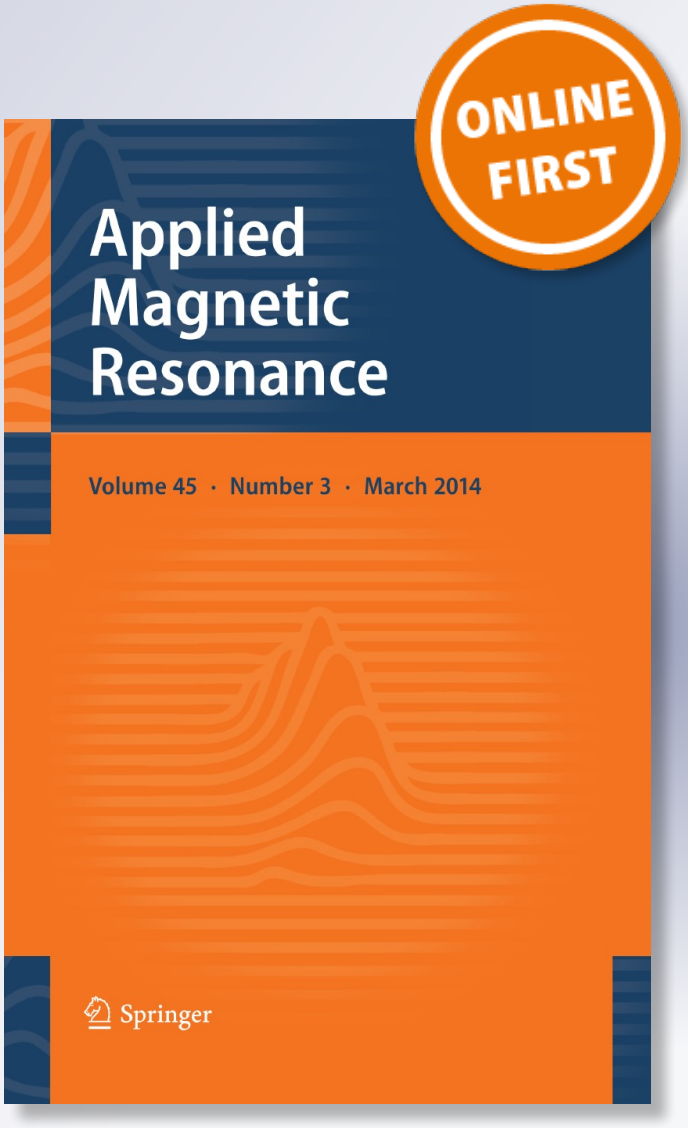
The Differential Diagnosis of Crohn's Disease and Celiac Disease Using Nuclear Magnetic Resonance Spectroscopy

Fariba Fathi, Laleh Majari Kasmaee, Kaveh Sohrabzadeh, Mohamad Rostami Nejad, Mohsen Tafazzoli & Afsaneh Arefi Oskouie

Applied Magnetic Resonance

ISSN 0937-9347

Appl Magn Reson
DOI 10.1007/s00723-014-0530-x



Springer

Appl Magn Reson
DOI 10.1007/s00723-014-0530-x

Applied
Magnetic Resonance

The Differential Diagnosis of Crohn's Disease and Celiac Disease Using Nuclear Magnetic Resonance Spectroscopy

Fariba Fathi · Laleh Majari Kasmaee · Kaveh Sohrabzadeh · Mohamad Rostami Nejad · Mohsen Tafazzoli · Afsaneh Arefi Oskouie

Received: 5 January 2014 / Revised: 8 March 2014

Abstract Crohn's disease and celiac disease belong to a group of autoimmune conditions that affect the digestive system, specifically the small intestine. They both attack the digestive tract and share many symptoms. Thus, the discovery of proper methods would be a major step toward differentiating celiac disease from Crohn's disease. The aim of this study was to search for the metabolic biomarkers to differentiate between these two diseases. Proton nuclear magnetic resonance spectroscopy (^{1}H NMR) was employed as the metabolic profiling method to look for serum metabolites that differentiate between celiac disease and Crohn's disease. Classification of celiac disease and Crohn's disease was done using random forest (RF). Based on RF results, Crohn's disease and celiac disease groups could be classified separately, using just two descriptors. The classification model showed 93 % correct classification of Crohn's disease and celiac disease subjects for the external test set. Based on feature importance of variables, isoleucine and lactate were selected. The findings of the present study reveal differentiating metabolites for Crohn's disease and celiac disease. These metabolites may provide diagnostic

