Talanta 84 (2011) 298–304

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

A metabonomic approach applied to predict patients with cerebral infarction

Zhiting Jiang[a,b,1], Jingbo Sun[c,1], Qionglin Liang[b], Yefeng Cai[c,**], Shasha Li[c], Yan Huang[c], Yiming Wang[b], Guoan Luo[a,b,*]

[a] School of Pharmacy, East-China University of Science & Technology, Shanghai, PR China
[b] Key Laboratory of Bioorganic Phosphorus Chemistry & Chemical Biology, Department of Chemistry, Tsinghua University, Beijing, PR China
[c] The Center of Encephalopathy, Guangdong Provincial Hospital of Traditional Chinese Medicine, Guangzhou 510120, PR China

ARTICLE INFO

Article history:
Received 20 October 2010
Received in revised form
21 December 2010
Accepted 6 January 2011
Available online 15 January 2011

Keywords:
Metabonomics
Cerebral infarction
UPLC–TOF MS
One-carbon cycle
Prediction model

ABSTRACT

Cerebral infarction is always of sudden onset, and usually leading to serious consequence. It is of therapeutic significance to develop fast and accurate diagnosis methods for cerebral infarction so that patients can be treated timely and properly. A metabonomic approach was then proposed to investigate the potential biomarkers and metabolic pathways associated with cerebral infarction and also establish a prediction model of cerebral infarction for the fast diagnosis. Serum metabolic profiling of sixty-seven cerebral infarction patients and sixty-two controls was obtained using UPLC–TOF MS. The resulting data were then processed by multivariate statistical analysis to graphically demonstrate metabolic variations. The PLS-DA model was validated with cross validation and permutation tests to assure the model's reliability, and significant difference was obtained between the original and hypothetical models ($p < 0.0001$). A series of endogenous metabolites in the one-carbon cycle, such as folic acid, cysteine, S-adenosyl homocysteine and oxidized glutathione, were determined as potential biomarkers of cerebral infarction. A prediction model developed using PLS–KNN algorithm was established to differentiate cerebral infarction patients from controls, and an average accuracy of 100% was obtained. In conclusion, metabonomic approach is a powerful tool to investigate the pathogenesis of stroke and is expected to be developed as a useful method for the fast diagnosis.

1. Introduction

Despite the past decade has witnessed intense medical advancement and tremendous achievement in the ability to diagnose stroke, the acute cerebrovascular problem still represents a leading cause of long-term disability and death. Cerebral infarction is a kind of acute ischemic stroke due to a disturbance in the blood vessels supplying blood to the brain [1,2]. As the only pharmacological intervention approved by the Food and Drug Administration for the treatment of acute ischemic stroke, tissue plasminogen activator must be administered within 3–6 h from symptom occurrence. However, only approximately 2–6% of the patients with stroke could be administered intravenous tissue plasminogen activator, due to the early diagnostic uncertainty of stroke [3]. Besides tissue plasminogen activator, other early management decisions regarding glycemic, blood pressure, and temperature control for fibrinolysis may also be effective when presenting in the 3-h window. Therefore, it is critical for stroke patients to be rapidly evaluated, diagnosed and considered for acute therapy.

There are quite a few ways to diagnose patients with stroke, such as CT angiography, CT perfusion techniques and MRI-based techniques [4]. Multimodal imaging techniques such as CT angiography and CT perfusion techniques may add diagnostic information regarding vascular occlusion, but they are insensitive to ischemic stroke and are not widely available on a timely basis at many institutions. MRI-based techniques have shown greatly enhanced sensitivity in the early diagnosis of stroke as well as the identification of intracerebral hemorrhage. However, as a practical issue, most hospitals do not have these specialized MRI services available in the acute setting, and these studies require a more advanced interpretation. Therefore, more and more scientists all over the world pay more attention on finding out pragmatic methods for diagnosis of ischemic stroke. Recently, a few of specific genes or proteins have been reported as useful biomarkers, such as matrix metalloproteinase 9, D-dimer, S100β and B-type natriuretic peptide [5–8]. However, only by these limited biomolecules, the availability and accuracy of diagnosis is hard to be determined. With the development of high throughput "omic" methods,

* Corresponding author. Department of Chemistry, Tsinghua University, Beijing, 100084, PR China. Tel.: +86 10 62781688.
** Corresponding author at: Guangdong Provincial Hospital of Traditional Chinese Medicine, Guangzhou, 510120, PR China. Tel.: +86 20 81887233.
E-mail addresses: caiyefeng@126.com (Y. Cai), luoga@mail.tsinghua.edu.cn (G. Luo).
[1] Both authors contributed equally to this work.

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.01.015

genomics and proteomics have been proved as powerful tools for better discrimination of stroke patients and aided diagnose of the disease [9,10].

Metabonomics, as another high throughput "omic" approach, can provide information distinctive from genomics and proteomics. As an integral component of systems biology investigations, metabonomics strategy makes use of multivariate statistical analysis, such as high performance liquid chromatography in tandem with mass spectrometry [11,12], gas chromatography in tandem with mass spectrometry [13,14] and nuclear magnetic resonance [15,16], to search for disease-related potential biomarkers and metabolic pathways. This approach has been successfully applied in revealing the pathophysiological perturbations of the diseases such as cardiovascular disease [17], type 2 diabetes [18] and renal failure [19].

In this study, an ultra high performance liquid chromatography in tandem with time of flight mass spectrometer (UPLC–TOF MS)-based metabonomic approach was used to acquire initial metabolic profiling of serum. Multivariate statistical analysis was then employed to investigate pathological variations of patients with cerebral infarction, and a prediction model was developed by PLS–KNN algorithm to determine the potential value of metabonomics in the prediction of cerebral infarction occurrence. The proposed metabonomic approach might give a new sight for early diagnosis and therapy of patients with cerebral infarction.

