Published in final edited form as:
Expert Syst Appl. 2008 October 1; 35(3): 967–975. doi:10.1016/j.eswa.2007.08.050.

Genetic algorithm-based feature selection in high-resolution NMR spectra

Hyun-Woo Cho[a], **Seoung Bum Kim**[b], **Myong K. Jeong**[a], **Youngja Park**[c], **Thomas R. Ziegler**[d], and **Dean P. Jones**[e]

Hyun-Woo Cho: hcho7@utk.edu; Seoung Bum Kim: sbkim@uta.edu; Myong K. Jeong: mjeong@utk.edu; Youngja Park: medyp@emory.edu; Thomas R. Ziegler: tzieg01@emory.edu; Dean P. Jones: dpjones@emory.edu

[a] Department of Industrial and Information Engineering, The University of Tennessee, Knoxville, TN 37996, USA

[b] Department of Industrial and Manufacturing Systems Engineering, The University of Texas at Arlington, Arlington, TX 76019, USA

[c] Clinical Biomarkers Laboratory, Department of Medicine, Emory University, Atlanta, GA 30322, USA

[d] Center for Clinical and Molecular Nutrition, Department of Medicine, Emory University, Atlanta, GA 30322, USA

[e] Clinical Biomarkers Laboratory, Center for Clinical and Molecular Nutrition, Department of Medicine, Emory University, Atlanta, GA 30322, USA

Abstract

High-resolution nuclear magnetic resonance (NMR) spectroscopy has provided a new means for detection and recognition of metabolic changes in biological systems in response to pathophysiological stimuli and to the intake of toxins or nutrition. To identify meaningful patterns from NMR spectra, various statistical pattern recognition methods have been applied to reduce their complexity and uncover implicit metabolic patterns. In this paper, we present a genetic algorithm (GA)-based feature selection method to determine major metabolite features to play a significant role in discrimination of samples among different conditions in high-resolution NMR spectra. In addition, an orthogonal signal filter was employed as a preprocessor of NMR spectra in order to remove any unwanted variation of the data that is unrelated to the discrimination of different conditions. The results of *k*-nearest neighbors and the partial least squares discriminant analysis of the experimental NMR spectra from human plasma showed the potential advantage of the features obtained from GA-based feature selection combined with an orthogonal signal filter.

Keywords

Metabolomics; Nuclear magnetic resonance (NMR); Feature selection; Discrimination; Genetic algorithm (GA); Orthogonal signal correction filter

1. Introduction

Metabolomics is developing as a major scientific discipline for predictive health research because it provides an opportunity to investigate dynamic and time-dependent metabolic

Correspondence to: Seoung Bum Kim, sbkim@uta.edu.

patterns in association with dietary, environmental and pathophysiologic stimuli as they occur in integrated biologic systems (Nicholson, Connelly, Lindon, & Holmes, 2002; Nicholson, Lindon, & Holms, 1999). However, in pre-morbid states, factors that cause or contribute to disease co-exist with early (pre-morbid) changes of disease. Consequently, predictive biomarkers obtained from cross-sectional studies often provide little or no insight into disease mechanisms. In principle, this limitation can be overcome by separating the search for biomarkers into two components, with the first being an effort to identify metabolic features linked to potential risk factors and the second being studies to explicitly test whether these features predict disease. Because there are only about 50 required nutrients, application of such an approach to each nutrient could provide a rigorous evaluation of specific dietary components as causative factors in both common and rare chronic diseases. Moreover, development of a dataset of metabolic features linked to specific nutrients would provide a rich resource for human systems biology research and also provide a novel basis for nutritional assessment. Specifically, such an approach would deviate from usual nutritional assessments which estimate intake or body contents of a nutrient and replace this with one which evaluates the integrated biologic response of the system to that nutrient. Such an approach critically depends upon improved methods to identify robust metabolic features and link these features with specific biologic metabolites.

A variety of techniques are available for studying metabolomics; of these, high-resolution NMR spectroscopy has an advantage because it requires minimal sample preparation, and is noninvasive or minimally invasive and can be adapted to a high-throughput format (Lindon, 2004). Fig. 1 shows an example of 158 spectra of human plasma obtained from a high-resolution proton (^{1}H)-NMR spectroscopy. The *x*-axis indicates the chemical shift with units in ppm, and the *y*-axis indicates the intensity values corresponding to each chemical shift. Peaks in this plot correspond to the specific resonance of chemical species in the sample. Such complexity in high-resolution NMR spectra poses a great challenge in terms of analytical and computational capabilities.

Statistical pattern recognition techniques are available to reduce complexity and thus help detect and recognize metabolomic changes in such complicated but information-rich spectra. Principal components analysis (PCA) and clustering analyses, as examples of unsupervised methods, have been widely used to facilitate the extraction of implicit patterns and elicit the natural groupings of the spectral dataset without prior information about the sample class (Beckonert et al., 2003; Jansen, Hoefsloot, Boelens, Greef, & Smilde, 2004; Solanky et al., 1993). Supervised classification methods have been applied to classify metabolic profiles according to biological/metabolic conditions (Beckonert et al., 2003; Holmes, Nicholson, & Tranterm, 2001; Wang et al., 2004). These include partial least square (PLS), *k*-nearest neighbors, and probabilistic neural networks. A comprehensive summary of statistical pattern recognition methods in metabolomics can be found in Holmes and Antti (2002) and Lindon, Holmes, and Nicholson (2001).

