Metabolomics (2011) 7:614–622
DOI 10.1007/s11306-011-0286-3

ORIGINAL ARTICLE

Application of L-EDA in metabonomics data handling: global metabolite profiling and potential biomarker discovery of epithelial ovarian cancer prognosis

Jing Chen · Yang Zhang · Xiaoyan Zhang · Rui Cao · Shili Chen · Qiang Huang · Xin Lu · Xiaoping Wan · Xiaohua Wu · Congjian Xu · Guowang Xu · Xiaohui Lin

Received: 26 November 2010 / Accepted: 28 January 2011 / Published online: 17 February 2011

Abstract Solution capacity limited estimation of distribution algorithm (L-EDA) is proposed and applied to ovarian cancer prognosis biomarker discovery to expatiate on its potential in metabonomics studies. Sera from healthy women, epithelial ovarian cancer (EOC), recurrent EOC and non-recurrent EOC patients were analyzed by liquid chromatography-mass spectrometry. The metabolite data were processed by L-EDA to discover potential EOC prognosis biomarkers. After L-EDA filtration, 78 out of 714 variables were selected, and the relationships among four groups were visualized by principle component analysis, it was observed that with the L-EDA filtered variables, non-recurrent EOC and recurrent EOC groups could be separated, which was not possible with the initial data. Five metabolites (six variables) with $P < 0.05$ in Wilcoxon test were discovered as potential EOC prognosis biomarkers, and their classification accuracy rates were 86.9% for recurrent EOC and non-recurrent EOC, and 88.7% for healthy + non-recurrent EOC and EOC + recurrent EOC. The results show that L-EDA is a powerful tool for potential biomarker discovery in metabonomics study.

Keywords Metabonomics · Ovarian cancer · Prognosis biomarker · Solution capacity limited EDA · Estimation of distribution algorithms

1 Introduction

Metabonomics is an approach that aims at comprehensively studying the endogenous small molecules (Nicholson et al. 1999). It has been widely applied in disease biomarker discovery (Xu et al. 2009; Wishart 2008),

Jing Chen, Yang Zhang contribute equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s11306-011-0286-3) contains supplementary material, which is available to authorized users.

J. Chen · S. Chen · Q. Huang · X. Lu · 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: xugw@dicp.ac.cn

Y. Zhang · X. Lin (✉)
School of Computer Science & Technology, Dalian University of Technology, Dalian 116024, China
e-mail: datas@dlut.edu.cn

X. Zhang · C. Xu (✉)
Obstetrics & Gynecology Hospital, Shanghai Medical School, Institute of Biomedical Science of Fudan University, Shanghai Key Laboratory of Female Reproductive Endocrine Related Diseases, Shanghai 200011, China
e-mail: xucongjian@gmail.com

R. Cao
Department of the Obstetrics and Gynecology Hospital, Dalian Medical University, Dalian 116033, China

X. Wan
The International Peace Maternity and Child Health Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200030, China

X. Wu
Department of Gynecologic Oncology, Cancer Hospital, Fudan University, Shanghai 200032, China

microorganisms, plant (Eisenreich and Bacher 2007), and nutrition (Kussmann et al. 2006) areas.

Data obtained from metabonomics studies are complex and multidimensional. Therefore, multivariate statistical analysis is often employed for biomarker discovery including principal component analysis (PCA) (Idborg-Bjorkman et al. 2003; Shi et al. 2004; Holmes et al. 2000; Chen and Hofestadt 2006; Choi et al. 2004a; Keun et al. 2002), hierarchical cluster analysis (HCA) (Holmes and Antti 2002; Lindon et al. 2004; Choi et al. 2004b; Shi et al. 2002a, b), nonlinear mapping (NLM) (Holmes and Antti 2002), partial least-squares discriminant analysis (PLS-DA) (Jonsson et al. 2004; Jiye et al. 2005; Azmi et al. 2002; Eriksson et al. 2004), orthogonal PLS (OPLS) (Wiklund et al. 2007), SIMCA (soft independent modeling of class analogy) (Shi et al. 2004; Holmes et al. 2000; Odunsi et al. 2005), and artificial neural network (ANN) (Yang et al. 2002; Beger et al. 2004; Holmes et al. 2001; Sugimoto et al. 2005). As an unsupervised multivariate data analysis method, PCA is useful to compress the multidimensional data into a few principal components and give an overview of the clustering trend in the data. However, PCA may fail to model the data with minor differences. The PLS based methods, OPLS and PLS-DA, calculate principal components in cooperation with the classification information and are more powerful to deal with complex multidimensional data, but overfitting often happens (Defernez and Kemsley 1997). To avoid this problem, in recent years, some new pattern recognition algorithms were employed in metabonomics studies, such as fuzzy c-means clustering (Li et al. 2009) and support vector machine (SVM) (Guan et al. 2009).

Estimation of distribution algorithms (EDAs) are a novel class of evolutionary algorithms and suitable for optimization. A most salient feature of EDAs is that they maintain a probabilistic model during the iterative search process (Larrañaga and Lozano 2002). The probabilistic model is modified in each iteration using a certain strategy and directs the search in next iteration in turn. The iterative search terminates when a termination condition is satisfied. EDAs have been successfully applied in solving bioinformatics problems (Saeys et al. 2003, 2007; Inza et al. 2001; Santana et al. 2007, 2008). Compared with traditional greedy methods, EDAs are more efficient in relevant feature subset selection for splice site prediction. Meanwhile the results got from EDA methods were comparable or even better (Saeys et al. 2003). A new randomized algorithm based on the EDA model, FSS-TREE, obtained the best average accuracy results for various classifiers among several compared feature subset selection techniques (Inza et al. 2001). In other studies, different variants of EDAs have been used for protein structure prediction in simplified models, and their use as a simulation tool for the analysis of the protein folding process was proposed (Santana et al. 2008). Besides, three different EDAs were applied to protein design by minimization of contact potentials (Santana et al. 2007).

