Metabolomics (2011) 7:549–558
DOI 10.1007/s11306-011-0274-7

ORIGINAL ARTICLE

A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection

Xiaohui Lin · Quancai Wang · Peiyuan Yin · Liang Tang · Yexiong Tan · Hong Li · Kang Yan · Guowang Xu

Received: 14 October 2010 / Accepted: 6 January 2011 / Published online: 20 January 2011

Abstract Metabolic markers are the core of metabonomic surveys. Hence selection of differential metabolites is of great importance for either biological or clinical purpose. Here, a feature selection method was developed for complex metabonomic data set. As an effective tool for metabonomics data analysis, support vector machine (SVM) was employed as the basic classifier. To find out meaningful features effectively, support vector machine recursive feature elimination (SVM-RFE) was firstly applied. Then, genetic algorithm (GA) and random forest (RF) which consider the interaction among the metabolites and independent performance of each metabolite in all samples, respectively, were used to obtain more informative metabolic difference and avoid the risk of false positive. A data set from plasma metabonomics study of rat liver diseases developed from hepatitis, cirrhosis to hepatocellular carcinoma was applied for the validation of the method. Besides the good classification results for 3 kinds of liver diseases, 31 important metabolites including lysophosphatidylethanolamine (LPE) C16:0, palmitoylcarnitine, lysophosphatidylethanolamine (LPC) C18:0 were also selected for further studies. A better complementary effect of the three feature selection methods could be seen from the current results. The combinational method also represented more differential metabolites and provided more metabolic information for a "global" understanding of diseases than any single method. Further more, this method is also suitable for other complex biological data sets.

Keywords Support vector machine · Genetic algorithm · Random forest · Liver diseases · Metabonomics · Metabolomics

1 Introduction

Metabonomics is an important platform of systems biology which provides holistic metabolic information of living systems for the clinic and pharmaceutical industry (Nicholson 2006). By the quantitative measurement of the metabolites and their dynamic changes in biological samples, metabonomics has been widely used in many areas such as drug toxic study and disease diagnosis (Yang et al. 2004). Usually the information-rich metabonomics data are acquired from NMR spectroscopy or mass spectrometry, therefore, chemometric methods are necessary for data analysis (Chan et al. 2007). A major goal of the metabonomic data analysis is to find out informative biomarkers for the subsequent study, e.g. diagnosis or prognosis of diseases (Kim et al. 2007; Pisitkun et al. 2006). Many multivariate analysis techniques, such as principal components analysis (PCA), partial least squares discriminant analysis (PLS-DA), and support vector machine recursive feature elimination (SVM-RFE) have

X. Lin (✉) · Q. Wang · H. Li · K. Yan
School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
e-mail: datas@dlut.edu.cn

P. Yin (✉) · G. Xu
CAS Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China
e-mail: yinpy@dicp.ac.cn

L. Tang · Y. Tan
International Cooperation Laboratory on Signal Transduction, Eastern Hepatobiliary Surgery Institute, The Second Military Medical University, Shanghai, China

been used for searching the potential biomarkers (Guyon et al. 2002; Maher et al. 2008; Righi et al. 2009).

PCA is an unsupervised method, it projects the raw data into a few principal components to keep the sufficient information of the data, and provides systemic views of the raw data (Jolliffe 2002). PLS-DA is a supervised method, it rotates the PCA components by using the response value to sharpen the separation between the classes of samples (Balding et al. 2007; Kima et al. 2010). Two methods reduce the dimension and summarize the original features into a few new components by calculating their loadings. They could measure the variable contribution for discriminating different groups by loadings (Ramadan et al. 2006). The variable influence on the projection (VIP) is a weighted sum of PLS loadings, it is quite common to be used to select the important features from metabonomic data (Bryan et al. 2008; Cho et al. 2008). PLS-DA also could evaluate the features by the sum of PLS-regression coefficients (Bryan et al. 2008). PCA and PLS-DA have been used in many metabonomics studies (Solank et al. 2003; Stella et al. 2006; Wanga et al. 2009).

Recently machine learning algorithms have shown to be an effective tool in analyzing genomics, proteomics and metabonomic data (Díaz-Uriarte and de Andrés 2006; Saeys et al. 2007; Trevino and Falciani 2006; Laxman et al. 2010). SVM is a very popular method and has been proved to be superior to PLS-DA in some cases (Mahadevan et al. 2008; Vapnik 1998). SVM-RFE is quite efficient in mining the most problem-related information involved in the data. It focuses on the samples on the margin, and evaluates the features by the support vectors. It can be extended to nonlinear cases with the help of kernels. Random forest (RF) applying an ensemble technique is an effective method in classification and feature selection (Breiman 2001). It is not sensitive to the data scaling and can evaluate the features by "permutation importance" (Strobl et al. 2007). As a supervised method, it has no problem of overfitting based on the large number theory (Breiman 2001). Genetic algorithm (GA) is a heuristic search that mimics the process of the natural selection and genetic. The appeal of GAs comes from their simplicity and elegance as robust search algorithms as well as from their power to discover good solutions rapidly for difficult high-dimensional problems (Li et al. 2001; Ooi and Tan 2003; Zou and Tolstikov 2008, 2009).

One major difference of SVM-RFE, GA and RF from PLS-DA is that they select the features by directly measuring their influence on the classification, while PLS-DA remaps the raw data to a small space and select the features by measuring their projection on the main components. Now more and more field scientists prefer to choose machine learning techniques, especially when PLS-DA fails to give credible or good solutions (Bhattacharyyas et al. 2006; Mahadevan et al. 2008; Man et al. 2004).

Chronic diseases are generally considered as systemic pathological changes, which may influence the metabonome in many aspects of different pathways. Therefore, the filtered metabolic markers should represent not only the differences, but also a comprehensive understanding of the metabolism disorders. Unfortunately, each feature selection technique measures features according to its own principle to get a good classification result, which may neglect some other important features that are out of the scope of the principle. In this case, the combination of these methods and development of new processing method would be important for metabonomics studies.

In this study, a new method based on the SVM classifier, feature selection methods including SVM-RFE, GA and RF was utilized for the filtration of metabolic differences in complex data. A data set which contains plasma metabolite data of rat liver diseases from hepatitis, cirrhosis to hepatocellular carcinoma was used to show the usefulness of the method.

