. This point demonstrates the reliability of these negative correlation coefficients mainly explained by the experimental design used to perform physiological studies.

Organization in a Common Hierarchical Model of the Structural Redundancy and Physiological Correlations. Three main results arise from clustering of variables: (i) 100% homology of sequences of variables inside each of the eight meta-clusters we have defined whatever the metric is used, which is a clue for a relevant informative system, (ii) nonsignificant aggregation of different meta-clusters underlining different levels in the information structuring, and (iii) existence of significant negative correlations. As there is no instrumental postulate leading to highly significant negative correlations in magnitude mode, the factorial structuring contained in the experimental design describing the hormonal treatments has necessarily modified in a coordinate way the composition in urinary analytes. So, since negative correlations related to biochemical pathways may occur in a context of physiological regulation,[26] we can infer thanks to NMR robustness that strong positive correlations close to 1 should correspond to variables that are structurally related. Consequently, our clustering model may oppose strong positive correlations corresponding to structural relationships between those highly correlated variables to weaker positive or negative correlations that could be explained as physiological links between such lower correlated variables. Therefore, each meta-cluster could include variables corresponding to coherent metabolic pathways or physiologically constrained metabolic networks that are relatively independent from other ones.

To confirm this hypothesis, we have characterized several metabolites from resonance assignments and examined the distribution of those variables inside the different meta-clusters. Resonance assignments were confirmed by a combination of chemical shift, spin−spin coupling patterns, coupling constants, and literature data.[7,57] To confirm structural identification (Table 1), complementary 2D experiments (TOCSY, HSQC, HMBC) were performed on some urine samples. Dimethylamine and trimethylamine *N*-oxide (TMAO) give sharp singlets for their $N-CH_3$ at 2.72 and 3.28 ppm, respectively. Creatine also gives sharp singlets at 3.01 and 3.90 ppm for its $N-CH_3$ and $N-CH_2$. They are connected in the HMBC spectrum. Allantoin gives a sharp singlet at 5.36 ppm for the CH and a large peak at 6.02 ppm for a NH.

[57] Willker, W.; Engelmann, J.; Brand, A.; Leibfritz, D. *J. Magn. Reson. Anal.* **1996**, *2*, 21−32.

Table 2. Urinary Metabolites Partially Assigned by ^{1}H, TOCSY, HSQC, and HMBC

		chemical shift					
compd	multiplicity	^{1}H	^{13}C	variables, **no.**{centroid δ^1H, centroid δ^{13}C}	meta-cluster	SCR$_{181}$	SCR$_{90}$
A	triplet	7.41	124.3; 130.2	**291**{7.41, 132.1}	2	2/2	2/2
	triplet	7.34	132.1; 138.2	**286**{7.34, 130.2}	2		
				285{7.41, 132.1}; **284**{7.41, 138.2}	3		
				156{7.34, 45.6}	3	4/18	4/44
	singlet	3.66	45.6	**190**{3.66, 177}	3		
				198{3.66, 138.2}; **199**{3.66, 132.1}	8	2/2	2/2
B	singlet	7.23	123.8; 132.7	**265**{7.23, 123.8}; **157**{7.23, 123.2}; **270**{7.23, 132.7};	3	5/22	5/41
			138.8; 151.6	**350**{7.23, 151.6}; **275**{7.23, 138.8}	3		
	singlet	2.31	23.2	**178**{2.31, 138.8}; **179**{2.31, 132.7}	8	2/2	2/2