Epithelium ovarian cancer (EOC) is the leading cause of death in gynecological cancer (Brown et al. 2002), and serous tumors are the most widespread forms of ovarian cancer (Williams et al. 2007). EOC can usually be diagnosed at the advanced stage, because only few symptoms of EOC can be noticed at the early stage. Moreover, EOC is likely to be recurrent after treatment, which leads to low survival rate (less than 30%) (Brown et al. 2002). Till now, biomarkers for EOC early diagnosis and prognosis have not been well developed. Cancer antigen 125 (CA125) is the commonly used biomarker for EOC diagnosis and prognosis (Jacobs and Menon 2004). However, levels of CA125 are also elevated in other cancers (Bast et al. 1998) and gynecological inflammations (An et al. 2006). Therefore, development of new sensitive and specific biomarkers for EOC early diagnosis or prognosis is quite necessary. Metabonomics (Odunsi et al. 2005; Guan et al. 2009; Denkert et al. 2006) and proteomics (Petricoin et al. 2002) have been used to study EOC to find potential biomarkers. The performance of pattern recognition methods including linear discriminant analysis, quadratic discriminant analysis, *k*-nearest neighbor classifier, bagging and boosting classification trees, support vector machine, and random forest were compared for the classification of ovarian cancer and control serum samples based on MS spectra (Wu et al. 2003).

In this study, an improved EDA method (L-EDA) was proposed and introduced to metabonomics study for mining specific different variables in potential EOC prognosis biomarker discovery. Sera from healthy, EOC, non-recurrent EOC and recurrent EOC were analyzed based on LC–MS platform. Specific variables reflecting the differences of EOC prognosis were selected by the L-EDA method, and visualized on PCA model. Furthermore, the specific metabolites from L-EDA filtered data were selected as potential biomarkers for EOC prognosis.

2 Experimental

2.1 Chemicals

Acetonitrile (HPLC grade) was purchased from Merck (USA). Formic acid (HPLC grade) was purchased from Tedia (USA). Distilled water was filtered through a Milli-Q system (Millipore MA). Cortisol and hypoxanthine were bought from Sigma–Aldrich (USA).

2.2 Serum collection and preparation

Twenty-four healthy, 21 EOC, 36 recurrent EOC and 25 non-recurrent EOC women with pathologic diagnosis were recruited by Obstetrics and Gynecology Hospital of Fudan University, and Obstetrics and Gynecology Hospital of Dalian Medical University. Informed consent was signed by each participant. This study was approved by the Ethics Committee of Obstetrics and Gynecology Hospital of Fudan University.

Sera were collected at 6–8 a.m. from the women without breakfast and immediately stored in −80°C refrigerator. Before analysis, serum was thawed at room temperature, and then 180 μl serum was mixed with 720 μl acetonitrile for protein precipitation. The mixture was vigorously extracted for 30 s, and centrifuged at 15,000×*g* for 10 min at 4°C. The supernatant was lyophilized and reconstituted in 150 μl water/acetonitrile solution (v/v 1/4).

2.3 Metabolic profiling analysis

Chromatographic separation was performed on an Agilent 1200 Rapid Resolution Liquid Chromatography system (Agilent, USA). A 50 mm × 2.1 mm, 1.7 μm BEH C18 column (Waters, USA) was used. The column was maintained at 50°C, and the sample manager was set to 4°C. Mobile phase A was water containing 0.1% formic acid and 2% acetonitrile, mobile phase B was acetonitrile. The flow rate was 0.35 ml/min. Metabolites were eluted with a 30 min gradient. The gradient was started at 5% B, changed to 35% B at 3 min and then to 80% B at 22 min. At 24 min, it was changed to 100% B and kept for 5 min, then quickly changed back to 5% B in 1 min and kept for another 5 min for column equilibrium. A 5 μl aliquot was injected onto the column.

Mass spectrometry was performed on an Agilent 6510 Q-TOF mass spectrometer (Agilent, USA), operating at positive ion mode. Parameters of the mass spectrometry were: capillary voltage 4 kV, fragmentor voltage 230 V and skimmer voltage 65 V, drying gas flow 11 l/min, gas temperature 350°C, nebulizer pressure 45 psig. 10 mM purine (*m*/*z* 121.0508) and 2 mM hexakis phosphazine (*m*/*z* 922.0097) were used as internal standards to ensure mass accuracy throughout the whole analysis. Data were collected in the centroid mode and mass ions at *m*/*z* 80–1000 were recorded. The scan time was set at 500 ms.

2.4 Data pretreatment and data analysis

The raw mass spectral data (retention time, *m*/*z*, abundance) were firstly analyzed by the Molecular Features Extraction software (Agilent), after that chemically qualified molecular features were output, and masses were finally grouped into ''compounds'' by their molecular features. Then, Genespring software (Agilent) was employed for peak alignment. Detected and matched peaks with retention time, *m*/*z* value and their corresponding peak area were listed to an Excel table, and then the peak area data were normalized to total area (=10,000). This final table (peak matrix) was used for the following analysis.

PCA and PLS-DA were performed by SIMCA-P software version 11.0 (Umetrics AB, Umeå, Sweden). Wilcoxon test was processed on SPSS 13.0 (SPSS, USA). Support Vector Machine (SVM) was adopted as the basic classifier of sevenfold cross validation, and it was implemented in C/C++ from LIBSVM (Chang and Lin 2001). L-EDA was written in C++.

