# Advantage of Applying OSC to ^{1}H NMR-Based Metabonomic Data of Celiac Disease

Mostafa Rezaei-Tavirani [1], Fariba Fathi [2], Fatemeh Darvizeh [3], Mohamad Reza Zali [4], Mohamad Rostami Nejad [5], Kamran Rostami [6], Mohsen Tafazzoli [2], Afsaneh Arefi oskouie [6*], Seyed AbdolReza Mortazavi-Tabatabaei [6]

[1] Proteomics Research Center, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, IR Iran
[2] Department of Chemistry, Sharif University of Technology, Tehran, IR Iran
[3] Department of Medicine, Debrecen Medical School, Debrecen, Hungary
[4] Research Center for Gastroenterology and Liver Disease, Shahid Beheshti University of Medical Sciences, Tehran, IR Iran
[5] Acute Medicine, Dudley Group of Hospital, Dudley, UK
[6] Department of Basic Science Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, IR Iran

| ARTICLE INFO | ABSTRACT |
|---|---|
| | ***Background***: Celiac disease (CD) is a disorder associated with body reaction to gluten. After the gluten intake, an immune reaction against the protein occurs and damages villi of small intestine in celiac patients gradually.<br>***Objectives:*** The OSC, a filtering method for minimization of inter- and intra-spectrometer variations that influence on data acquisition, was applied to biofluid NMR data of CD patients.<br>***Patients and Methods:*** In this study, metabolites of total 56 serum samples from 12 CD patients, 15 CD patients taking gluten-free diet (GFD), and 29 healthy cases were analyzed using nuclear magnetic resonance (NMR) and associated theoretical analysis. Employing ProMetab (version ProMetab_v3_3) software, data obtained from NMR spectra were reduced and orthogonal signal correction (OSC) effect on celiac disease metabonomics before and after the separation by principle component analysis (PCA) was investigated.<br>***Results:*** The three groups were separated by OSC and findings were analyzed by partial least squares discriminant analysis (PLS-DA) method. Root mean square error of calibration (RMSEc) and correlation coefficient of calibration (Rc) for PLS-DA referred to an efficient group separation filtered by OSC.<br>***Conclusions:*** The applied leave-one-out cross-validation to PLS-DA method performed along with OSC confirmed validation of data analysis. Finally four metabolites are introduced as CD biomarkers.<br><br>Publish by *Kowsar Corp*. All rights reserved. |

▶ *Implication for health policy/practice/research/medical education:*
^{1}H NMR-Based Metabonomic showed that four metabolites were introduced to differentiate between celiac patients on GFD, celiac patients without specific diet, and healthy people. The results of present study could help someone who desire to focus on the Investigation of complex metabolic systems such as disease mechanisms, toxic reactions, and genetic manipulations.

* *Corresponding authors:* Mohsen Tafazzoli, Department of Chemistry, Sharif University of Technology, Tehran, IR Iran. Tel.: +98-2166165305; fax: +98-2166012983;P.O. Box 11155- 9516, E-mail Tafazzoli@sharif.edu; Afsaneh Arefi oskouie, Department of Basic Science Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, P.O. Box: 19395-4618, Tehran, IR Iran. Tel +98-2122718505, Fax: +98-2166012983, *E-mail:* a.arefi@sbmu.ir

## 1. Background

Celiac desease (CD) is a disorder caused essentially by body reaction to the gluten. Intake of foods containing gluten promotes an immune reaction against the

protein that damages villi of small intestine in celiac patients gradually. Consequently, lack of vitamins, minerals, and other essential nutrients occurs. Therefore, celiac patients are at the risk of malnutrition, , anemia and osteoporosis from which the anemia attributes to iron deficiency and results in declining red blood cell efficiency, and the osteoporosis may represent as fragile bones caused by lack of calcium (1).

Full analysis of a living organism can be achieved by an integrated set of 'omics' approaches including metabonomics, genomics, transcriptomics, and proteomics in order to increase data complexity. Metabonomics generates sufficient quantitative or qualitative metabolic data for analytical studies of biological systems (2-7) and is originally defined as 'the quantitative measurement of the dynamic multi-parametric metabolic response of living systems to pathophysiological stimuli or genetic modification'(5).

Investigation of complex metabolic systems such as disease mechanisms, toxic reactions, and genetic manipulations requires full analytical data sets. Nuclear magnetic resonance (NMR) and mass spectrometry (MS) are the most conventional techniques for metabolic profiling (8-10).