The spin system of the aromatic protons of hippurate (7.83/7.62/7.53 ppm) can be detected in the ^{1}H$-$^{1}H TOCSY experiment. Hippurate gives also a singlet at 3.98 ppm for the aliphatic CH_2. Peaks arising from two major spin systems are also observed, but these have not yet been assigned (A and B) (Table 2). These analytes have chemical shifts at 7.41/7.34/3.66, 7.23/2.31, and 2.94/3.89 ppm, respectively. However, in the HMBC spectrum, since the suppression of correlations via ^1J(C, H) is not perfect, some residual cross-peaks caused by ^1J(C, H) coupling constants are still observable for allantoin (var158{5.54;66.5} and var159{5.22;66.5}), creatine (var119{3.17;40} and var120{2.89;40}), and hippurate (var63{4.12;46.5} and var64{3.83;46.5}). We have not applied ^{13}C decoupling, so that the spectrum displays in F_2 doublets with the spin coupling constant ^1J(C, H) and cross-peaks caused by ^1J and $^{2/3}J$(C, H) can be distinguished. Nevertheless, these correlations were considered as supplementary variables that are coherently clustered in meta-clusters 8, 7, and 2 for allantoin, creatine, and hippurate, respectively.

The variables issued from citrate and allantoin (Table 1) consecutively aggregate in their respective meta-clusters 1 and 8, which corroborates our hypothesis. Nevertheless, the hippurate and creatine variables do not perfectly match this model. Some variables from hippurate are aggregated in meta-cluster 2, but other variables aggregated in the same way belong also to meta-clusters 5 and 8. The same observation arises from creatine with two variables in meta-cluster 7 and one variable in meta-clusters 2 and 3.

Such mismatches raise the question of the residual variances seen in ANOVA that reveal significant imprecision in measurement of variables in addition to the inherent biological variability existing among individuals. The consequence is a detrimental effect on both the clustering method and the hierarchical model for aggregation of variables concerning a putative organization of variables between different structural sets on one hand and metabolic relationships on the other hand (Figure 4 and Table 1). We can hypothesize that residual variances could be significantly decreased by increasing the degrees of freedom of the data set. So, when we performed HCT variable aggregation sequences with half the data set (one aliquot per animal—90 DFs—instead of two), the performance is severely decreased (Table 1). To quantify the differences in aggregation we obtained, we defined a sequence contiguity ratio (SCR) that corresponds to the number of variables belonging to a metabolite (m) over the length of the minimal sequence of consecutive variables ($l_{\min}$) including the m metabolite variables: $\mathrm{SCR}_{\mathrm{DFs}} = m/l_{\min}$. For the variables characterizing hippurate in meta-cluster 2, $\mathrm{SCR}_{181} = 5/7$ with 181 DFs, which is better than $\mathrm{SCR}_{90} = 5/10$ with half the data set. The same observation is made for hippurate variables belonging to meta-cluster 5 with, respectively, 5/6 and 5/28 for SCR$_{181}$ and SCR$_{90}$ and for citrate: $\mathrm{SCR}_{181} = 7/22$, $\mathrm{SCR}_{90} = 7/98$. Compounds A and B (Table 2), which are not structurally characterized, and creatinine also follow this rule. Then, we can hypothesize that computing the correlation matrix with more DFs, i.e., one or a few thousands of individuals instead of hundreds, may lower considerably the influence of this “statistical noise” and reach a better fit of the hierarchical model of correlations running from structural links to physiological relationships. Therefore, we can assume that, in large databases containing 2D NMR fingerprints, such a factorial structuring of the data set should reveal chemical structures of analytes explaining metabolic contrasts linked to biological disruptions.

As a complementary approach, we computed PCA and projected variables belonging to each meta-cluster on principal components (PC) (plots not shown). As anticipated by variable clustering (Figure 4), most of the variables belonging to meta-clusters 5, 1, 4, and 6 are respectively attributed to PC 2, 3, 4, and 5. Yet, variables from other meta-clusters can be also projected on the same PC. This result corroborates conclusions obtained above from HCT concerning the nonperfect segregation of variable groups explained by analytes or structurally related compounds.

Nevertheless, if we take into account the experimental design to segregate the different groups of individuals, then correlation-based HCT on metabolites should provide suitable dendrograms for analysis of physiological effects as has been shown in functional genomics.[56] Yet, the existence of such an information redundancy highlights the need to get a filtration of variables prior to performing classification of individuals.