2.5 Solution capacity limited EDA (L-EDA)

Estimation of distribution algorithms (EDAs) are evolutionary algorithms that maintain a probabilistic model during the iterative search procedure (Larrañaga and Lozano 2002). In each iteration, EDAs generate multiple candidate solutions (each solution represents a selected variable subset) according to the current probabilistic model, and employ an objective criterion to evaluate each candidate solution. A proportion of best evaluated solutions are derived and used to update the probabilistic model. The learning ratio controls how much the best solutions contribute to the model. Then, the updated probabilistic model directs the search in turn. The iterative search terminates until a termination condition is satisfied.

If a solution contains hundreds of variables, there is a large chance that some non-informative ones are evaluated quite well with informative ones. Besides, a learning method is usually used to evaluate each candidate solution, if the solution capacity is small, less time is consumed on the evaluation. Hence we propose an improved EDA (L-EDA) which limits each candidate solution (variable subset) to a fixed, relatively small capacity to select the most discriminative variables (potential biomarkers) from the complex metabolite data matrix.

Meanwhile the strategy of updating the probabilistic model is very crucial in EDAs. Just as EDAs, L-EDA uses the best evaluated solutions to update the probabilistic model. In order to reward the informative variables, occurrence frequency of each variable and the average occurrence frequency of all variables in the best evaluated solutions are used as a measure of updating the probabilistic model. Since the average occurrence frequency reflects the average performance of all variables, only those competitive variables with discriminative ability should be rewarded. Therefore, if the occurrence frequency of variable m ($f[m]$) is larger than the average frequency (average), the probabilistic model ($p[m]$) is rewarded using

formula (1), otherwise, it is punished according to formula (2).

$$p[m] = (1 - r) * p[m] + r * (1 - p[m]) * (f[m] - \text{average}) \quad (1)$$

$$p[m] = (1 - r) * p[m] + r * p[m] * (f[m] - \text{average}) \quad (2)$$

When the procedure terminates, L-EDA ranks the variables by the probabilistic model in a descending order. The top-ranked variables are selected for the subsequent potential biomarker discovery. See Appendix A, B, C in the supplementary materials for details of the algorithm.

L-EDA applies SVMs to evaluate the candidate solutions. For each candidate solution, an SVM model is constructed. The sevenfold cross validation accuracy rate is calculated. In sevenfold cross validation, all the samples are divided into seven disjoint subsets, each subset has (nearly) the same sample distribution among all the groups as the global sample distribution. Six out of seven subsets are used as the training set, and the remaining subset is used as the test set. All the subsets are used as the test set once and correctly classified sample number divided by total sample number is calculated as the cross validation accuracy.

In the implementation of L-EDA, the learning ratio, best solution ratio and maximum iteration number were set to 0.3, 0.2 and 100, respectively.

3 Results and discussion

The study was initiated because there is imperative need for diagnosis of ovarian cancer recurrence, until now the 5-years postoperative survival rate is not satisfactory. Hence, the detection and identification of metabolite biomarkers in serum may greatly improve the diagnostic options of ovarian cancer.

High resolution separation of serum metabolites was acquired based on LC–MS system. After peak alignment by using the Genespring software, a peak matrix containing 714 variables were generated. Nine representative peaks were chosen to evaluate the deviation of the LC–MS method. It was found that relative standard deviations (RSD) of peak area were 1.5–5.9% in 14 QC samples which show acceptable reproducibility according to FDA guidance (Guidance for industry, bioanalytical methods validation 2001).

3.1 PLS-DA modeling

The data consisting of 714 variables were firstly analyzed by PCA and PLS-DA. Figure 1 shows the score plots of PCA and PLS-DA models. PCA can be used to describe the relationship among groups without artificial intervention, but four groups in our study can not be well separated.

In PLS-DA, eight principal components were calculated via sevenfold cross validation, the cumulative R^2Y was 0.79 while Q^2 was 0.462. It is observed that healthy women were well separated from the cancer groups (EOC and recurrent EOC), but the PLS-DA model failed to distinguish the non-recurrent and recurrent EOC groups.

PLS-DA is a supervised method, and the possibility of overfitting should not be neglected. Two hundred times permutations on the class labels were conducted, R^2Y-intercept was 0.419 and Q^2-intercept was −0.678. According to the previous research (Eriksson et al. 2001), the R^2Y-intercept should not exceed 0.4 and Q^2-intercept should not exceed 0.05 for a valid model. Hence the PLS-DA model overfitted the data, and the results from the model might be unreliable.

3.2 L-EDA filtration

Because all EOC women took or are taking the different therapies, the metabolic differences between the non-recurrent and the recurrent EOC groups can be easily concealed by the factors caused by chemotherapy and radiotherapy etc., while the raw data contained influence of all factors. In fact, many variables in the data matrix are not positively correlated with classification, they should be removed before multi-variable analysis. Variable selection methods are helpful for reducing the possibility of overfitting for supervised methods (Li et al. 2010; Ma and Huang 2008).

Here, L-EDA was used to select the important variables that could reflect the differences between the non-recurrent EOC and the recurrent EOC groups. Due to the metabolite data from LC–MS were normalized to total peak area (=10,000), a cutoff peak area was set to 1. If the normalized peak area of a variable in over 80% samples were higher than 1 in any group, the corresponding variable was kept, otherwise excluded (Bijlsma et al. 2006). Finally, 394 variables were kept for the sub-sequential analysis.