F. Fathi (✉) · L. M. Kasmaee · M. Tafazzoli
Department of Chemistry, Sharif University of Technology, P.O. Box 11155-9516, Tehran, Iran
e-mail: f.fathi63@yahoo.com; f.fathi63@gmail.com

K. Sohrabzadeh
Department of Electrical Engineer, Payam Nonprofit Higher Education Institution, Golpayegan, Iran

M. R. Nejad
Research Center for Gastroenterology and Liver Disease, Shahid Beheshti University of Medical Sciences, Tehran, Iran

A. A. Oskouie (✉)
Department of Basic Science Faculty of Paramedical, Shahid Beheshti University of Medical Sciences, P.O. Box 19395-4618, Tehran, Iran
e-mail: a_arefioskouie@yahoo.com

Published online: 26 March 2014

biomarkers and/or monitoring tools as well as insight into potential targets for prevention and disease therapy.

1 Introduction

Crohn's disease (CrD) is an inflammatory bowel disease that, along with ulcerative colitis and some other conditions, causes inflammation (swelling) of the gastrointestinal tract [1]. Celiac disease (CD) is an autoimmune disorder caused by a permanent sensitivity to gluten protein in genetically predisposed individuals [2]. Diagnostic and monitoring tools for CrD are currently inadequate. Although several serological biomarkers have been proposed for the diagnosis of CrD [3], none of them can function as a stand-alone biomarker for clinical applications so currently their combination is employed as supplement to endoscopy. Therefore, non-invasive and accurate methods for early diagnosis of CrD, which can be used in place of endoscopy, are desirable.

Since CD shares symptoms with some other diseases, it is difficult to diagnose. It may be misdiagnosed as irritable bowel syndrome, iron-deficiency anemia caused by menstrual blood loss, CrD, diverticulitis, intestinal infections, and chronic fatigue syndrome, and as a result it is too often under diagnosed [4]. There are similarities between CD and CrD; they can cause weight loss, fatigue, diarrhea, skin problems, anemia, acid reflux and just plain discomfort in the digestive tract and/or intestines. In this regard, metabonomics is a valuable technique for detection of metabolites to diagnose two diseases from each other [5]. Metabonomics is defined as "the quantitative measurement of the dynamic multi-parametric response of living systems to path physiological stimuli or genetic modification" [6]. This approach combines higher resolution nuclear magnetic resonance (NMR) spectroscopic profiling of biological fluids with multivariate analysis as a technique to identify the metabolites that correlate with changes of physiological conditions. Also, metabonomics provides a global quantitative description of hundreds of low molecular endogenous metabolites present in a biological sample, such as urine, plasma, or tissue [7]. ^{1}H NMR provides an approach to evaluate metabolomes, and to determine endpoints of metabolic processes in biological systems [8, 9].

Proton signals of various metabolites in the form of complex matrices compose high-resolution ^{1}H NMR spectroscopy datasets. To select the relevant variables in the ^{1}H NMR dataset to build the pattern recognition and classification models, random forest (RF) can be used as a chemometrics method. RF is an ensemble of classification trees, and was developed by Leo Breiman [10]. Application of RF to the dataset reduces the variance and provides a low-bias estimate of the prediction accuracy. It has been used as a classification and feature selection method in omics science such as metabonomics. In many biological studies, RF has been used as a classification and regression method [11–16].

In a quest to differentiate CD and CrD patients using serum biomarkers, in this study, we investigated the correlation between specific metabolites and these diseases.

