OMICS A Journal of Integrative Biology
Volume 8, Number 3, 2004

Development of Biomarkers Based on Diet-Dependent Metabolic Serotypes: Characteristics of Component-Based Models of Metabolic Serotypes

UGO PAOLUCCI,[1] KAREN E. VIGNEAU-CALLAHAN,[2] HONGLIAN SHI,[1] WAYNE R. MATSON,[2] and BRUCE S. KRISTAL[1,3]

ABSTRACT

Our research seeks to identify a serum profile, or serotype, that reflects the systemic physiologic modifications resultant from dietary restriction (DR), in part such that this knowledge can be applied for biomarker studies. Direct comparison suggests that component-based classification algorithms consistently out-perform distance-based metrics for studies of nutritional modulation of metabolic serotype, but are subject to over-fitting concerns. Inter-cohort differences in the sera metabolome could partially obscure the effects of DR. Further analysis now shows that implementation of component-based approaches (also called projection methods) optimized for class separation and controlled for over-fitting have >97% accuracy for distinguishing sera from control or DR rats. DR's effect on the metabolome is shown to be robust across cohorts, but differs in males and females (although some metabolites are affected in both). We demonstrate the utility of projection-based methods for both sample and variable diagnostics, including identification of critical metabolites and samples that are atypical with respect to both class and variable models. Inclusion of non-statistically different variables enhances classification models. Variables that contribute to these models are sharply dependent on mathematical processing techniques; some variables that do not contribute under one paradigm are powerful under alternative mathematical paradigms. In practical terms, this information may find purpose in other endeavors, such as mechanistic studies of DR. Application of these approaches confirms the utility of megavariate data analysis techniques for optimal generation of biomarkers based on nutritional modulation of physiological processes.

INTRODUCTION

DIETARY RESTRICTION (DR) exerts protective effects against cancer (Yu, 1994; Weindruch and Walford, 1988) that appears dominant to genetics (Fernandes et al., 1995), carcinogen exposure (Kritchevsky et al., 1984), and specific components of the diet (Klurfeld et al., 1987). Although the majority of past research on DR has focused on elucidating the underlying mechanisms by which it influences systemic processes of an organism, our fundamental aim has been to characterize a predictive model of DR through

[1]Dementia Research Service, Burke Medical Research Institute, White Plains, New York.
[2]ESA, Inc., Chelmsford, Massachusetts.
[3]Departments of Biochemistry and Neuroscience, Cornell University Medical College, New York, New York.

analysis of the metabolome, in part so that these profiles could be used as biomarkers. Advances in high-throughput screening techniques have facilitated the generation of large volumes of quantitative data. A practical application with respect to analysis of proteins is exemplified by studies that illustrate that pathological changes in ovarian cancer can be detected through algorithmic analysis of proteomic patterns in serum (Petricoin et al., 2002). Recent data suggest that, at least at the population level, markers associated with DR define segments of the population with distinctive longevity (Roth et al., 2002).

Given that DR is a macronutrient effect, we have predicted, and then provided support for, the concept that DR will be associated with shifts in the sera metabolome. Serotypes are being identified and characterized using HPLC coupled with coulometric electrochemical array detectors (Vigneau-Callahan et al., 2001; Shi et al., 2002a,b). Previous research has identified analytically valid metabolites (Vigneau-Callahan et al., 2001), demonstrated proof of principle (classification accuracy in the cohorts in which the markers were developed) (Shi et al., 2002b), and validated these markers in independent cohorts (Shi et al., 2002a). The accompanying studies identified and partially eliminated inter-cohort differences in these markers and demonstrated that these markers could comprise useful expert systems, ie they could be used to develop algorithms capable of objective prediction (Shi et al., 2004; Paolucci et al., 2004). Perhaps not surprisingly, work presented in these last two reports showed that the effects of DR on the metabolome are modulated by the gender of the animal and by environmental factors (strictly speaking, we have shown the double negative, that these markers are insufficiently robust as to be unaffected by these factors). Thus, although the DR signal can be readily perceived mathematically in mixed derivation datasets, it is already apparent that more defined sample sets will yield stronger data.

Earlier studies revealed the utility of component-based mathematical approaches (e.g., principal components analysis, PCA), also called projection methods, for a better generation, understanding, and utilization of metabolic serotypes that reflect nutritional modulation (Shi et al., 2002a,b). Projection methods have several qualities that suggest their potential utility for the studies we wish to conduct (Eriksson et al., 2001; Manly, 2000). First, they compress large and complex datasets by identifying linear mathematical combinations of variables that are capable of replacing the majority of a dataset. Second, they enable objective diagnostics of both observations (e.g., which samples are typical, which are not) and variables (e.g., which variables behave consistently across datasets, and which do not). Third, projection methods can be used to build and/or optimize classification models, and these models can enable objective evaluation of both observations (e.g., which samples fit predicted categories and which do not) as well as variables (e.g., which variables contribute to distinguishing class, and which do not). Projection methods do, however, also have potential drawbacks, including their potential to over-fit models. (Soft Independent Modeling of Class Analogy [Wold, 1976], a supervised version of PCA, constructs models based on pre-defined [i.e., pre-assigned, supervised] samples, eg, an AL rat. While PCA calculates principal components on a whole data set, SIMCA generates principal component models for each training set class. Whereas PCA can only describe a dataset and thus provide visual information as to the relationship between a new sample and members of the training sets, SIMCA predicts class membership of a new sample, or indicates that a new sample is not a member of the training classes.) In the accompanying study (Shi et al., 2004), we showed that SIMCA-based models could distinguish sera derived from *ad libitum* (AL) fed animals from those undergoing DR, but that it also tended to derive models that are too specifically based on training set data (termed "overfitting" the models).

Thus, we now address the question of whether component based approaches can be successfully optimized for generating metabolic serotypes that reflect DR and that could then be evaluated as biomarkers. This problem has three aspects that must be addressed: (1) Can we avoid over-fitting concerns without compromising the model? (2) Can we optimize the use of component-based models for classification purposes? (3) Can we deal with the inter-cohort and sex-specific differences that distinguish cohorts and lead to over-fitting?

MATERIALS AND METHODS

All aspects of the animal husbandry and basic HPLC analysis in this work have been previously presented (Vigneau-Callahan et al., 2001; Shi et al., 2002a,b). Data analysis presented was conducted using SIMCA-P9/P10 (Umetrics, Inc., Kinnelon, NJ).

RESULTS AND DISCUSSION

Identification of serotypes to distinguish caloric intake

In our initial studies, we focused on testing the effects of DR on the metabolome of each gender independently. Single-cohort, single-gender studies were used to optimize power to identify those metabolites that might help distinguish AL and DR serotypes. For long-term use, however, we require serotypes with robust ability to distinguish caloric intake. To approach this problem, we took the previously identified male and female serotypes and merged them, then removed extraneous metabolites based on overlap between the serotypes (Vigneau-Callahan et al., 2001; Shi et al., 2002a,b). We further identified and removed metabolites that were essentially represented twice in the dataset due to their dual reactivity on multiple channels of the coulometric array (termed back-waves; Paolucci et al., 2004). We further attempted to normalize each sample within each of these datasets to eliminate previous observed sensitivity to specific cohorts (the profiles and approaches used were sufficiently sensitive so as to separate not only AL from DR, but also each of the cohorts from which they were derived; Paolucci et al., 2004). Normalization to serum tyrosine concentration eliminated ~50% of the inter-cohort variation (Paolucci et al., 2004).

