Nonnegative Least Squares Approach to Quantification of 1H Nuclear Magnetic Resonance Spectra of Human Urine

Kopriva, Ivica; Jerić, Ivanka; Popović Hadžija, Marijana; Hadžija, Mirko; Vučić Lovrenčić, Marijana

Source / Izvornik: **Analytical Chemistry, 2020, 93, 745 - 751**

Journal article, Published version
Rad u časopisu, Objavljena verzija rada (izdavačev PDF)

https://doi.org/10.1021/acs.analchem.0c02837

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:264:969781

analytical chemistry

ACS AUTHORCHOICE

pubs.acs.org/ac

Article

Non-negative Least Squares Approach to Quantification of ^{1}H Nuclear Magnetic Resonance Spectra of Human Urine

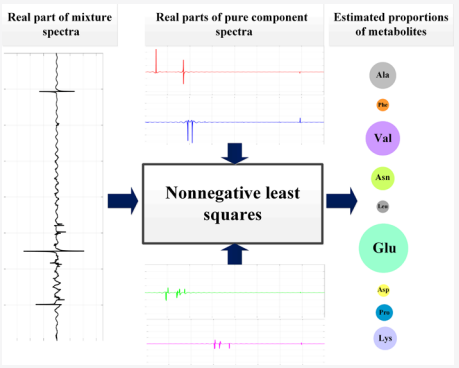
Ivica Kopriva,* Ivanka Jerić, Marijana Popović Hadžija, Mirko Hadžija, and Marijana Vučić Lovrenčić

Cite This: *Anal. Chem.* 2021, 93, 745−751

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Because of its quantitative character and capability for high-throughput screening, ^{1}H nuclear magnetic resonance (NMR) spectroscopy is used extensively in the profiling of biofluids such as urine and blood plasma. However, the narrow frequency bandwidth of ^{1}H NMR spectroscopy leads to a severe overlap of the spectra of components present in the complex mixtures such as biofluids. Therefore, ^{1}H NMR-based metabolomics analysis is focused on targeted studies related to concentrations of the small number of metabolites. Here, we propose a library-based approach to quantify proportions of overlapping metabolites from ^{1}H NMR mixture spectra. The method boils down to the linear non-negative least squares (NNLS) problem, whereas proportions of the pure components contained in the library stand for the unknowns. The method is validated on an estimation of the proportions of (*i*) the 78 pure spectra, presumably related to type 2 diabetes mellitus (T2DM), from their synthetic linear mixture; (*ii*) metabolites present in 62 ^{1}H NMR spectra of urine of subjects with T2DM and 62 ^{1}H NMR spectra of urine of control subjects. In both cases, the in-house library of 210 pure component ^{1}H NMR spectra represented the design matrix in the related NNLS problem. The proposed method pinpoints 63 metabolites that in a statistically significant way discriminate the T2DM group from the control group and 46 metabolites discriminating control from the T2DM group. For several T2DM-discriminative metabolites, we prove their presence by independent analytical determination or by pointing out the corresponding findings in the published literature.



INTRODUCTION

Metabolomics refers to the identifying and quantifying small-molecule components (a.k.a. metabolites or pure components) in complex biological mixtures.[1] Detection and identification of metabolites that discriminate between healthy and disease states are the primary purposes of metabolomics.[2] Due to its quantitative and non-destructive character, time efficiency, and robustness, ^{1}H nuclear magnetic resonance (NMR) spectroscopy is extensively used for metabolic profiling of biofluids such as plasma,[3] urine,[4] or cerebral spinal fluid.[5] However, the narrow frequency bandwidth of ^{1}H NMR spectroscopy causes overlapping of the many resonances of metabolites present in the mixtures. That represents a severe challenge to chemometrics methods used in the identification of metabolites.[6−12] Because the area under the NMR signal amplitude spectra is directly proportional to the concentration, many methods for detection and identification of metabolites are based on fitting NMR signal spectra with a sum of model spectra of metabolites known, or expected, to be present in a mixture.[4,5,7,8,10] Such an approach presumes that amplitude NMR mixture spectra are a sum of amplitude NMR spectra of the pure components present in the mixture. That, however, is correct only when the spectra of the pure components do not overlap. In the case of ^{1}H NMR spectroscopy, that is true only at the selected spectral windows where the small number of metabolites is present. As emphasized in ref 5, that is why the described approach is used for metabolic profiling of the cerebrospinal fluid, and it is not preferred for metabolic profiling of urine or blood. Instead of a single NMR mixture spectra, it is sometimes possible to rely on multiple mixture spectra and use multivariate data analysis methods such as partial least squares,[11] or blind source separation (BSS).[12] However, multivariate methods, such as refs 11 and 12, presume a linear mixture model (LMM) of non-negative amplitude NMR component spectra. That is incorrect for the overlapping spectra.[13] Thus, the BSS methods for separation of the pure components are sensitive to the overlap between their spectra.[13−15] That is expected in the case of metabolic profiling of biological samples where the number of metabolites can reach several hundreds. As an example, there are 458 metabolites identified in urine and 309 in cerebrospinal fluid.[6]