Multivariate statistical approaches are also suitable methods for extraction metabolic data associated with each level of dynamic processes. Since NMR method produces various and highly correlated data, employing multivariate methods such as PCA prior to any further data analysis is essential to choose efficient number of descriptors (11).

## 2. Objectives

In this study OSC, a filtering method for minimization of inter- and intra-spectrometer variations that influence on data acquisition, was applied to biofluid NMR data of CD patients.

## 3. Patients and Methods

### 3.1. Blood Samples

Using syringes, two milliliters of blood samples were drawn from antecubital vein of each case through a single puncture followed by immediate serum separation via centrifugation, and storage at -20°C. Experimental data was obtained by placing serum samples in 5mm NMR tubes. 600 μL serum samples each were diluted by 100 μL $D_2O$ to provide a field-frequency lock.

The experimental data consisted of 56 serum samples including 12 samples from celiac patients on gluten-free diet, 15 samples from patients without a specific diet, and 29 samples from healthy cases.

### 3.2. $^1$H NMR Spectroscopy

A BrukerAvance DRX 600 spectrometer operating at 500 MHz at 300 K using the Carr–Purcell–Meiboom–Gill (CPMG) spin–echo sequence (12) with pre saturation was employed for recording $^1$H NMR spectra. QNP probe was used in this experiment.

Spin–echo loop time (2$nt$) and relaxation delay for recorded spectra were 43.9 ms and 2s, respectively. A total of 128 transients were collected into 32k data points using a spectral width of 8389.26 Hz and an acquisition time of 1.95 s. Prior to Fourier transformation, an exponential line broadening function of 0.30 Hz was applied to the Free Induction Decay (FID).

### 3.3. Data Pre-Processing

Employing Prometab software (version Prometab_v3_3), correction of baseline was carried out according to 4, 4-Dimethyl-4-Silapentane-1-Sulfonic acid (DSS) as reference. The number of variables was reduced to 245 in each spectrum by integrating spectral intensity in regions of equal width [0.04 parts per million (ppm) over the range 0.2–10.0 ppm]. In addition, the region of ($\delta$ = 4.50–5.98) was excluded from the analysis to avoid unauthentic effects of variability in the suppression of water resonance. Therefore, the number of variables was reduced to 205. All spectra were normalized to constant total intensity.

### 3.4. Statistical Analysis

### 3.4.1. OSC for Classification Modeling

A response matrix ($Y$) describing variation between defined sample classes was constructed and the variation in $X$ orthogonal to $Y$ was removed by applying OSC. The filtered data matrix, $X_{osc}$, that included class-correlated variations was modeled by subsequent multivariate modeling methods such as PCA or PLS-DA(14). OSC Modeling eliminates variables that are not correlated with desirable features from $^1$H NMR biofluid data; as a result calculated values in multivariate models motivate only class separation. In this study, OSC was applied to biofluid NMR data prior to chemometric analysis to minimize influences of inter- and intra-spectrometer variations during data acquisition. Moreover, it was employed to eliminate physiological variation from data sets (13-17).

### 3.4.2. Principal Components Analysis (PCA)

PCA is a conventional technique for multivariate analysis and mainly employed for multivariate data representation in a low-dimensional space that means it describes maximum possible number of variables with minimum possible number of principal components (PCs). In PCA, principal components are named consecutively starting from $PC_1$ until total variance is defined. $PC_1$ or first principal component is a line that goes through the points in a variable space and best conserves relevant distances between objects and defined by a loading vector as follows. $PC_1$ is a hidden variable and has maximum variance of the scores. Scores are predicted data values on the hid-

den variable (18).

All calculations were performed in MATLAB6.5 and PCA was implemented with the PLS-Toolbox Version 3.0. Both the graphical ('pcagui') and the command-line ('pca') versions of PCA have been employed (19).

### 3.4.3. Partial Least Squares Discriminant Analysis (PLS-DA)

PLS regression was directed by a response data set Y to derive components from descriptor data set X that best describe specified Y structure, as it maximizes the covariance that expresses common structure between X and Y(20-22). PLS is also divided into regression as well as discriminant analysis (PLS-DA). Classification by DA assigns the samples to proper separate classes which are represented by using so-called 'dummy' variables (23).