Once completed, combining the data and normalization yield a profile composed of 93 metabolites that had been scored in both male and female rats from three independent cohorts. We further defined profiles based on whether metabolites had p values of <0.2 in males or females. This led to the creation of three profiles: (1) 93 metabolites, containing all metabolites originally identified as significant in either male or female rats; (2) 56 metabolites, containing the subset of the 93 metabolite profiles that had p values of <0.2 in the combined female dataset. This data set is only used in studies of female rats; (3) 37 metabolites, containing the subset of the 93 metabolite profiles that had p values of <0.2 in the combined male dataset. This data set is only used in studies of male rats.

Cohort-specific effects

Neither the 93 metabolite data set, or the two subsets optimized for gender-specific studies (37, 56 metabolites), were free of cohort specificity (Fig. 1). Thus, neither normalization nor optimization eliminates the noise in the analysis introduced by cohort specific-effects on the metabolome. This noise is highlighted in Figure 1, in which we clearly see the ability of PCA to distinguish the individual cohorts. This suggests that we need to turn to methods designed to focus on the separation between AL and DR.

To ensure that the inter-cohort variability did not completely obscure the AL-DR signal, we re-examined male only and female only sets using PCA. These data clearly reveal that PCA is capable of partially distinguishing groups in both male and female rats, but that the increased noise from the different cohorts does considerably weaken the resolving power of this approach as compared with the previous studies that were based on single sexes and single cohorts. Furthermore, the use of the smaller "optimized" datasets did not aid in separation (not shown).

Gender-specific effects

A second issue arose as we looked in more depth at the relationships between gender and diet. As shown in the accompanying report, many (40/93, 43%) of the markers that we study differ by $p < 0.01$ between male and female animals when the animals are fed AL, but fewer (12/93, 13%) differ between the sexes when the animals are maintained on DR feeding regimens. Coupled with other data presented in the accompanying paper, this result suggests that the ideal case—identifying truly sex-independent models for DR, is not feasible (or arguably, is not optimal), at least at this point. This point was addressed directly using PCA on combined data sets. These datasets could not be used to distinguish diet (not shown), because the mathematical planes of separation between AL and DR feeding groups are different in males and females (not shown). For this reason, the use of sex-blind (i.e., mixed sex) datasets will result in significant loss of resolving power.

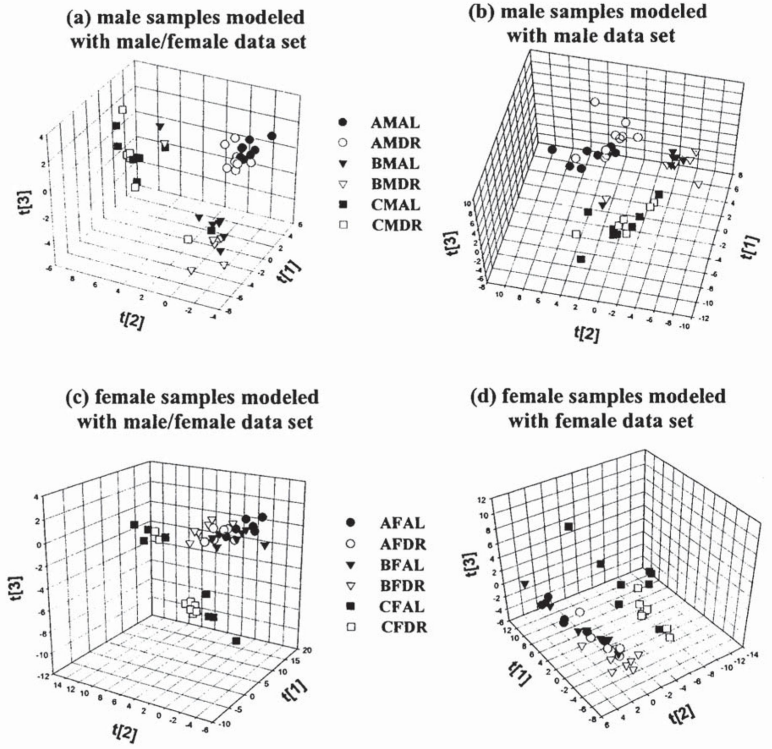

FIG. 1. Principal components analysis (PCA) of male and female rats readily distinguish the three cohorts. PCA plots of the first three components based on 93 metabolites. Models were built on data sets comprised of combined male and female samples (**a,c**), male samples only (**b**) or female samples only (**d**). Rotated to show cohort distinctions. Letter labels refer to cohort of origin. In the legend, first letter designates cohort (A, B, C), second letter designates gender (M, F), the third and fourth letters refer to *ad libitum* (AL) or dietary restriction (DR).

Mathematical pre-processing fails to eliminate cohort specificity

We then sought to determine if we could eliminate the noise through mathematical means, specifically pre-processing. Pre-processing is a means for attempting to ensure "proper" (a context dependent issue) weighting for the variables. For the more basic levels of analysis, we examined this issue previously (Shi et al., 2002a,b), and found that *Autoscaling* the data [mean centering (subtracting the mean), followed by variance scaling (dividing by the standard deviation)] worked well. This approach was also followed in Figure 1 of this manuscript. It is possible, however, that other preprocessing mechanisms might help reduce the cohort problem. These methods might be sufficient, for example, if a major aspect of the difference be-

tween cohorts was simply a change in the variability of the overall cohort relative to a mean that was equivalent to that of the other cohorts. Such an increase in variability might, for example, result from an increase in daily disturbances noise in the animal colony. We therefore systematically examined two aspects of preprocessing—scaling and distribution.

Scaling. We examined three scaling methods: no scaling, Pareto scaling, and unit variance (UV) scaling. No scaling is, as it sounds, using the raw numbers directly in the analysis. As the concentrations for each given metabolite (as determined by HPLC) are already standardized to the concentration of that given metabolite in a pooled serta sample used as a standard, the values being used are already somewhat standardized. In UV scaling, SIMCA-P9's equivalent of the *Autoscale* preprocessing described previously (from Pirouette), a given variable is centered on its mean and then divided by its standard deviation computed around that mean. Pareto scaling, which provides a level of manipulation that is in between no scaling and UV scaling, gives the variable a variance equal to its standard deviation instead of unit variance. Specifically, the variable is centered and divided by the square root of the standard deviation around the mean (Eriksson et al., 2001). No major differences were observed with the three scaling techniques (not shown). Specifically, we were essentially unable to separate AL and DR rats using PCA.