Variable Selection Algorithm Specificities. Informative variables were selected from the whole fingerprint and redundant variables were rejected to avoid pseudosingularity of the variance−covariance matrix in LDA. Different ways for selecting variables from the meta-clusters were compared in order to perform a statistically convenient variable filtration. First, we randomly picked out variables. Second, we performed ANOVA on the different variables and kept the most affected. We applied a third method consisting of a stepwise selection of variables[51] within each meta-cluster. To evaluate the efficiency of these three modes of selection of variables, we computed LDA CV error rates taking

Table 3. Summary of LDA Cross-Validation Error Rates on Selected Variables from Meta-Clusters with Several Criteria[a]

no. of variables/ meta- cluster	error rates (%) calculated with different selection criteria									
			stepwise with the following β values							
	random	ANOVA	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3
1	39.0	31.9	37.9	32.4	32.4	30.2	30.2	28.6	28.6	28.6
2	30.8	27.5	25.3	28.0	31.3	25.3	25.3	26.4	26.4	28.0
3	31.9	26.9	26.3	25.8	23.6	22.0	22.0	23.1	23.1	NC
4	31.3	23.6	20.3	22.0	18.1	17.0	16.5	16.5	16.5	NC
5	30.8	31.3	22.5	23.1	17.6	17.6	20.3	20.3	20.3	NC
6	26.9	28.0	17.6	19.2	14.8	16.5	19.8	18.1	18.1	NC
7	29.1	29.1	17.6	19.2	16.5	17.0	18.7	17.6	NC	NC
8	30.8	31.3	16.5	19.8	17.6	21.4	20.9	NC	NC	NC
9	26.9	31.9	15.4	16.5	19.2	23.6	NC	NC	NC	NC
10	30.2	33.0	20.9	21.4	19.2	21.4	NC	NC	NC	NC
min−max	26.9−39.0	23.6−33.0	15.4−37.9	16.5−32.4	14.8−32.4	16.5−30.2	16.5−30.2	16.5−28.6	16.5−28.6	28.0−28.6
mean ± SD	30.8 ± 3.4	29.5 ± 2.9	22.0 ± 6.7	22.7 ± 4.8	21.0 ± 6.1	21.2 ± 4.4	21.7 ± 4.3	21.5 ± 4.6	22.2 ± 4.7	28.3 ± 0.4

[a] From 1 to 10 variables were selected within each cluster following random, ANOVA, or stepwise introduction criteria. Error rates are given in percent. SD, standard deviation; nc, not computable for this condition.

up to the top 10 variables per meta-cluster for each criterion (Table 3). Because of the great number of combinations of individuals to be tested, a 10-fold CV was retained as recommended recently,[52] avoiding time-consuming calculations involved in the simpler leave-one-out CV. Furthermore, the error rates are reduced and then become more reliable.[52] The “random” criterion gave a minimum of 27% error rate with six variables, whereas the ANOVA method reached 24% error rate with the first four affected variables per meta-cluster. The stepwise introduction of variables led to 15% error rate with six variables selected with a robustness parameter of 0.8. Highest standard deviations resulting from the stepwise criterion indicate that this method should potentially enable lower error rates. Random and ANOVA criteria for the selection of one variable per cluster were advised by Tate.[27] Stepwise introduction without clustering of variables was recommended by Carlier,[51] but the main disadvantage of this method is to enable the selection of nonsignificant variables, which could be interpreted as a noisy information introduction.

Finally, as an alternative to PLS methods[58] that handle multicollinearity, we used a two-step algorithm for selecting variables before performing classification methods. This filtering step of variables was done from their supposed biological and spectrochemical characteristics as discussed above, and two constraints were used: (i) variables must be informative, and (ii) variables shall not introduce redundancy in the data set (Figure 1). In this algorithm, variables are no longer selected from inside the different meta-clusters but from the entire space of variables. This algorithm, combining an ANOVA on the different variables included in the initial data set and a stepwise introduction of informative variables, leads to a selection of the most informative variables and a parallel throwing out of variables correlated to those that were retained.