2 Theory

2.1 SVM-RFE

SVM-RFE is a feature selection method based on SVM, which searches for the hyper-plane that maximizes the distance between each class (Vapnik 1998). Let (x_i, y_i) $(i = 1, 2,\ldots,n)$ denote a sample pair in the training set, where $x_i \in R^k$, y_i is the class label of sample x_i. The support vector machines require the solution of the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \qquad (1)$$

subject to $y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0$. Here $C > 0$ is the penalty parameter which is a tradeoff between the training accuracy and predictive ability (Vapnik 1998).

SVM-RFE performs feature selection in a backward sequence (Guyon et al. 2002). Firstly, a SVM classifier is constructed using all the input variables; the classification accuracy rate and the squared coefficient of the weight of each feature are computed. Then the features with the smallest squared coefficient of the weight are deleted from the feature set. This procedure is repeated until the feature set is empty.

In each iteration, the accuracy rate of the SVM learning model is calculated through 7-fold cross-validation, the maximal classification accuracy rate is kept and the corresponding feature subset is the selected feature subset (Xue et al. 2006).

2.2 Genetic algorithm for feature selection

GAs are randomized search and optimization algorithms inspired by Darwin's theory of evolution (Holland 1992). In GAs, each chromosome represents a potential solution (a candidate feature subset, the number of the features in it is denoted as the chromosome size) and the fitness value associated with each chromosome stands for the degree of its goodness to the best solution. The fitness criterion can be changed flexibly according to the goal of the optimization problems, such as the classification accuracy rate or other criterion. GAs start with an initial population which consists of some randomly generated chromosomes and a criterion function to evaluate the fitness of the chromosomes. Then by applying evolutionary operations, an offspring population with the same number of individuals is produced. The genetic operations which contain selection, crossover and mutation are repeated in the process of evolution until some chromosome achieves the goal fitness. The goal fitness is defined to ensure the average of a reasonable number of generations could be achieved (Zou and Tolstikov 2008).

A discriminating feature subset is generated after one GA evolution stops. By performing the genetic algorithm many times, a number of subsets of features that potentially discriminate the three stages of liver disease have been selected. The frequencies of all the features can be counted through the statistical analysis. The feature frequency is correlated with the relative predictive ability. Then the features are ranked according to their frequencies and the top ranked features are selected for the analyzing of the liver disease.

2.3 Random forest

Random forest is an ensemble method consisting of many tree classifiers. For each classification tree, a bootstrap sample is drawn from the original samples (Breiman 2001). At each non-leaf node of a classification tree, the best split feature is selected from a small random subset of the original features. When the forest receives an input vector, each classification tree casts a unique vote, the final prediction is determined by the majority votes of all the trees in the random forest.

Since the bootstrap sample is drawn with replacement, the samples which are not in the bootstrap samples are called out-of-bag (OOB) data (Breiman 2001). The OOB data can be used to estimate the prediction error of the random forest.

Random forest can measure variable importance through permutation (Archer and Kimes 2008; Strobl et al. 2007). For a variable (feature) a_j, random permutation is conducted on the OOB data, their original association with the class label is broken. When the permuted feature together with all the other features is used to predict the OOB data, the prediction accuracy ratio will decrease dramatically if the permuted feature is highly related to the class label. Hence the degree of the accuracy decreasing is used to evaluate the importance of the feature. All the features are ranked according to their importance and the top ranked features are the most informative ones according to random forest. Feature's relevance has influence on feature's score, therefore, we use feature's relevance to modify feature's score (Lee et al. 2008).

It can be known from the above theories that the three techniques select the meaningful features from different points of view. SVM-RFE conducts the feature selection based on support vector machines to find a hyper-plane to separate the largest interval categories, and thus get a generalization. In each iteration, it removes the variable which has little impact on the margin. On the other hand, because a set of variables (arranged in chromosomes) are tested in combination to get the fitness during the selection process, GA takes into consideration the interaction among features and the synergy between metabolites (Trevino and Falciani 2006). RF evaluates each feature by testing its own impact on the classification. Hence combination of the three methods will be possible to obtain the information related to the margin, combination of the features and independent features.

3 Experimental

3.1 Animals

Male rats at 6-week-old Sprague–Dawley (SD) ($n = 80$, 120–150 g) were supplied by the Shanghai Experimental Animal Center, Chinese Academy of Science (Shanghai, China). The rats were maintained in pathogen-free environmental conditions with a regular 12 h light/dark cycle, a room temperature of 22°C and a relative humidity of 55%. Rats were randomly assigned to two groups. 10 SD rats in the model group received intraperitoneal injections of DEN at 70 mg/kg body weight once a week for 10 weeks. And 10 rats in the control group received intraperitoneal injections of sodium chloride. Plasma were collected after 8 h fasting once every 2 weeks and stored at −80°C until analysis.

Animal experiments conformed to the Guide for the Care and Use of Laboratory Animals from the Second Military Medical University. The protocols of the experiments were approved by the Ethical Committee of the Second Military Medical University.

The plasma was thawed before analysis. Acetonitrile (200 μl) was added to plasma (50 μl) and shaken vigorously, then the mixture was laid at 4°C for 10 min, centrifuged at 15,000×g for 10 min at 4°C. The supernatant (4 μl) was injected into the column.

3.2 Metabolic profiling

Agilent 1200 Rapid Resolution Liquid Chromatography (RRLC) system (Agilent, USA) was used for metabolic profiling. A reversed phase C18 column (10 cm × 2.1 mm 1.8 μm, ZORBAX TM SB-AQ Agilent, USA) was used for the separation. Mobile phase A was water with 0.1% formic acid (v/v), and mobile phase B was acetonitrile. An Agilent 6510 quadrupole-time of flight mass spectrometer (Q-TOF MS, Agilent, USA) was used as detector. The detail settings of the chromatography and mass spectrometer were similar and reported in detail in our former work (Yin et al. 2009).

Molecular Feathers Extraction (Agilent, USA) was used for peak detection, and the results were exported to Genespring (Agilent) for peak alignment. The retention time (tr) and m/z data were exported to Excel table (Microsoft, USA) for data analysis.