Solution number and solution capacity are two main parameters in L-EDA. A larger solution number may increase the diversity and raise the possibility that better solutions are generated to some extent. Meanwhile, a larger solution capacity may increase the number of shared features among the top-ranked solutions. However, if the two parameters become too large, the chance that irrelevant or noisy variables sampled into the best solutions are mistaken as relevant ones could also increase, the probabilistic model becomes not accurate. In the meantime, the evaluation of each solution will be more time-consuming as the solution capacity becomes larger.

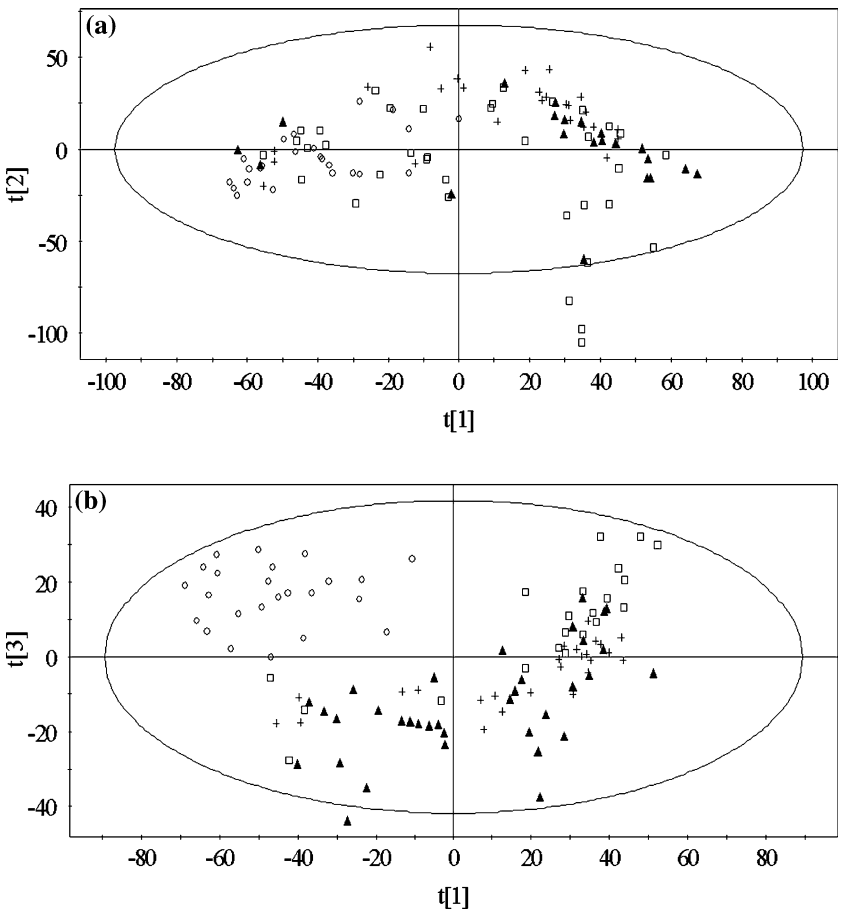

Fig. 1 PCA (**a**) and PLS-DA (**b**) score plots based on original data. Healthy women (*open circle*), non-recurrent EOC (*plus*), recurrent EOC (*closed triangle*) and EOC (*open square*) groups are displayed. In PCA three principle components were calculated, R^2X was 0.433. In PLS-DA eight principal components were calculated automatically by cross validation

To evaluate the effects of the above two parameters of L-EDA, sevenfold cross validation accuracy of SVM model and percentage of overlapping genes-related (POGR) values (Zhang et al. 2009) were measured based on the 20% top-ranked variables selected by L-EDA. In the following studies, the solution number was set to {400, 700, 1000}, and the solution capacity was limited to {40, 70, 100}. It was worth mentioning that a too small solution capacity might hurt the reliability of the probabilistic model, so the minimum value was set to 40, about 10% of the total variable number. Table 1 shows the sevenfold cross validation accuracy rates under different settings of solution number and solution capacity. The high accuracy rates in Table 1 indicate the selected variable subsets under different parameter settings are discriminative. POGR measures the similarity from a variable subset to another and ranges from 0 to 1. The POGR values between the variable subsets selected by L-EDA under different parameter settings are shown in Table 2. It is observed that even at significantly different parameter settings, the POGR values between two selected variable subsets are still acceptable (around 0.75), which indicates that the selected variable subsets are similar to each other and further demonstrate stability of L-EDA in variable selection.

Table 1 Cross validation* accuracy rates of the L-EDA under different parameter settings

	Solution capacity		
	40 (%)	70 (%)	100 (%)
Solution number			
400	95.3	99.1	99.1
700	97.2	99.1	98.1
1000	98.1	98.1	99.1

* 7-cross-validation accuracy rate was calculated

When the solution number and solution capacity were set to 700 and 70, respectively, 78 variables (20% top-ranked variables) were selected by L-EDA for further analysis. Figure 2 displays the accuracy rate variation with top-ranked variable numbers selected by traditional EDA and L-EDA. It is observed that L-EDA clearly outperformed traditional EDA by approximate 10% increase in cross validation accuracy using different number of selected variables. For L-EDA, the accuracy rate curve rises dramatically when the variable number was very small (less than 10), which indicates the very top-ranked features are of prominent discriminative capacity. Then, the small fluctuation of the curve indicates that different amount of

Table 2 POGR* of top 20% features selected by L-EDA under different parameter settings