2. Materials and methods

2.1. Chemicals

Formic acid (HPLC grade) and leucine enkephalin (HPLC grade) were purchased from Sigma–Aldrich (St. Louis, MO, USA). HPLC grade acetonitrile and methanol used for MS analysis were purchased from J.T. Baker (Phillipsburg, NJ, USA). Ultrapure water (18.2 MΩ) was prepared using a Milli-Q water purification system (Millipore, France). Folic acid, cysteine, S-adenosyl-homocysteine, oxidized glutathione, hydroxyeicosatetraenoic acid, adenosine, aldosterone, hydroxyoctadecadienoic acid and deoxocathasterone standards were purchased from Sigma–Aldrich (St. Louis, MO, USA).

2.2. Subjects

Sixty-seven patients were collected from Guangdong Province Hospital of Traditional Chinese Medicine, China for the period of June 1, 2008 to December 30, 2009. Inclusion criteria are set as: (1) patients' age between 40 and 75 years; (2) the acute onset of neurological deficit lasting less than 6 h; (3) being the first time of stroke attack; (4) patients were conscious so that clinical data could be collected. Exclusion criteria are that patients had cancer, cardiac insufficiency, hepatosis, renal inadequacy, respiratory failure, alimentary tract hemorrhage, or other diseases that will affect the clinical observations and biological indicators. The subjects were confirmed according to the part of "acute ischemic stroke" in the International Classification of Diseases (ICD). Source documentation used for disease validation included emergency room records, admission history, physical examinations, neurology consultations, and CT or MRI reports. CT or MRI reports were available for all patients to differentiate hemorrhagic stroke from ischemic stroke, and only cases of cerebral infarction were included. The control group consisted of blood samples from 62 individuals who came to the hospital for routine physical check-up, and their age, gender matched with that of the stroke patients. The experimental protocol was reviewed and approved by the Local Committee of Medical Ethics. Written informed consents were obtained from all subjects.

2.3. Sample preparation

Thawed serum samples (100 μL) were mixed with methanol (400 μL) by vortexing for 2 min, and then centrifuged at 4 °C for 15 min at 13,000 rpm. The supernatant was transferred to a 1.5 mL polypropylene tube, diluted with 500 μL of ultrapure water, and then filtered through a syringe filter (0.22 μm) for UPLC–TOF MS analysis.

2.4. Chromatographic condition

Serum metabolite profiling was performed with UPLC–MS. Chromatography was carried out with an ACQUITY BEH C_{18} chromatography column (2.1 mm × 100 mm, 1.7 μm). The column temperature was maintained at 50 °C, and then binary mobile phase was composed of phase A (water with 0.1% formic acid) and phase B (acetonitrile).

The gradient for the serum sample was: 0–3 min, 5–50% B; 3–7 min, 50–70% B; 7–12 min, 70–95% B; 12–14 min, followed by washing with 95% B. The proportion of phase B returned to 5% in 1 min, and the column was allowed to re-equilibrate for 5 min before the next injection. The flow rate was 0.4 mL/min, and 4 μL was injected into the column.

2.5. TOF MS condition

For MS we used the W mode of operation and negative ion electrospray mode. The capillary voltage was set at 2300 V, and cone voltage at 35 V. Nitrogen was used as the dry gas, the desolvation gas flow rate was set at 700 L/h, and cone gas flow was maintained at 35 L/h. The desolvation temperature was set at 350 °C, and source temperature at 120 °C. The scan time and inter-scan delay were set to 0.1 s and 0.02 s, respectively. The TOF data were collected from m/z 50–1000. All the data were acquired using an independent reference lock mass via the LockSpray™ interface to ensure accuracy and reproducibility. Leucine enkaphalin was used as the reference compound ($[M-H]^-$ = 554.2615) at a concentration of 50 pg/μL under a flow rate of 10 μL/min. The data were collected in the centroid mode, and the LockSpray frequency set at 10 s and averaged over 10 scans for correction.

2.6. Data processing

Data were collapsed into a single matrix by aligning peaks with the retention time-exact mass pair together from each data file along with their associated intensities using the MakerLynx Applications Manager v4.1 (Waters, MS Technologies). The parameters included retention time (Rt) range from 0 to 20 min, mass range from 50 to 1000 Da, and mass tolerance was 0.02 Da. The minimum intensity was set at 15% of base peak intensity, maximum mass per Rt was set at 6, and Rt tolerance was 0.02 min. Prior to multivariate statistical analysis, data of each chromatogram were normalized to a constant integrated intensity of the number of peaks to partially compensate for the concentration bias of each sample. The between-subject data X is then Pareto-scaled to facilitate analysis of the major effect in the data. The Pareto-scaling procedure is given:

$$\tilde{X}_{ij} = \frac{X_{ij} - \bar{X}_j}{\sqrt{S_j}}$$

where $\tilde{X}_{ij}$ is the intensity of the Pareto-scaled UPLC–MS signal j; $\bar{X}_j$ is the mean signal j of X, and S_j is the standard deviation of signal j.

Table 1
The basic characteristics of subjects.