3.3 Data analysis

The PLS-DA model was implemented with the help of SIMCA-P 11.0 (Umetrics AB, Umea, Sweden) software. The algorithms of SVM-RFE and RF were written in C/C++. SVM-RFE adopted a linear SVM with the penalty parameter $C = 1$ and the implementation of SVM is from LIBSVM (available at http://www.csie.ntu.edu.tw/~cjlin/libsvm). The implementation RF is available at our web site (http://www.402.dicp.ac.cn/download_Random%20forest%20program.htm). The size of the random forest was set to 100. For each classification tree, its training set was selected with replacement and *mtry* was set as the empiric value, sqrt(m), where m is the number of the features. The implementation of GA was from the R package GALGO (Trevino and Falciani 2006) (available at http://www.bip.bham.ac.uk/vivo/galgo/AppNotesPaper.htm#_Toc12 7597486) and a SVM classifier was used to evaluate the fitness of the chromosomes. In GALGO, the GA was run 1000 times. The goal fitness parameter was set as 0.96. The chromosome size was set as 5. The other parameters were set as their default values in the package. SVM-RFE, GA and RF were performed on the computer: Pentium (R) Dual-Core CPU @ 2.93 GHz and 1.99G memory. They run times on our data are 65 s, 67,139 s and 3 s, respectively.

4 Results and discussion

The data of liver diseases contained the plasma metabonomic profiles of the 10 model rats at 6 w (the 6th week), 8 w,…,18 w and 20 w. According to the clinical diagnosis, the model rats at 6 w and 8 w are hepatitis, at 12 w and 14 w are cirrhosis and at 18 w and 20 w are hepatocellular carcinoma (HCC). After peak alignment, the collected data contained 1459 ion features. Screening of significantly changed metabolites in various stages of liver disease development is the target of this work. In order to compare the change of the selected metabolites between the controls and models, the plasma metabonomic profiles of the 10 normal rats at 6 w, 8 w, 10 w,…,18 w and 20 w were also collected.

Because there are many noises included in the profiling data, analysis of variance (ANOVA) were conducted to filter out some unimportant features. In general, the significant variation has the P-value smaller than 0.05 ($P < 0.05$). It was found that there are 489 variables left by filtering out the features with $P \geq 0.05$, 970 variables were deleted, the remaining data were used for the following feature selection.

4.1 PLS-DA

Firstly, PLS-DA was used to analyze the data. The R^2Y value is the variation of the Y variable that can be explained by the selected component. The Q^2 value is a cross-validated measure of R^2. According to the regular index of PLS-DA, a PLS-DA model's cumulative R^2Y and cumulative Q^2 values should be large, and Q^2 of each new component should be larger than a significant limit (Eriksson et al. 2006). Figure 1 gives the detail information about the PLS-DA model. With the increase of the component number, the R^2 and Q^2 values increase. When five significant components were chosen, its R^2Y and Q^2 were 0.949 and 0.825, respectively. It seems that the model has a good interpretation and predictive ability.

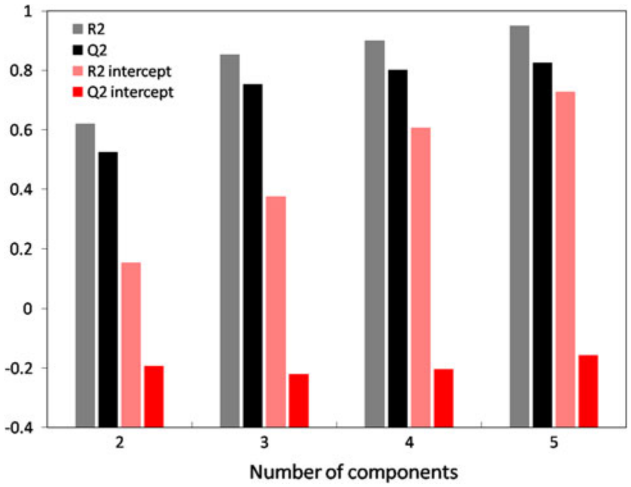

Fig. 1 PLS-DA model parameters with number of different components

Since PLS-DA is a supervised technique, the possibility of overfitting should never be neglected. Overfitting could distort real group relationships and impair the reliability of multivariable analysis (Defernez and Kemsley 1997). Therefore, the permutation test and randomization test were adopted to validate the model. In the R^2 and Q^2 validation plot, by randomly permuting the class labels of the samples while keeping the samples data intact, different models can be constructed using the same method. The corresponding R^2 and Q^2 values can be calculated. Hence, if the learning method has a good predictive ability for new samples, all the Q^2 and R^2 values on the permuted data sets should be lower than the values on the actual data set. It can be seen that the increasing of R^2 intercept may bring overfitting risks. When the number of the components increases to 5, the model is overfitted. With 2 components, the intercepts of R^2 and Q^2 are 0.154 and −0.194, respectively, and the model is without overfitting. However, its Q^2 value is 0.525, which indicates it has a poor predictive ability. Considering the goals of the data handling, only when a good model with satisfied classification ability is established, the consequent feature selection could be effective and meaningful. Thus, in the following study, new classification model and feature selection methods were applied.

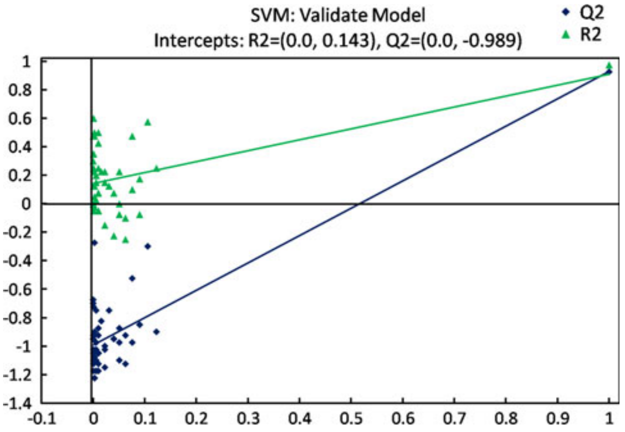

Fig. 2 Permutation test plot of SVM

4.2 SVM and SVM-RFE

For SVM, the visualization of the classification is difficult, especially for the multi-class problem. Here, we adopted the 7-fold cross validation and leave-one-out cross validation (LOOCV) to validate the model. The classification accuracy rates of 7-fold cross validation and LOOCV of the SVM model (where a linear kernel function was adopted) for the metabolic data of liver diseases are 95% and 96.67%, respectively. This indicates that the SVM model has a good discriminative ability for different liver diseases. The R^2 and Q^2 validation plot given in Fig. 2 shows that the model is credible without overfitting (Mahadevan et al. 2008).