Despite extensive research in using supervised and unsupervised methods in metabolomics, few attempts have been made to identify the major chemical species that have different spectral patterns among conditions. PCA has been widely used in metabolomics because PCA has the capability to reduce the high dimensionality for visualization using a few variables obtained through a combination of the original variables. The reduced dimensions, called principal components (PCs) are uncorrelated with each other; typically, the first few PCs are sufficient to account for most of the variability in the original high-dimensional space. Although PCA is a useful technique for visualization purposes, it may be inappropriate for feature selection because the reduced dimensions (i.e., PCs) are linear combinations of the large number of original variables. PC loading can be used to identify which chemical species separate the samples seen in the score plot. However, an approach using PC loadings provides limited information because these loadings are the contribution of each metabolite within the first few PC dimensions. Another major drawback of PCA is that the selected variables may not always produce maximum discrimination between classes.

The present research presents the use of a genetic algorithm (GA) in combination with an orthogonal signal correction (OSC) filter for feature selection in high-resolution ^{1}H-NMR spectra of plasma obtained from individuals under different levels of sulfur amino acid intake. GA was recently shown to be efficient in feature selection in a PLS environment (Esteban-Díez, González-Sáiz, Gómez-Cámara, & Pizarro, 2006; Gourvénec, Capron, & Massart, 2004; Leardi & Gonzalez, 1998). A major advantage of GA for spectral data, illustrated in the application to near-infrared spectral data (Esteban-Díez et al., 2006), is that selected features, i.e., wavelengths, are not scattered throughout the spectrum. Thus, if wavelength n is selected, wavelengths $n - 1$ and $n + 1$ are likely to be selected by GA. On the other hand, the selection of features can be considered an optimization problem. In this respect, GA is a very efficient technique for feature selection because the size of the search domain ($2^N - 1$ combinations for N variables) is enormous and many local optima are present.

This work also demonstrates the advantage of using an OSC filter in the GA-based feature selection process. The OSC can selectively remove the largest variation of predictor variable **X** that is orthogonal or unrelated to response variable **Y**. This is possible because an OSC uses the response **Y** to construct a kind of signal filter for **X**. For the classification problem of this work, the OSC method is modified for discrimination purposes so that it is implemented as in a PLS-discriminant analysis (PLS-DA). As a result, the **Y** matrix in an OSC contains class information of samples among varying conditions in NMR spectra. The efficiency of GA as a technique for feature selection is enhanced with the addition of an OSC in the GA-based feature selection process.

The remainder of this paper is organized as follows. First, experiment data are described in Section 2. Then, a brief review of PLS, PLS-DA, OSC, and GA is given in Section 3, which is followed in Section 4 by results, and then by concluding remarks.

2. Sample collection and preprocessing of NMR spectra

The experimental dataset is available from Emory University Center for Clinical and Molecular Nutrition to show the effectiveness of the GA algorithm for feature selection. We used plasma samples obtained from healthy subjects under controlled metabolic conditions in the Emory General Clinical Research Center (GCRC). The subjects signed an informed consent approved by the Emory Institutional Review Board and were screened prior to admission with a physician-performed medical history and physical examination, plasma chemistry profile, complete blood count and urinalysis. During the 12-day GCRC admission, the subjects consumed defined diets at standardized intervals. For the first two days (equilibration), the subjects consumed a balanced meal plan with foods selected to ensure adequate energy, protein and sulfur amino acid (SAA) intake (SAA at 19 mg/kg/day). After this phase, subjects were placed on constant semi-purified diets designed to alter SAA intake. The diets provided adequate energy and amino acid nitrogen to meet estimated maintenance needs of individual subjects. The semi-purified diet was provided in the form of cookies and beverages containing L-amino acids, sherbert, corn oil, butter, sugar and corn starch prepared in the GCRC metabolic kitchen. Daily micronutrient needs were provided in the form of standardized oral doses of multivitamin-mineral supplements, choline, sodium chloride, potassium, and magnesium. The L-amino acid component of the diet was altered to provide zero sulfur amino acids during the initial 5 days and 117 mg/kg/day during the latter

5 days of the GCRC stay. Blood was drawn serially 34 times from four subjects over 10 days and ^{1}H-NMR spectra were obtained by a Varian INOVA 600 MHz instrument. During the first 17 time points, blood was collected from subjects consuming zero SAA and 117 mg/kg/day SAA during the latter 17 time points.

NMR spectra require preprocessing steps before statistical analysis in order to detect subtle variations from metabolic profiles. In general, preprocessing steps involve phase/baseline correction, spectra alignments, redundant region elimination, and normalization of signal to an internal standard. Phase and baseline corrections were done using NUTS software (Acorn NMR Inc., Livermore, CA). Variations in spectra because of instrumental and environmental instabilities affect the spectra alignment and thus can interfere with direct comparison between samples. We used a beam search algorithm that determines the best alignment between the spectra by maximizing their correlation (Lee & Woodruff, 2004). Further, redundant regions (e.g., water and the regions containing no significant metabolite signals) were removed. Normalization was done by dividing each spectral point by the area of the internal standard. A spectrum after removal of its redundant regions is displayed in Fig. 2a. To further reduce complexity and remove noises, a spectrum was segmented into 0.01 ppm chemical shift bins. An NMR spectrum was reduced to 574 bins of equal width (0.01 ppm). The spectral area within a bin was integrated using MATLAB (MathWork Inc., Natick, MA). The reduced (binned) spectrum is displayed in Fig. 2b, showing that it maintains similar spectral structure in the original spectrum. This simplification was further justified because the PCA score plots of both the original and the reduced data show essentially identical patterns, differing only in the sign of PC1 (Fig. 3).