	Case	Control	*p* value [a]
No. of subjects	67	62	
Age, mean ± SD, years	62.85 ± 8.45	60.65 ± 7.31	0.393
Gender, male/female, no.	34/33	32/30	0.921
Body mass index	25.72 ± 2.85	25.83 ± 2.63	0.524
Medical history (%)			
Diabetes mellitus	11(16.4)	10(16.1)	0.964
Hypertension	34(50.7)	30(48.3)	0.789
Hyperlipidemia	23(34.3)	21(33.9)	0.956
Smoking	25(37.3)	25(40.3)	0.725
Glycosylated hemoglobin, %	6.49 ± 1.89	6.18 ± 1.77	0.741
Serum glucose, mmoL/L	7.25 ± 3.94	6.95 ± 2.21	0.493
Creatinine, μmoL/L	96.33 ± 48.74	80.15 ± 20.34	0.018
Cholesterol, mmoL/L	5.11 ± 2.25	4.08 ± 0.68	0.001
Triglyceride, mmoL/L	1.75 ± 1.09	1.32 ± 0.55	0.006
HDL, μmmoL/L	1.24 ± 1.30	1.23 ± 0.20	0.721
LDL, μmmoL/L	3.15 ± 0.94	2.53 ± 0.73	0.005

[a] Unpaired *t*-tests and χ^2 tests were used to assess the difference between cases and controls.

2.7. Prediction

K-Nearest Neighbor (KNN) has long been used in pattern recognition, exploratory data analysis, and data mining problems, which is a similarity learning approach that can be easily understood and programmed [20]. It can be simply described that an unknown sample in the test set is classified into the class according to the majority belongs of *k* neighbors in the training set which are closest to this test sample. Parameter *k* in the algorithm is usually optimized via leave-one-out (LOO) cross validation. Routine of PLS–KNN for prediction was coded using Matlab (version 2008a, MathWorks).

2.8. Statistical analysis

Statistical analysis was performed using the SPSS software (Version 16.0, USA) and statistical significance was set at $p < 0.05$. Data are presented as means ± standard deviation for continuous measures, while counts and percentages for categorical variables. Differences in study variables were compared using unpaired *t*-test for continuous measures and χ^2 test for categorical variables.

Multivariate statistical analysis was performed by the software package SIMCA-P 11.5 (Umetrics AB, Umeå, Sweden). The PLS-DA model was then validated using cross model validation and permutations. Routines of cross model validation and permutation tests were in-house written and performed using Matlab 2008a.

3. Results and discussion

3.1. Clinical characteristics of study population

Sixty-seven patients and sixty-two controls were included in this study. The characteristics of cerebral infarction patients and controls were listed in Table 1. Comparison of the demographic characteristics (such as age, gender, and body mass index) and medical histories of the study population did not reveal significant differences between patients and controls. It is indicated that the intra-group differences were mainly due to cerebral infarction-related pathological variations. Many biochemical detections found that the levels of creatinine, cholesterol, triglyceride and low-density lipoprotein in cerebral infarction patients were higher than that in controls, whereas there was no significant difference in glycosylated hemoglobin, serum glucose and high-density lipoprotein levels. Lipid metabolism dysfunction is known as a primary cause of ischemic stroke, and it is lipid peroxidation that increased free radical generation, which may lead to oxidative stress in acute ischemic stroke [21].

3.2. Method validation of UPLC/MS

Precision of injection was carried out by the continuous analysis of six injections of the same samples. Stable retention time and intensity of the peaks were observed. And the relative standard deviations (RSD) of the peak area value were between 2.3 and 4.2%. The precision of injection reflected the stability of the analysis, which was of great importance to guarantee the reliability of the acquired metabonomic data.

To evaluate the influence of sample preparation on the stability of the data, six parallel samples were prepared using the same preparation protocol, and the resulting data showed that the repeatability of sample preparation met the requirements for metabonomic analysis: the retention time of the peaks kept almost unchanged, and the RSD values of the peak intensity for the main peaks were less than 4.8%.

3.3. Serum metabonomic study of patients and controls

Serum metabolic profiling was established to explore other important compounds and metabolic pathways related to cerebral infarction and developed a prediction model for aided diagnosis of the disease. The typical UPLC–MS base peak ion current (BPI) chromatograms of patients with cerebral infarction and controls were compared visually (Fig. 1).

To acquire as many features as possible in a single injection, UPLC–TOF MS operating condition was optimized in advance. Profiles obtained from positive and negative ion mode were carefully compared and the negative mode was finally adopted for it could provide much more metabolic information with less background intensities. Using the optimized UPLC–MS analysis protocol and subsequent processes, such as baseline correction, peak deconvolution, alignment and normalization, we obtained a three-dimensional matrix, including data file name, retention time-exact mass pair and normalized peak areas. Overall 7926 retention time-exact mass pairs were determined in each sample profile. Since PLS-DA is a scale-dependent method, the use of an appropriate scaling technique to pre-treat the chromatographic data sets is essential. The Pareto-scaling technique is known as a method emphasizes the contribution of lower concentration metabolites but not to such an extent where noise produces a large contribution. This process facilitates the detection of metabolites consistently present in the biological samples. Hence, the data were standardized using Pareto-scaling technique which has been described in Section 2.6.

To evaluate the capability of the UPLC–MS based metabonomic approach is useful to differentiate cerebral infarction patients from controls, principal component analysis (PCA) and partial least squares-discriminant analysis (PLS-DA) were carried out in the study. PCA was used firstly to investigate general interrelation between groups, including clustering and outliers among the samples. And then, PLS-DA was used to maximize the difference of metabolic profiles between control and cerebral infarction groups. Fig. 2 illustrates the PLS-DA score plot ($R2X = 0.0991$, $R2Y = 0.998$, $Q2 = 0.947$), which showed evident clustering of controls and cerebral infarction patients.