Received: July 3, 2020
Accepted: November 16, 2020
Published: December 7, 2020

https://dx.doi.org/10.1021/acs.analchem.0c02837
Anal. Chem. 2021, 93, 745−751

Herein, we propose the pure component library-based non-negative least squares (NNLS) approach to estimate the metabolites' proportions present in a ^{1}H NMR mixture spectrum. The approach is motivated by the fact that the identification of metabolites requires knowledge about pure components expected to be present in the mixture.[4,5,7−10,12−15] Hence, we have to assume that metabolites present in the mixture are contained in the library. Due to its linear response,[12] ^{1}H NMR mixture signal is a linear combination of ^{1}H NMR signals associated with pure components. In particular, as shown in Section 2, estimation of the proportions boils down, in the absence of noise, to solving the system of linear equations. Due to the presence of noise and non-negativity of proportions, their estimation becomes the NNLS problem.[16] To validate the proposed approach, we built the in-house library composed of 210 ^{1}H NMR pure components spectra. Thereby, approximately 70 of them were expected to be present in the urine of the subjects with the type 2 diabetes mellitus (T2DM) and approximately another 50 of them are urine associated. Table S1 in the Supporting Information provides detailed information about the library content. Due to the nonnegativity of the proportions, only additive combinations of pure components are allowed to model the experimental spectra. That is a strong constraint, and it virtually eliminates the selection of the pure components from the library that are not present in the mixture. The proposed method is validated on the estimation of the proportions of: (*i*) the 78 pure ^{1}H NMR spectra of metabolites, presumed to be related to T2DM, from their synthetic linear mixture contaminated with the additive white Gaussian noise (AWGN) of controlled power; (*ii*) metabolites present in 62 ^{1}H NMR spectra of urine of the subjects with T2DM and in 62 ^{1}H NMR spectra of urine of the control subjects. The proposed method emphasized 63 metabolites in samples from the subjects with T2DM. According to the Student's *t*-test, $p < 0.05$, they had proportions statistically significantly higher from those contained in the samples of the control subjects. The proposed method also emphasized 46 metabolites in samples from control subjects to have proportions statistically significantly higher from those contained in the samples of the subjects with T2DM.

THEORY

Linear Mixture Model of Multicomponent ^{1}H NMR Spectra. Due to its linear response, ^{1}H NMR mixture signal is linear combination of the pure components ^{1}H NMR signals. Thus, the model in the Fourier (chemical shift) domain in the absence of additive noise reads out as

$$\mathbf{x} = \mathbf{S}\mathbf{a} \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^{T\times 1}$ represents one complex ^{1}H NMR mixture signal comprising values at T frequencies and $\mathbb{C}$ stands for the set of complex numbers. $\mathbf{a} \in \mathbb{R}_{0+}^{M\times 1}$ stands for a vector of proportions of the M pure components contained in the library, and $\mathbb{R}_{0+}$ denotes the set of non-negative real numbers. $\mathbf{S} \in \mathbb{C}^{T\times M} =: \{\mathbf{S}_m \in \mathbb{C}^{T\times 1}\}_{m=1}^{M}$ is a library with the columns representing ^{1}H NMR signals of the metabolites in the Fourier domain, where "=:" means "by definition." We name non-negative elements of the mixing vector $\mathbf{a}$ in 1 proportions as opposed to concentrations in ref 12. As it is discussed in refs 8 and 10, concentrations of the pure components are directly proportional to their amplitude spectra. However, LMM 1 does not hold in the amplitude domain when the pure components' spectra overlap.[13] We assume that $\mathbf{S}_I \subseteq \mathbf{S}$, where $I \subseteq \{1, ..., M\}$ is a set of indexes of metabolites present in the mixture $\mathbf{x}$. Using the real parts of $\mathbf{x}$ and $\mathbf{S}$ in 1, we obtain the real system of linear equations where the elements of $\mathbf{a}$ stand for the unknown proportions

$$\mathrm{Re}(\mathbf{x}) = \mathrm{Re}(\mathbf{S})\mathbf{a} \tag{2}$$

$\mathrm{Re}(\mathbf{x}) \in \mathbb{R}^{T\times 1}$ represents the real part of ^{1}H NMR mixture signal, and $\mathrm{Re}(\mathbf{S}) \in \mathbb{R}^{T\times M}$ represents real part of the library. The system of linear eq 2 is exact, that is no approximations were made. Thus, when the condition 3 is fulfilled

$$\mathrm{rank}(\mathrm{Re}(\mathbf{S})) = M \tag{3}$$

the solution of 2 is unique and it is obtained as: $\mathbf{a} = [\mathrm{Re}(\mathbf{S})]^{\dagger}\mathbf{x}$, where $[\mathrm{Re}(\mathbf{S})]^{\dagger}$ stands for the pseudoinverse of $\mathrm{Re}(\mathbf{S})$. Model 2 holds when pure components and mixtures are acquired under the same conditions. Otherwise, peak shifts will occur and the model is not valid. Since in the reported experiments, the urine samples and pure components were prepared and acquired following the same protocol, described in detail in the Experimental Section, the model 2 was valid. Thus, there was no need for the peak alignment steps prior to the analysis. Provided that the library is built taking into account information related to the human metabolome,[17] for example urine-related metabolites identified by the human metabolome project,[18] it is realistic in untargeted metabolic studies to at least approximately satisfy condition 3 (see Supporting Information, S11−S13, for a more detailed discussion).