By considering the roles of the features in SVM classification, SVM-RFE was conducted to measure the importance of the ion features and select the meaningful information for discriminating different stages of liver disease. SVM-RFE is a backward elimination procedure, which starts with the entire data set. In each iteration, it deletes one bottom ranked feature depending on the feature weights evaluated by SVM model, and the corresponding accuracy rate is calculated. Figure 3 shows the accuracy rate varying with the feature deletion. It can be observed that the classification accuracy rate of SVM learning model with all input variables is 95%. As the noisy and nonrelated features are deleted, the accuracy rate increases. When the feature number is reduced to 12, the classification accuracy rate arrives at 100%. And if one more feature is deleted, the rate will apparently decrease. Hence the 12 features (see Table 1) are the least number of the most significant metabolites related to the disease pathology according to SVM-RFE. The R^2, Q^2, R^2 intercept and Q^2 intercept (see Table 2) of the SVM model based on these

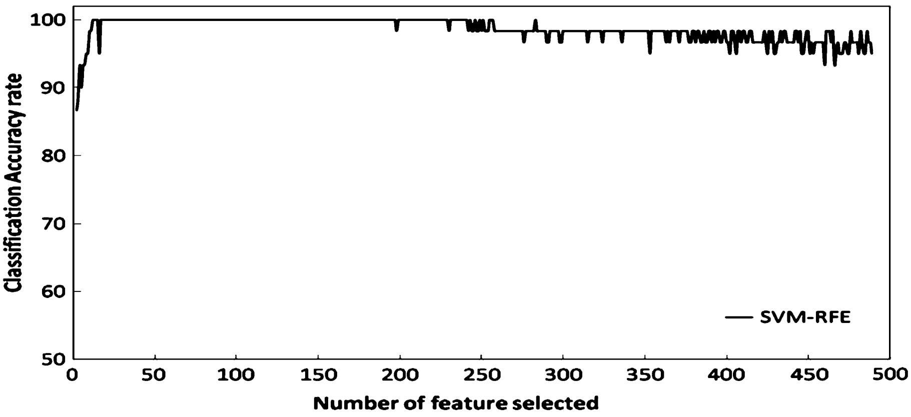

Fig. 3 Feature selection processes of SVM-RFE

Table 1 Significantly changed metabolites selected by SVM-RFE, GA and RF, respectively

ID	Mass	RT	*P*-value (a)			*P*-value (b)		
			6–8 w	12–14 w	18–20 w	(6–8 w)–(12–14 w)	(12–14 w)–(18–20 w)	(18–20 w)–(6–8 w)
1[a,b,c]	399.34	13.499	7.80E−03	6.45E−01	6.92E−01	2.51E−4	6.19E−06	2.58E−10
2[b,c]	453.29	14.026	4.06E−03	1.70E−10	7.90E−08	2.40E−08	8.30E−01	5.97E−06
3[a,c]	481.32	16.287	7.94E−03	9.51E−10	9.40E−09	7.87E−08	8.09E−01	2.05E−06
4[a]	159.12	0.683	1.28E−01	5.18E−03	9.31E−02	4.18E−05	6.50E−02	4.71E−03
5[a,c]	234.12	8.306	3.18E−02	4.77E−01	3.19E−01	2.69E−09	6.37E−01	1.07E−10
6[a]	272.20	21.111	9.08E−01	9.01E−03	9.68E−01	1.69E−05	7.87E−02	3.35E−03
7[a,c]	276.21	17.325	1.31E−02	1.37E−01	4.69E−01	5.48E−03	3.15E−05	2.58E−13
8[b,c]	674.53	23.492	3.86E−01	4.27E−01	2.75E−01	2.10E−02	7.09E−04	3.40E−06
9[b,c]	310.23	19.548	7.55E−02	1.59E−01	2.19E−01	2.22E−03	1.58E−09	4.06E−08
10[b]	312.27	14.038	1.56E−02	1.03E−09	2.55E−08	7.56E−08	8.51E−01	1.92E−06
11[b,c]	328.24	19.551	1.64E−02	1.87E−01	3.48E−02	5.01E−05	4.25E−08	5.92E−11
12[b]	357.29	10.493	1.89E−02	9.78E−06	3.44E−05	4.67E−03	2.68E−03	7.79E−01
13[a,b,c]	435.27	14.026	6.81E−03	6.41E−10	4.14E−08	9.90E−09	8.05E−01	1.16E−06
14[a,b,c]	497.35	12.68	2.63E−01	5.09E−02	8.38E−02	2.21E−02	7.63E−02	4.27E−06
15[a,b,c]	544.63	13.554	5.04E−07	3.07E−04	3.99E−01	5.91E−04	4.07E−04	8.27E−10
16[b,c]	330.26	20.288	3.31E−03	1.20E−02	3.28E−02	3.64E−04	3.86E−07	1.06E−09
17[a,b]	569.35	14.647	7.77E−01	2.55E−01	2.95E−04	3.93E−01	3.24E−05	1.64E−05
18[a]	592.23	12.678	2.88E−01	3.42E−04	4.20E−01	7.45E−01	1.13E−03	1.60E−03
19[a]	618.29	16.437	6.93E−01	4.82E−01	9.27E−02	5.62E−01	2.48E−02	5.62E−02
20[b]	453.29	13.619	6.34E−04	2.56E−10	4.32E−07	8.83E−07	5.96E−01	2.60E−04
21[c]	616.40	7.01	–	2.66E−01	3.95E−01	4.37E−04	–	1.12E−04
22[c]	302.22	18.465	2.75E−03	1.44E−03	1.38E−03	2.83E−02	9.61E−02	1.59E−06
23[c]	281.27	14.027	4.73E−01	4.45E−08	3.35E−07	8.42E−06	4.27E−06	6.69E−06
24[c]	309.30	16.288	1.41E−01	4.03E−06	8.57E−07	6.95E−07	0.416953	3.08E−06
25[c]	467.34	13.601	4.40E−01	3.06E−02	4.50E−02	3.05E−04	3.74E−01	6.32E−08
26[c]	732.49	8.622	2.19E−02	1.56E−02	4.17E−01	1.05E−05	1.37E−03	6.23E−06
27[c]	551.39	17.161	–	4.34E−01	2.46E−01	4.53E−04	1.92E−01	4.66E−03
28[c]	330.18	11.786	1.29E−01	1.95E−01	2.00E−02	3.62E−05	1.64E−01	1.99E−02
29[c]	413.35	14.354	2.40E−06	1.78E−01	1.02E−01	2.47E−08	8.16E−04	1.11E−08
30[c]	344.20	7.447	7.76E−05	3.46E−01	2.11E−02	6.94E−04	2.36E−01	1.8E−07
31[c]	279.26	10.348	6.08E−03	1.29E−04	9.08E−04	2.84E−04	3.75E−04	3.22E−06