	B: Solution number–solution capacity								
	400–40	700–40	1000–40	400–70	700–70	1000–70	400–100	700–100	1000–100
A: Solution number–solution capacity									
400–40	1.000	0.936	0.936	0.897	0.936	0.910	0.859	0.859	0.846
700–40	0.910	1.000	0.936	0.923	0.923	0.936	0.897	0.910	0.872
1000–40	0.923	0.974	1.000	0.923	0.962	0.962	0.910	0.936	0.910
400–70	0.846	0.897	0.872	1.000	0.923	0.910	0.910	0.974	0.897
700–70	0.859	0.885	0.885	0.897	1.000	0.910	0.885	0.910	0.872
1000–70	0.846	0.897	0.885	0.897	0.910	1.000	0.923	0.910	0.885
400–100	0.808	0.859	0.872	0.897	0.897	0.923	1.000	0.949	0.910
700–100	0.731	0.808	0.808	0.872	0.833	0.846	0.885	1.000	0.859
1000–100	0.744	0.795	0.795	0.846	0.846	0.846	0.859	0.897	1.000

* POGR (A, B) reflects the similarity from variable set A to variable set B (Zhang et al. 2009)

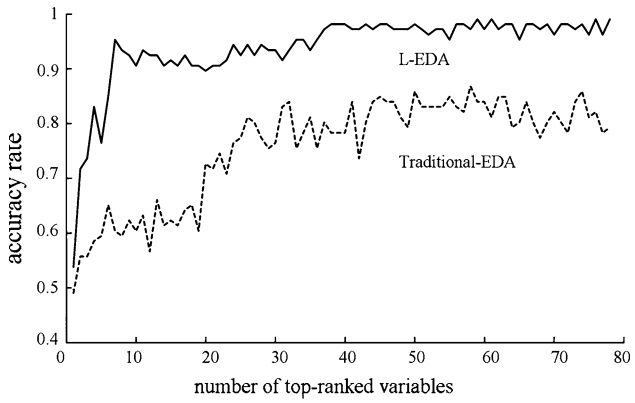

Fig. 2 Accuracy rate of cross validation with top-ranked variables

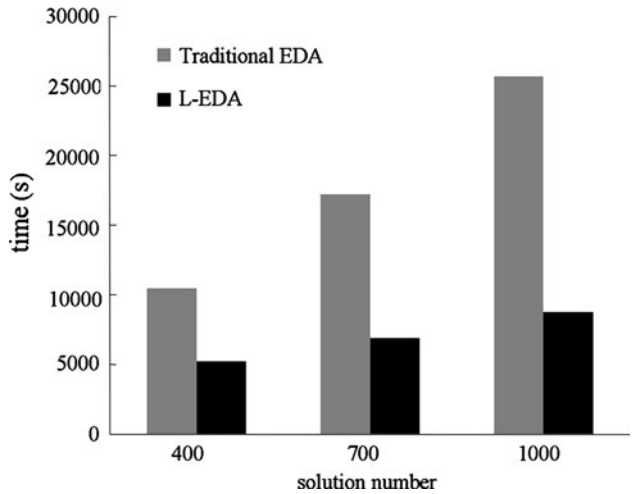

Fig. 3 Times consumed by traditional EDA and L-EDA

top-ranked variable can achieve satisfactory and stable classification results. The times consumed by traditional EDA and L-EDA are shown in Fig. 3. Under various settings of the solution number, L-EDA saved about 50–65% time compared to traditional EDA. Note that, the solution capacity for L-EDA was set to 70, if the solution capacity was set to a smaller number it was expected that more time could be saved.

The discrimination ability of 78 variables was visualized by PCA model (Fig. 4). Recurrent EOC and EOC women were clustered in the same dimensional space, the non-recurrent EOC group was separated from them, in the middle of the healthy women group and EOC groups. This is very different from Fig. 1a. With the original data, PCA cannot distinguish four groups. The data processed by L-EDA reasonably describe the metabolite profiling of recurrent and non recurrent EOC. Without L-EDA process, the metabolic differences of EOC prognosis are easily covered by the factors induced by chemotherapy and radiotherapy etc., non-recurrent EOC and recurrent EOC are not possible to be resolved. L-EDA is proved to be a powerful tool in metabonomics data mining. After L-EDA filtration, the data matrix has been largely simplified, and the selected variables can represent the metabolic differences of recurrent EOC and non-recurrent EOC.

In order to further validate the performance of L-EDA, 1/3 hold-out validations were carried out 10 times using L-EDA. In each hold-out cross validation, 2/3 samples of each group were used for variable selection and the left-out 1/3 samples were used as new samples for test. With top 78 variables in the validations, the accuracy rate, sensitivity, and specificity were 92.9% ± 3.0%, 93.2% ± 5.6% and 92.7% ± 6.2% (mean ± standard deviation), respectively. The results indicated that L-EDA selected variables exhibited satisfactory prediction performance.