Estimation of selection algorithm efficiency (Figure 1) and its accurate optimization was still assessed by LDA CV after the selection step of variables on this data set. Error rates were computed for each combination of the three parameters (P, β, n) and their influence on CV error rates was modeled, to find optimal sets of parameters. Preliminary results show that the most salient differences occur for ANOVA cutoff probabilities comprised between 10^{-8} and 10^{-1} and for β robustness values adjusted between 0.6 and 1. For $\beta < 0.6$, the selection is so robust that too few variables are selected. A rapid control of the results showed that LDA can even reach a 0% error rate with triplet (P, β, n) having (10^{-2}, 1, 100−115) or (10^{-1}, 0.9, 65−83) values. However, we rather targeted optimal parameters regions under a given error rate, i.e., 0−5%, than combinations providing suitable but local minimal error rates.

We first evaluated the dependence of the error rate on the number of selected variables (Figure 1). Fixing the other parameters P and β, we incremented n and calculated LDA error rates for each value. With (P, β) = (10^{-6}, 1), the LDA CV error rates are greatly affected by n (Figure 5a). In the first decreasing part of the curve, each added marginal variable increases the performance of the pattern recognition. There is not enough information to correctly fit the model and then to distinguish the different hormonal treatments. The system is clearly underinformed. In the second part, each extra variable increases the error rate. The information is precise enough to discriminate the different groups, but the model only fits the training set and it is not adaptive enough to correctly classify individuals of the test set. Such a system is overinformed. With (P, β) fixed to (10^{-6}, 1), n values comprised between 29 and 92 give an error rate lower than 15%, but the triplet (P, β, n) = (10^{-6}, 1, 29−92) cannot be considered as a sufficient optimal parameter set, regardless of β.

So, we examined the dependence of the error rate on the interaction between n and P. When fixing β, error rates are fitted by local quadratic regression, and response surface can be deduced. With $\beta = 0.8$, following a contour line gives the (P, n) doublets needed for the corresponding error rate calculated by LDA (Figure 5b). For instance, the doublets (10^{-1}, 30), (10^{-4}, 50), and (10^{-5}, 50) provide a LDA CV error rate of 5%. It appears quite paradoxical to introduce nonsignificantly affected variables (P = 0.1) in the discrimination and reach good results for classification.

[58] Tenenhaus, M. *La régression PLS. théorie et pratique*; Technip: Paris, 1998.