The first three columns mean feature's ID, fragment ion mass and retention time, respectively. (a) *P*-value is calculated between control group and model group in the same time point lasting 2 weeks using *t*-test. (b) *P*-value is calculated in the model group using *t*-test, column 7 to column 9 mean the comparison of hepatitis and cirrhosis, cirrhosis and HCC, hepatitis and HCC, respectively. [a]Marks the features selected by SVM-RFE, [b] Marks the features selected by GA, [c] Marks the features selected by RF. "–" means that its denominator is zero for calculating *P*-value, in other words, the values in all the samples are zero in hepatitis, cirrhosis or HCC group for this feature

Table 2 Overfitting tests for the models constructed based on selected features

Method	Classification accuracy rate (%)	Feature number	R^2	Q^2	R^2 intercept	Q^2 intercept
SVM-RFE	100	12	1	1	−0.192	−0.98
GA	100	13	1	1	−0.254	−1.187
RF	100	23	1	0.975	0.245	−1.00
Combination	100	31	1	1	0.252	−1.05

12 features imply that the model is reliable without overfitting.

Among the 12 ion features, two are identified as palmitoylcarnitine and lysophosphatidylethanolamine (LPE) C18:0, their mass and retention time (RT) are (339.34, 13.49) and (481.32, 16.28), respectively. Figure 4 shows the changes of LPE C18:0 in different periods of the liver diseases and the control group. The metabolite contents were stable in the control group, whereas increased significantly in the model group, especially in the stages of cirrhosis and HCC.

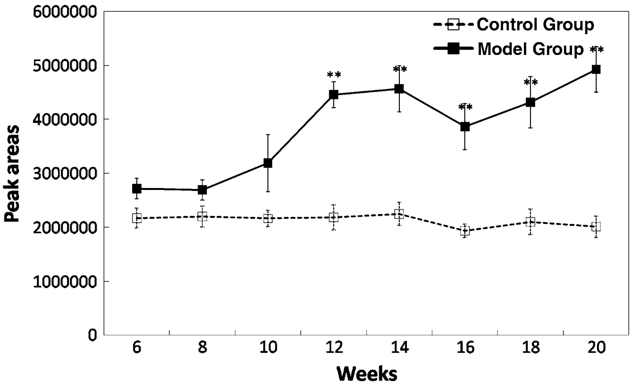

Fig. 4 Lysophosphatidylethanolamine C18:0 temporal changes in the model (*solid line*) and control group (*dashed line*)

4.3 GA and RF

SVM-RFE is an efficient feature selection technique, which has been widely used in processing biological data. However, it only focuses on the support vectors and puts little consideration on features relevance. To find out more important differential metabolites and gain a better "systemic" understanding of the organism, different methods of feature selection are intensively required.

Thus GA and RF were recruited as complementary methods. GA takes the interaction among features and the synergy between metabolites into consideration (Trevino and Falciani 2006), whereas RF considers the function of each variable in all the data samples. Hence combination of the three methods will increase the possibility to find out more information related to the margin, combination of the features and independent features.

GA was conducted 1000 times, and 1000 feature subsets were generated. Each feature subset contained 5 features. All the features were ranked according to their occurrence frequencies in these subsets. Then the top ranked features were inserted one by one into the SVM model and the corresponding accuracy rate was calculated. The least top ranked features which could achieve the maximum accuracy rate were selected (Fig. 5a). It could be seen that the top 13 ranked features are the most significantly related to discriminating the three different stages of the liver disease. No overfitting could be found based on the parameters of the corresponding SVM model (Table 2).

A random forest with 100 tree classifiers was also built. The features were measured by "permutation importance", and then all the features were ranked. SVM was also used as that in the GA method to determine how much the top ranked features were selected. It can be observed from Fig. 5b that the top 23 ranked features are the most significantly related to discrimination among the three different stages of the liver disease. Also the parameters of the corresponding SVM model in Table 2 prove the credibility of the selected features.