3. Methods

3.1. Partial least squares and orthogonal signal correction

PLS was developed to model the relation between a predictor matrix **X** and a response matrix **Y**. It seeks to find a set of latent variables that maximizes the covariance between **X** ($n \times N$) and **Y** ($n \times M$). PLS decomposes **X** and **Y** into the form as follows:

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{\mathrm{T}}+\mathbf{E}, \quad Y=\mathbf{U}\mathbf{Q}^{\mathrm{T}}+\mathbf{F}, \tag{1}$$

where **T** and **U** are ($n \times A$) matrices of the extracted A score vectors, **P** ($N \times A$) and **Q** ($M \times A$) loading matrices, and **E** ($n \times N$) and **F** ($n \times M$) residual matrices (Kourti, 2005). The PLS method based on a nonlinear iterative partial least squares (NIPALS) algorithm Wold, Geladi, Esbensen, and Öhman (1987) searches for weight vectors **w** and **c** that maximize the sample covariance between **t** and **u**. The NI-PALS algorithm repeats a sequence of the following steps until convergence:

1. $\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{u}/\mathbf{u}^{\mathrm{T}}\mathbf{u}$,
2. $\|\mathbf{w}\| \to 1$
3. $\mathbf{t} = \mathbf{X}\mathbf{w}$
4. $\mathbf{c} = \mathbf{Y}^{\mathrm{T}}\mathbf{t}/\mathbf{t}^{\mathrm{T}}\mathbf{t}$
5. $\mathbf{u} = \mathbf{Y}\mathbf{c}/\mathbf{c}^{\mathrm{T}}\mathbf{c}$
6. Repeat steps (1) through (5).

By regressing **X** on **t** and **Y** on **u** after convergence, loading vectors **p** and **q** can be obtained as follows:

$$\mathbf{p}=(\mathbf{t}^{\mathrm{T}}\mathbf{t})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t}, \quad \mathbf{q}=(\mathbf{u}^{\mathrm{T}}\mathbf{u})^{-1}\boldsymbol{Y}^{\mathrm{T}}\mathbf{u}. \tag{2}$$

Then, a PLS regression model can be expressed using regression coefficients **B** and residual matrix **G**:

$$\boldsymbol{Y}=\mathbf{XB}+\mathbf{G}, \quad \mathbf{B}=\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{C}^{\mathrm{T}}. \tag{3}$$

It was shown that the weight vector **w** corresponds to the first eigenvector of the following eigenvalue problem of $\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}w = \lambda\mathbf{w}$ (Hoskuldsson, 1988).

OSC is a preprocessing technique for removal of undesirable systematic variation in data. It was first developed in Wold, Antti, Lindgren, and Öhman (1998) to remove systematic variation from the predictor **X** that is orthogonal to the response **Y**. The largest variation of **X** having zero correlation with **Y** is selectively removed from **X**. The first step of an OSC is to calculate the first PC score vector **t** from **X**. The score vector **t** is then orthogonalized with respect to **Y**, producing the following actual correction vector $\mathbf{t}^*$:

$$\mathbf{t}^{*}=\{\mathbf{I}-\boldsymbol{Y}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})^{-1}\boldsymbol{Y}^{\mathrm{T}}\}\mathbf{t}. \tag{4}$$

Then PLS weight vector **w** is computed such that $\mathbf{Xw} = \mathbf{t}^*$ followed by the calculation of a new score vector $\mathbf{t} = \mathbf{Xw}$. These processes are repeated until **t** has converged. Finally, a loading vector **p** is computed and the correction term $\mathbf{tp}^{\mathrm{T}}$ is subtracted from **X**, giving the residual. The next components can be calculated in a similar way.

After the introduction of the OSC by Wold et al. (1998) several OSC algorithms have been reported (Fearn, 2000; Sjoblom, Svensson, Josefson, Kullberg, & Wold, 1998; Westerhuis, de Jong, & Smilde, 2001) Here a direct orthogonal signal correction (DOSC) algorithm was used because of its high reliability (Westerhuis et al., 2001). The DOSC algorithm was modified and implemented in a PLS-discriminant analysis (PLS-DA) environment for classification. PLS-DA is the classical PLS algorithm applied to classification problems (Barker & Rayens, 2003). One common way to use PLS in classification problems is to introduce a coding in which each column in **Y** contains information about the class memberships of samples. When K is the number of classes in data, each row of **Y** in PLS-DA has the following structure:

$$y_k^{\mathrm{T}}=\begin{cases}1 & \text{if sample belongs to class } k,\\ 0 & \text{otherwise,}\end{cases}$$

where y_k is the kth column of **Y**, $k = 1, \ldots, K$.

3.2. Genetic algorithms-based feature selection

GA is a problem-solving method inspired by evolution theory that simulates a natural evolution process. According to this principle, the best individuals are the most likely to survive for reproduction, thus propagating their genome. A population evolves in such a way that its fitness to its environment is usually better than the previous generation's. The sources of variation derive from mutations or cross-over to form a new population. GA has been used successfully in solving many problems such as molecular modeling, curve fitting, classification, and feature selection for calibration (Blommers, Lucasius, Kateman, &

Kaptein, 1992; DeWeijer et al., 1994; Hunger & Huttner, 1999; Kemsley, 2001). In particular, GA has been shown to be a very efficient tool for feature selection in calibration (Leardi, Seasholtz, & Pell, 2002). A detailed review of GA is available elsewhere (Hibbert, 1993; Leardi, 2001). In this work, GA is incorporated with PLS-DA and used as an optimization procedure to select the major metabolite features that discriminate among different classes in spectral data. Each feature is represented by a binary code 0 or 1 in a vector, called chromosome. Then a chromosome is perturbed randomly to create the initial population. Only a subset of original features is selected in each chromosome. A fitness function is evaluated for all chromosomes. Because variables are initially selected at random, the fit-ness function may be low so that mutations and cross-over are performed to improve the chromosomes. This procedure is repeated until a convergence criterion has been reached.

