Metabonomics based NMR in Crohn’s disease applying PLS-DA

Fariba Fathi1, Afsaneh Arefi Oskouie2, Mohsen Tafazzoli1, Nosratollah Naderi3, Kaveh Sohrabzedeh4, Soraya Fathi5, Mohsen Norouzinia5, Mohammad Rostami Nejad6

1Department of Chemistry, Sharif University of Technology, Tehran, Iran
2Department of Basic Science, Faculty of Paramedical, Shahid Beheshti University of Medical Sciences, Tehran, Iran
3Gastroenterology and Liver Diseases Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran
4Department of Electrical Engineering, Golpayegan Payam University, Tehran, Iran
5Department of Mathematics, Teacher Training University, Tehran, Iran

ABSTRACT

Aim: The aim of this study was to search for metabolic biomarkers of Crohn’s disease (CD).

Background: Crohn’s disease (CD) is a type of inflammatory bowel disease that causes a wide variety of symptoms. CD can influence any part of the gastrointestinal tract from mouth to anus. CD is not easily diagnosed because monitoring tools are currently insufficient. Thus, the discovery of proper methods is needed for early diagnosis of CD.

Patients and methods: We utilized metabolic profiling using proton nuclear magnetic resonance spectroscopy (1H-NMR) to find the metabolites in serum. Classification of CD and healthy subject was done using partial least squares discriminant analysis (PLS-DA).

Results: According to PLS-DA model, we concluded just by using one descriptor CD and control groups could be classified separately. The level of lipid in blood serum of CD compared to healthy cohorts was decreased. For the external test set, the classification model showed a 94% correct classification of CD and healthy subject.

Conclusion: The result of classification model presents that NMR based metabonomics is a key tool as well as insight into potential targets for disease therapy and prevention.

Keywords: (1H-NMR), Metabonomics, Crohn's disease, PLS-DA.

Introduction

Crohn's disease (CD) is a type of inflammatory bowel disease (IBD) that is found in the last part of the small intestine and the first part of the large intestine (1-3). Both the host genotype and environmental factors play a role in etiology of CD. Also, presence of bacteria is needed for disease induction.

In order for an exact diagnosis of CD, several serological biomarkers have been proposed. However, they are used in conjunction and as a supplement to endoscopy. Thus, metabonomics as a monitoring tool are needed for early diagnosis (4). Metabonomics is defined as “the quantitative measurement of the dynamic multi-parametric response of living systems to pathophysiological stimuli or genetic modification” (5). One of the most commonly applied metabonomics is proton nuclear magnetic resonance (1H-NMR). This analytical technology has several advantages,
which is as follow: It provides quantitative and reproducible information with little sample preparation. Hence, it is widely used to build metabolic profiles in diverse metabolic studies. 1H-NMR spectroscopy of biofluid and tissue samples has been applied to the investigation and diagnosed many diseases of gastrointestinal tract (6-8).

So far several study has investigated metabolite profiles in serum of patients with CD in the literature (9). In the study by Schicho et al. metabolites in serum and plasma of IBD subjects were analysed. They obtained regular one-dimensional proton NMR spectra using a standard pulse sequence (Bruker pulse program prnoesy1d) (9). In another survey Fathi et al. applied classification and regression tree (CART) to explore the metabolic biomarker causes of CD compare to control group (10).

In this study, to determine the metabolites, we used 1H-NMR spectroscopy and performed quantitative analysis of metabolites in the serum of patients with active CD. PLS-DA was employed as a powerful classification method. Our aim was to search for metabolic biomarkers of CD to classify the control and CD groups.

**Patients and Methods**

**Sample collection**

Twenty-six adult patients with the mean age (± standard deviation) of 33.6±11.3 years diagnosed with Crohn's disease and twenty-nine healthy subjects with the mean age (± standard deviation) of 34.7±12.2 years were recruited from Gastroenterology and Liver Disease Research Center, Shahid Beheshti University of Medical Sciences. To avoid the affect of aging and gender influence on metabonomics, the healthy subjects were matched with CD subjects (9). Experienced gastroenterologists made diagnosis of CD on the basis of radiographic, clinical findings and often colonoscopy criteria.

Both CD and healthy cohort who entered the study had not significant other past medical history including hypertension, diabetes mellitus or hyperlipidemia. Serum samples were collected in the morning after a 12-hour fast and stored at −70 °C till measurement. 1H-NMR spectroscopy and data preprocessing were thoroughly explained in our previous study (11).

**Statistical analysis**

Partial least squares discriminate analysis (PLS-DA) was employed using PLS-Toolbox method (Version 2, Eigenvector Research Inc., Manson, WA) within MATLAB (version 6.5.1, The Mathworks, Cambridge, U.K.). This method is a regression technique adapted to a supervised classification task. Thus, it is a frequently used for classification method. It is on the basis of the partial least squares (PLS) approach (12). The standard PLS algorithm can be utilized and class labels can be used for the dependent Y vector. Usually in the two-class case, the values of the Y are given 1 for one class and 0 or -1 for the other class. Using this supervised analysis technique; we can identify those metabolites which show a discrepancy between diagnostic groups (13). In this study, 1H-NMR data and class labels were used as x matrix and y matrix respectively. The dataset was divided into two parts training set and test set. Training set was used to build a model and identify the most relevant metabolites and in order to test predict ability of the classification model, test set was employed.