Proportion Estimation and the NNLS Problem. Experimental recordings of the ^{1}H NMR mixture spectra include additive noise. Thus, 2 becomes

$$\mathrm{Re}(\mathbf{x}) = \mathrm{Re}(\mathbf{S})\mathbf{a} + \mathbf{n} \tag{4}$$

where $\mathbf{n} \sim N(0,\sigma^2)$ denotes the zero mean AWGN. Hence, the solution of 4 is approximate. It is obtained by solving the NNLS problem

$$\hat{\mathbf{a}} = \min_{a\geq 0} \frac{1}{2} \|\mathrm{Re}(\mathbf{x}) - \mathrm{Re}(\mathbf{S})\mathbf{a}\|_2^2 \tag{5}$$

where $\hat{\mathbf{a}}$ stands for the estimate of $\mathbf{a}$ and $\|\circ\|_2^2$ denotes the square of the l_2 norm. Problem 5 is a convex quadratic programming problem and has a globally optimal solution. When the number of metabolites present in the mixture, $\#I$, is expected/known to be smaller than the library size M, it is justified to further impose the sparsity constraint on $\mathbf{a}$. By using the l_1-norm as a measure of sparsity, we obtain the l_1 -regularized NNLS problem[19]

$$\hat{\mathbf{a}} = \min_{a\geq 0} \frac{1}{2} \|\mathrm{Re}(\mathbf{x}) - \mathrm{Re}(\mathbf{S})\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1 \tag{6}$$

where $\lambda \geq 0$ is the regularization parameter. Instead of the l_1 -norm, the l_0-quasi-norm of $\mathbf{a}$ can be used as a measure of sparsity, that is $\|\mathbf{a}\|_0 = \#\{a_m \neq 0, m = 1,..., M\}$. The fast non-negative orthogonal matching pursuit (FNNOMP) algorithm,[21] with the Matlab code available at ref 22, solves the l_0-regularized optimization problem

$$\hat{\mathbf{a}} = \min_{a_I\geq 0} \frac{1}{2} \|\mathrm{Re}(\mathbf{x}) - \mathrm{Re}(\mathbf{S}_I)\mathbf{a}_I\|_2^2 \tag{7}$$

https://dx.doi.org/10.1021/acs.analchem.0c02837
Anal. Chem. 2021, 93, 745−751

where I stands for the support of $\mathbf{a}^*$ assumed to be known a priori. Proportions in percentage are obtained by scaling solution of 5−7 by its l_1-norm

$$\hat{\mathbf{a}} \leftarrow \frac{\hat{\mathbf{a}}}{\|\hat{\mathbf{a}}\|_1} \tag{8}$$

As it is pointed out in ref 20, the nonnegativity constraint on $\mathbf{a}$ in 5 is similarly effective as combined nonnegativity constraint and explicit l_1 regularization on $\mathbf{a}$ in 6. Regarding robustness to noise, the most influential factors are the correlations of the columns of Re($\mathbf{S}$) and the amount of regularization.

EXPERIMENTAL SECTION

We recorded the in-house library comprising ^{1}H NMR spectra of 210 pure components expected to correspond with metabolites. Thereby, around 120 pure components are related to urine, whereas around 70 of them to the urine of patients with T2DM. The library content is presented in Table S1 of the Supporting Information. The library comprising the first 160 pure components was described previously in ref 9. The correlation structure of the library is discussed in the Results section. The NNLS methods for proportions estimation were applied to ^{1}H NMR spectra of urine obtained from 62 diabetic patients and to ^{1}H NMR spectra of urine collected from 62 healthy controls. All of the experiments were executed on a PC running under a 64-bits Windows 10 operating system with 256 GB of RAM using Intel Xeon CPU E5-2650 v4 2 processors and operating with a clock speed of 2.2 GHz. For the reason of comprehensiveness, we briefly describe the protocol related to the recording of the ^{1}H NMR spectra of the pure components, collection of urine samples, their preparation, and acquisition of their ^{1}H NMR spectra. The more detailed description is presented in the Supporting Information (Pages S34−S36).

Recording of ^{1}H NMR Spectra of 210 Pure Components. All measurements were performed on a Bruker AVANCE 600 MHz spectrometer, operating at 298 K. 5−10 mg of pure components were used for generating the library. Samples were dissolved in 700 μL of phosphate buffer (100 mM, pH 7.2 prepared with D_2O) prior to NMR measurements. 3-(Trimethylsilyl)-1-propane sulfonic acid sodium salt was used as an internal standard. We tested two water suppression methods, WATERGATE and Water Suppression by Excitation Sculpting[23] on metabolites from different classes (amino acids, carbohydrates, and nucleic acids) and also urine samples. Revealed analysis of the obtained spectra brought us to the selected Suppression by Excitation Sculpting method.