One of the major concerns with using GA is the problem of overfitting in which noise instead of information is modeled. This usually occurs when the features are noisy or the dimension of the features is high. The performance of the GA-based feature selection algorithm has been found to diminish when more than 200 features are used (Gourvénec et al., 2004). In this case, an iterative feature selection procedure is recommended (Leardi et al., 2002). It also is interesting to note that several researchers have developed GA-based feature selection methods, each of them using a different GA structure. We selected the method developed by Gourvénec et al. (2004) for this study because it has been successfully applied to the analysis of spectral datasets (Gourvénec et al., 2004; Leardi et al., 2002). GA was implemented in MATLAB 6.5 (MathWork Inc., Natick, MA), and the parameters used in implementing GA are as follows:

- Population size: 30 chromosomes.
- Probability of mutation: 1%.
- Probability of cross-over: 50%.
- Deletion groups: 5.
- Number of runs: 100.
- Maximum number of features selected in the same chromosome: 30.
- Average number of features per chromosome in the original population: 5.
- Maximum number of components: determined by cross-validation.
- Response: cross-validated % explained variance.

4. Results

4.1. OSC filtering of ^{1}H-NMR spectral data

Before undertaking analysis and feature selection, OSC was performed to improve the separability of two different classes by removing unwanted variations of data that do not contribute to discrimination of two SAA dietary phases. Table 1 summarizes the modeling results of two PLS-DA models. Three measures of a model’s ability to fit data and a predictive power, i.e., R^2X (cum), R^2Y (cum), and Q^2 (cum) were computed. Here R^2X(cum) and R^2Y (cum) represent a cumulative sum of the squares of **X** and **Y** explained, respectively. On the other hand, Q^2 (cum) represents a cumulative fraction of the total variance of **Y** that can be predicted by extracted components. For example, “PLS-DA with OSC” used 59.6% of the variation of **X** to explain 89.3% of the variation in **Y**. This model has a predictive power of Q^2 (cum) of 88.6%. As shown in Table 1, it turned out that the

"PLS-DA with OSC" model has better predictive power (i.e., more discriminative in this case) than the "PLS-DA without OSC" model: 0.886 vs. 0.225.

4.2. GA-based feature selection

To handle the previously mentioned overfitting problem common to GA, an iterative feature selection procedure was adopted. First, the average of three consecutive points of the NMR spectra was taken out of 574 metabolite features (viz., 191 "new" features generated using window size 3), and a GA-based feature selection was performed, choosing 183 features (viz., 61 "new" features). When a large window size is used for averaging, it may be possible to lose metabolite features that are potentially important to discrimination. To avoid this possibility, a GA-based feature selection was repeated with a smaller window (i.e., window size 0 at this time) and selection of 52 features of the preselected 183 features. In next section we will describe the results obtained at the last iteration of the GA-based feature selection.

An overfitting problem in a GA-based feature selection can be addressed using a randomization test (Leardi & Gonzalez, 1998). This test provides a metric to assess the risk of overfitting and is executed by randomizing **Y** relative to **X** and building a calibration model. The order of the elements of **Y** is randomized so that each row of **X** corresponds to the randomized **Y**, not to its own elements of **Y**. Thus, there is no information in such a dataset. It means that the calibration model constructed from a randomization test should have no significant prediction ability. If such a prediction ability is found, this finding indicates the presence of overfitting.

Fig. 4a shows the results of randomization tests. Here 50 GA runs were performed on randomized datasets, and in each run 100 chromosomes were evaluated. The fitness function of this randomization test is the average percentage of explained variances in cross-validation. The average value from the randomization test is 2.02, the progress of which in each run is shown in Fig. 4a as bold dashed lines. In general, the better or more reliable a dataset, the lower this average value. As a rule of thumb, GA can be applied safely without an overfitting problem when an average percent variance value is less than 10 (Leardi & Gonzalez, 1998).

A critical decision in implementing GA is to determine when to stop a GA run. Thus, one needs to select the optimal number of evaluations to perform in each GA run to obtain a good calibration model without an overfitting problem. Performing too many evaluations in each GA run means that noise in data is modeled. To determine the optimal number of evaluations, a total of 40 runs, each with 200 evaluations, were performed: The first 20 runs were based on "real" **Y** values and the last 20 runs were based on "randomized" **Y** values. Then, the difference between the averages of "real" and "randomized" runs was obtained as a function of the number of evaluations, which is shown in Fig. 4b. The optimal number of evaluations was chosen as 120, after which no significant increase in the degree of overfitting was observed.

4.3. Results of feature selection and discrimination

For feature selection, GA was performed on the datasets by using the optimal number of evaluations determined earlier (i.e., 120 evaluations in each GA run). To verify the robustness of the feature selection results, a GA-based feature selection procedure was run five times to produce five different sets of selected features. The main reason for this procedure is that the features selected from each GA model will not be identical because of the stochastic nature of GA. If some features are chosen only from one GA model, the possibility exists that they may have been selected merely by chance.