2 Materials and Method

2.1 Sample Collection

Twenty-six adult patients (12 males and 14 females with mean age of 34 ± standard deviation 11 years) diagnosed with CD (Marsh Stage 3) participated in this study. Each patient's diagnosis was based on positive serology and confirmed by histological examination of small bowel biopsy taken from the second part of the duodenum at the Research Center for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences. Also, 26 adult patients (11 males and 15 females with mean age of 33 ± standard deviation 10 years) diagnosed with CrD (Moderate to severe CrD), referred to the Research Center for Gastroenterology and Liver Disease, Shahid Beheshti University of Medical Sciences, participated in this study. The diagnosis of CrD had been made by established radiographic, experimental, and often colonoscopy criteria. After full explanation, all patients agreed to participate. Both CrD and CD cohort evaluated in this study had no other significant past medical history including hypertension, diabetes mellitus, or hyperlipidemia. The blood samples were collected in Eppendorf tubes. Immediately after the collection, the blood samples were placed in a sterile stoppered test tube and were allowed to coagulate for 20 min. Then the tubes were centrifuged at 2,500 rpm for 10 min to separate the sera, and stored at −80 °C until NMR spectroscopic analysis was carried out.

2.2 NMR Spectra Acquisition

^{1}H NMR experiments were carried out on a Bruker DRX500 MHz spectrometer operating at 500.13 MHz, equipped with 5 mm high-quality NMR tubes (Sigma Aldrich., RSA). 100 µl of D_2O (Deuterium oxide, 99.9 %D, Aldrich Chemicals Company) provided NMR lock signal for the NMR spectrometer.

The broad signals of high molecular weight species such as lipoproteins are overlaid with sharper resonances caused by the low molecular weight species (amino acids, carboxylic acids). To reduce the broadness of the signals, Carr–Purcell–Meiboom–Gill (CPMG) spin echo pulse sequent with spin echo sequence $\pi/2$–t_D–π–t_D was carried out [11]. Acquisition parameters were: spectral width 8,389.26 Hz, time domain points 32 K, relaxation delay 2 s, number of scans 154, acquisition time 1.95 s, spectrum size 32 K. Prior to Fourier transformation, an exponential line broadening function of 0.3 Hz was applied to the free induction decay. All serum ^{1}H NMR spectra were manually phased and baseline corrected by means of XWINNMR software (version 3.5, Bruker Spectrospin Ltd). Also by means of this software, the serum spectra were with reference to the methyl doublet of lactate ($\delta = 1.33$ ppm) [12].

2.3 Data Pre-Processing

Each NMR spectrum was reduced to smaller number of variables, calculated by integrating regions of equal bucket size of 0.04 ppm within ProMetab software

(version prometab_v3_3) in MATLAB (version 6.5.1, The Mathworks, Cambridge, UK). Data were reduced into 205 spectral integral regions corresponding to the chemical shift range of δ 0.2–10 ppm [13–15]. Since the spectra in the region of δ 4–5.5 ppm correspond to the effects of variations in the pre-saturation of the water resonance, they were removed [16]. To reduce variation in the sample concentrations, the integral values of each spectrum were normalized to a constant sum of all integrals in a spectrum [17, 18]. Before any statistical analysis, NMR spectral variables were mean centered [19].

2.4 Random Forest

Random forest (RF) is one of the most successful ensemble learning techniques [10]. RF is employed to construct a collection of individual decision tree classifiers, which utilize the classification and regression trees (CART) algorithms [20]. The simplest random forest with random features is created. By means of arbitrary selecting, at each node, a small group of input variables is split. While the forest is growing, the size of the group is fixed. Each tree is grown without pruning. To classify x, following formation of the forest, a case with input x is dropped into the forest for each tree. The class that has the majority vote is selected for x by the forest. About one-third of the cases were not used in the construction of the k-th tree. They are called out-of-bag (OOB), and are left out of the bootstrap sample. After the construction of the trees, by entering each OOB case into its relevant k-th tree, the classification ability and accuracy of the k-th tree are estimated [21, 22].

3 Result

3.1 Classification Using Random Forest

To identify the signals in the serum spectra, chemical shifts in the spectrum in Fig. 1 were compared with established libraries reported in the literature [23–25] and the Human Metabolome Data Base (HMDB) [26]. After signal identification process, RF was employed to classify the CD and CrD samples.