3.4. Model validation

Typical applications of two class classifications using PLS-DA show a score plot of the classified samples with a $Q2$ value indicating the validity of the discrimination. However, a score plot illustrating separation between two groups does not have reliable meaning as similar plots can be obtained when random data are classified [22]. Furthermore, due to the low number of individuals compared to the large number of variables, it can easily lead to

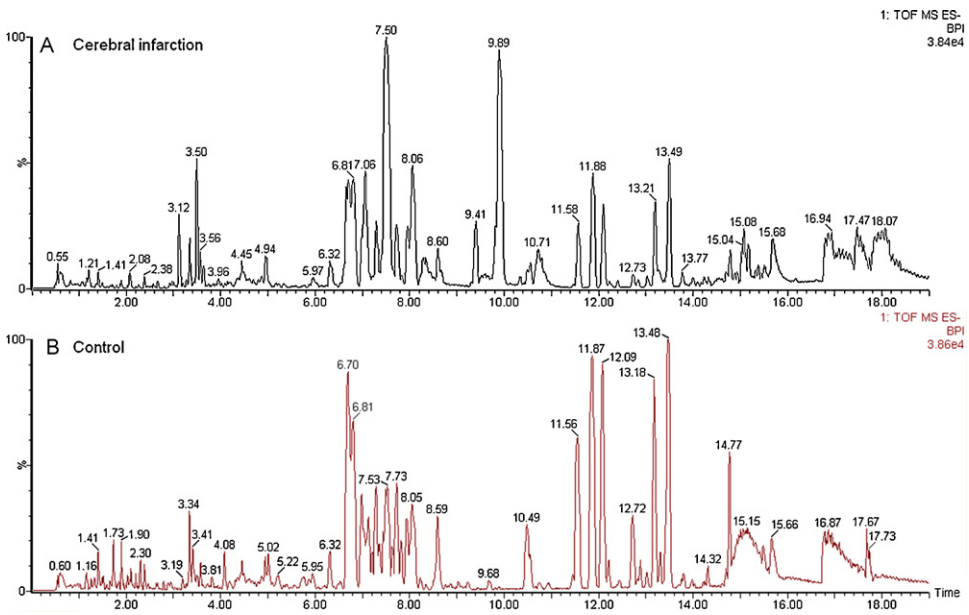

Fig. 1. Representative UPLC–MS base peak ion current (BPI) chromatograms of patients with cerebral infarction and controls.

chance classifications. It is also not clear how to evaluate the separation with *Q*2, and there is no readily criteria to compare the *Q*2 value against for evaluation purpose, so the value of this number is meaningless. Numerous studies were dedicated to proper validation method, because more and more researchers are aware of the importance of PLS-DA model validation [23].

There are several parameters reported to quantify a certain classification, including *Q*2 value, number of misclassifications, true/false positive/negative and the area under curve of a receive operating characteristic curve (AUROC). However, these parameters have not given a measure for statistical significance (such as *p*-value) between the original model and hypothetical model (after changing class labels of the samples) [24]. Here, a validation method with cross model validation and permutations were used in this study. Cross validation is to validate a classification model for the low number of samples obtained. For a cross validation, the total data were divided into a training set, a validation set and a test set. A model was established and optimized by the training and validation set, while the test set is used to test the model performance. The permutation test is carried out to evaluate if the specific classification of the samples in the two designed groups is significantly better than any other random classification in two arbitrary groups [25]. The class labels of case and control are permuted randomly and the classification models were calculated. If there is no difference between the two classes by repeating the permutation process many times, the H0 distribution is expected not to be significant.

In this study, permutation of the class labels should in theory lead to an average number of 65 misclassifications, which is approximately by 50% of data set. To verify this hypothetical prediction error as well as the distribution width, the cross validation prediction errors from 1000 different permutations were collected. The method to determine the cross validation prediction error in each

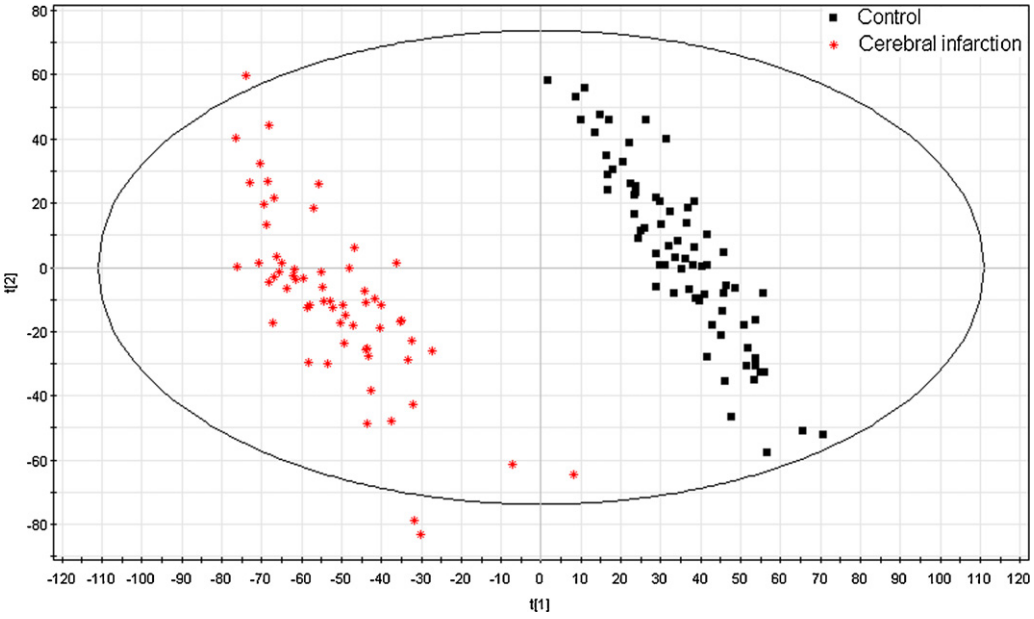

Fig. 2. Score plot of partial least square-discriminant analysis (PLS-DA) performed on the UPLC–TOF MS profiles of serum from patients with cerebral infarction and controls.

permutation is the same as the one for the original classification model. In each permutated data set, 10 cross validations were subjected and a mean cross validation prediction error was estimated. The distribution of prediction errors were obtained from the 1000 permutation considered as the H0 distribution of no effect. Prediction error of the original model was 0, and a p-value was much smaller than 0.0001 when comparing the cross validation prediction error of the original model with the permutations under the H0 distribution (Fig. 3). Finally, the experimentally obtained H0 distribution exactly met the requirement, so the PLS-DA model is tough enough for use.