We extracted common information from the results of five GA models to select the most informative metabolite features. This information can be obtained from the frequency with which each feature is selected. Consequently, we selected those features that are consistently selected (i.e., high selection frequency) in the five GA models. Fig. 5a shows, as an example, a bar plot of the cumulative frequency of selections obtained from one of the five GA models. It shows which features are selected most often and which ones are rarely or never selected. The horizontal line of Fig. 5a represents the threshold values for selecting informative metabolite features on the basis of the cumulative frequency of selection, which was calculated by an *F*-test (Hunger & Huttner, 1999). The metabolite features selected, the full spectrum consisting of 183 features, and 136 observations are also shown in Fig. 5b. Here the regions corresponding to the selected metabolite features are denoted as a broken line at the bottom of Fig. 5b.

The number of features selected from a GA model was determined by investigating two plots of explained variance and root mean squared error in cross-validation (RMSECV) as a function of the number of selected features. For example, the two plots corresponding to Fig. 5a and b are shown in Fig. 5c and d. By referencing two threshold values (denoted as asterisk points) given at 60 in Fig. 5c and d, a total of 60 features were selected from this GA model. Similarly, the number of features to be selected was determined for the other four GA models. Fig. 6 shows the selected metabolite features obtained from the five GA models. The features selected are represented as five broken lines at the bottom of Fig. 6, indicating that the similar features are consistently chosen. These common features are the final solution for the GA-based feature selection, and consequently we selected a total of 52 metabolite features.

To investigate the potential advantage of using the GA-based feature selection combined with an OSC filter, classification was performed based on the metabolite features selected using a *k*-nearest neighbors algorithm. The experimental data were split into four groups corresponding to four individuals; each group includes the two different dietary phases (either zero SAA or supplemented SAA). Then classification results were obtained for each group: As a result, the use of the selected features yielded a zero average misclassification rate. In contrast, classification performance deteriorated when a full set of the original variables (i.e., 574) and their PCs (i.e., 23) were used to classify the same datasets. Average misclassification rates were 35.3% and 44.0%, respectively. In *k*-NN, different values of neighborhood parameters *k* were examined so that *k* = 4 was found to have the minimum misclassification rate. These classification results are consistent with score plots of PCA and PLS-DA as shown in Fig. 7. Compared with the score plots of Fig. 7a and c obtained using “all features,” those corresponding to “selected features” showed a clear discrimination of the NMR spectral data between the zero dietary SAA (sample numbers 1 through 68) and SAA-supplemented diet phases (sample numbers 69 through 136) (Fig. 7b and d).

5. Conclusion

This study presented the use of a combination of GA-based feature selection and an OSC filter as a method for analyzing high-resolution NMR spectra. A GA-based feature selection was performed to identify key metabolite features that retain most of information of the original data relevant to discrimination of different experimental conditions. The exclusion of noninformative metabolite features made it possible to produce better discrimination and classification results using fewer selected. In addition, an OSC was used to process the NMR data to remove unwanted variation of the data that was either orthogonal or unrelated to the discrimination of the two classes. The GA-based feature selection combined with an OSC filter produced better discrimination with no classification errors. The effectiveness of the presented approach was demonstrated using real NMR spectra in human plasma in which the ultimate goal is to characterize metabolic patterns attributed to SAA intake and identify important metabolite features that contribute to the distinction of the zero dietary SAA and SAA-supplemented diet phases.

Acknowledgments

This study was supported by grants from the National Institutes of Health: R03 DK066008, R03 ES012929, National Science Foundation CMMI-0644830 and the Emory General Clinical Research Center grant M01 RR00039.

References

Barker M, Rayens W. Partial least squares for discrimination. Journal of Chemometrics. 2003; 17:166–173.

Beckonert O, Bollard ME, Ebbels TMD, Keun HC, Antti H, Holmes E, et al. NMR-based metabonomic toxicity classification: Hierarchical cluster analysis and *k*-nearest-neighbour approaches. Analytica Chimica Acta. 2003; 490:3–15.

Blommers MJJ, Lucasius CB, Kateman G, Kaptein R. Conformation analysis of a dinucleotide photodimer with the aid of the genetic algorithm. Biopolymers. 1992; 32:45–52. [PubMed: 1617149]

DeWeijer AP, Lucasius CB, Buydens LMC, Kateman G, Heuvel HM, Mannee H. Curve fitting using natural computation. Analytical Chemistry. 1994; 66:23–31.

Esteban-Díez I, González-Sáiz JK, Gómez-Cámara D, Pizarro MC. Multivariate calibration of near infrared spectra by orthogonal WAVElet correction using a genetic algorithm. Analytica Chimica Acta. 2006; 555:84–95.

Fearn T. On orthogonal signal correction. Chemometrics and Intelligent Laboratory Systems. 2000; 50:47–52.

Gourvénec S, Capron X, Massart DL. Genetic algorithms (GA) applied to the orthogonal projection approach (OPA) for variable selection. Analytica Chimica Acta. 2004; 519:11–21.

Hibbert DB. Genetic algorithms in chemistry. Chemometrics and Intelligent Laboratory Systems. 1993; 19:277–293.