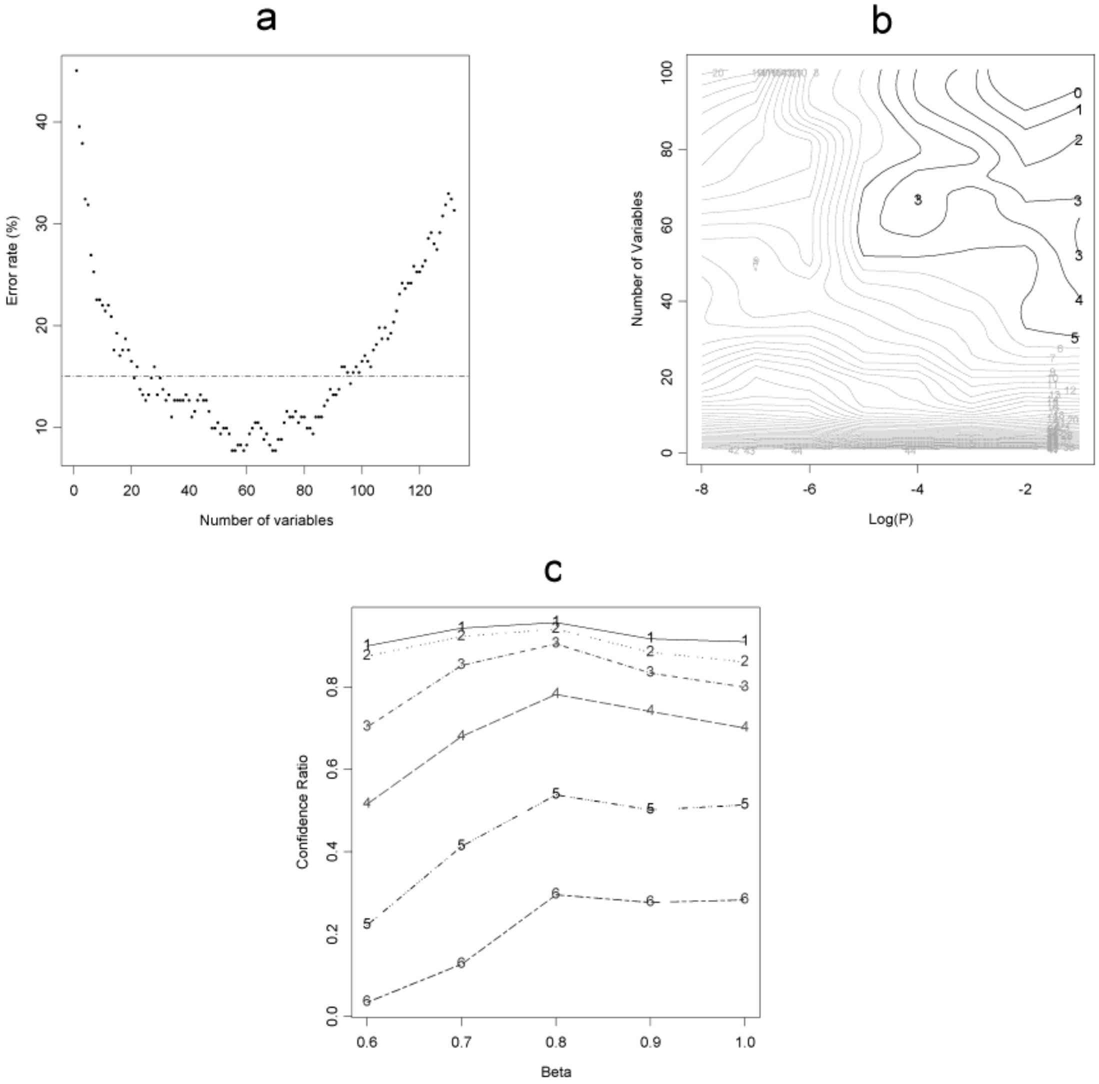

Figure 5. Optimization of parameters (P, β) for selecting variables (n) by a pattern recognition procedure. (a) Dependence of the LDA error rate on the number n ($1 \leq n \leq k_{\max}$) of the first correlated variables introduced in the system for (P, β) = (10^{-6}, 1) (an optimal region is defined under the 15% error rate dashed line). (b) Dependence of the LDA error rate on P and n. Surface responses were calculated by local regression models for $\beta = 0.8$. Note the upper right corner region under 5% error rate (black lines), which is considered for defining the optimal values of parameters P and n. (c) Optimization of filtration estimated by confidence ratios (CR_α) for LDA. The α-risk values taken for CR_α calculations are 30 (1), 25% (2), 20 (3), 15 (4), 10 (5), and less than 5% (6). Pattern recognition confidence ratios are plotted for each β value, with different error rate thresholds. Maximal values mean the greatest proportion of area under the threshold error rate and correspond to optimum parameters.

But for more restrictive conditions, i.e., 1% error rate, optimum discrimination is located in the optimal region given by the doublet (P, n) = (10^{-2}, 100). Use of more significantly affected variables ($P < 10^{-3}$) does not lower significantly the LDA CV error rates, which is also paradoxical but mostly observed in practice.[59] Consequently, the P and n values giving LDA CV error rates lower than 5% define an optimal region in the upper right corner of the surface response plot reaching 29.6% of the total area (Figure 6b). This percentage defines a supplementary parameter we have called *response surface confidence ratio for 5% error rate* (CR_5). It can be considered as a reliability estimator of the β value.