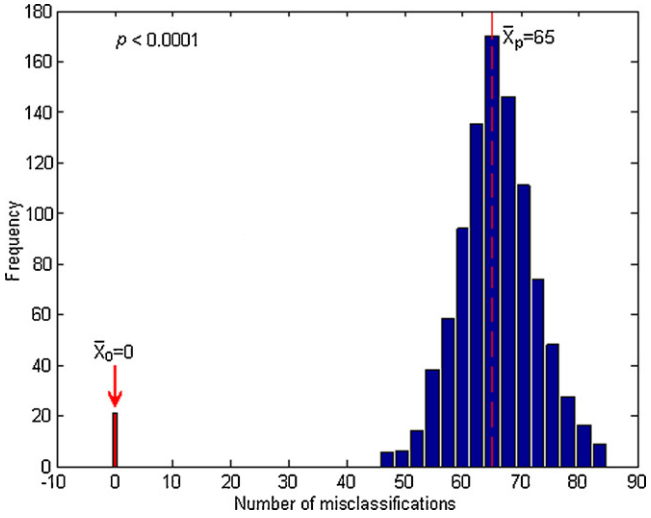

Fig. 3. The mean cross validation error of original PLS-DA model ($\bar{X}_o$) is compared with the H0 distribution of no effect ($\bar{X}_p$). The H0 distribution is based on the cross validation error of 1000 permutations.

3.5. Cerebral infarction-related metabolites

The VIP (variable importance in the projection) value of each variable in the model was ranked according to its contribution to the classification. The VIP list of retention time-exact mass pairs was obtained from PLS-DA by the software MassLynx v4.1 (Waters, England). Some parameters, such as deviation from calculated mass (mDa or ppm), double bond equivalent (DBE), and i-fit value (the isotopic pattern of the selected ion) were used to evaluate the accuracy of possible formulas. Here, the biomarker with Rt-m/z of 0.62–383.1145 in negative ion mode was detailed as an example to illustrate the identification process. Using a mass tolerance of 5 mDa, $C_{14}H_{19}N_6O_5S$ was located as the candidate because of its high mass accuracy (−0.2mDa or −0.5 ppm) and low i-fit value (=4.4) among the possible chemical formulas. And then, $C_{14}H_{19}N_6O_5S$ was input in the KEGG ligand (http://www.genome.jp/kegg/ligand.html) for possible compound, and S-adenosyl-homocysteine (SAH) was finally emerged, which was further confirmed by comparing it to its authentic standards. Typical extracted ion chromatograms of each compound were listed in Fig. 4. Finally, twelve compounds were identified as potential biomarkers of cerebral infarction (shown in Table 2).

For these potential biomarkers, folic acid (FA), tetrahydrofolic acid (THF), cysteine (Cys), S-adenosylhomocysteine (SAH) and oxidized glutathione (GSSG) are involved in the “folate cycle” and in the conjoined “activated one-carbon cycle”. They are important in proteins and DNA stabilization, synthesizing other molecules, and counteracting the toxicity of reactive oxygen metabolites [26]. One of the proposed risk factors for stroke is serum total homo-