To develop novel biomarkers for monitoring ovarian cancer prognosis, specific metabolites ought to be identified. According to the clinical needs, the level of an EOC prognosis biomarker should be significantly different

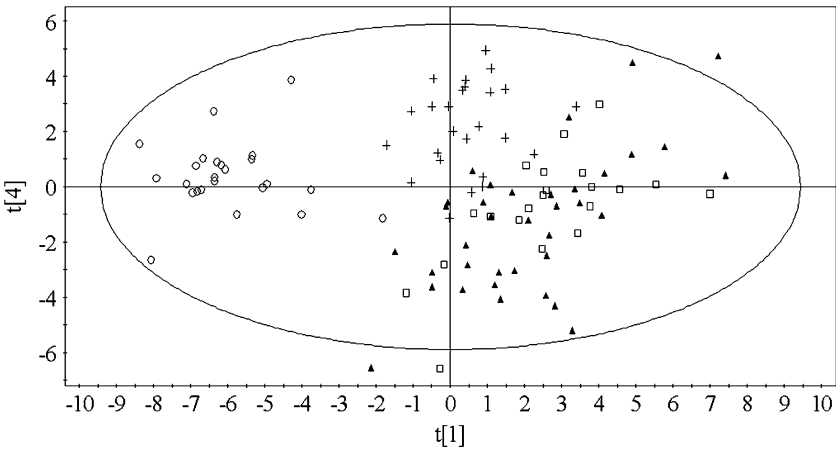

Fig. 4 PCA score plot based on 78 features selected by L-EDA, healthy women (*open circle*), non-recurrent EOC (*plus*), recurrent EOC (*closed triangle*) and EOC (*open square*) groups are displayed. Six principal components were calculated automatically, and R^2X was 0.585

Table 3 Potential biomarkers of EOC prognosis

t_R	*m/z*	Compound	Normalized peak area				Wilcoxon *P*			
			Healthy women (1)	Non-recurrent EOC (2)	EOC (3)	Recurrent EOC(4)	(1) vs. (3)	(1) vs. (2)	(2) vs. (3)	(2) vs. (4)
0.55	137.05	Hypoxanthine[a]	3.09 ± 1.23	3.71 ± 1.83	6.08 ± 4.64	7.58 ± 5.07	0.0032	4.65E-10	0.0342	3.99E-06
0.56	176.07	Guanidinosuccinic acid[b]	10.17 ± 3.70	9.81 ± 5.05	12.67 ± 4.54	16.39 ± 11.76	0.0095	0.0067	0.0157	0.0031
6.26	363.22	Cortisol[a]	6.33 ± 2.15	6.23 ± 2.75	10.40 ± 4.53	9.35 ± 4.30	0.0009	0.0009	0.0009	0.0015
13.51	526.29	Lyso PE(22:6)[c]	34.78 ± 12.45	45.57 ± 12.32	61.48 ± 20.17	60.00 ± 20.12	6.22E-08	1.24E-07	0.0028	0.0056
13.52	385.27	Lyso PE(22:6) fragment[c]	0.85 ± 0.41	1.21 ± 0.30	1.64 ± 0.51	1.57 ± 0.68	3.68E-07	2.99E-06	0.0003	0.0016
13.58	521.35	Lyso PC(18:2)[c]	490.20 ± 56.86	517.98 ± 129.35	304.68 ± 192.73	323.48 ± 209.79	9.45E-05	0.0008	2.36E-05	2.12E-05

Compound(s) labeled with "a" were validated by commercial standard samples; with "b" was identified based on the accuracy mass and the hmdb (www.hmdb.ca); with "c" were identified based on the retention, fragmentation mechanism and our previous work

(Wilcoxon $P < 0.05$) between each group with EOC (EOC or recurrent EOC group) and the group without EOC (healthy or non-recurrent EOC group). Besides, the change tendency in EOC and recurrent EOC should be the same (both up-regulated or both down-regulated) compared with the healthy women and non-recurrent EOC patients. As a result, six variables from L-EDA filtered data were screened out and given in Table 3. The structure identification followed our previous work (Chen et al. 2008) in which the main steps included detecting the quasi-molecular ion, acquiring accuracy mass, studying LC retention behavior and mass fragmentation mechanism, matching database, and final confirmation with standard samples. The molecular structures of the six variables were elucidated and five metabolites were identified. Based on the five metabolites, the classification accuracy using SVM was calculated as 86.9% for non-recurrent EOC and recurrent EOC groups, and 88.7% for healthy + non-recurrent EOC and EOC + recurrent EOC groups. Two hundred times permutation tests were executed and the R^2-intercept and Q^2-intercept (Mahadevans et al. 2008) were −0.601 and −1.079 for non-recurrent EOC and recurrent EOC groups, and −0.729 and −1.172 for healthy + non-recurrent EOC and EOC + recurrent EOC groups, respectively. These parameters indicate that the SVM models avoided overfitting and yielded reliable results.

4 Conclusions

In this study, L-EDA method was proposed and proved to be a powerful tool in discovering differences covered by other factors, such as chemotherapy and radiotherapy. Serum metabolome data from healthy women, EOC, recurrent EOC, and non-recurrent EOC patients obtained on LC–MS platform were used to demonstrate the applicability of L-EDA in metabonomic studies. L-EDA outperformed traditional EDA method in discriminative variables selection and efficiency, and its stability was validated by cross validation accuracy rate and POGR values under different parameter settings. By L-EDA filtration, 78 variables were

selected in which the differences between non-recurrent EOC and recurrent EOC groups could be easily visualized in PCA model. Five metabolites with a significant difference according to Wilcoxon test were finally identified, and their classification capacity was 86.9% for non-recurrent EOC vs. recurrent EOC, and 88.7% for healthy + non-recurrent EOC vs. EOC + recurrent EOC. The application of L-EDA in metabonimics has great prospects in efficient potential biomarker discovery, improved visualization, and accurate classification.

Acknowledgments This study was supported by grants from the National High-tech R&D Program (863 Program) (Project Number: 2006AA02Z342), the National Basic Research Program of China (No. 2007CB914701) from the State Ministry of Science & Technology of China, National Natural Science Foundation of China (No. 20835006), the State Key Science & Technology Project for Infectious Diseases (2008ZX10002-019).