Ultimately, using the CR_α synthetic parameter, we optimized filtration parameters for LDA. Response surface CR_α [54,55] calculated for each β value, summarizing ~10 000 CV analyses (Figure 5c). LDA confidence ratios are over 0.8 for CR_{25} and CR_{30}. It means that more than 80% of the (P, β, n) conditions tested gave error rates lower than 25 or 30% by LDA CV. Whatever the error rates, CR_α reaches a maximum for $\beta = 0.8$ (Figure 5c).

A lot of classification methods such as multiple logistic regression (MLR),[52,53] k-nearest neighbors (k-nn),[53] learning vector quantization (LVQ),[60] neural networks (NN),[61] and classification and regression trees (CART)[52,53,62] have been ascribed in order to predict class of unknown individuals. As LDA, some of these classifiers are parametric methods (MLR, CART) when LVQ and k-nn are considered as nonparametric ones. In most cases, nonlinear methods such as NN and LVQ are the most adaptive.[52,63] In an extended study taking LDA as a reference, we optimized

[59] McLachlan, G. J. *Discriminant analysis and statistical pattern recognition*; Wiley-Interscience: New York, 1992.

[60] Kohonen, T. *Proc. IEEE* **1990**, *78*, 1464−1480.

[61] McCulloch, W. S.; Pitts, W. *Bull. Math. Biophys.* **1943**, *5*, 115−133.

[62] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and regression trees*; Wadsworth and Brooks/Cole: Monterrey, CA, 1984.