Distributions. We next systematically examined six approaches designed to help the distribution: no adjustments, log transformation, winsorizing at two standard deviations, winsorizing at three standard deviations, log transformation coupled with winsorizing at two standard deviations, and log transformation coupled with winsorizing at three standard deviations. (Winsorizing is replacing a value that is beyond a mathematically defined limit with a predetermined, specific value. For example, replacing all observations greater than three standard deviations from the mean with the value at three standard deviations. This has the effect of reducing the impact of severe outliers on a model.) Log transformations improve signal distribution across datasets that are skewed or which have large minimum/maximum ratios. There were no major effects observed with the different methods of manipulating the distribution (not shown). We have also looked further, looking at all possible combinations with other scaling approaches. Overall, these means failed to remove the noise introduced by inclusion of multiple cohorts.

Summary. The inability to correct the cohort separations using mathematical manipulations such as scaling, transforms, normalizations, and winsorizing suggest that the cohort separations are biological, not analytical, in nature.

Evaluating mathematical proprocessing and its effects on PCA-based analyses

Subsequent analysis showed that no additional scaling was needed to develop models that were effective at describing metabolite concentrations (i.e., regenerating the values [concentrations] in the x-block [independent variables, metabolites]; Figs. 2 and 3). In Figure 2, we consider all statistically valid PCA components. This approach overfits the x-block with scaled data, but models the x-block well with no scaling, generating models with both good descriptive and predictive power. (The descriptive power of the model can be described by several terms, most directly [within SIMCA P] by R^2X and R^2Xcum. These terms are defined within SIMCA-P as the fraction of the Sum of Squares [SS] of all the X's explained by the current component [R^2X] or the cumulative SS of all the X's explained by all extracted components R^2Xcum. These terms serve as the direct measure of the percentage of the total variability in the independent variables that is *captured* by the model. Likewise, Q^2X and Q^2Xcum serve as a measure of the fraction of the total variability of the independent variables that can be *predicted* by a given component [Q^2X] or by the overall model [Q^2Xcum]. Q^2[cum] is computed as: $Q^2cum = [1 - \Pi(PRESS/SS)a]_{[a = 1, \ldots A]}$, where Π[PRESS/SS]a = the product of PRESS/SS for each individual component a, and the the prediction error sum of squares [PRESS] is the squared differences between observed and predicted values for the data kept out of the model fitting.) When the model is limited by only including components that increase predictive power (ie, $Q^2X>0$,) strong models are built with all scaling techniques, and near perfect models are built with no scaling. Note that descriptive mathematical models of the metabolome are essentially independent of the

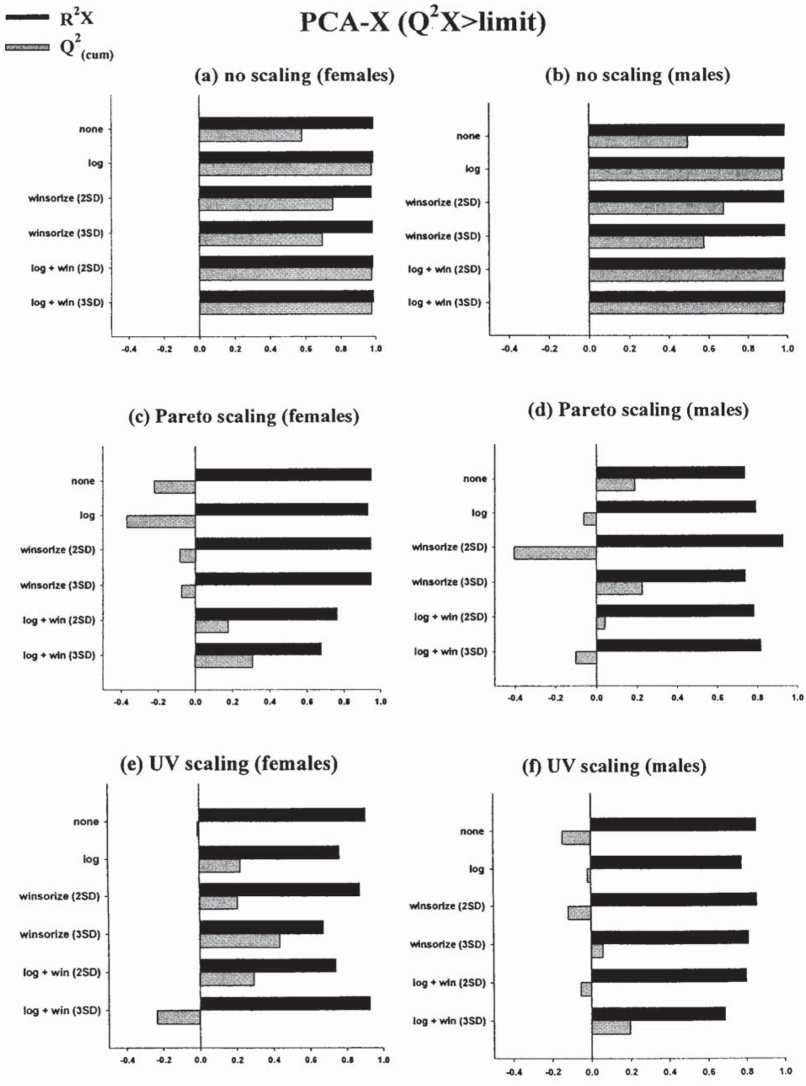

FIG. 2. Unscaled PCA models out-perform scaled models. Parameters for all valid PCA components (Q^2X>limit) (Eriksson et al., 2001). Plots shows males and females, all scaling and distribution methods.

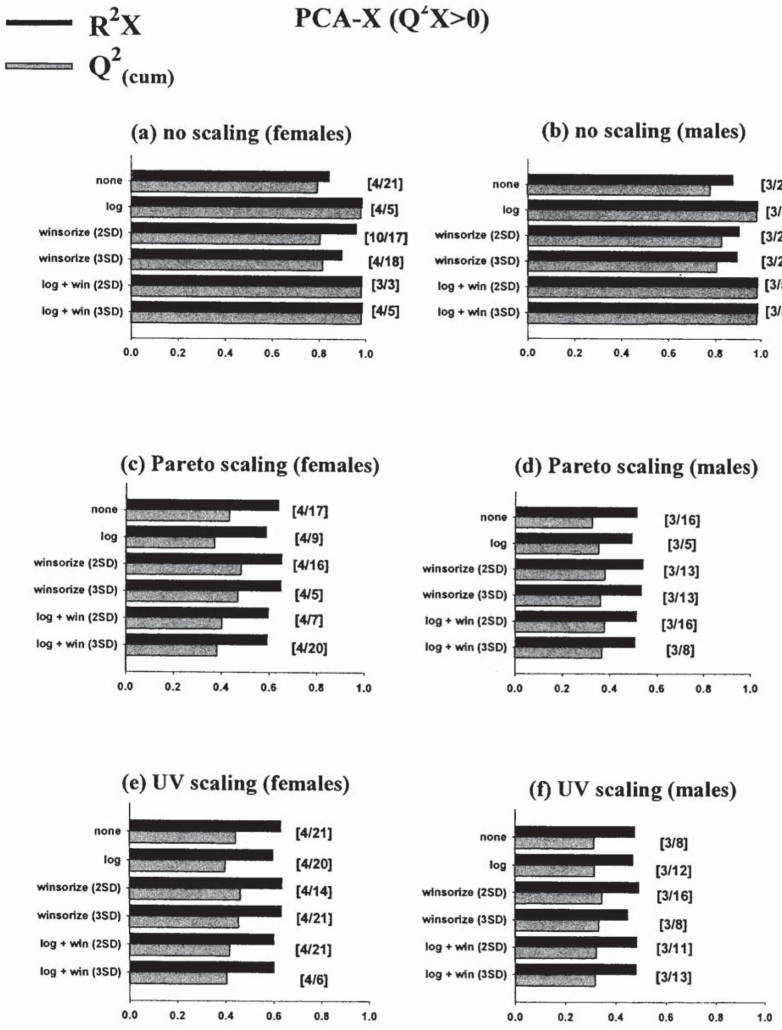

FIG. 3. Unscaled PCA models still outperform scaled models when only components that add predictive power are considered. Parameters for all valid PCA components ($Q^2X>0$) (Eriksson et al., 2001). Plots shows males and females, all scaling and distribution methods. Numbers in brackets refer to those components having $Q^2X>0$ relative to the total number of valid components ($Q^2X>$limit).

choice of distribution modification (e.g., log transformation, winsorization) when only components having $Q^2X>0$ are used.