## 4. Results

As shown in *Figure 1a,* data analysis leads to introduce two groups of control and celiac cases. Since separation of GDF group from other two groups was an important index in this study, more analytical efforts were effected by using OSC in order to achieve better transparency between sample groups. The useful role of OSC in separation groups is reported in several studies (24). By using OSC, 95.26% total variability was distributed between three PCs as $PC_1$ (91.11%), $PC_2$ (2.88%), and $PC_3$ (1.47%) (See *Figure 1b*). After successful division of samples into three groups, PLS-DA as regression extension of PCA was applied to maximize the separation. It is reported that PLS-DA is useful tool for maximizing covariance between measured data (x) and response variable (y) (25). PLS-DA findings introduced three most influential metabolites that play important role in separation of groups. Serum level alteration of these metabolites in CD group compared to control group is shown in *Table 1*.

The percentage of captured variance by PCA model is a suitable index for data validation (26); accordingly, three times $PC_s$ and their associated variance amounts were calculated and shown in *Table 2* with total variance equal to 92.22% out of which the percentage variance in step 1 was equal to 84.38%, and in second and third steps were 4.74 and 3.12, respectively. . Since 92.24% is close to 1, we may conclude that PCA model could separate the three groups in a real and valid manner.

Score plots of PLS-DA without and with application of OSC for three groups are shown in *Figure 2*. The useful

**Table 2.** Percent Variances Captured by PCA Model. First column corresponds to steps of model application and the second and third columns refer to variance of each step and total variance, respectively.

| Number | % Variance | % Variance Total |
|---|---|---|
| 1 | 84.38 | 84.38 |
| 2 | 4.74 | 89.12 |
| 3 | 3.12 | 92.24 |

**Table 3.** Summary of PLS-DA and OSC- PLS-DA models of NMR spectra. LVs refer to number of latent variances and Vx and Vy correspond to NMR data matrix and respond matrix, respectively. Vx (cuml, %) and Vy show total variances.

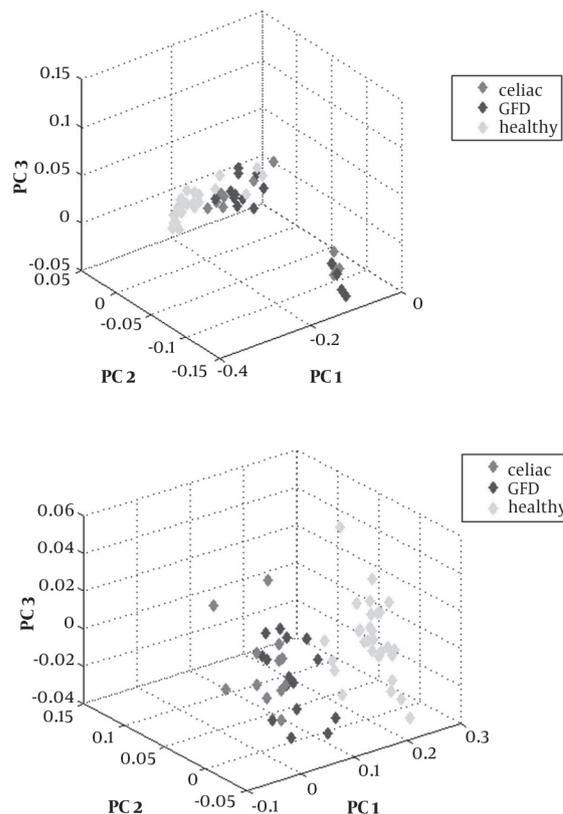| LVs | $V_X$ (%) | $V_X$ (cuml, %) | $V_Y$ (%) | $V_Y$ (cuml, %) |
|---|---|---|---|---|
| | | **PLS-DA Model** | | |
| 1 | 63.96 | 63.96 | 44.07 | 44.07 |
| 2 | 9.05 | 73.01 | 16.34 | 60.41 |
| 3 | 2.54 | 75.55 | 8.24 | 66.73 |
| | | **OSC- PLS-DA Model** | | |
| 1 | 75.25 | 75.25 | 100.00 | 100.00 |
| 2 | 6.11 | 81.36 | 0.00 | 100.00 |
| 3 | 5.46 | 86.82 | 0.00 | 100.00 |



**Figure 1.** Score Plot PCA a) Without OSC, and b) With OSC

**Table 1.** PCA-Detected 1H NMR Spectral Regions that Separate Significantly CD and Control Groups Based on Metabolites Levels