Holmes E, Antti H. Chemometric contributions to the evolution of metabonomics: Mathematical solutions to characterising and interpreting complex biological NMR spectra. Analyst. 2002; 127:1549–1557. [PubMed: 12537357]

Holmes E, Nicholson JK, Tranterm G. Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks. Chemical Research in Toxicology. 2001; 14:182–191. [PubMed: 11258967]

Hoskuldsson A. PLS regression methods. Journal of Chemometrics. 1988; 2:211–228.

Hunger J, Huttner G. Optimization and analysis of force field parameters by combination of genetic algorithms and neural networks. Journal of Computational Chemistry. 1999; 20:455–471.

Jansen JJ, Hoefsloot HCJ, Boelens HFM, Greef JVD, Smilde AK. Analysis of longitudinal metabolomics data. Bioinformatics. 2004; 20:2438–2446. [PubMed: 15087313]

Kemsley EK. A hybrid classification method: Discrete canonical variate analysis using a genetic algorithm. Chemometrics and Intelligent Laboratory Systems. 2001; 55:39–51.

Kourti T. Application of latent variable methods to process control and multivariate statistical process control in industry. International Journal of Adaptive Control and Signal Processing. 2005; 19:213–246.

Leardi R. Genetic algorithms in chemometrics and chemistry: A review. Journal of Chemometrics. 2001; 15:559–569.

Leardi R, Gonzalez AL. Genetic algorithms applied to feature selection in PLS regression: How and when to use them. Chemometrics and Intelligent Laboratory Systems. 1998; 41:195–207.

Leardi R, Seasholtz MB, Pell RJ. Variable selection for multivariate calibration using a genetic algorithm: Prediction of additive concentration in polymer films from Fourier transform-infrared spectral data. Analytica Chimica Acta. 2002; 461:189–200.

Lee GC, Woodruff DL. Beam search for peak alignment of NMR signals. Analytica Chimica Acta. 2004; 487:189–199.

Lindon JC. Metabonomics – Techniques and applications. Business Briefing: Future Drug Discovery. 2004:1–6.

Lindon JC, Holmes E, Nicholson JK. Pattern recognition methods and applications in biomedical magnetic resonance. Progress in Nuclear Magnetic Resonance Spectroscopy. 2001; 39:1–40.

Nicholson JK, Connelly J, Lindon JC, Holmes E. Metabonomics: A platform for studying drug toxicity and gene function. Nature Reviews Drug Discovery. 2002; 1:153–161.

Nicholson JK, Lindon JC, Holms E. Metabonomics: Understanding the metabolic responses of living systems to patho-physiological stimuli via multi-variate statistical analysis of biological NMR spectroscopic data. Xenobiotica. 1999; 29:181–1189.

Sjoblom J, Svensson O, Josefson M, Kullberg H, Wold S. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. Chemometrics and Intelligent Laboratory Systems. 1998; 44:229–244.

Solanky KS, Bailey NJC, Beckwith-Hall BM, Davis A, Bingham S, Holmes E, et al. Application of biofluid ^{1}H nuclear magnetic resonance-based metabonomic technique for the analysis of the biochemical effects of dietary isoflavones on human plasma profile. Analytical Biochemistry. 1993; 323:197–204. [PubMed: 14656525]

Wang Y, Holmes E, Nicholson JK, Cloarec O, Chollet J, Tanner M, et al. Metabonomic investigations in mice infected with Schistosoma mansoni: An approach for biomarker identification. Proceedings of the National Academy of Sciences USA. 2004; 101:12676–12681.

Westerhuis JA, de Jong S, Smilde AK. Direct orthogonal signal correction. Chemometrics and Intelligent Laboratory Systems. 2001; 56:13–25.

Wold S, Antti H, Lindgren F, Öhman J. Orthogonal signal correction of near-infrared spectra. Chemometrics and Intelligent Laboratory Systems. 1998; 44:175–185.

Wold S, Geladi P, Esbensen K, Öhman J. Multi-way principal components-and PLS-analysis. Journal of Chemometrics. 1987; 1:41–56.