References

An, H. J., Miyamoto, S., Lancaster, K. S., Kirmiz, C., Li, B., Lam, K. S., et al. (2006). Profiling of glycans in serum for the discovery of potential biomarkers for ovarian cancer. *Journal of Proteome Research, 5*(7), 1626–1635.

Azmi, J., Griffin, J. L., Antti, H., Shore, R. F., Johansson, E., Nicholson, J. K., et al. (2002). Metabolic trajectory characterisation of xenobiotic-induced hepatotoxic lesions using statistical batch processing of NMR data. *Analyst, 127*(2), 271–276.

Bast, R. C., Xu, F. J., Yu, Y. H., Barnhill, S., Zhang, Z., & Mills, G. B. (1998). CA 125: The past and the future. *International Journal of Biological Markers, 13*(4), 179–187.

Beger, R. D., Harris, S., & Xie, Q. (2004). Models of steroid binding based on the minimum deviation of structurally assigned C-13 NMR spectra analysis (MiDSASA). *Journal of Chemical Information and Computer Sciences, 44*(4), 1489–1496.

Bijlsma, S., Bobeldijk, I., Verheij, E. R., Ramaker, R., Kochhar, S., Macdonald, I. A., et al. (2006). Large-Scale human metabonomics studies: A strategy for data (pre-) processing and validation. *Analytical Chemistry, 78*(2), 567–574.

Brown, M. L., Riley, G. F., Schussler, N., & Etzioni, R. (2002). Estimating health care costs related to cancer treatment from SEER-Medicare data. *Medical Care, 40*(8), 104–117.

Chang, C.C., & Lin, C. J. (2001). LIBSVM: A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, M., & Hofestadt, R. (2006). A medical bioinformatics approach for metabolic disorders: Biomedical data prediction, modeling, and systematic analysis. *Journal of Biomedical Informatics, 39*(2), 147–159.

Chen, J., Zhao, X. J., Fritsche, J., Yin, P. Y., Schmitt-Kopplin, P., Wang, W. Z., et al. (2008). Practical approach for the identification and isomer elucidation of biomarkers detected in a metabonomic study for the discovery of individuals at risk for diabetes by integrating the chromatographic and mass spectrometric information. *Analytical Chemistry, 80*(4), 1280–1289.

Choi, H. K., Choi, Y. H., Verberne, M., Lefeber, A. W., Erkelens, C., & Verpoorte, R. (2004a). Metabolic fingerprinting of wild type and transgenic tobacco plants by H-1 NMR and multivariate analysis technique. *Phytochemistry, 65*(7), 857–864.

Choi, Y. H., Kim, H. K., Hazekamp, A., Erkelens, C., Lefeber, A. W. M., & Verpoorte, R. (2004b). Metabolomic differentiation of *Cannabis sativa* cultivars using 1H NMR spectroscopy and principal component analysis. *Journal of Natural Products, 67*(6), 953–957.

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.

Denkert, C., Budczies, J., Kind, T., Weichert, W., Tablack, P., Sehouli, J., et al. (2006). Mass spectrometry-based metabolic profiling reveals different metabolite patterns in invasive ovarian carcinomas and ovarian borderline tumors. *Cancer Research, 66*(22), 10795–10804.

Eisenreich, W., & Bacher, A. (2007). Advances of high-resolution NMR techniques in the structural and metabolic analysis of plant biochemistry. *Phytochemistry, 68*(22–24), 2799–2815.

Eriksson, L., Antti, H., Gottfries, J., Holmes, E., Johansson, E., Lindgren, F., et al. (2004). Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Analytical and Bioanalytical Chemistry, 380*(3), 419–429.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., & Wold, S. (2001). *Multi-and megavariate data analysis*. Umetrics.

FDA. (2001). *Guidance for industry, bioanalytical method validation, Food and Drug Administration: A guidance*. Rockville, MD: Centre for Drug Valuation and Research (CDER).

Guan, W., Zhou, M. S., Hampton, C. Y., Benigno, B. B., Walker, L. D., Gray, A., et al. (2009). Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics, 10*, 259.

Holmes, E., & Antti, H. (2002). Chemometric contributions to the evolution of metabonomics: Mathematical solutions to characterising and interpreting complex biological NMR spectra. *Analyst, 127*(12), 1549–1557.

Holmes, E., Nicholls, A. W., Lindon, J. C., Connor, S. C., Connelly, J. C., Haselden, J. N., et al. (2000). Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chemical Research in Toxicology, 13*(6), 471–478.

Holmes, E., Nicholson, J. K., & Tranter, G. (2001). Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks. *Chemical Research in Toxicology, 14*(2), 182–191.

Idborg-Bjorkman, H., Edlund, P. O., Kvalheim, O. M., Schuppe-Koistinen, I., & Jacobsson, S. P. (2003). Screening of biomarkers in rat urine using LC/electrospray ionization-MS and two-way data analysis. *Analytical Chemistry, 75*(18), 4784–4792.

Inza, I., Merino, M., Larranaga, P., Quiroga, J., Sierra, B., & Girala, M. (2001). Feature subset selection by genetic algorithms and estimation of distribution algorithms—A case study in the survival of cirrhotic patients treated with TIPS. *Artificial Intelligence in Medicine, 23*(2), 187–205.

Jacobs, I. J., & Menon, U. (2004). Progress and challenges in screening for early detection of ovarian cancer. *Molecular & Cellular Proteomics, 3*(4), 355–366.

Jiye, A., Trygg, J., Gullberg, J., Johansson, A. I., Jonsson, P., Antti, H., et al. (2005). Extraction and GC/MS analysis of the human blood plasma metabolome. *Analytical Chemistry, 77*(24), 8086–8094.