Thus, we have explored PCA-based analyses and shown that they are (1) more robust for these studies than distance based metrics such as clustering-based algorithms (Shi et al., 2002a,b, 2004); (2) robust across scaling techniques (Vigneau-Callahan et al., 2001; Shi et al., 2002a,b; and above data); and (3) robust across different means of manipulating distributions when $Q^2X>0$ (Fig. 3). Despite these advantages, PCA appears unable to effectively separate AL and DR sera in sets from multiple cohorts. Likewise, we have determined that there are considerable differences in the effects of AL and DR feeding on male and female rats (Paolucci et al, 2004). Finally, while SIMCA-based expert system analysis consistently outperforms clustering-based k-nearest neighbor analyses, we have clear evidence that SIMCA also is weakened by the inter-cohort differences (Shi et al, 2004). We therefore make two decisions. First, we will continue (for now) to use the same set of metabolites to analyze male and female sera, but we will build both predictive and descriptive models on each gender separately. Second, we will turn away from PCA to a projection method optimized for class separation.

Partial least squares projection to latent structures-discriminant analysis (PLS-DA)

PLS-DA versus SIMCA. Functionally, PLS-DA (Sjöström et al., 1986; Ståhle and Wold, 1987) may be considered the equivalent of finding a rotation of a multidimensional PCA space (or arguably, specifically creating the multidimensional space) so as to optimize the separation between classes of interest. This aspect of PLS-DA differentiates it from SIMCA, which focuses on the independent variables whose variability most underlies variation within the dataset, but which does not necessarily coincide with the maximum separation directions between the classes of interest. SIMCA builds models one class at a time (and thus classifies unknowns by whether or not they fit a given class), whereas PLS-DA builds models by identifying major differences that account for maximal co-variation between the X-block (metabolites/variables) and the "Y" values (i.e., class membership; and thus classifies by class distinctions). Each has distinct advantages. For example, SIMCA can build more complex mathematical models. This feature enables SIMCA to correctly identifying group members when, for example, one group is entirely enveloped within another within a principal component space. PLS-DA's advantages include several that are directly applicable to the next stage of the analysis. These advantages, include strong over-fit diagnostics (categorical analysis programs can, as shown above, find solutions that are appropriate only for the training set); diagnostics present in the program used (SIMCA-P, Umetrics) readily enable us to monitor (and avoid) this potential concern. Secondly, specific variables can be readily monitored in PLS-DA for their importance in distinguishing groups, and prediction sets can be readily generated and tested in a more appropriate way (an entire set can be tested and failures noted) than in SIMCA (in which only single sets can be tested). Finally, PLS-DA attempts to assign all variables to a class, whereas SIMCA is more likely to leave observations as unclassified, especially in early stages of modeling. For this reason, the next stage of analysis was conducted with PLS-DA.

We first examined the preprocessing and distribution considerations as done for PCA (Fig. 4). While "no scaling" continued to model the X component well, it became obvious that both Pareto and UV scaling better model the Y component (i.e., class, which are described as SIMCA-P by the terms R^2Y and Q^2Y, which are determined as for their X equivalents, except looking at the fraction of Y that is modeled; Eriksson et al., 2001), and that log transformation yields only non-reproducible (i.e., non-predictive) models for non-scaled data (Fig. 4). In particular, UV or Pareto scaling, coupled with log transformation and with winsorizing at 2 or 3 SD, seems very robust. To avoid losing the bottom of our distributions (those compounds present in very low concentrations relative to the mean concentration of that metabolite in the overall sample set) while capturing more of the variability at higher metabolite concentrations, we have chosen to UV scale and winsorize at 3 SD on the upper limit and 2 SD on the lower limit. These parameters are expected to provide a robust platform for further studies.

PLS-DA models of AL versus DR serotype. To examine the potential methods in more detail, we now examine seven models built with the previous described data. Models 1 (Fig. 5) and 2 (Fig. 6) are, respec-

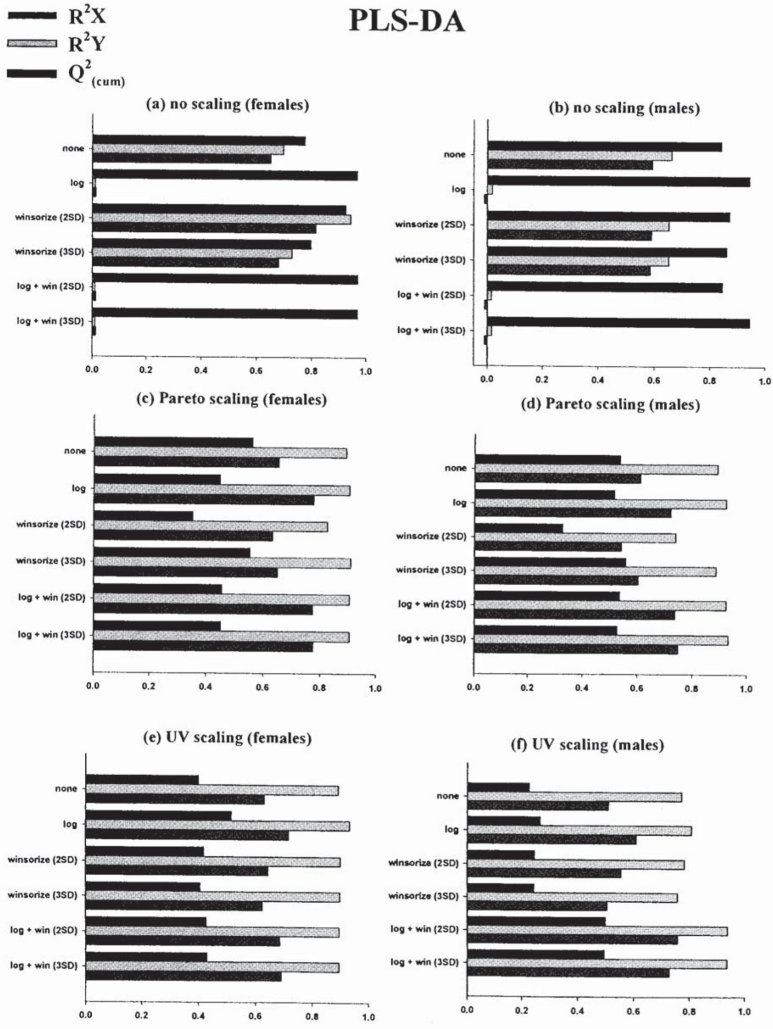

FIG. 4. PLS-DA models built on scaled data capture class identity. Parameters for all valid PLS-DA components ($Q^2X>0$) (Eriksson et al., 2001). Plots shows males and females, all scaling and distribution methods.