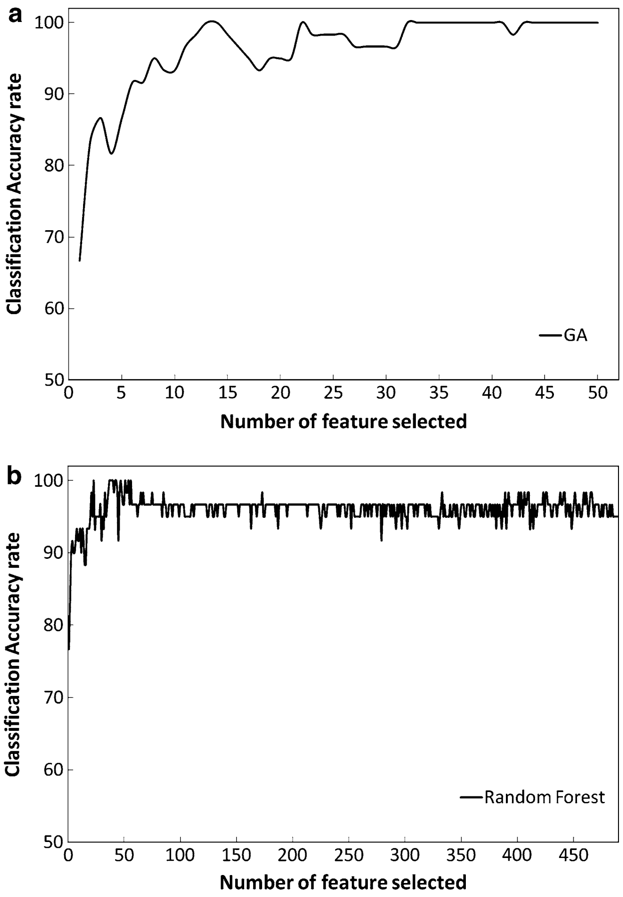

Fig. 5 Feature selection process of GA (**a**) and RF (**b**)

4.4 Comparison of features selected by SVM-REF, RF and GA

All of the features selected by SVM-RFE, GA and RF are given in Table 1. Based on these 31 ion features from 3 methods, a SVM classification model was constructed again to validate their ability of classification. The R^2, Q^2, R^2 intercept and Q^2 intercept (Table 2) are 1, 1, 0.252 and −1.05, respectively, indicating that no overfitting occurs.

Though three techniques adopt different strategies, some most importantly related features are found by all or two of them, such as (435.2742, 14.026), LPC (20:4) (544.6331, 13.554) and LPC (22:5) (569.3475, 14.647) (Fig. 6). (435.2742, 14.026) can significantly discriminate between liver disease and control group. LPC (20:4) (544.6331, 13.554) shows significant difference between models and controls in hepatitis and cirrhosis stages, while LPC (22:5) (569.3475, 14.647) shows significant difference between liver cancer and controls.

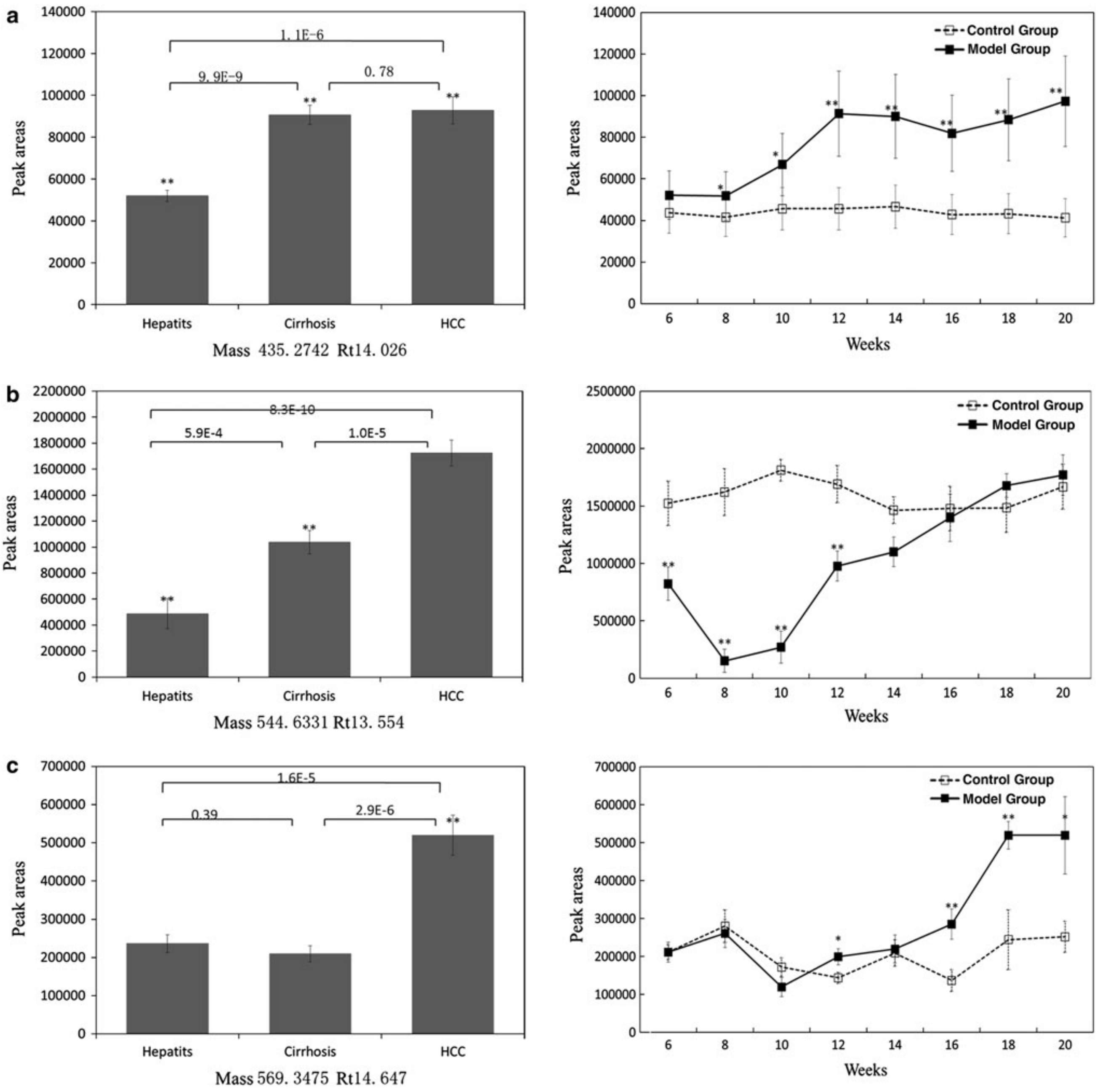

Fig. 6 Three common significantly changed metabolites identified by SVM-RFE, GA and RF. Left: group averages and standard deviations of three potential biomarkers in the model group. Right: temporal changes between the model and the control group of three potential biomarkers **a** feature (435.2742, 14.026); **b** feature (544.6331, 13.554); **c** feature (569.3475, 14.647). "*" and "**" represent $0.01 \leq P < 0.05$ and $P < 0.01$ in the *t*-test between model and control groups, respectively

Besides these shared metabolites which could be filtered by multiple methods, some other metabolites were also selected characteristically by single method.