Jonsson, P., Gullberg, J., Nordstrom, A., Kusano, M., Kowalczyk, M., Sjostrom, M., et al. (2004). A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. *Analytical Chemistry, 76*(6), 1738–1745.

Keun, H. C., Ebbels, T. M., Antti, H., Bollard, M. E., Beckonert, O., Schlotterbeck, G., et al. (2002). Analytical reproducibility in H-1 NMR-based metabonomic urinalysis. *Chemical Research in Toxicology, 15*(11), 1380–1386.

Kussmann, M., Raymond, F., & Affolter, M. (2006). OMICS-driven biomarker discovery in nutrition and health. *Journal of Biotechnology, 124*(4), 758–787.

Larrañaga, P., & Lozano, J. A. (2002). *Estimation of distribution algorithms: A new tool for evolutionary computation*. Dordrecht: Kluwer Academic Publishers.

Li, X., Lu, X., Tian, J., Gao, P., Kong, H. W., & Xu, G. W. (2009). Application of fuzzy c-means clustering in data analysis of metabonomics. *Analytical Chemistry, 81*(11), 4468–4475.

Li, Z., Zhou, X., Dai, Z., & Zou, X. (2010). Classification of G-protein coupled receptors based on support vector machine with maximum relevance minimum redundancy and genetic algorithm. *BMC Bioinformatics, 11*, 325.

Lindon, J. C., Holmes, E., & Nicholson, J. K. (2004). Metabonomics: Systems biology in pharmaceutical research and development. *Current Opinion in Molecular Therapeutics, 6*(3), 265–272.

Ma, S., & Huang, J. (2008). Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics, 9*(5), 392–403.

Mahadevans, 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.

Nicholson, J. K., Lindon, J. C., & Holmes, E. (1999). 'Metabonomics': Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica, 29*(11), 1181–1189.

Odunsi, K. R., Wollman, M., Ambrosone, C. B., Hutson, A., McCann, S. E., Tammela, J., et al. (2005). Detection of epithelial ovarian cancer using H-1-NMR-based metabonomics. *International Journal of Cancer, 113*(5), 782–788.

Petricoin, E. F., I. I. I., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., et al. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet, 359*(9306), 572–577.

Saeys, Y., Degroeve, S., Aeyels, D., Van de Peer, Y., & Rouzé, P. (2003). Fast feature selection using a simple estimation of distribution algorithm: A case study on splice site prediction. *Bioinformatics, 19*(2), 179–188.

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

Santana, R., Larranaga, P., & Lozano, J. (2007). The role of a priori information in the minimization of contact potentials by means of estimation of distribution algorithms. In *Proceedings of the fifth european conference on evolutionary computation, machine learning and data mining in bioinformatics, Lecture Notes in Computer Science,* Vol. 4447, pp. 247–257.

Santana, R., Larranaga, P., & Lozano, J. (2008). Protein folding in simplified models with estimation of distribution algorithms. *IEEE Transactions on Evolutionary Computation, 12*(4), 418–438.

Shi, H. L., Paolucci, U., Vigneau-Callahan, K. E., Milbury, P. E., Matson, W. R., & Kristal, B. S. (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*(3), 197–208.

Shi, H. L., Vigneau-Callahan, K. E., Shestopalov, A. I., Milbury, P. E., Matson, W. R., & Kristal, B. S. (2002a). Characterization of diet-dependent metabolic serotypes: Primary validation of male and female serotypes in independent cohorts of rats. *Journal of Nutrition, 132*(5), 1031–1038.

Shi, H. L., Vigneau-Callahan, K. E., Shestopalov, A. I., Milbury, P. E., Matson, W. R., & Kristal, B. S. (2002b). Characterization of diet-dependent metabolic serotypes: Primary validation of male and female serotypes in independent cohorts of rats. *Journal of Nutrition, 132*(5), 1039–1046.

Sugimoto, M., Kikuchi, S., Arita, M., Soga, T., Nishioka, T., & Tomita, M. (2005). Large-scale prediction of cationic metabolite identity and migration time in capillary electrophoresis mass spectrometry using artificial neural networks. *Analytical Chemistry, 77*(1), 78–84.

Wiklund, S., Johansson, E., Sjoestroem, L., Mellerowicz, E. J., Edlund, U., Shockcor, J. P., et al. (2007). Visualization of GC/TOF-MS-based metabonomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry, 80*(1), 115.

Williams, T. I., Toups, K. L., Saggese, D. A., Kalli, K. R., Cliby, W. A., & Muddiman, D. C. (2007). Epithelial ovarian cancer: Disease etiology, treatment, detection, and investigational gene, metabolite, and protein biomarkers. *Journal of Proteome Research, 6*(8), 2936–2962.

Wishart, D. S. (2008). Applications of metabonomics in drug discovery and development. *Drugs in R&D, 9*(5), 307–322.

Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., et al. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics, 19*(13), 1636–1643.

Xu, E. Y., Schaefer, W. H., & Xu, Q. (2009). Metabonomics in pharmaceutical research and development: Metabolites, mechanisms and pathways. *Current Opinion in Drug Discovery & Development, 12*(1), 40–52.

Yang, J., Xu, G. W., Kong, H. W., Zheng, Y. F., Pang, T., & Yang, Q. (2002). Artificial neural network classification based on high-performance liquid chromatography of urinary and serum nucleosides for the clinical diagnosis of cancer. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences, 780*(1), 27–33.

Zhang, M., Zhang, I., Zou, J., Tao, C., Xiao, H., Liu, Q., et al. (2009). Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics, 25*(13), 1662–1668.