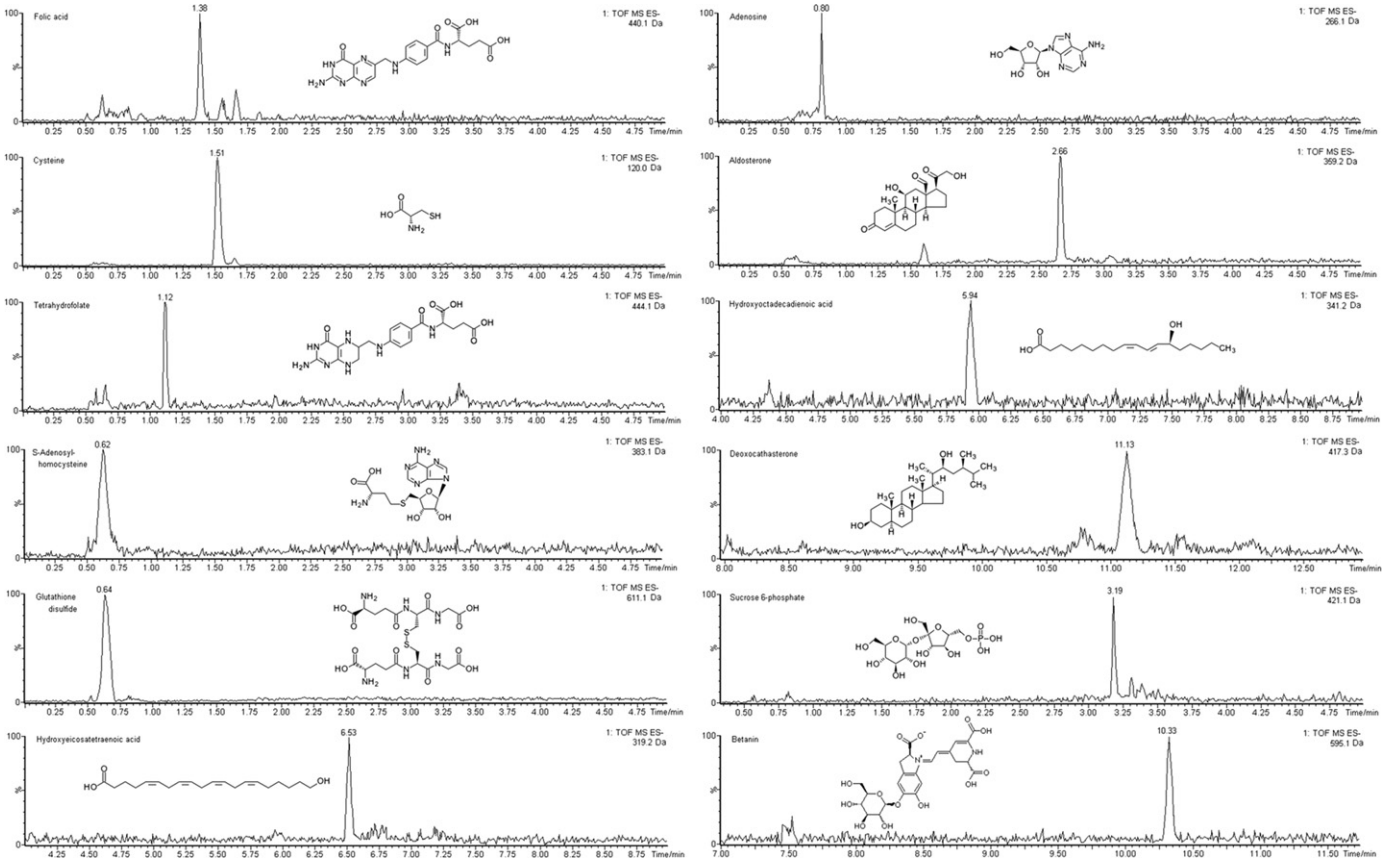

Fig. 4. Typical extracted ion chromatograms of 12 potential biomarkers identified from the negative ion mode in serum.

Table 2
Potential biomarkers identified in serum according to the results of PLS-DA comparing the cerebral infarction group and control group.

No.	Rt	Ion	m/z	Formula	Identified potential biomarker	Related metabolism	Concentration for patients and controls [a,b]	
							Controls	Patients
1	1.38	$[M-H]^-$	440.1329	$C_{19}H_{19}N_7O_6$	Folic acid[c]	Folic acid metabolism	5.86 ± 3.33	1.42 ± 0.31*
2	1.51	$[M-H]^-$	120.0130	$C_3H_7NO_2S$	Cysteine[c]	Cysteine metabolism	2.96 ± 1.40	31.32 ± 13.06**
3	1.12	$[M-H]^-$	444.1639	$C_{19}H_{23}N_7O_6$	Tetrahydrofolate	Folic acid metabolism	15.52 ± 7.90	2.77 ± 1.18**
4	0.62	$[M-H]^-$	383.1145	$C_{14}H_{20}N_6O_5S$	S-Adenosyl-homocysteine[c]	Cysteine metabolism	3.62 ± 2.27	5.35 ± 2.43*
5	0.64	$[M-H]^-$	611.1450	$C_{20}H_{32}N_6O_{12}S_2$	Oxidized glutathione[c]	Glutathione metabolism	2.78 ± 0.61	9.23 ± 1.82**
6	6.53	$[M-H]^-$	319.2276	$C_{20}H_{32}O_3$	Hydroxyeicosatetraenoic acid[c]	Arachidonic acid metabolism	10.48 ± 2.68	15.60 ± 8.86*
7	0.80	$[M-H]^-$	266.0899	$C_{10}H_{13}N_5O_4$	Adenosine[c]	Purine metabolism	23.10 ± 12.03	2.56 ± 2.09**
8	2.66	$[M-H]^-$	359.1865	$C_{21}H_{28}O_5$	Aldosterone[c]	Steroid hormone biosynthesis	191.78 ± 40.03	1.33 ± 0.29**
9	5.94	$[M+COOH]^-$	341.2332	$C_{18}H_{32}O_3$	Hydroxyoctadecadienoic acid[c]	Linoleic acid metabolism	1.49 ± 0.57	96.14 ± 44.69**
10	11.13	$[M-H]^-$	417.3743	$C_{28}H_{50}O_2$	Deoxocathasterone[c]	Steroid hormone biosynthesis	5.71 ± 3.54	1.65 ± 0.52*
11	3.19	$[M-H]^-$	421.0755	$C_{12}H_{23}O_{14}P$	Sucrose 6-phosphate	Glycometabolism	3.24 ± 1.26	1.09 ± 0.22*
12	10.33	$[M+COOH]^-$	595.1536	$C_{24}H_{26}N_2O_{13}$	Betanin	Betaine metabolism	3.18 ± 0.97	1.02 ± 1.06*

Abbreviations: Rt, retention time; *m/z*, mass to charge ratio.
[a] Relative peak area of the potential biomarkers for different groups, which was presented as mean ± SD.
[b] Variation trends of the potential biomarkers for different groups were statistically tested by *t*-test.
[c] Metabolites were confirmed by Rt and *m/z* with authentic chemicals.
* $p < 0.05$.
** $p < 0.01$.