To examine the predictability of the models, the CD and CrD datasets were divided into training and test sets. The division in test and training set was performed using the Duplex algorithm. It was chosen that the test sets would contain about 30 % of the samples.

At first, two objects in the data matrix were chosen using the Duplex algorithm. Selection of these two objects is based on their Euclidean distance. These objects are put into the training set. The next step was to select the two objects farthest from each other among the remaining candidates and put them into the test subset. Then, consecutive objects are selected and put alternatively in the training and test sets. Application of this method assures us that no significant extrapolation of the training dataset takes place in the test set, and the test and training points were evenly distributed throughout the data space [27, 28].

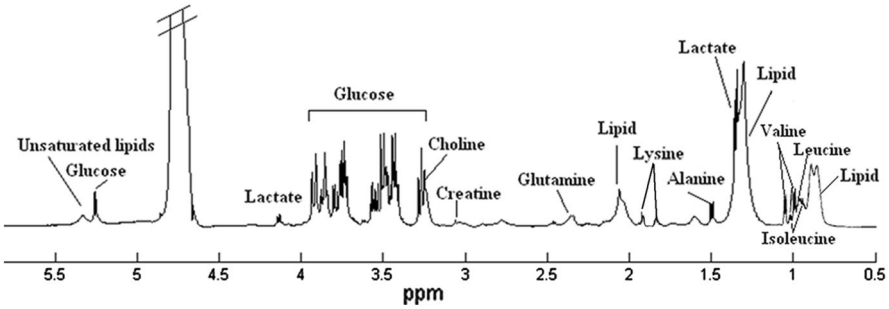


Fig. 1 Typical 500 MHz ^{1}H NMR spectra of control human blood serum. Eleven metabolites viz., lipid, leucine/isoleucine, valine, lactate, alanine, glutamine, glutamate, citrate, creatine, and glucose were identified

Using the samples in the training set, 500 trees were grown with RF application. The OOB data were used to estimate the prediction accuracy of classification. Figure 2 presents the OOB error rate. To estimate the significance of a variable, the RF algorithm considers the amount of increase in the prediction error when OOB data for that variable are permuted while all other variables are left unchanged. Isoleucine and lactate were selected according to feature significance of variables. Distributions of two metabolites are considerably different between CD and CrD cohorts ($p < 0.006$). Confusion matrix of the RF classification model for the training and test sets is shown in Table 1. Other classification parameters are shown in Table 2. These results show that RF classification model has great chance to diagnose CD and CrD cohorts.

Receiver operating characteristic (ROC) curve represents the performance of the decision-making algorithm with regard to the decision parameter. An Area under the ROC Curve (AUC) is traditionally used in medical diagnosis systems [29]. It can also evaluate the predictability of learning algorithms. Figure 3 present AUC for test set. The obtained values of AUC for training and test sets are 1 and 0.94, respectively. Based on the high AUC scores of the proposed model for the samples in the external test set, it is evident that our RF model is highly capable of CD and CrD diagnosis.

4 Discussion

In accordance with previous findings, the significance of the metabolites chosen by RF and their concentration changes in CD and CrD patients are scrutinized below. Based on the area under the curve for metabolite concentration, isoleucine and lactate levels in CD are lower than in CrD. Isolecine is an essential amino acid so cannot be produced in the body, and therefore should be obtained from the diet. Along with leucine and valine, it is a branched chain amino acid (BCAA). Generally, BCAAs aid muscle recovery after physical exercise. Isolecine specifically helps with blood sugar and energy regulation, hemoglobin formation, and