tively models of the male and female serotypes built on the markers having $p \leq 0.2$ in the cognate gender (37 markers in males; 56 markers in females). Models 5 and 6 (Figs. 7 and 8) use all 93 markers that were defined in either sex using an equivalent number of components (all of which have $Q^2Y>0$), so that models built with the larger dataset can be directly compared to those built with the smaller dataset. Note the broader spread of useful metabolites in the larger models (i.e., the less significant peaks by univariate criteria add considerably to multivariate analyses) Figure 9 uses a different preprocessing approach to show the existence of previously unidentified information in our metabolomics datasets.

Each model (Figs. 5–9) is described/presented in the context of six panels. Panels A and B present data relevant to the strength of the models. Panel A shows the ability of the statistically valid components to capture the variability that distinguishes AL and DR serotypes. The first bar in each pair describing the component (termed R^2Ycum by SIMCA-P) is a quantitative measure of the goodness of fit, i.e., R^2Ycum is the overall proportion of the variation in the Y variable (here, AL vs. DR) that is explained by a model containing this and all previous components (R^2Y is that proportion explained by the specific component alone). Thus, the first bar in the second pair is a measure of the ability of this two-component model to explain the overall variation in the outcome variable (class, AL or DR). The second bar in each pair describing the component (termed Q^2cum by SIMCA-P) is a quantitative measure of the predicted goodness of fit, by cross-validation, made by a model containing this and all previous components (Q^2Y is that proportion explained by the specific component alone). (Cross-validation [CV] is implemented in SIMCA as follows: Parts of the data [as run, 1/7 of data per cycle] are kept out of model development, and then predicted by the model, and compared with the actual values. Specifically, the prediction error sum of squares [PRESS] is the squared differences between observed and predicted values for the data kept out of the model fitting. This procedure is repeated [seven times, as run] until every data element has been kept out once and only once. The final PRESS then has contributions from all data. For every dimension, SIMCA computes the overall PRESS/SS, where SS is the residual sum of squares of the previous dimension. SIMCA also computes [PRESS/SS]k for each *X* variable [*xk*]. A component [model dimension] is considered significant if PRESS/SS is statistically smaller than 1.0.) Essentially by definition, R^2Ycum will increase with increasing components, whereas Q^2Ycum will reach a threshold and then decline as the models become overfit. For PLS studies, a component is considered significant when $Q^2Y>0$ (i.e., when the component adds to the predictive strength of the model, termed rule R1, the initial testing rule). All components in all models shown are considered significant by rule R1. Panel B presents the permutation validation of the model. This validation test consists of randomizing the Y variables (i.e., class membership), building a new PLS-DA model, and re-evaluating accuracy. In every case, we conducted 100 permutation tests, and in all 100 tests of each model our model based on actual data was better than all 100 permutated models, essentially giving a p value of $p < 0.01$ to each model. We note that another criteria is that models are considered valid for their ability to describe variation when the *Y* intercept of R^2Y, the top line, is <0.3–0.4. The four component models are considerably overfit with respect to R^2Y (not shown), and some smaller models approach or slightly exceed overfit conditions. This does suggest a need for caution in interpreting fit of a specific rat to a specific class. However, the more general predictive parameter, i.e., the ability of the model to predict group/class membership (the more important determinant for our purposes) is considered valid when

FIG. 5. PLS-DA based classification of males based on the 37 metabolite dataset. PLS-DA models were built on UV scaled data winsorized at 2 SD (lower limit) and 3 SD (upper limit). (**A**) Parameters of the model. R^2Y is the amount of variability of the Y variable (i.e., of class, AL vs. DR) of the overall dataset explained. Q^2Y is the predicted accuracy of the model based on coefficients of variations from a series of 7 jack-knifing (multiple sample leave out) analyses. (**B**) Validate plot, results of 100 permutations of the Y variable. Correlation with actual dataset on X-axis: R^2Y and Q^2Y for each permutation on Y-axis. Shown with regression lines. (**C**) Plot based on the one statistically significant component. Dashed line added to emphasize AL/DR distinction. (**D**) Weights plot showing the direction in which different metabolites pull the model. This plot is the PLS-DA equivalent of a loadings plot in PCA. (**E**) variable importance on projection. X-axis lists metabolites by retention time. (**F**) Predicted vs. actual class assignments. Overall accuracy was >97%.

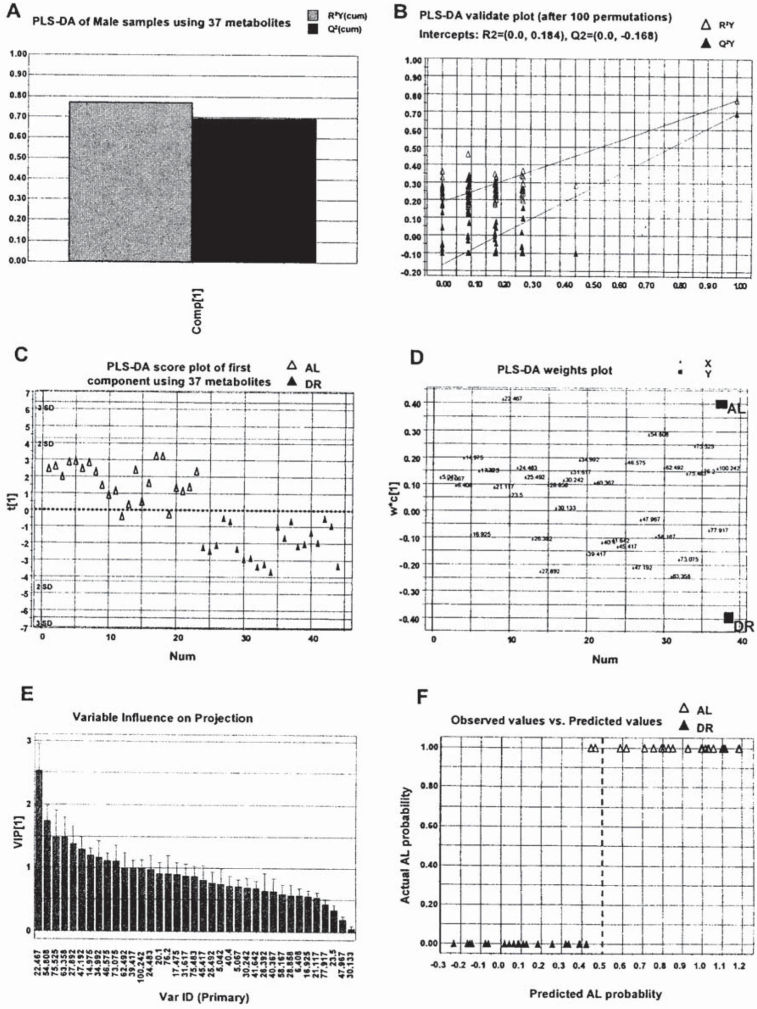
A
PLS-DA of Male samples using 37 metabolites
R²Y(cum)
Q²(cum)
Comp[1]
B
PLS-DA validate plot (after 100 permutations)
Intercepts: R2=(0.0, 0.184), Q2=(0.0, -0.168)
R²Y
Q²Y
C
PLS-DA score plot of first component using 37 metabolites
AL
DR
t[1]
Num
D
PLS-DA weights plot
X
Y
w*c[1]
AL
DR
Num
E
Variable Influence on Projection
VIP[1]
Var ID (Primary)
F
Observed values vs. Predicted values
AL
DR
Actual AL probability
Predicted AL probablity

FIG. 5.