cysteine (tHcy), an intermediate product in the metabolism of one-carbon cycle. However, trials aimed at lowering tHcy levels failed to reduce the risk of stroke suggesting that tHcy increase should be an acute phase-reactant of brain ischemia [27]. To testify the tHcy increase as merely the result rather than the cause of stroke onset, the relationship of stroke and other intermediates in the one-carbon related cycle was explored. Folic acid plays an important role as a methyl donor in the remethylation of homocysteine back to methionine. Indeed, low serum or dietary folic acid is associated with serum tHcy concentration increase. Furthermore, there are several findings of an inverse relationship between dietary or serum folic acid and risk of stroke [28]. Cys was known as a neurotoxic amino acid, and its level was found significantly increased in the patients with stroke. It is reported that high Cys may be translated into increased production of H_2S, which mediates tissue injuries through the NMDA receptors [29]. It is entirely possible that the Cys levels rises because of other comorbidities that may be present even before the stroke. SAH is a strong inhibitor to the S-adenosylmethionine dependent methylation reactions, which is found as a more important compound compared to tHcy in our precious study on diabetic nephropathy [30]. GSSG, a kind of oxidized GSH, is used as a biomarker of oxidative stress. GSSG level is raised in the cerebral infarction group showed that unbalanced production of reactive oxygen species (ROS), which is related to the pathogenesis of ischemic stroke.

Besides, the levels of compounds in the fatty acid metabolism (hydroxyoctadecadienoic acid and hydroxyeicosatetraenoic acid) and steroid hormone metabolism (aldosterone and desoxycortone) were found significantly different between patients and controls. As a biomarker of the oxidative stress status, hydroxyoctadecadienoic acid was reported associated with the pathogenesis of atherosclerosis [31]. Hydroxyeicosatetraenoic acid, formed via ω-hydroxylation of arachidonic acid, has been a focus of recent investigations regarding its function as a vasoconstrictor in cerebral and mesenteric arteries [32].

Our present study regarding the relationship between these potential biomarkers and cerebral infarction are still at the initial stage, and their specificity and sensitivity in indicating the disease are to be further validated by clinical trials. However, the metabonomic result itself is in well agreement with the related biochemical study and thus provides credible basis for further extensive investigation.

3.6. Predict model based on PLS–KNN

KNN can work well in many situations however it does have some requirement on data for accuracy concern. First, the numbers in each class of the training set should be approximately equal, or otherwise the judgment will be biased towards the class with most representatives. Second, ambiguous and outlying samples should be avoided in the training set, which is crucial to the performance of classification. Third, the feature variables used in the KNN algorithm should be much less than the number of samples [33]. Therefore, appropriate feature extraction techniques are usually employed to reduce the dimensions of data for improving the classification accuracy of KNN. PLS is a widely used supervised linear subspace method of reducing the dimensionality of the data, which extracts independent features by iteratively maximizing the covariance of the deflated observation vectors and their labels. Moreover, compared with PCA, another frequently used linear subspace method, PLS would have components that would maximize the discriminative power of the classification in KNN [34]. PLS decreases a large degree of redundant information to get useful feature variables. Hence, an approach combined of PLS and KNN was adopted to explore the cerebral infarction-related prediction model.

All the samples in the data set are randomly divided into the training set and test set. The training set consisted of 103 samples, and the test set 26 samples. The training set was used to establish the prediction model while the test set was responsible for evaluating the model performance. Since the first principle component (PC1) denoted more than 90% of the all variables after PLS, it was input to the KNN algorithm. In KNN, the 'Euclidean' distance of an unknown sample to all members of the training set was calculated and these were ranked in order. Pick the 3 smallest distances ($k = 3$) and see what classes the unknown in closest to, and take the 'majority judgment' and use it for classification. To reduce the possible bias and variability, the random division of training set and test set was carried out in twenty trials, and means and standard deviations of prediction accuracy were calculated. As a result, PLS–KNN method obtained an average prediction accuracy of 100% in discriminating cerebral infarction patients from controls. Therefore, metabonomic approach combined with multivariate data analysis is able to accurately predict the cerebral infarction at the early stage of onset.

4. Conclusion

Taking advantage of metabonomic study and multivariate statistical analysis has shown great promise in simultaneously monitoring multi-metabolic pathways related to cerebral infarction. The PLS-DA model of case and control established in the study was validated by cross model validation and permutation tests, and the model is significantly different from the random model, which showed the model is reliable enough for screening biomarkers. Some metabolites in the one-carbon cycle, such as folic acid, cysteine, S-adenosyl homocysteine and oxidized glutathione, were found out as potential biomarkers. The result is in agreement with the current understanding that folic acid is effective in preventing stroke, and provides the basis for the relationship between one-carbon metabolism and stroke. A prediction model was developed by PLS–KNN algorithm to indicate cerebral infarction, and an accuracy of 100% was obtained. It is suggested that metabonomic approach is highly effective in aiding cerebral infarction identification and thus implies a new strategy in early diagnosis of stroke.

Acknowledgements

The investigation received financial assistance from the National Basic Research Development Program of China (973 Program, Nos. 2005CB523503 and 2007CB714505) and National Natural Science Foundation of China (No. 20805026).