Urine Sample Collection, Preparation, and ^{1}H NMR Spectroscopy Measurements. Urine aliquots were obtained from the residual routine samples from 62 unrelated patients with T2DM (age range: 30−84 years). They were collected in the morning, during the regular outpatient checkup in the clinical laboratory affiliated to the tertiary-level diabetes clinic. Patients were categorized and treated according to the current World Health Organization (WHO) recommendations at the University Clinic Vuk Vrhovac, Zagreb. The institutional Ethics Committee approved the study protocol, and patients gave their written consent for the usage of their residual samples. The group of control subjects included 62 healthy, unrelated consenting adult volunteers, matched for age and sex to diabetic subjects. For each of them, the glucose level was measured before taking urine, and they were all normoglycemic. All study subjects were Caucasians. Morning urine samples were stored at −200 °C until the clean-up procedure, that is performed by C^{18} SampliQ Solid Phase Extraction (Agilent Technologies, USA). C^{18} polymer sorbents were first conditioned by passing MeOH (3 × 5 mL) and then equilibrated by passing QH_2O (3 × 5 mL). Each urine sample (3 × 5 mL) was loaded into the column, and a fraction was collected after cleaning in separate tubes. All of the steps were performed at a flow rate of 1 mL min^{-1}. After that, samples were frozen by immersion in liquid nitrogen followed by evaporation in the vacuum chamber of a freeze dryer to dryness (under controlled temperature and reduced pressure). 10 mg of each dry sample was further used for spectroscopic analysis. The NMR urine spectra were recorded, as described in the previous section, related to the recording of pure component spectra.

RESULTS

NNLS-Based Estimation of Proportions from a Synthetic ^{1}H NMR Mixture Spectrum. To select the most suitable solver(s) for the problem at hand, we designed the synthetic mixture such as 4. Thereby, the library $\mathbf{S}$ comprised 210 ^{1}H NMR spectra of pure components with the content presented in Table S1. Proportions of 78 pure components, that according to Table S1 were expected to be T2DM relevant, were generated randomly according to the uniform distribution on the (0, 1] interval. The number of pairs of the subset of 78 spectra with the normalized correlation coefficient (NCC) greater than or equal to: 0.2 was 82, 0.3 was 38, 0.5 was 12, and 0.7 was 2. The numbers of pairs with the corresponding values of the NCC for the whole library were 381, 123, 29, and 5. Thus, many pure components in the library were structurally similar, with the overlapping spectra. In accordance with 4, the AWGN was generated with the signal-to-noise (SNR) ratio in dB: SNR ∈ {0, 10, 20, 30, 40, 50, 60, "inf"}, where "inf" stands for no AWGN. We have tested 11 solvers for the NNLS problem 4. They are elaborated in the Supporting Information (Pages S13−S17). Matlab code for most of these methods was downloaded from ref 24. For each SNR value, we generated 100 realizations and estimated the following figure of merits for each solver of the NNLS problems 5−7: sensitivity, specificity, balanced accuracy, F_1 score, positive predicted value (PPV), relative proportions error, and the total fit of the spectrum. Mean values (±standard deviations) of the logarithm of the relative proportion errors and percentage of total fit of synthetic spectrum and residuals are shown in Figures S1 to S3 in the Supporting Information. Mean values (±standard deviations) of the estimates of balanced accuracy, sensitivity, specificity, F_1 score, and PPV are shown in Figures S4 to S8 in the Supporting Information. It can be seen that for SNR ≥ 10 dB, the Lawson Hanson (LH) method,[16] the positive modification of the LARS (PLARS) algorithm,[19,20] the projected quasi-Newton (PQN) algorithm,[25] and the NNPINV + FNNOMP yield consistent and accurate results by achieving less than 10% of relative error and total fit greater than 95%. Figure 1 shows the total fit for four mentioned NNLS methods as a function of SNR. As shown in Figures 2 and S3, for SNR = 10 dB residuals between the real part of clean synthetic spectrum and approximations based on LH, PLARS, PQN, and NNPINV + FNNOMP methods are 50 times smaller in the amplitude range than that of the spectrum itself. As shown in Table S2,

https://dx.doi.org/10.1021/acs.analchem.0c02837
Anal. Chem. 2021, 93, 745−751

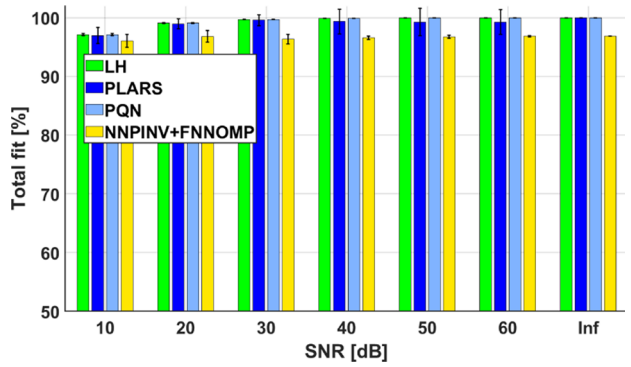


Figure 1. Mean values of the percentage of the total fit of the synthetic spectrum (±standard deviations) vs the SNR values.