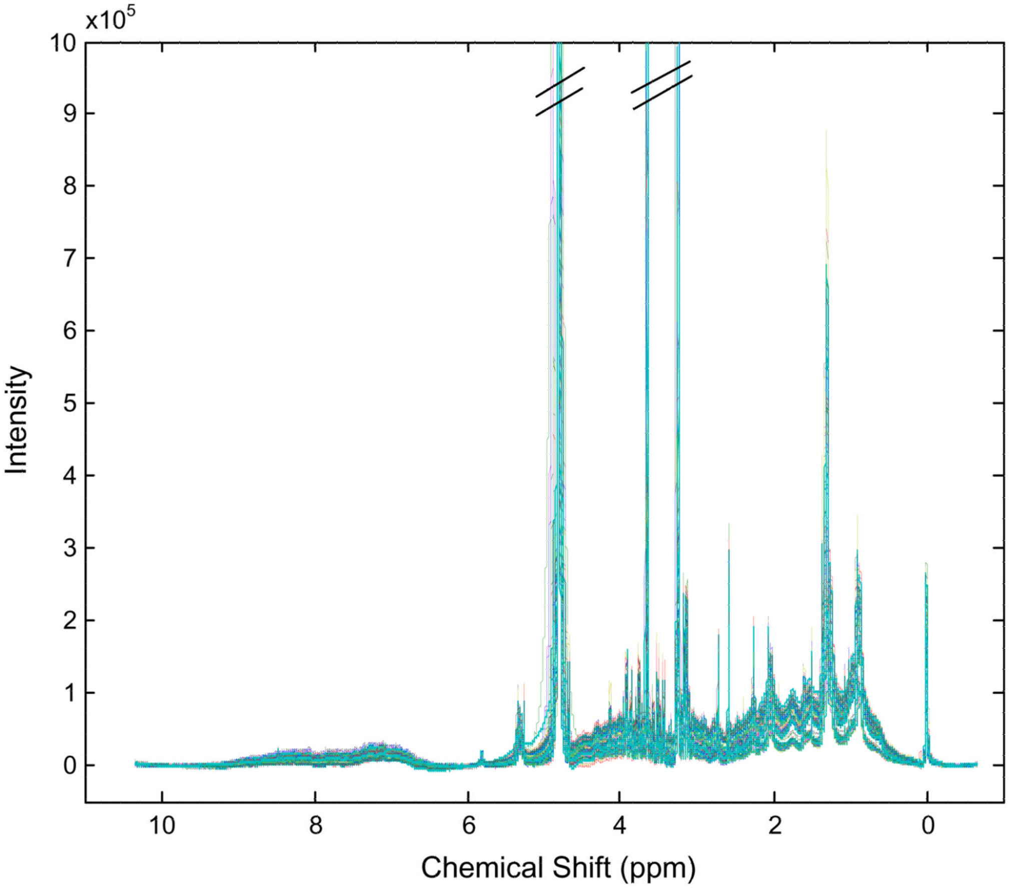

Fig. 1.
Multiple ^{1}H-NMR spectra of human plasma obtained by a 600 MHz NMR spectroscopy.

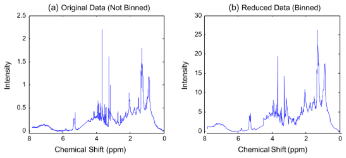

Fig. 2.

The 600 MHz ^{1}H-NMR spectra of human plasma: (a) original spectrum after preprocessing steps (number of features = 8445) and (b) reduced spectrum by binning (number of features = 574).

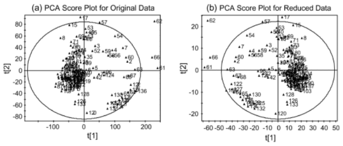

Fig. 3.

Two-dimensional PCA score plots for (a) original spectrum and (b) reduced spectrum.

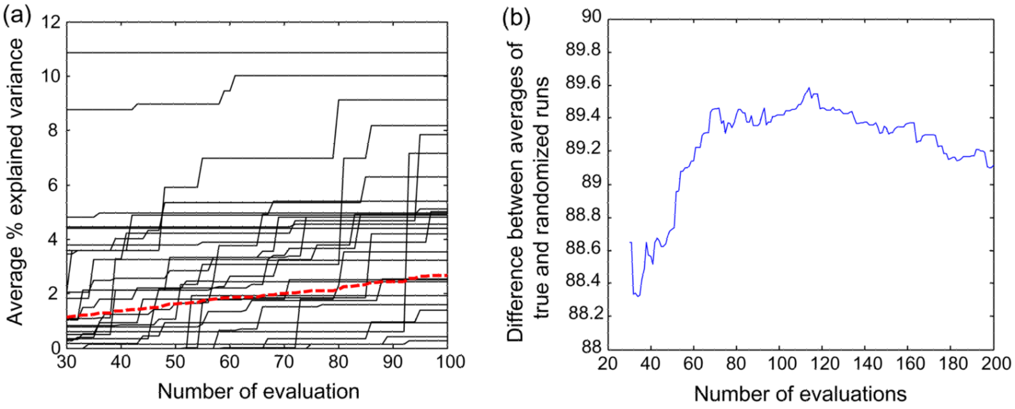

Fig. 4.

Plots for (a) randomization test and (b) optimal number of evaluations.