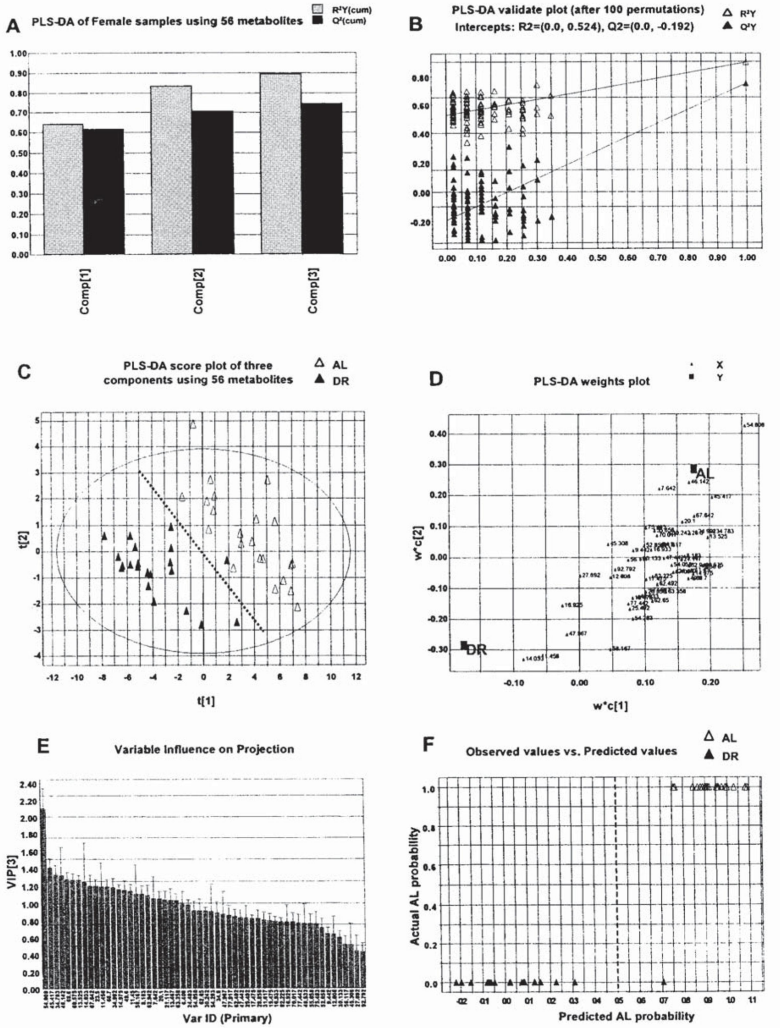

FIG. 6. PLS-DA based classification of females based on the 56 metabolite dataset. See legend to Figure 5. Note that in plot in panel C for this figure, the figure is based on two statistically significant components.

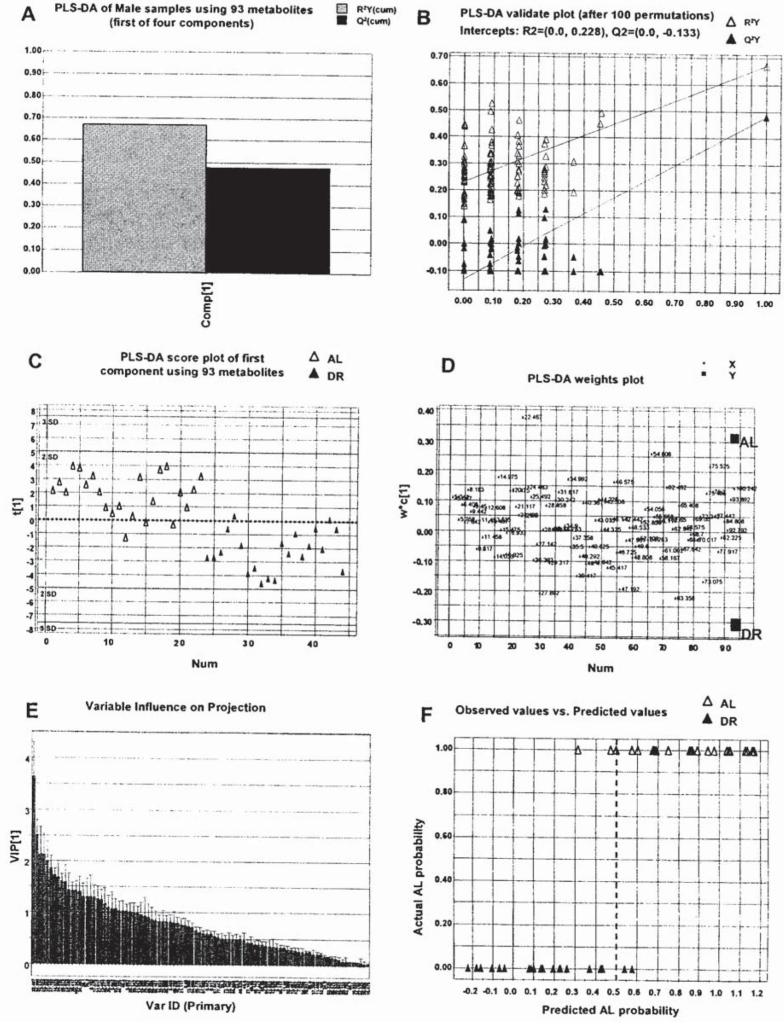

FIG. 7. PLS-DA based classification of males based on the first component of 93 metabolite dataset. See legend to Figure 5. Overall accuracy >97%.