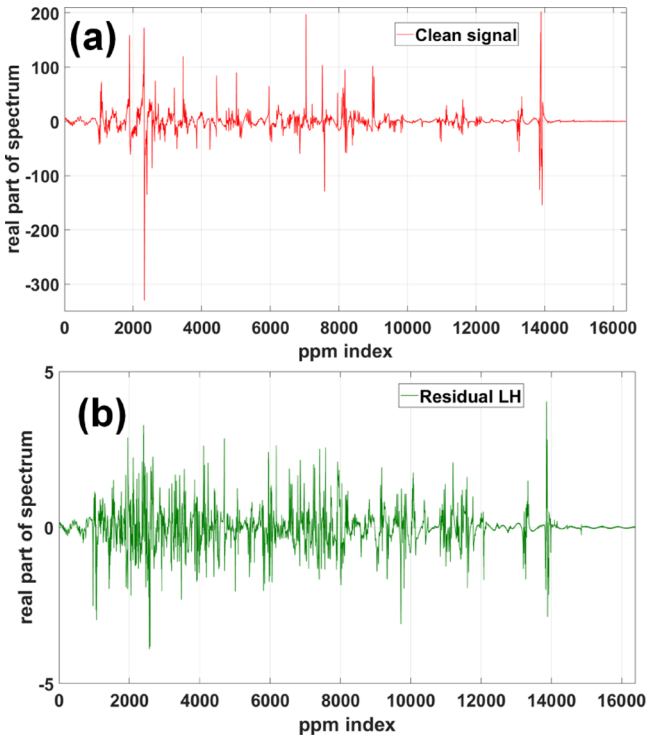


Figure 2. (a) Real part of the clean synthetic spectrum. (b) Residual between the clean and fitted real part of synthetic spectra using the LH NNLS algorithm for SNR = 10 dB.

the LH, the PQN, and the PLARS algorithms are also computationally efficient (0.76, 0.83, and 0.91 s on the described synthetic problem). Thus, the LH, the PQN, the PLARS, and the NNPINV + FNNOMP algorithms were selected for metabolic profiling of the ^{1}H NMR spectra of urine of the subjects with T2DM as well as of the control subjects.

NNLS-Based Estimation of Proportions from a ^{1}H NMR Spectrum of Urine of T2DM and Control Subjects. In accordance with the problems 5−7, we applied the LH, the PLARS, the PQN, and the NNPINV + FNNOMP algorithms to estimate proportions from 62 ^{1}H NMR spectra of urine of patients with T2DM as well as from 62 ^{1}H NMR spectra of urine of healthy controls. Afterward, estimated vectors of proportions were scaled in accordance with 8. Scaled vectors of proportions were stored column-wise in non-negative matrices with dimensions 210 × 62: $\hat{\mathbf{A}}^d_{\mathrm{LH}}$, $\hat{\mathbf{A}}^c_{\mathrm{LH}}$, $\hat{\mathbf{A}}^d_{\mathrm{PLARS}}$, $\hat{\mathbf{A}}^c_{\mathrm{PLARS}}$, $\hat{\mathbf{A}}^d_{\mathrm{PQN}}$, $\hat{\mathbf{A}}^c_{\mathrm{PQN}}\hat{\mathbf{A}}^d_{\mathrm{NNPINV\ +\ FNNOMP}}$, and $\hat{\mathbf{A}}^c_{\mathrm{NNPINV\ +\ FNNOMP}}$. From diabetes-related matrices, we selected components with the mean proportion greater than the mean proportion in corresponding control-related matrices. Thus, the index sets (I^d_{LH}, I^d_{PLARS}, I^d_{PQN}, and $I^d_{\mathrm{NNPINV\ +\ FNNOMP}}$) of these components are obtained according to

$$I^d_{\mathrm{method}} = \{m = 1, \ldots, 210:\ \mathrm{mean}(\hat{A}^d_{\mathrm{method}}(m, :)) > \mathrm{mean}(\hat{A}^c_{method}(m, :))\}$$
$$\mathrm{method} \in \{\mathrm{LH, PLARS, PQN, NNPINV + FNNOMP}\} \quad (9)$$

From control-related matrices, we selected components with the mean proportion greater than the mean proportion in corresponding T2DM-related matrices. The index sets (I^c_{LH}, I^c_{PLARS}, I^c_{PQN}, and $I^c_{\mathrm{NNPINV\ +\ FNNOMP}}$) of these components were obtained analogously to 9. The Matlab notation and indexing scheme have been assumed in 9. We are interested in finding out whether pure components indexed by sets I^d_{methods} discriminate the T2DM group from the control group in a statistically significant way. Likewise, we are interested in finding out whether pure components indexed by sets I^c_{methods} discriminate the control group against the T2DM group in a statistically significant way. Thus, we applied the two-sample Student's *t*-test, implemented with the Matlab function ttest2. The function returns a decision for the null hypothesis that proportions of the selected pure components from the T2DM and control groups come from normal distributions with equal means and unequal and unknown variances. Pure components discriminate T2DM against the control group, and vice versa, in a statistically significant way if the test rejects the null hypothesis at the 5% significance level. Afterward, we performed Benjamini−Hochberg correction, ref 26, of obtained *p*-values with the false discovery rate (FDR) set to 0.05. The overall number of the null hypothesis equals to the cardinalities of the sets I^c_{LH}, I^c_{PLARS}, I^c_{PQN}, and $I^c_{\mathrm{NNPINV\ +\ FNNOMP}}$ and sets I^d_{LH}, I^d_{PLARS}, I^d_{PQN}, and $I^d_{\mathrm{NNPINV\ +\ FNNOMP}}$. Tables S3 to S6 in the Supporting Information list the pure components discriminative of the T2DM and control groups. Thereby, proportions were respectively estimated with the LH, the PLARS, the PQN, and the NNPINV + FNNOMP methods. As can be seen, results obtained by the four methods are very consistent. The final result, presented in Tables 1 and S7, is obtained as the intersection of the corresponding results in Tables S3 to S6. Herein, the number of the null hypothesis used in the Benjamini−Hochberg correction equals the cardinalities of the intersection sets. The T2DM-discriminative group shown in Table 1 contains 63 (out of 210) pure components. The control-discriminative group shown in Table S7 contains 46 (out of 210) pure components.