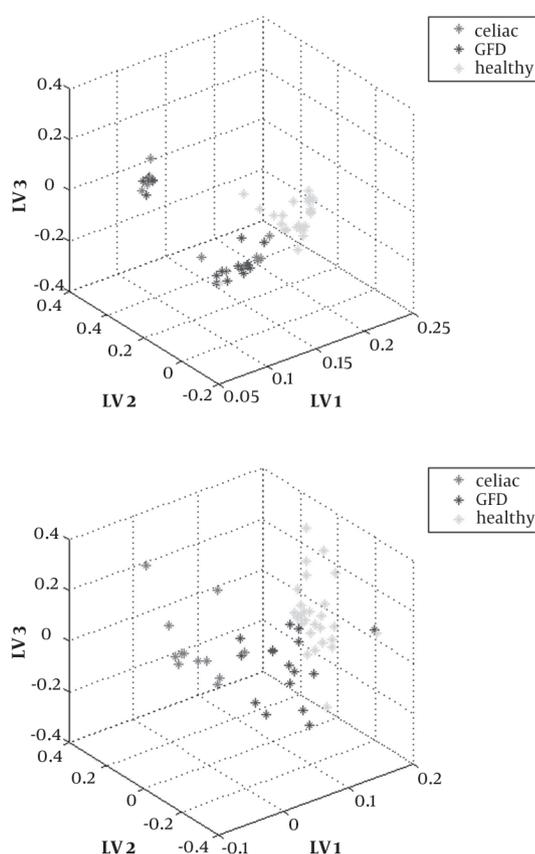| Chemical Shift (ppm) | Metabolite | Alteration of Metabolite Level in CD Group Compared to Control Group |
|---|---|---|
| 0.86 | lipid (mainly VLDL) | Decrease |
| 1.26 -1.3 | Lipid | Decrease |
| 1.34 | Lactate | Decrease |
| 3.22 | Choline | Increase |

**Figure 2.** Score Plot PLS-DA a) Without OSC, and b) With OSC

parameter to analyze PLS-DA findings is latent variable variance (LV) (16). LVs of represented data in *Figure 2* for three steps of PLS-DA were calculated and shown in *Table 3*. There are several evidences implicate that total variance value above 80% is an acceptable index for data validation. Accumulated PLS-DA score plot by using OSC for three LV variances was 86.82% (*see Table 3*) that revealed a clear separation between the three groups; however this procedure without applOSC application (identified by an accumulated score plot of 75.55%) could not be considered as a suitable analytical method.

RMSEc and Rc values for PLS-DA calibration without applying OSC were equal to 0.6682 and 0.6348, respectively. By using OSC, these values changed to 0.4226 and 0.8629, respectively. Amounts of Rc close to 1 correspond to acceptable data analysis (27) that means the calculated Rc value of 0.8629 refers to application of a proper analytical method for group separation. Besides this, OSC application reduced RMSEc value, that means an error reduction occurred. For better clarity, the applied OSC model was performed accompanied by leave-one-out cross-validation. The characterized findings were 0.5175 and 0.8510 for RMSEcv and Rcv, respectively. The low amount of error and Rcv close to 1, confirm sufficient validity of findings.

## 5. Discussion

According to orthogonal theory in mathematics, OSC can eliminate data in X matrix that are orthogonal with response Y matrix. The response Y matrix is class variable. The Y matrix variables determined as control, CD, and GFD groups were assigned to as 0 , 1, and 2, respectively. Orthogonal component with eigenvalues greater than 1 were eliminated. In order to perform metabonomics analysis of recorded ¹H NMR spectra of healthy, celiac, and GFD groups, two different recognition pattern methods before and after OSC were applied. Score plots of PCA are drawn using the MATLAB software. *Figure 1a* and *Figure 1b* demonstrate PCA score plots of NMR spectra without and with OSC, respectively.

On the basis of this study, four metabolites were introduced to differentiate between celiac patients on GFD, celiac patients without specific diet, and healthy people. We hope that further investigations lead to determine exact metabonomic pattern of CD.

## References

1. Nejad MR, Rostami K, Pourhoseingholi MA. Atypical presentation is dominant and typical for coeliac disease. *J Gastrointestin Liver Dis*. 2009;**18**: 285 – 91.
2. Greef JVd, Tas A, Bouwman J, Debrauw M, Schreurs W. Evaluation of field-desorption and fast atom-bombardment massspectrometric profiles by pattern-recognition techniques. *AnalChim Acta*. 1983;**150**:45–52.
3. Kaddurah-Daouk R, Kristal B, Weinshilboum R. Metabolomics: a global biochemical approach to drug response and disease. *Annu Rev Pharmacol*. 2008;**48**:653–83.
4. Nicholson J, Lindon J. Systems biology – metabonomics. *Nature*. 2008;**445**:1054–6.
5. Nicholson J, Lindon J, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic
data. *Xenobiotica*. 1999;**29**:1181–9.
6. Nicholson J, Wilson I. High resolution proton magnetic resonance spectroscopy of biological fluids. *Prog Nucl Magn Reson Spectrosc*. 1989;**21**:449–501.
7. Oliver S, Winson M, Kell D, Baganz F. Systematic functional analysis of the yeast genome. *Trends Biotechnol*. 1998;**16**: 373-8.
8. Lindon J, Nicholson J. Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics.