References

[1] C. Sarti, D. Rastenyte, Z. Cepaitis, M.J. Tuomilehto, Stroke 31 (2000) 1588–1601.
[2] W. Rosamond, K. Flegal, G. Friday, K. Furie, A. Go, K. Greenlund, N. Haase, M. Ho, V. Howard, B. Kissela, S. Kittner, D.J. Jones, M. Mcdermott, J. Meigs, C. Moy, G. Nichol, Circulation 115 (2007) E69–E171.
[3] D.Z. Wang, J.A. Rose, D.S. Honings, D.J. Garwacki, J.C. Milbrandt, Stroke 31 (2000) 77–81.
[4] J.B. Fiebach, P.D. Schellinger, O. Jansen, M. Meyer, P. Wilde, J. Bender, P. Schramm, E. Juttler, J. Oehler, M. Hartmann, S. Hahnel, M. Knauth, W. Hacke, K. Sartor, Stroke 33 (2002) 2206–2210.
[5] N. Eldrup, M.L.M. Gronholdt, H. Sillesen, B.G. Nordestgaard, Circulation 114 (2006) 1847–1854.
[6] M. Barber, P. Langhorne, A. Rumley, G.D.O. Lowe, D.J. Stott, Stroke 37 (2006) 1113–1115.
[7] D. Brea, T. Sobrino, M. Blanco, I. Cristobo, R.G. Raquel, R.Y. Yanez, O. Moldes, J. Agulla, R. Leira, J. Castillo, Clin. Chem. Lab. Med. 47 (2009) 1513–1518.
[8] T. Takahashi, M. Nakamura, T. Onoda, M. Ohsawa, K. Tanno, K. Itai, K. Sakata, M. Sakuma, F. Tanaka, S. Makita, Y. Yoshida, A. Ogawa, K. Kawamura, A. Okayama, Atherosclerosis 207 (2009) 298–303.
[9] T.L. Barr, Y. Conley, J. Ding, A. Dillman, S. Warach, A. Singleton, M. Matarin, Neurology 75 (2010) 1009–1014.
[10] A. Minagar, J.S. Alexander, R.E. Kelley, M. Harper, M.H. Jennings, J. Mol. Neurosci. 38 (2009) 182–192.
[11] Y. Wang, J.S. Wang, M. Yao, X.J. Zhao, J. Fritsche, S.K. Philippe, Z.W. Cai, D.F. Wan, X. Lu, S.L. Yang, J.R. Gu, H.U. Haring, E.D. Schleicher, R. Lehmann, Anal. Chem. 80 (2008) 4680–4688.
[12] J.J. Pesek, M.T. Matyska, S.M. Fischer, T.R. Sana, J. Chromatogr. A 1204 (2008) 48–55.
[13] M.Z. Ding, J.S. Cheng, W.H. Xiao, B. Qiao, Y.J. Yuan, Metabolomics 5 (2009) 229–238.
[14] S. Trenkamp, P. Eckes, M. Busch, A.R. Fernie, Metabolomics 5 (2009) 277–291.
[15] Z. Ramadan, D. Jacobs, M. Grigorov, S. Kochhar, Talanta 68 (2006) 1683–1691.
[16] M. Coen, E. Holmes, J.C. Lindon, J.K. Nicholson, Chem. Res. Toxicol. 21 (2008) 9–27.
[17] F.X. Zhang, Z.H. Jia, P. Gao, H.W. Kong, X. Li, J. Chen, Q. Yang, P.Y. Yin, J.S. Wang, X. Lu, F.M. Li, Y.L. Wu, G.W. Xu, Talanta 79 (2009) 836–844.
[18] Y. Gu, Y.F. Zhang, X.Z. Shi, X.Y. Li, J. Hong, J. Chen, W.Q. Gu, X. Lu, G.W. Xu, G. Ning, Talanta 81 (2010) 766–772.
[19] X.M. Lu, Z.L. Xiong, J.J. Li, S.N. Zheng, T.G. Huo, F.M. Li, Talanta 83 (2011) 700–708.
[20] T.M. Cover, P.E. Hart, IEEE Trans. Inform. Theory 13 (1967) 21–27.
[21] F. van Kooten, G. Ciabattoni, C. Patrono, D.W.J. Dippel, P.J. Koudstaal, Stroke 28 (1997) 1557–1563.
[22] O. Cloarec, M.E. Dumas, A. Craig, R.H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J.C. Lindon, E. Holmes, J.K. Nicholson, Anal. Chem. 77 (2005) 1282–1289.
[23] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duynhoven, Metabolomics 4 (2008) 81–89.
[24] E.J.J. van Velzen, J.A. Westerhuis, J.P.M. van Duynhoven, F.A. van Dorsten, H.C.J. Hoefsloot, D.M. Jacobs, S. Smit, R. Draijer, C.I. Kroner, A.K. Smilde, J. Proteom. Res. 7 (2008) 4483–4491.
[25] P. Golland, F. Liang, S. Mukherjee, D. Panchenko, Lect. Notes Comput. Sci. 3559 (2005) 501–515.
[26] M. Rafii, R. Elango, J.D. House, C.M. Glenda, P. Darling, L. Fisher, P.B. Pencharz, J. Chromatogr. B 877 (2009) 3282–3291.
[27] E. Lonn, S. Yusuf, M.J. Arnold, P. Sheridan, J. Pogue, M. Micks, M.J. McQueen, J. Probstfield, G. Fodor, C. Held, J. Genest, N. Engl. J. Med. 354 (2006) 1567–1577.
[28] X.B. Wang, X.H. Qin, H. Demirtas, J.P. Li, G.Y. Mao, Y. Huo, N.L. Sun, L.H. Liu, X.P. Xu, Lancet 369 (2007) 1876–1882.
[29] P.T.H. Wong, K. Qu, G.N. Chimon, A.B.H. Seah, H.M. Chang, M.C. Wong, Y.K. Ng, H. Rumpel, B. Halliwell, J. Neuropathol. Exp. Neurol. 65 (2006) 109–115.
[30] Z.T. Jiang, Q.L. Liang, G.A. Luo, P. Hu, P. Li, Y.M. Wang, Talanta 77 (2009) 1279–1284.
[31] Y. Yoshida, E. Niki, Biofactors 27 (2006) 195–202.
[32] J.H. Capdevila, J.R. Falck, J.D. Imig, Kidney Int. 72 (2007) 683–689.
[33] R.G. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, John Wiley, Chichester, 2003.
[34] G.Z. Li, H.L. Bu, M.Q. Yang, X.Q. Zeng, J.Y. Yang, BMC Genom. 9 (2008) S24–S39.