Figure 3 shows the mean value (±standard deviation) of the total fit of urine spectra for the T2DM group and the control group for four selected NNLS methods. The total fits are highly consistent and approximately amount to 12.44% ± 5.37% for the diabetic group and 8.40% ± 3.2% for the control group. As described in Table S1, our in-house library contains 210 pure components, where 78 of them are expected to be T2DM relevant. As shown in Tables 1 and S7, the four NNLS methods agreed on 63 components from the library to be T2DM-discriminative and on 46 components to be control-discriminative. Given that the human metabolome project identified 458 metabolites in the urine,[18] 63 T2DM-discriminative components contribute to 13.76% of the overall

https://dx.doi.org/10.1021/acs.analchem.0c02837
Anal. Chem. 2021, 93, 745−751

Table 1. List of Pure Components (Metabolites) That in a Statistically Significant Way Discriminate T2DM from Control Group[a]

pure component	*p*-value	proportion [%] mean ± std	pure component	*p*-value	proportion [%] mean ± std
2′-deoxyuridine	2.19×10^{-9}	1.34 ± 1.26	glycogen	2.26×10^{-4}	2.14 ± 1.44
glucose-1-phosphate	9.61×10^{-9}	1.55 ± 1.54	D-mannose[27,28]	2.51×10^{-4}	0.34 ± 0.53
2-furoic acid	4.77×10^{-8}	0.14 ± 0.17	1,3-dihydroxyacetone	2.60×10^{-4}	0.21 ± 0.17
2,5-diaminopimelic acid	5.55×10^{-8}	0.99 ± 0.86	imidazole	3.45×10^{-4}	0.17 ± 0.23
3-hydroxy-5-methohymandelic acid	1.21×10^{-7}	0.32 ± 0.25	1,3-dihydroxyacetone	2.60×10^{-4}	0.21 ± 0.17
Z-prolyl-phenylalanine	1.75×10^{-7}	1.35 ± 1.62	creatine[40]	4.16×10^{-4}	1.10 ± 1.46
quercetin	7.76×10^{-7}	0.24 ± 0.22	beta alanine[37]	4.77×10^{-4}	0.35 ± 0.48
asparagine[37]	1.28×10^{-6}	0.53 ± 0.72	cinnamaldehyde	5.89×10^{-4}	0.61 ± 0.83
4-methyl-2-oxovalenic-acid	2.23×10^{-6}	0.08 ± 0.10	D-arabinose [5]	6.75×10^{-4}	0.28 ± 0.38
ethanol	2.31×10^{-6}	0.23 ± 0.32	Z-valyl-glycyl-glycine	7.38×10^{-4}	0.30 ± 0.49
thiamine*	2.51×10^{-6}	0.03 ± 0.03	thymol	8.29×10^{-4}	0.03 ± 0.031
3-methylsalicyclic acid	5.63×10^{-6}	0.15 ± 0.10	formic acid	8.45×10^{-4}	0.64 ± 0.77
2′-deoxyinosine	7.11×10^{-6}	1.04 ± 1.14	glutamine[37]	8.90×10^{-4}	0.64 ± 0.67
4-aminobutyric acid (GABA)[28]	7.28×10^{-6}	0.42 ± 0.62	biotin	1.46×10^{-3}	0.39 ± 0.50
alloxan	1.26×10^{-5}	0.11 ± 0.12	*β* fructose	1.66×10^{-3}	0.43 ± 0.74
Z-Ser-OH	1.31×10^{-5}	1.23 ± 1.37	D-glucose[27,28]	1.66×10^{-3}	0.29 ± 0.85
hippuric acid[27,28] *	1.88×10^{-5}	0.07 ± 0.14	2-oxogularic acid	2.42×10^{-3}	0.16 ± 0.38
vanillylmandelic acid	2.33×10^{-5}	0.52 ± 0.29	3-methyladipic acid	2.70×10^{-3}	0.178 ± 0.24
acetophenon	2.82×10^{-5}	1.38 ± 1.89	carnitine	3.36×10^{-3}	0.15 ± 0.20
cholic acid	2.96×10^{-5}	0.71 ± 0.67	naphthoic acid	3.88×10^{-3}	0.94 ± 1.12
Leu−Trp	5.75×10^{-5}	0.761 ± 0.71	3-methylxanthine	4.15×10^{-3}	0.53 ± 0.95
glycylglycine[35]	6.84×10^{-5}	0.11 ± 0.22	3-aminoisobutyric acid	4.41×10^{-3}	1.37 ± 2.20
ethionine	7.99×10^{-5}	0.368 ± 0.45	*S*-methyl-L-cysteine	4.99×10^{-3}	0.12 ± 0.20
betaine[27,28]	8.15×10^{-5}	0.54 ± 0.80	threonine	6.37×10^{-3}	0.327 ± 0.35
3-phenylpropionic acid	8.28×10^{-5}	0.38 ± 0.69	D-glucosamine	6.46×10^{-3}	1.07 ± 1.74
2,5-dihydroxybenzoic acid*	8.49×10^{-5}	0.67 ± 0.70	urea	1.02×10^{-2}	1.16 ± 1.78
maltose	1.08×10^{-4}	0.88 ± 1.59	aspartic acid[37]	1.21×10^{-2}	0.61 ± 0.98
Uric acid[37]	1.14×10^{-4}	0.56 ± 0.54	2,3,4,6-tetramethyl-D-glucose	1.45×10^{-2}	0.87 ± 0.96
allantoin[27,28]	1.20×10^{-4}	0.03 ± 0.06	ribose	1.70×10^{-2}	0.30 ± 0.50
2-ketobutyric acid	1.51×10^{-4}	0.24 ± 0.25	4-acetamiclophenol	1.84×10^{-2}	0.21 ± 0.33
adip acid	1.57×10^{-4}	0.12 ± 0.16	*p*-hydrophenylpyric acid	1.95×10^{-2}	0.29 ± 0.32
caprylic acid	2.21×10^{-4}	0.45 ± 0.69	alpha-methylserine	2.05×10^{-2}	0.13 ± 0.15