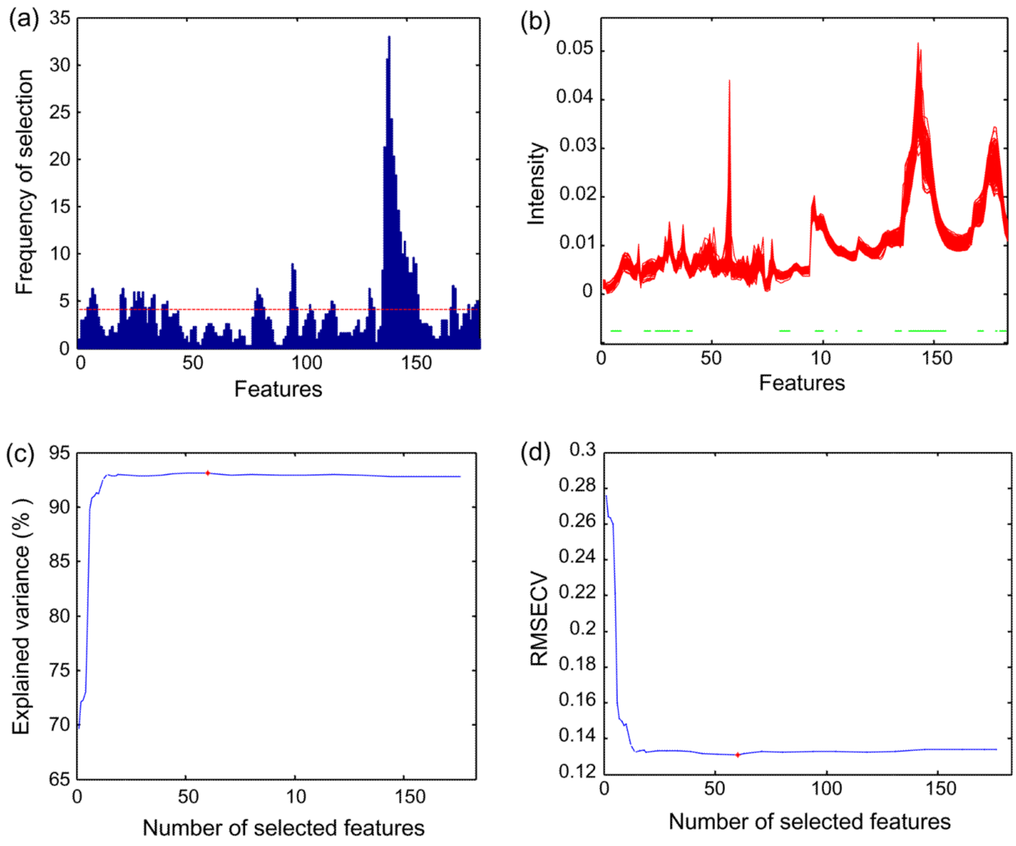

Fig. 5.

Feature selection plots: (a) cumulative selection frequency, (b) selected spectral regions, (c) explained variance (%), and (d) RMSECV.

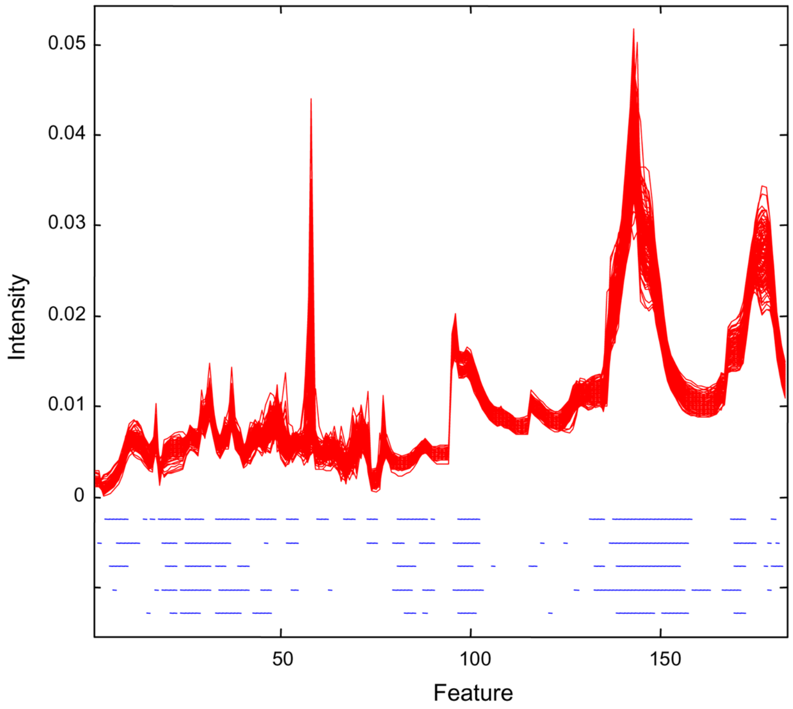

Fig. 6.

Plots of selected metabolite features obtained from the final iteration of GA-based feature selection.

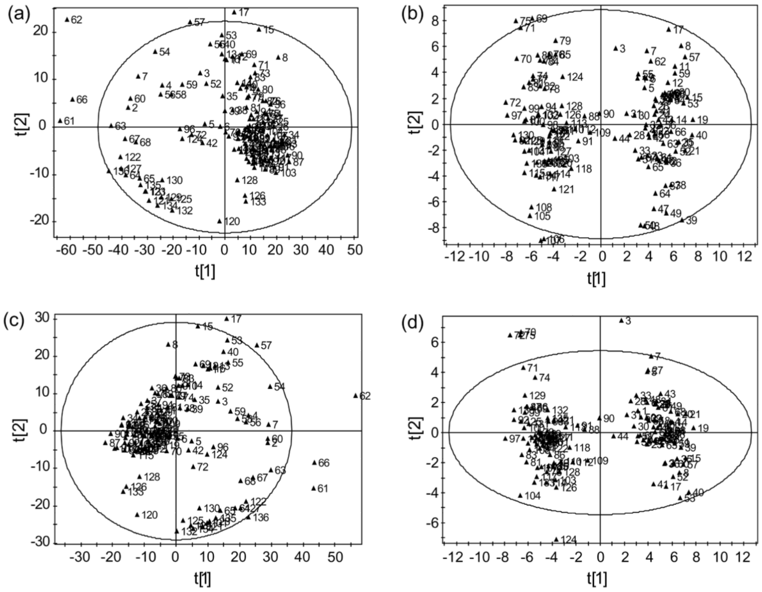

Fig. 7.

Score plots of (a) PCA using all variables, (b) OSC-PCA using selected variables, (c) PLS-DA using all variables and (d) OSC-PLS-DA using selected variables.

Table 1

Overview of PLS-DA models with/without OSC preprocessing

A[a]	PLS-DA without OSC			PLS-DA with OSC		
	R^2X (cum)	**R^2Y (cum)**	**Q^2 (cum)**	**R^2X (cum)**	**R^2Y (cum)**	**Q^2 (cum)**
1	0.580	0.130	0.120	0.339	0.824	0.822
2	0.830	0.240	0.225	0.596	0.893	0.886

[a] Indicates the number of latent components retained in the PLS-DA models.