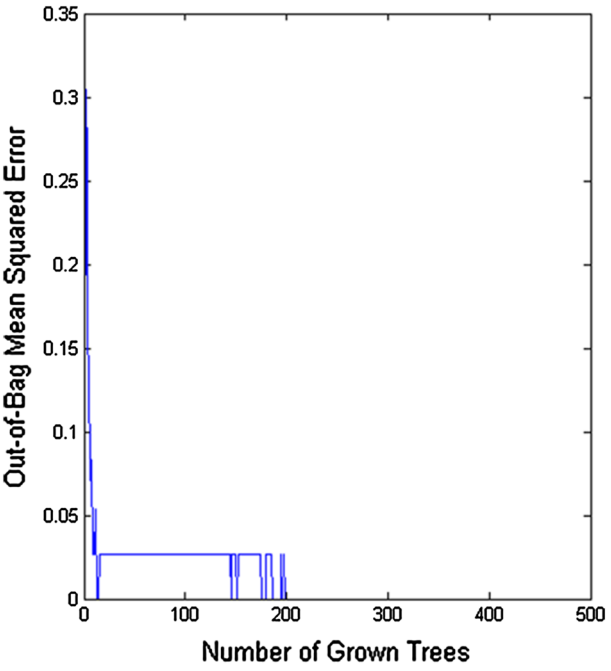


Fig. 2 Plot of OOB error for RF classification of PD and control group. To estimate the prediction accuracy of classification, the OOB data were applied

Table 1 Confusion matrix for training and test set

	Predicted		
	Observed	CrD class	CD class
Training set	CrD class	18	0
	CD class	0	19
Test set	CrD class	7	1
	CD class	0	7

Table 2 Calculated error and non-error rates of the classification index and the classification performances of training and test sets

	Error rate	Non-error rate	Specificity	Sensitivity	Accuracy
Training set	0.0	1	1	1	1
Test set	0.07	0.93	0.88	1	0.93

blood clotting. Isoleucine is the degradation product of 3-methyl-2-oxovalerate. In enzymology, a valine-3-methyl-2-oxovalerate transaminase is an enzyme that catalyzes the chemical reaction between valine and 3-methyl-2-oxopentanoate. This

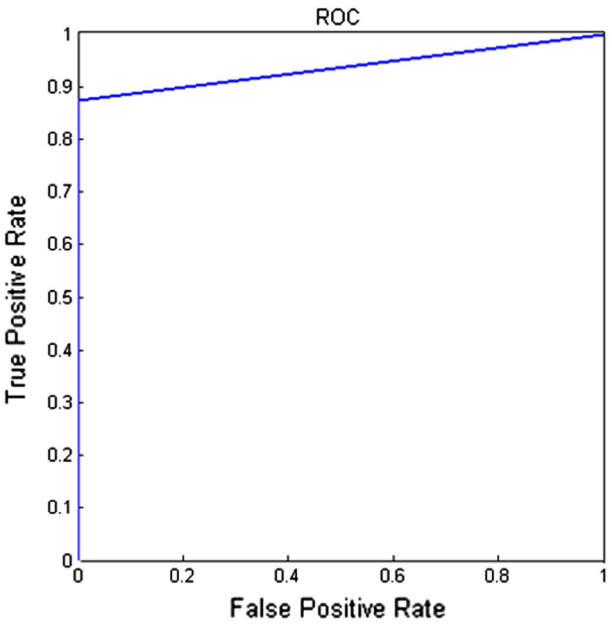


Fig. 3 High AUC for test set suggests that the model is able to accurately predict the value of an observation's response

enzyme belongs to a group of transferases family that transfers nitrogenous groups. According to literature, the level of transaminase is elevated in 40–50 % of CD patients [30]. CrD can affect an amino acid, such as isoleucine, metabolism [31].

Schicho and co-authors showed the metabolic differences between the CrD and control subjects, and concluded that amino acids such as isoleucine were the most prominently increased metabolites in serum of CrD patients, compared to the healthy individuals [31]. In other study, Hernanz and co-authors investigated plasma precursor amino acids of central nervous system monoamines in children with CD [32]. They found plasma concentrations of neutral amino acids such as valine, isoleucine, and leucine were significantly decreased in CD. These decreased levels of larger neutral amino acids could explain the normal plasma tyrosine ratio found in CD group compared with the control group [32].

Our findings show high levels of lactate in CrD patients' plasma samples, although, in the literature high lactate levels have been preferably linked to severe ulcerative colitis [33]. In CrD patients, there is a decline in Krebs cycle intermediates, such as succinate and citrate, as well as decrease in species involved in energy metabolism (such as acetate). This may be an indication of higher demand for these metabolites in CrD patients, and fast utilization of them in energy producing pathways [31].