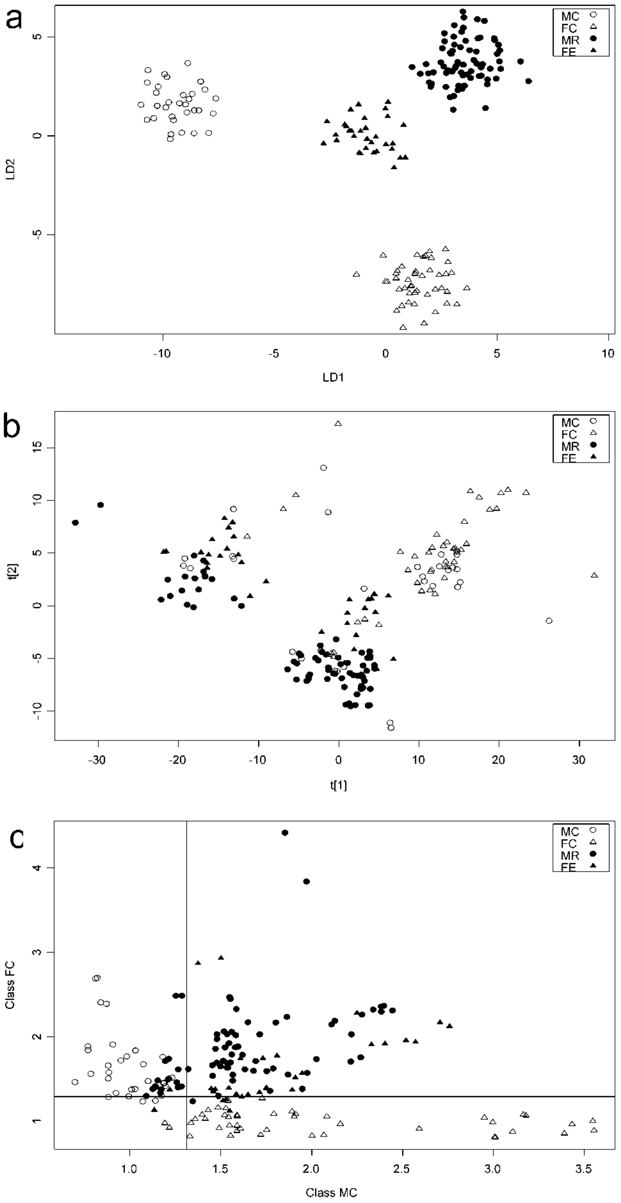

Figure 6. Classification performances by LDA on selected variables and PLS-DA. (a) LDA plot with $(P, \beta, n) = (10^{-2}, 0.9, 105)$, proportion of variance on the two first discriminant axes: 44 (LD1) and 35.2% (LD2). (b) PLS-DA plot, proportion of variance on the two first discriminant axes: 44.5 (t[1]), and 9.7% (t[2]). (c) Cooman plot for sex (castrated males and females) of treated animals (steers and cows) for identifying unknown classes. Hotelling's T^2 5% boundaries define four quadrants: upper left corner for class MC, lower right corner for class FC, lower left corner for both classes, and upper right corner nonmembership to both classes. Control males (MC), control females (FC), males treated with Revalor-S implants (MR), and females treated with testosterone enanthate (FE).

the efficiency of the algorithm and compared it to these other classical pattern recognition methods (not shown). None of these classifiers had better performances than LDA, and these are sorted by decreasing order as follows: LDA > MLR > NN > *k*-nn > LVQ > CART. LDA clearly passes NN and MLR parametrized by neural networks, which are generally considered as suitable benchmarks for performance assessment between discriminating methods.[52,63] Then, *k*-nn and LVQ are only suited for high error rates with unfortunately weak performances. For CART, which is made by a single split of distribution over a single variable at each node, the great number of variables decreases classification performances. In fact, there are too few independent variables for a suitable discrimination when the CART method is used. For all these methods, optimal parameters belong to the triplet $(P, \beta, n) = (10^{-1}-10^{-2}, 0.8-0.9, 80-110)$ and do not seem to depend on the pattern recognition device. Parametrization by minimizing the CV error rate is asymptotically equivalent to a parametrization obtained by minimizing the Akaike's information criterion (AIC).[64] The accuracy of this approach can be advantageously observed by the regular distribution of selected variables inside the different meta-clusters drawn in the classification tree (Figure 4a).

Comparison of LDA on Selected Variables versus PLS Methods. Contrary to LDA, PLS manages multicollinearity and performs data reduction without any prior variable filtration step.[29] The closest analysis to LDA is PLS-DA, which is PLS2 multiregression on dummy variables.[58] LDA provides seemingly better plots (Figure 5a) than PLS-DA (Figure 5b) for their ability to discriminate the different groups. But PLS-DA reveals a within-group heterogeneity linked to information concerning time-delayed collection of urine (10th, 23rd, or 90th day) and variation in the number of implants administered (1, 2, or 4) that are completely lost in LDA. In fact, within-group covariance matrices are considered as identity in LDA[53] whereas PLS do not formulate such a hypothesis.[65] This lack of interpretability has been already pointed out[66] and may be related to the nonlinear modeling involved by PLS.[65] In terms of model parsimony, LDA summarizes between-group variance (LD1 44%, LD2 35.2%) in a more efficient way than PLS-DA (t[1] 44.5%, t[2] 9.7%) on the first two discriminant axes. SIMCA independently models each group and tests the membership of a sample to each modeled class. For more than two classes, interpretation of Cooman plots becomes more complex, but segregating control males from control females seems to be quite efficient (Figure 5c). From the Cooman plot, we can consider that treated animals belong to neither the control male group nor to female group. LDA plots clearly show its ability to segregate groups when PLS-DA underlines especially the within-variance specificity, SIMCA being the only one to point out membership to none of the modeled classes. Yet, some treated animals are incorrectly predicted as controls and therefore illustrate the false negative problem of classification of unknown individuals as it appears with LDA if the training set do not incorporate fully characterized positive individuals.

In a last estimation of classification performance, 10-fold CV error rates for PLS-DA and SIMCA were computed using the same partitioning as in LDA CV error rate calculation. LDA reaches 0.6% error rate whereas PLS-DA and SIMCA are less efficient with respectively 19.8 and 13.9% error rates. In LDA, CR_{20} corresponds to confidence ratios comprised between 0.7 and 0.9, depending on β (Figure 5c), which means that 70−90% of the (P, β, n) combinations tested during the filtration step generate lower error

(63) Reibnegger, G.; Weiss, G.; Werner-Felmayer, G.; Judmaier, G.; Wachter, H. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 11426−11430.