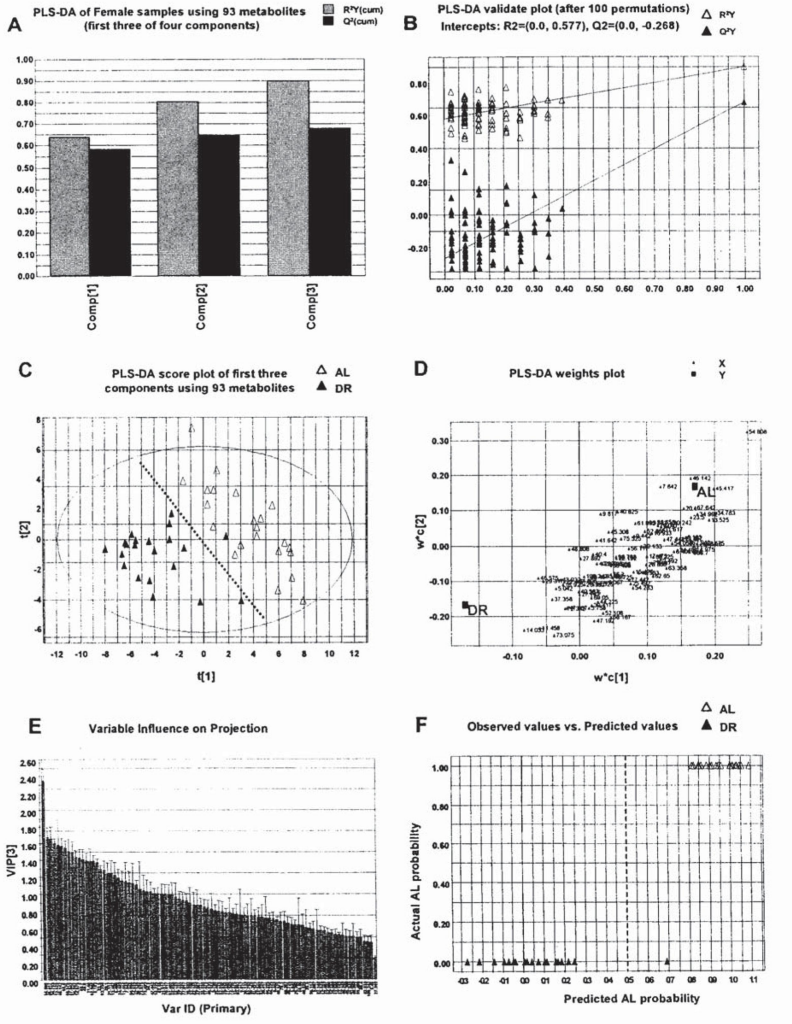

FIG. 8. PLS-DA based classification of females based on the first three components of the 93 metabolite dataset. See legend to Figure 5. Overall accuracy >97%.

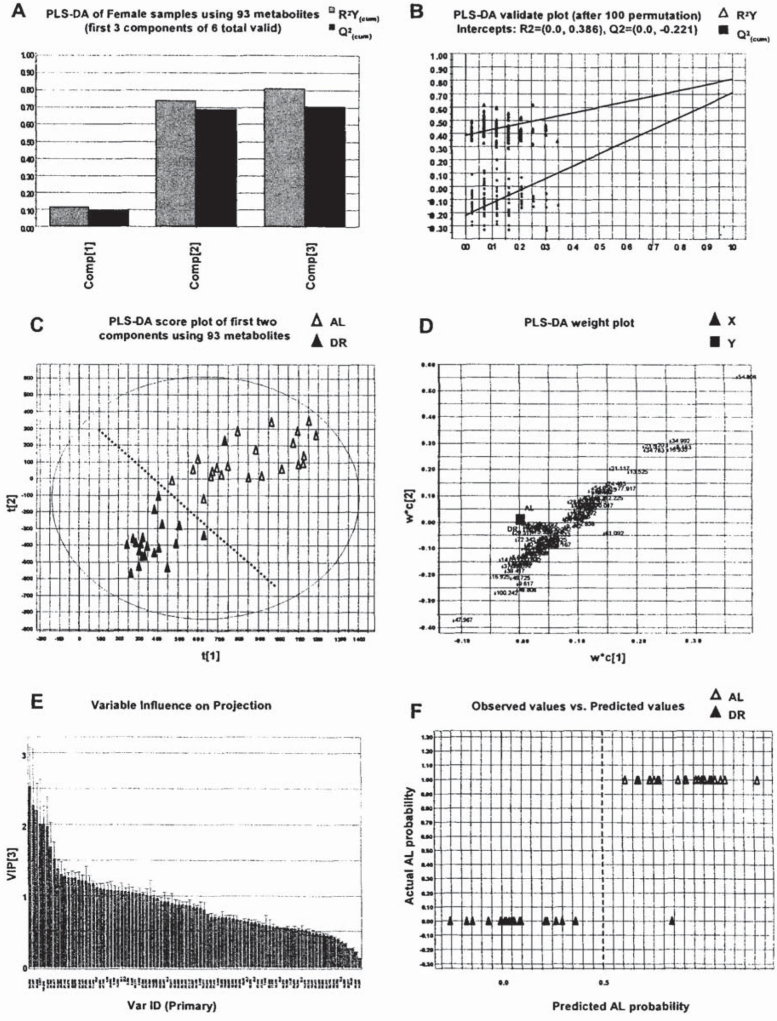

FIG. 9. PLS-DA based classification of females based on the 93 metabolite dataset (unscaled). See legend to Figure 5. Overall accuracy >97%. Note, however, the poor separation between AL and DR classes, which are approximately located at the origin.

the y intercept of Q^2Y, the bottom line, is <0.05. All models meet this criteria, thus no models are considered overfit with respect to their predicted classification power, Q^2Y.

Panel C graphically presents the location in the model space of each animal with respect to component 1 and 2 (panel C). These graphs highlight the separability of AL and DR rats.

Panels E and F present information on the metabolites themselves. The data in Panel D shows how each of the individual analytes affects, or "pulls" the model. Metabolites whose influence maps in the same direction (from the origin) as the diet's serotype pull the model in that direction as they become present in greater quantities. Distance from origin indicates the strength of that pull. A complementary plot is shown in Panel E (variable importance on projection, VIP).* VIP is a marker of how well each analyte separates the groups of interest (AL and DR) across the sum of all active components. Analytes having VIPs>1 are considered the most important for separtions. Note that this analysis suggests 24 analytes contain the bulk of classification power in males and 34 in females. There are 11 overlaps between these two groups (not significant by consideration of binomial distribution).

Panel H shows class membership. The score on the X-axis is predicted membership; the score on the Y-axis is actual membership. Formally, the test is membership or not for the tested class (the class on the top having a value of ~1 on the upper axis).

The results obtained from these models may be summarized as follows:

1. The models produced are mathematically robust/statistically valid.
2. The data processing approaches (scaling and distribution) are valid.
3. PLS-DA is superior to PCA for these studies because of its ability to ignore interference resulting from inter-cohort variation.
4. PLS-DA generates strong, statistically valid models that have >97% accuracy in distinguishing AL and DR sera.
5. Relatively few variables have VIP scores >1, suggesting the possibility that we can shrink the datasets being analyzed (but see below).

We note that both the male and female dataset display disproportionately powerful variables (one each in males and females). This observation raises a potential issue as to whether the presence of these variables is the sole determinant of the differences between AL and DR serotypes. To address this issue, we excluded these variables and re-analyzed the data. The models remained capable of distinguishing AL and DR rats (not shown).

At first glance, the combination of (A) few variables having high VIP scores and (B) the ability to remove the highest VIP score without compromising separation appear potentially mutually exclusive. Further study of the VIP scores derived from the study in which the strongest marker was removed suggested that other markers now distinguished AL and DR serotypes.