To GA, (357.2868, 10.493) and (453.2856, 14.026) are meaningful (Fig. 7). According to the concentration of (357.2868, 10.493), cirrhosis is significantly different from hepatitis and HCC. The metabolite (453.2856, 14.026) was identified as lysophosphatidylethanolamine C16:0. Its concentration in the control group is relatively stable and much lower than that in the model group (Fig. 7b), which indicates that it can significantly discriminate between the models and the controls. Besides, it also well distinguishes hepatitis from cirrhosis and HCC. This metabolite could not only present classification for liver diseases, but also demonstrate different metabolism of phosphatidylethanolamine among the patients.

(330.1825, 11.786) (Fig. 8) is informative in random forest. It provides suggestion of the hepatits stage, which

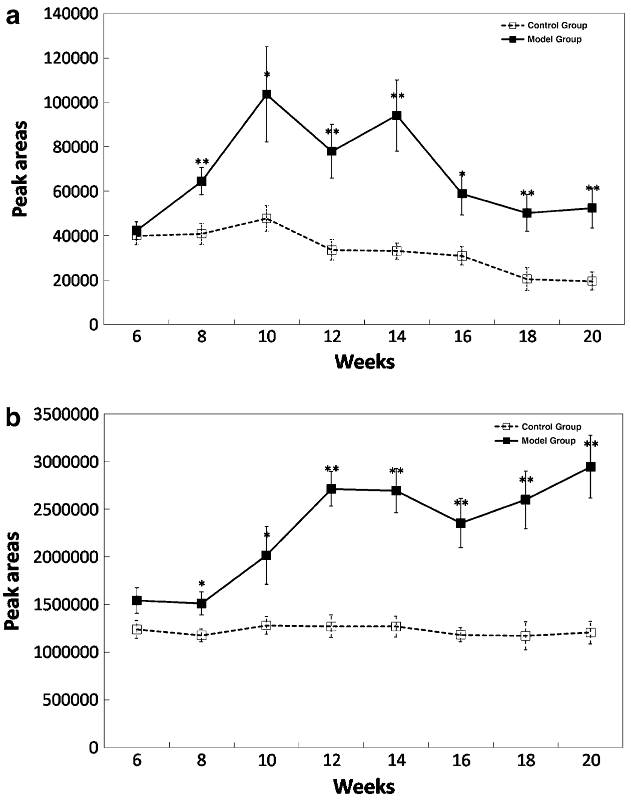

Fig. 7 Temporal changes of two potential biomarkers only found by GA. **a** feature (357.2868, 10.493); **b** feature (453.2856, 14.026). Both **a** and **b** have significant difference between model and control group in cirrhosis and HCC stages. Others are the same as Fig. 6

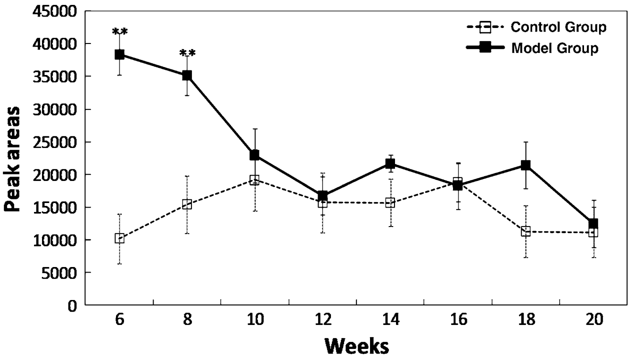

Fig. 8 Temporal changes of potential biomarkers only found by RF. Feature (330.1825, 11.786) has a significant difference between model and control group in hepatitis stage. Others are the same as Fig. 6

also indicates the potential pathological changes in the hepatitis rats.

Based on the combination of the above three feature selection methods, more significant metabolites could be found than any single method. More importantly, the specific metabolites selected by different methods could also represent distinct aspects of the metabolic deregulations although they perhaps have some relevance. And the complementary results may also help the systemic understandings of diseases.

5 Conclusions

In this paper, we propose a new method for feature selection, which could be applied to analyze complex biological data. The core of the method is to combine different algorithms instead of a single chemometrics technique to find more important metabolites. Based on the analysis of a test dataset, the combination of SVM-RFE with GA and RF was demonstrated to be effective for finding more significantly changed metabolites of liver diseases. By taking the advantages of each method, shared and specific differential metabolites could be explored. The method provides more comprehensive information for the systemic understanding of the 'omics' data.

Acknowledgments The study has been supported by the State Key Science & Technology Project for Infectious Diseases (2008ZX10002-019, 2008ZX10002-017), National Basic Research Program of China (2007CB914701) and National Key Project of Scientific and Technical Supporting Programs (2006038079037) from State Ministry of Science & Technology of China, and the foundation (No. 20835006, 90713032) from National Natural Science Foundation of China.

References

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Journal Computational Statistics & Data Analysis, 52*(4), 2249–2260.

Balding, D. J., Bishop, M., & Cannings, C. (2007). *Handbook of statistical genetics*. England: John Wiley & Sons, Ltd.

Bhattacharyyas, S., Epstein, J., & Suval, J. (2006). Biomarkers that discriminate multiple myeloma patients with or without skeletal involvement detected using SELDI-TOF mass spectrometry and statistical and machine learning tools. *Disease Markers, 22*(4), 245–255.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Bryan, K., Brennan, L., & Cunningham, P. (2008). MetaFIND: A feature analysis tool for metabolomics data. *BMC Bioinformatics, 9*, 470.

Chan, E. C. Y., Yap, S., Lau, A., Leow, P., Toh, D., & Koh, H. (2007). Ultra-performance liquid chromatography/time-of-flight mass spectrometry based metabolomics of raw and steamed Panax notoginseng. *Rapid Communications in Mass Spectrometry, 21*, 519–528.