(64) Celeux, G. In *Analyse discriminante sur variables qualitatives*; Celeux, G., Nakache, J. P., Eds.; Polytechnica: Paris, 1994; Chapter 1.

(65) Frank, I. E.; Friedman, J. H. *Technometrics* **1993**, *35*, 109−135.

(66) Mello, K. L.; Brown, S. D. *J. Chemom.* **1999**, *13*, 579−590.

rates and better discriminations than PLS-DA (19.8% errors). In the same way, depending on the β value, from 20 to 75% of filtration combinations give equivalent or better CV values than SIMCA (13.9% error rate).

In this CV procedure, variables were selected from the complete data set, whereas the number of individuals used to perform classifications is lower in the CV calibration step. In a generalization of the CV procedure with larger data sets, one should perform selection of variables just before LDA inside each partition. Performances of CV classification are surprising because PLS methods are often supposed to be more efficient than classical ones in terms of model parsimony and efficiency of prediction.[65] Indeed, PLS models are fitted on the predicted variance (Q^2), which is determined by CV rather than on the observed variance (R^2) explained by the model.[58] In fact, sets of accurate filtration parameters are chosen by extensive use of CV on LDA, and parameters are tested through their ability to explain not only the observed but also the predicted variance: sets of accurate parameters are kept when predicting efficiently the variance and groups. Then, LDA is successfully performed on selected variables, with an approach close to Wold or Akaike's methodologies.[28,29,64]

One could define a typical context for using these complementary methods. For classification purposes performed on between-group variance, only LDA enables the best segregation of groups. However, PLS-DA can underline in addition some heterogeneity within groups. Last, SIMCA fills the lack issued from these two classification methods, by identifying individuals belonging to none of the modeled classes and might be helpful to detect new anabolic treatments or unknown physiological disruptions.

CONCLUSION

This work validates spectroscopic and statistical methods that can be used for a metabonomic purpose. Despite longer acquisition times than with ^{1}H NMR, ^{1}H−^{13}C and ^{1}H−^{1}H 2D NMR can be easily involved in metabonomic studies aside from 1D NMR, especially ^{1}H−^{13}C HMBC. 1D rapid acquisition and SIMCA analyses can be used as efficient diagnostic tools, especially to underline nonmodeled classes as shown recently.[7,8,31] ^{1}H−^{13}C HMBC-NMR enables a like spectroscopic deconvolution of variables and therefore enhances their ability to underline a physiological contrast. Analysis of statistical correlations by variable clustering in a disturbed physiological system evidences a highly structured organization of metabolic variables in meta-clusters. Providing further studies involving thousands of samples, meta-clusters should easily reveal endogenous metabolite structures from their statistical redundancies. The algorithm for selecting variables proposed here allows LDA to achieve better performances than PLS-DA or SIMCA in most cases. PLS-DA underlines small variations within groups whereas LDA rather seems to be a method of choice for graphical interpretations of groups and decision quantification. This tool is designed for biological situations where an understanding of the metabolic impairment in response to endocrine disruptions is awaited. ^{1}H−^{13}C HMBC-NMR combined with variable filtration algorithms and pattern recognition devices should provide useful interpretations in the analysis of metabolic homeostasis modifications during hormonal treatment with anabolics and more generally in the field of long-lasting weak physiological variations. Therefore, in management of the doping control, this method could constitute a valuable complementary tool to the available techniques devoted to direct detection of hormone traces[41] in biofluids.

ACKNOWLEDGMENT

The authors thank Dr. D. Maume and Dr. B. Le Bizec for providing samples and Dr. O. Convert, Dr. D. Bertrand, and Dr. J. Mallard for helpful discussions. INRA and ENV-Nantes supported this research through a grant from European Commission (B6-7920/98/000823).

Received for review November 14, 2001. Accepted January 31, 2002.

AC0156870