This result suggested to us that there might be markers that could help differentiate AL and DR, but whose effect was masked by the power of other markers. This hypothesis was confirmed by changing the scaling (from UV to no scaling), and redoing the PLS-DA (Fig. 9). Under these conditions, PLS-DA still distinguished AL and DR female rats, but the distribution of VIP values was markedly different, and the second and third strongest markers in Figure 9 had VIP values <1 in Figure 8; indeed, marker 48.808 was actually the least valuable marker in Figure 8. This finding suggests that hidden information within the dataset exists and will require further analysis.

*"SIMCA computes the influence on Y of every term (xk) in the model called VIP. VIP is the sum over all model dimensions of the contributions VIN (variable influence). For a given PLS dimension, **a,** (VINak2 is equal to the squared PLS weight (wak)2 of that term, multiplied by the explained SS of that PLS dimension. The accumulated (over all PLS dimensions) value, $VIP_{ak} = \Sigma_a VIN_k$ is then divided by the total explained SS by the PLS model and multiplied by the number of terms in the model. The final VIP is the square root of that number. The Sum of squares of all VIP's is equal to the number of terms in the model hence the average VIP would be equal to 1. One can compare the VIP of one term to the others. Terms with large VIP, larger than 1, are the most relevant for explaining Y."

CONCLUSION

In this report, we turn away from component based approaches that describe either the entire dataset (e.g., PCA) and/or specific classes (SIMCA), to a modeling approach that instead focuses on identifying the components that best distinguish two defined classes (in this report, we focus on PLS-DA). The strength of PCA as a data analysis approach is that it can, relatively objectively, describe the major sources of variation in a population. The weighting of these variables on the components of a PCA-based representation, however, does not necessarily coincide with an outcome variable of interest. The great strength of SIMCA as a data analysis approach is, in contrast, that it builds a series of models each based on one class. By focusing on one class at a time, SIMCA thus allows more appropriate description on each class. While such an approach has advantages, it also has a major disadvantage in that, because it considers the groups independently, SIMCA can thus often underweight variables that can be particularly useful for classification because they don't explain a lot of variation within their own class. A complementary approach, termed PLS-DA, focuses instead on the analytes that are best able to distinguish classes. Because PLS-DA takes the entire dataset into consideration, but then focuses on the changes that reflect class, it can theoretically be used to see a specific signal (e.g., AL-DR serotype differences) in the context of noise (e.g., male-female differences, cohort distinctions). The use of PLS-DA in the study presented here thus shows that DR does exert different effects on the sera metabolome of males and females. Furthermore, PLS-DA analysis identifies a clear signal representing the effects of DR on the sera metabolome, which is robust across cohorts.

In summary, we return to the three questions raised at the end of the introduction: (1) Can we avoid overfitting concerns without compromising the model? (2) Can we optimize the use of component-based models for classification purposes? (3) Can we deal with the inter-cohort and sex-specific differences that distinguish cohorts and lead to over-fitting? In all three cases, PLS-DA solves the problems resulting from inter-cohort variation, and allows us to build strong models defining metabolic serotype. This conclusion, however, was not readily apparent at the onset of analysis and required initial exploration the data set by means of PCA. Only after identification of the shortcomings inherent in PCA when applied to metabolomics data were we able to demonstrate the requirement for implementing a technique (i.e., PLS-DA) that would remedy these failures.

ACKNOWLEDGMENTS

We thank Dr. Tom Vogl for his many critical discussions and comments on the manuscript.

REFERENCES

ERIKSSON, L., JOHANSSON, E., KETTANEH-WOLD, N., et al. (2001). *Multi- and Megavariate Analysis* (Umea, Sweden, Umetrics).

FERNANDES, G., CHANDRASEKAR, B., TROYER, D.A., et al. (1995). Dietary lipids and calorie restriction affect mammary tumor incidence and gene expression in mouse mammary tumor virus/v-Ha-ras transgenic mice. Proc Natl Acad Sci USA **92,** 6494–6498.

KLURFELD, D.M., WEBER, M.M., and KRITCHEVSKY, D. (1987). Inhibition of chemically induced mammary and colon tumor promotion by caloric restriction in rats fed increased dietary fat. Cancer Res **47,** 2759–2762.

KRITCHEVSKY, D., WEBER, M.M., and KLURFELD, D.M. (1984). Dietary fat versus caloric content in initiation and promotion of 7,12-dimethylbenz(a)anthracene-induced mammary tumorigenesis in rats. Cancer Res **44,** 3174–3177.

MANLY, B.F.J. (2000). *Multivariate Statistical Methods* (Boca Raton, FL, CRC Press).

PAOLUCCI, U., VIGNEAU-CALLAHAN, K.E., SHI, H., et al. (2004). Development of biomarkers based on diet-dependent metabolic serotypes: characteristics of component-based models of metabolic serotypes. OMICS **8,** 209–220.

PETRICOIN, E.F., ARDEKANI, A.M., HITT, B.A., et al. (2002). Use of proteomic patterns in serum to identify ovarian cancer. Lancet **359,** 572–577.

ROTH, G.S., LANE, M.A., INGRAM, D.K., et al. (2002). Biomarkers of caloric restriction may predict longevity in humans. Science **297,** 811.

SHI, H., VIGNEAU-CALLAHAN, K.E., SHESTOPALOV, A.I., et al. (2002a). Characterization of diet-dependent metabolic serotypes: primary validation of male and female serotypes in independent cohorts of rats. J Nutr **132,** 1039–1046.

SHI, H., VIGNEAU-CALLAHAN, K.E., SHESTOPALOV, A.I., et al. (2002b). Characterization of diet-dependent metabolic serotypes: Proof of principle in female and male rats. J Nutr **132,** 1031–1038.

SHI, H., PAOLUCCI, U., VIGNEAU-CALLAHAN, K.E., et al. (2004). Development of biomarkers based on diet-dependent metabolic serotypes: practical issues in development of expert system-based classification models in metabolomic studies. OMICS **8,** 197–208.

VIGNEAU-CALLAHAN, K.E., SHESTOPALOV, A.I., MILBURY, P.E., et al. (2001). Characterization of diet-dependent metabolic serotypes: analytical and biological variability issues in rats. J Nutr **131,** 924S–932S.

SJÖSTRÖM, M., WOLD, S., and B. SÖDERSTRÖM, B. (1986). PLS discriminant plots. In *Pattern Recognition in Practice II.* E.S. Gelsema and L.N. Kanal, eds. (Elsevier, Amsterdam), p. 486.

STÅHLE, L., and WOLD, S. (1987). Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. J Chemometrics **1,** 185–196.

WEINDRUCH, R., and WALFORD, R. (1988). *The Retardation of Aging and Disease by Dietary Restriction* (St. Louis, Charles C. Thomas).

WOLD, S. (1976). Pattern recognition by means of disjoint principal components models. Pattern Recognition **8,** 127–139.

YU, B.P. (1994). *Modulation of Aging Processes by Dietary Restriction* (Boca Raton, FL, CRC Press).

Address reprint requests to:
Dr. Bruce S. Kristal
Dementia Research Service
Burke Medical Research Institute
785 Mamaroneck Ave.
White Plains, NY 10605

E-mail: Bkristal@burke.org