Cho, H., Kim, S. B., Jeong, M. K., Park, Y., Miller, N., Ziegler, T., et al. (2008). Discovery of metabolite features for the modeling and analysis of high-resolution NMR spectra. *International Journal of Data Mining and Bioinformatics, 2*(2), 176–192.

Defernez, M., & Kemsley, E. K. (1997). The use and misuse of chemometrics for treating classification problems. *TrAC Trends in Analytical Chemistry, 16*(4), 216–221.

Díaz-Uriarte, R., & de Andrés, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics, 7*, 3.

Eriksson, L., Johansson, E., Kettaneh-wold, N., Trygg, J., Wikstrom, C., & Wold, S. (2006). *Multi- and megavariate data analysis principles and applications-principles and applications*. Umetrics AB: Umeå.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*, 389–422.

Holland, J. H. (1992). *Adaptation in natural and artificial systems* (2nd ed.). Cambridge, MA: MIT Press.

Jolliffe, I. T. (2002). *Principal component analysis*. New York: Springer.

Kim, Y., Park, I., & Lee, D. (2007). Integrated data mining strategy for effective metabolomic data analysis. In *The First International Symposium on Optimization and Systems Biology (OSB'07)*, Beijing, China.

Kima, S. H., Kima, D. H., Parka, J., Choia, E. J., Parkb, S., Leec, K. Y., et al. (2010). Discrimination of *Scrophularia* spp. according to geographic origin with HPLC-DAD combined with multivariate analysis. *Microchemical Journal, 94*(2), 118–124.

Laxman, Y., Jarkko, T., & Jaakko, H. (2010). Functional prediction of unidentified lipids using supervised classifiers. *Metabolomics, 6*, 18–26.

Lee, S. S. F., Sun, L., Kustra, R., & Bull, S. B. (2008). EM-random forest and new measures of variable importance for multi-locus quantitative trait linkage analysis. *Bioinformatics, 24*(14), 1603–1610.

Li, L., Darden, T. A., Weingberg, C. R., Levine, A. J., & Pedersen, L. G. (2001). Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Combinatorial Chemistry & High Throughput Screening, 4*(8), 727–739.

Mahadevan, S., Shah, S. L., Marrie, T. J., & Slupsky, C. M. (2008). Analysis of metabolomic data using support vector machines. *Analytical Chemistry, 80*(19), 7562–7570.

Maher, A. D., Crockford, D., Toft, H., Malmodin, D., Faber, J. H., Mccarthy, M. I., et al. (2008). Optimization of human plasma ^{1}H NMR spectroscopic data processing for high-throughput metabolic phenotyping studies and detection of insulin resistance related to type 2 diabetes. *Analytical Chemistry, 80*, 7354–7362.

Man, M. Z., Dyson, G., Johnson, K., & Liao, B. (2004). Evaluating methods for classifying expression data. *Journal of Biopharmaceutical Statistics, 14*(4), 1065–1084.

Nicholson, J. K. (2006). Global systems biology, personalized medicine and molecular epidemiology. *Molecular Systems Biology, 2*, 52.

Ooi, C. H., & Tan, A. P. (2003). Genetic algorithms applied to multiclass prediction for the analysis of gene expression data. *Bioinformatics, 19*(1), 37–44.

Pisitkun, T., Johnstone, R., & Knepper, M. A. (2006). Discovery of urinary biomarkers. *Molecular & Cellular Proteomics, 5*, 1760–1771.

Ramadan, Z., Jacobs, D., Grigorov, M., & Kochhar, S. (2006). Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms. *Talanta, 68*(5), 1683–1691.

Righi, V., Durante, C., Cocchi, M., Calabrese, C., Difebo, G., Lecce, F., et al. (2009). Discrimination of healthy and neoplastic human colon tissues by ex vivo HR-MAS NMR spectroscopy and chemometric analyses. *Journal of Proteome Research, 8*(4), 1859–1869.

Saeys, Y., Lnza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507–2517.

Solank, K. S., Bailey, N. J. C., Holmes, E., Lindon, J. C., Davis, A. L., Mulder, T. P. J., et al. (2003). NMR-based metabonomic studies on the biochemical effects of epicatechin in the rat. *Journal of Agricultural and Food Chemistry, 51*, 4139–4145.

Stella, C., Beckwith-hall, B., Cloarec, O., Holmes, E., Lindon, J. C., Powell, J., et al. (2006). Susceptibility of human metabolic phenotypes to dietary modulation. *Journal of Proteome Research, 5*, 2780–2788.

Strobl, C., Boulesteix, A., Zeileis, A., & Hothornt, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*, 25.

Trevino, V., & Falciani, F. (2006). GALGO-an R package for multivariate variable selection using genetic algorithms. *Bioinformatics, 22*(9), 1154–1156.

Vapnik, V. (1998). *Statistical learning theory*. New York: John Wiley and Sons.

Wanga, Y., Taoa, Y., Lina, Y., Liangb, L., Wub, Y., Qua, H., et al. (2009). Integrated analysis of serum and liver metabonome in liver transplanted rats by gas chromatography coupled with mass spectrometry. *Analytica Chimica Acta, 633*(1), 65–70.

Xue, Y., Li, H., Ung, C. Y., Yap, C. W., & Chen, Y. Z. (2006). Classification of a diverse set of tetrahymena pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods. *Chemical Research in Toxicology, 19*, 1030–1039.

Yang, J., Xu, G., Zheng, Y., Kong, H., Pang, T., Lv, S., et al. (2004). Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases. *Journal of Chromatography B, 813*(1–2), 59–65.

Yin, P., Wan, D., Zhao, C., Chen, J., Zhao, X., Wang, W., et al. (2009). A metabonomic study of hepatitis B-induced liver cirrhosis and hepatocellular carcinoma by using RP-LC and HILIC coupled with mass spectrometry. *Molecular Biosystems, 5*(8), 868–876.

Zou, W., & Tolstikov, V. V. (2008). Probing genetic algorithms for feature selection in comprehensive metabolic profiling approach. *Rapid Communications in Mass Spectrometry, 22*(8), 1312–1324.

Zou, W., & Tolstikov, V. V. (2009). Pattern recognition and pathway analysis with genetic algorithms in mass spectrometry based metabolomics. *Algorithms, 2*(2), 638–666.