Lack of pyruvate is the underlying reason for low lactate levels in CD patients. Glycolysis is the anaerobic metabolism of glucose, and produces pyruvate. Upregulation of glucose intake at the cellular level, caused by alteration of the

lipid-to-protein ratio of the microvillus membrane, can be the reason for higher levels of glucose. In CD patients, a decrease of pyruvate in serum level is the result of impairment in glycolysis due to reduction of glucose intake at cellular level. Impairment of glycolysis explains both a lowering of pyruvate and lactate levels, and an increase of glucose levels in blood [34].

The accumulation of undigested carbohydrate in the small intestine causes the microbial multiplication. Short-chain organic acids such as lactic and acetic acids, as well as gas produced by bacteria can damage the small intestine. In addition, lactic acid produced from fermentation in the intestine causes abnormal brain function and behavior that often accompany intestinal disorders [35]. Moreover, increased amount of gas, acids, and other fermentation products destroys the enzymes on the intestinal cell surface and prolongs malabsorption.

To control the extent of undigested carbohydrate fermentation and the growth of lactic acid amount in the small intestine, a special carbohydrate diet is prescribed.

As a consequence of carbohydrate malabsorption, the undigested carbohydrates undergo further fermentation, which produces excessive mucus. This phenomenon triggers a self-defense mechanism in the body; increased mucus production is an attempt by small intestine to lubricate itself against injury caused by microbial toxins, acids, and the presence of incompletely digested and unabsorbed carbohydrates.

We suggest that further studies be conducted to investigate the correlation between serum metabolites and severity, and different stages of CD and CrD. This has been out of the scope of this study.

In conclusion, our findings indicate that quantitative metabolite analysis of serum from CD and CrD patients is an effective technique to discriminate between these two diseases. Also, based on the results, metabolic profiling can be applied to detect intestinal inflammation, and therefore may be employed to manage CD and CrD, and to explore disease pathogenesis in clinical studies.

References

1. F. Fathi, A. Kyani, M.R. Nejad, M. Rezaye-Tavirani, N. Naderi, M.R. Zali, M. Tafazzoli, Health-MED. **6**, 3577–3584 (2012)
2. M. Rezaei-Tavirani, F. Fathi, F. Darvizeh, M.R. Zali, M.R. Nejad, K. Rostami, M. Tafazzoli, A.A. Oskouie, S.A. Mortazavi-Tabatabaei, Int. J. Endocrinol. Metab. **10**, 548–552 (2012)
3. J.D. Lewis, Gastroenterology **140**, 1817–1826 (2011)
4. F. Fathi, F. Ektefa, A.A. Oskouie, K. Rostami, M. Rezaei-Tavirani, A.H.M. Alizadeh, M. Tafazzoli, M.R. Nejad, Gastroenterol. Hepatol. Bed. Bench **6**, 190–194 (2013)
5. J.B. German, B.D. Hammock, S.M. Watkins, Metabolomics **1**, 3–9 (2005)
6. J. Nicholson, J. Lindon, E. Holmes, Xenobiotica **29**, 1181–1189 (1999)
7. J. Nicholson, J. Connelly, J. Lindon, E. Holmes, Nat. Rev. Drug Discov. **1**, 153–161 (2002)
8. R. Hewer, J. Vorster, F.E. Steffens, D. Meyer, J. Pharma, Biomed. Anal. **41**, 1442–1446 (2006)
9. I. Barba, R. Fernandez-Montesinos, D. Garcia-Dorado, D. Pozo, J. Cell Mol. Med. **12**, 1477–1485 (2008)
10. L. Breiman, Mach. Learn. **45**, 5–32 (2001)
11. G. Zhang, G. Hirasaki, J. Magn. Reson. **163**, 81–91 (2003)