[a]Reported *p*-values and proportions are based on estimations obtained by the LH algorithm. *p*-values were corrected according to the Benjamini−Hochberg test with FDR = 0.05. [xx] indicates reference confirming corresponding metabolite. [*] indicates confirmation by HPLC-MS based independent analytical determination.

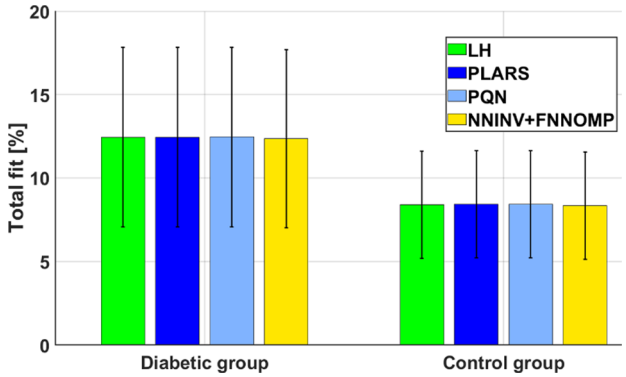


Figure 3. Total fit of urine spectra for the T2DM group and control group for four selected NNLS algorithms.

number of urine relevant metabolites. Likewise, 46 control-discriminative components contribute to 10.04% of the overall number of urine relevant metabolites. Thus, both numbers are within the corresponding ranges of the estimated total fits of the urine spectra. In other words, the estimated total fits explain the urine spectra in the amounts possible with the current version of our in-house library. For non-targeted metabolomics analysis of urine, a library comprising 458 metabolites identified in the human metabolome project has to be built.

DISCUSSION

Our in-house library-based NNLS approach to non-targeted metabolic profiling led us to the list of 68 components that in a statistically significant way discriminate T2DM against control group (see Table 1). For several T2DM-discriminative metabolites, we proved their presence by independent analytical determination or by pointing out relations with the corresponding findings in the published literature. In ref 27, ASICS analysis of ^{1}H NMR spectra of human urine of 50 patients with the T2DM and 84 healthy volunteers confirmed results from Table 1 for allantoin, betaine, D-glucose, D-mannose, 4-aminobutyric acid (GABA), and hippuric acid. Furthermore, in ref 28, BATMAN and BAYESIL analysis of ^{1}H NMR spectra of urine of rats with the T2DM confirmed results from Table 1 for allantoin, betaine, choline, D-glucose, D-mannose, and urea. To this end, HPLC-MS analysis of urine samples of patients with T2DM confirmed the presence of the metabolites: thiamine, 2,5-dihydroxybenzoic acid (DHBA), and hippuric acid (see Pages S35−S42 in the Supporting