**Results**

In classification part, PLS-DA was used to classify CD and healthy samples. Based on a randomly choice, approximately 30% of samples are left out to form the test set. Consequently, the training set was included of 39 1H-NMR spectra and the test set was composed of 16 spectra. By means of this procedure we can reduce the risk of over-fitting and avoid any possibility that the best classification
models selected had a chance correlation to peculiarities in the capacity of the test set (10). Figure 1 and Figure 2 present the score plots of the related PLS-DA of the 1H-NMR spectra of the serum for training and test sets respectively.

In figures 1 and 2, the first two latent variables LV1 versus LV2 map the representative points of the serum samples in the space were spanned. Also, scores plot was elucidating a logical clustering appearance according to class (CD and healthy subject). To explore exactly which metabolites have caused the separation between CD and healthy subjects, the loadings plot of PLS-DA model is depicted in Figure 3. Table 1 reveals the metabolite was the most prominent in serum separation with P-value < 0.00001.

A confusion matrix includes knowledge about the number of correct and incorrect predictions compared to the real outcomes by a classification model (17). Performance of such systems is commonly evaluated using the data in the matrix. The table 2 shows the confusion matrix for a two-class classifier. As it is clear from table 2, PLS-DA model has an accuracy of 0.94 in detecting CD patients for external test set.

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Assignment</th>
<th>1H chemical shift (ppm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipid</td>
<td>CH₂CH₂(CH₂)₉</td>
<td>1.26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observation</th>
<th>CD class</th>
<th>Healthy class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>Test</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

A summary of the classification parameters is shown in table 3. These results show that PLS-DA classification model has great chance in the
diagnosis of CD. Area under ROC curve is often used as a measure of quality of the classification models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1. The obtained values of AUC for training and test set are 1 and 0.94, respectively. The high AUC score of the proposed model for the samples in the external test set is another evidence for capability of PLS-DA model in CD detection.

Table 3. The calculated error and non-error rates of the classification index and the classification performances of training and test sets

<table>
<thead>
<tr>
<th></th>
<th>Error rate</th>
<th>Non-error rate</th>
<th>specificity</th>
<th>sensitivity</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Test</td>
<td>0.06</td>
<td>0.94</td>
<td>0.89</td>
<td>1</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Discussion

Investigation of the selected variable revealed the selected chemical shift corresponds to the NMR spectrum of lipid (14-16). Consequently, it can be stated the discrimination of CD and control samples by the PLS-DA model, based on NMR data, is on the basis of different amounts of lipid in the two groups. This result shows the reduction of lipid level in blood serum of patients compared to healthy individuals. Using PLS-DA method, the most important metabolite lipid was identified. The results of Kuroki and Hrabovsky et al. for lipid level in CD are similar to our study. They found that the serum levels of total lipids and total cholesterol were reduced in patients with CD (18, 19).

Lipids are molecules that include fat-soluble vitamins (such as vitamins A, D, E and K), monoglycerides, diglycerides, phospholipids, and others. They are necessary elements, which act as structural components of cell membranes. The main energy supply in cells and tissues is the role of lipids. Also it has been proved that the serum vitamin E level associates almost with serum lipid levels (20, 21). In the study by Fernandez-Banares et al. and Kuroki et al. (22, 23), serum concentrations of vitamin E have also been found to be lower in CD patient.

Fernandez-Banares et al. stated that the pathophysiological and clinical implications of the suboptimal vitamin status observed in acute CD are unknown. Moreover, Kuroki and colleagues suggested that there is a variety of vitamin deficiencies in CD prior to treatment that may reflect the severity of the disease (23).

Vitamin E is a fat-soluble antioxidant. It can be effective at scavenge free radicals at the cellular level and can be also prevent a great deal of the resultant scarring during CD. The result of inflammation in the gut mucosa and decreased oral intake can be cause of a risk for vitamin and mineral deficiencies in CD. Probably the reduction of lipid levels may relate to the remission of body level of Vitamin E. Classification model prediction procedure permits demonstration of the classification obtained by the PLS-DA technique is good enough to execute classification of unknown samples.

In the present study, a 1H-NMR based metabonomics approach was given as evidence for the existence of clear metabolic differentiation between two groups (CD and control group). Since 1H-NMR based metabonomics is effective to monitor the progression of disease, and helpful to discover biomarkers of CD, we can suggest that, NMR based on metabonomics can provide the possibility for assisting in early diagnosis of CD. Therefore, further investigations are required to establish its real usefulness in clinical practice.

Acknowledgements

The authors would like to express there thanks to the Gastroenterology and Liver Diseases Research Center and School of Paramedical
Science, Shahid Beheshti University of Medical Sciences, for their valuable collaboration in this study. We gratefully acknowledge financial support from Iran National Science Foundation (INSF), Sharif University of Technology and Shahid Beheshti University of Medical Sciences. This article was part of the PhD thesis of Fariba Fathi.

References