12. K.A. Verwaest, T.N. Vu, K. Laukens, L.E. Clemens, H.P. Nguyen, B.V. Gasse, J.C. Martins, A.V. Derlinden, R. Dommisse, Biochim. Biophys. Acta **1812**, 1371–1379 (2011)
13. E. Holmes, P.J. Foxall, J.K. Nicholson, G.H. Neild, S.M. Brown, C.R. Beddell, B.C. Sweatman, E. Rahr, J.C. Lindon, M. Spraul, P. Neidig, Anal. Biochem. **220**, 284–296 (1994)
14. C.L. Gavaghan, E. Holmes, E. Lenz, I.D. Wilson, J.K. Nicholson, FEBS Lett. **484**, 169–174 (2000)
15. E. Holmes, J.K. Nicholson, A.W. Nicholls, J.C. Lindon, S.C. Connor, S. Polley, J. Connelly, Chemom. Intell. Lab. Syst. **44**, 245–255 (1998)
16. M. Coen, P. Kuchel, CiA **71**, 13–17 (2004)
17. E. Holmes, A.W. Nicholis, J.C. Lindon, S. Ramos, M. Spraul, P. Neidig, S.C. Connor, J. Connelly, S.J.P.H. Damment, J.K. Nicholson, NMR Biomed. **11**, 235–244 (1998)
18. N.J. Waters, E. Holmes, A. Williams, C.J. Waterfield, R.D. Farrant, J.K. Nicholson, Chem. Res. Toxicol. **14**, 1401–1412 (2001)
19. S.A. Mortazavi-Tabatabaei, F. Fathi, F. Ektefa, M. Tafazzoli, A.A. Oskouie, M. Rezaie-Tavirani, M.R. Zali, M.R. Nejad, K. Rostami, J. Paramed. Sci. **4**, 2–10 (2013)
20. L. Breiman, J. Friedman, R. Olshen, C. Stone, *Classification and Regression Trees* (Chapman and Hall, London, 1984)
21. J.H. Barrett, D.A. Cairns, Stat. Appl. Genet. Molec. Biol. **7**, 1–29 (2008)
22. M.W. Mitchell, Open J. Stat. **1**, 205–211 (2011)
23. J.K. Nicholson, P.J. Foxall, M. Spraul, R.D. Farrant, J.C. Lindon, Anal. Chem. **67**, 793–811 (1995)
24. M. Fan, Prog. Nucl. Magn. Reson. Spectrosc. **28**, 161–219 (1996)
25. J.C. Lindon, Annu. Rep.NMR Spectrosc. **38**, 1–88 (1999)
26. D.S. Wishart, D. Tzur, C. Knox, R. Eisner, A.C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, Nucleic Acids Res. **35**, D521–D526 (2007)
27. R.D. Snee, Technometrics **19**, 415–428 (1977)
28. E. Deconinck, P.Y. Sacré, D. Coomans, J. Debeer, J. Pharm. Biomed. Anal. **57**, 68–75 (2012)
29. F. Fathi, A. Kyani, F. Darvizeh, M. Mehrpour, M. Tafazzoli, G. Shahidi, Appl. Magn. Reson. **44**, 721–734 (2013)
30. S. Korpimäki, K. Kaukinen, P. Collin, A.M. Haapala, P. Holm, K. Laurila, K. Kurppa, P. Saavalainen, K. Haimila, J. Partanen, M. Mäki, M.L. Lähdeaho, Am. J. Gastroenterol. **106**, 1689–1696 (2011)
31. R. Schicho, R. Shaykhutdinov, J. Ngo, A. Nazyrova, C. Schneider, R. Panaccione, G.G. Kaplan, H.J. Vogel, M. Storr, J. Proteome Res. **11**, 3344–3357 (2012)
32. A. Hernanz, I. Polanco, Gut **32**, 1478–1481 (1991)
33. P. Vernia, R. Caprilli, G. Latella, F. Barbetti, F.M. Magliocca, M. Cittadini, Gastroenterology **95**, 1564–1568 (1988)
34. P. Bernini, I. Bertini, A. Calabro, G.L. Marca, G. Lami, C. Luchinat, D. Renzi, L. Tenori, J. Proteome Res. **10**, 714–721 (2011)
35. E. Gottschall, *Breaking the Vicious Cycle: Intestinal Health through Diet* (Kirkton Press, Baltimore, 1994)