*Annu Rev Anal Chem*. 2008;**1**:45–69.

9.  Ren Y, Wang T, Peng Y, Xia B, Qu L-J. Distinguishing transgenic from non-transgenic Arabidopsis plants by 1H NMR-based metabolic fingerprinting. *J Genet Genomics*. 2009;**36**:621-8.

10. Werner E, Heilier J, Ducruix C, Ezan E, Junot C, Tabet J. Mass spectrometry for the identification of the discriminating signals from metabolomics: current status and future trends. *J Chromatogr B*. 2008;**871**:143–63.

11. Keun C, Ebbels TMD, Antti H, Bollard ME, Beckonert O, Holmes E, *et al.* Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling Hector *Anal Chim Acta*. 2003;**490**:265–76.

12. Meiboom S, Gill D. Modified spin-echo method for measuring nuclear relaxation time. *Rev Sci Instrum*. 1958;**20**:688–91.

13. Beckwith-Hall BM, Brindle JT, Barton RH, Coen M, Holmes E, Nicholson JK, *et al.* Application of orthogonal signal correction to minimise the effects of physical and biological variation in high resolution 1H NMR spectra of biofluids. *Analyst*. 2002;**127**:1283–8.

14. Gavaghan C, Wilson I, Nicholson J. Physiological variation in metabolic phenotyping and functional genomic studies: use of orthogonal signal correction and PLS-DA. *FEBS Lett*. 2002;**530**:191–6.

15. Mao H, Xu M, Wang B, Wang H, Deng X, Lin D. Evaluation of filtering effects of orthogonal signal correction on metabonomic analysis of healthy human serum 1H NMR spectra. *ActaChimSinica*. 2007;**65**:152–58.

16. Rantalainen M, Cloarec O, Beckonert O, Wilson I, Jackson D, Tonge R, *et al.* Statistically integrated metabonomicproteomic studies on a human prostate cancer xenograft model in mice. *J Proteome Res*. 2006;**5**:2642–55.

17. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J Chemom*. 2002;**16**:119–28.

18. Berrueta LA, Alonso-Salces RM, Héberge K. Supervised pattern recognition in food analysis. *J Chromatogr A*. 2007;**1158**:196-214.

19. Groot PJd, Postma GJ, Melssen WJ, Buydens LMC, Deckert V, Zenobi R. Application of principal component analysis to detect outliers and spectral deviations in near-field surface-enhanced Raman spectra. *Anal Chim Acta*. 2001;**446**:71–83.

20. skuldsson HA. PLS regression methods. *J Chemom*. 1988;**2**:211–28.

21. Wold H. Operative aspects of econometric and sociological models current developments of FP (fix-point) estimation and NIPLS (nonlinear iterative partial least squares) modelling. *Econ Appl*. 1973;**26**:385–421.

22. Wold S, Ruhe A, Wold H, Dunn W. The collinearity problem in linear regression – the partial least squares (PLS) approach to generalized inverses. *SIAM J Sci Stat Comput*. 1984;**5**:735–43.

23. Fonville JM, Richardsa SE, Bartona RH, Boulangea CL, Ebbels TMD, Nicholson JK, *et al.* The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *J Chemometrics*. 2010;**24**:636–49.

24. Svensson O, Kourti T, MacGregor JF. An investigation of orthogonal signal orrection Algorithms and their characteristics. *Chemometrics*. 2002;**16**(4):176-88.

25. Son HS, Kim KM, Berg FVD, Hwang G-S, Park W-M, C-H. L. 1H Nuclear Magnetic Resonance-Based Metabolic Characterization of Wines by Grape Varieties and Production Areas. *J Agric Food Chem*. 2008;**56**:8007–16.

26. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom Intell Lab Syst*. 1987(2):37–52.

27. Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, E.J.J v-V, *et al.* Assessment of PLS-DA cross validation. *Metabolomics*. 2008;**4**:81–9.