https://dx.doi.org/10.1021/acs.analchem.0c02837
Anal. Chem. 2021, 93, 745−751

Information for more details). An important approach to diabetes treatment involves the regulation of postprandial hyperglycemia by delaying glucose release into the bloodstream using inhibitors for carbohydrate digesting enzymes such as maltase.[29] Therefore, this could explain the high proportion of maltose found in the urine of diabetic patients involved in our study.[30] Glycogen serves as energy storage in living organisms. It is cleaved by the enzyme glycogen phosphorylase responsible for the production of glucose-1-phosphate. Because of its polarity, glucose-1-phosphate cannot cross cell membranes and must be involved in catabolism within cells.[31] Since glycogenolysis occurs in diabetes, its product glucose-1-phosphate is expected to be associated with diabetes in our study. DHBA is one of the tyrosine metabolism products detected in our study with a significantly higher proportion in the diabetic group. It has been shown that DHBA inhibits low-density lipoprotein oxidation in hyperglycemic conditions describing its action as a free-radical scavenger.[32] By the method presented here, the remarkable portion of phenylalanine-pathway metabolites (phenylalanine, Z-prolyl-phenylalanine, hippuric acid, glycylglycine, and *p*-hydrophenylpyric acid) is detected in urine of T2DM patients. It follows the fact that the products of this metabolic pathway can be excreted by the urine, indicating dysfunction in this metabolic pathway and predicting both diabetes risk and chronic kidney failure.[33] Metabolites involved in the phenylalanine pathway might also be mammalian-microbial co-metabolites, such as is 4-hydroxyphenylpyruvic acid transformed by intestinal microorganisms.[34] An HPLC-MS study of 188 individuals with T2DM and 181 healthy controls confirmed a statistically significant presence of glycylglycine in urine samples of a T2DM group.[35] In our study, several metabolites of purine and pyrimidine pathways (3-methylxanthine, 2-deoxyuridine, and so forth) were detected to differentiate the T2DM group from the control group significantly. Purine catabolism is an essential component of the homeostatic response of mitochondria to oxidative stress. It has been proved to be altered in the liver mitochondria of diabetic rats.[36] Glutamine, uric acid, and asparagine are additional metabolites of purine/pyrimidine pathway pointed as markers for discriminating the T2DM group in our study. The study very relevant to our results has been reported in ref 37, where the amino acid concentrations were measured in blood and urine collected from 100 patients with diabetes and compared with the 100 healthy subjects. Urinary amino acids with statistically significantly higher concentrations in the diabetic group[37] that coincide with the acids reported in our study (Table 1) were asparagine, aspartic acid, glutamine, and beta-alanine. In recent years, many experimental and clinical data have accumulated the effects of the flavonoid quercetin on the treatment of diabetes[38] since humans can absorb significant amounts of quercetin from food or supplements.[39] The significantly higher level of urinary creatine in patients with T2DM was detected here by NNLS, reported in our previous publication,[9] and in ref 40.

CONCLUSIONS

Due to its quantitative character and time efficiency, ^{1}H NMR spectroscopy is used to investigate metabolic profiling of biofluids. However, the narrow frequency bandwidth of ^{1}H NMR spectroscopy leads to the severe overlap of the metabolites' spectra in the complex mixtures such as urine or blood plasma. Nevertheless, non-targeted metabolic profiling of structurally similar metabolites from a complex mixture is of potentially high clinical relevance. Driven by this motivation, this paper presented a method for estimating the proportions of the metabolites present in the ^{1}H NMR spectrum of a complex mixture. The method relies on a library of pure component ^{1}H NMR spectra and boils down to the NNLS problem. As opposed to approaches that estimate concentrations from the amplitude spectra, the proposed method is, in principle, insensitive to the overlapping of the spectra of the pure components. In addition to the synthetic mixture, the method was tested on metabolic profiling of 62 urine samples collected from the subjects with T2DM as well as from 62 urine samples collected from the healthy controls. Thereby, the in-house built library comprising 210 pure component ^{1}H NMR spectra was used as a design matrix in the related NNLS problem. The proposed method emphasized 63 metabolites in samples from the subjects with T2DM to have proportions statistically significantly higher, $p < 0.05$ according to the Student's *t*-test, than those in the samples from the control subjects. The proposed method also emphasized 46 pure components in samples from control subjects to have proportions statistically significantly higher than those in the samples from the subjects with T2DM. In both cases, discriminative components were discovered consistently by the four NNLS solvers. For the many of the prominent metabolites in the urine of patients with T2DM, we proved their presence by independent analytical determination, discussed their metabolic interpretation, or pointed out the corresponding findings in the published literature.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c02837.

Content of the in-house library, discussion of the uniqueness of conditions of proportion estimation, NNLS formulation of proportion estimation from the synthetic mixture, results related to the estimation of proportions of metabolites from the synthetic mixture, results related to the estimation of proportions of metabolites from urine spectra, experiments and materials, and independent analytical determination of T2DM-discriminative metabolites using HPLC-MS analysis of urine samples (PDF)

AUTHOR INFORMATION

Corresponding Author

Ivica Kopriva − *Division of Electronics, Ruđer Bošković Institute, HR-10000 Zagreb, Croatia;* orcid.org/0000-0002-8610-8877; Phone: +385-1-4571-286; Email: ikopriva@irb.hr; Fax: +385-1-4680-104

Authors

Ivanka Jerić − *Division of Organic Chemistry and Biochemistry, Ruđer Bošković Institute, HR-10000 Zagreb, Croatia;* orcid.org/0000-0001-9245-3530

Marijana Popović Hadžija − *Division of Molecular Medicine, Ruđer Bošković Institute, HR-10000 Zagreb, Croatia*

Mirko Hadžija − *Division of Molecular Medicine, Ruđer Bošković Institute, HR-10000 Zagreb, Croatia*

Marijana Vučić Lovrenčić − *Department of Medical Biochemistry and Laboratory Medicine, University Hospital Merkur, HR-10000 Zagreb, Croatia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c02837

Author Contributions

I.K. developed models for NNLS approach to the estimation of proportions of metabolites in ^{1}H NMR urine spectra, implemented NNLS algorithms in MATLAB, and conducted validation. I.K. also wrote the initial draft of the paper. I.J. prepared the in-house library comprising 210 ^{1}H NMR spectra of pure components, synthesized laboratory mixtures for method validation, coordinated acquisition of ^{1}H NMR spectra of laboratory mixtures and urine samples of patients with T2DM and of control subjects, and performed HPLC-MS based independent analytical confirmation of several T2DM-discriminative metabolites. M.P.H. and M.H. coordinated the collection of urine samples of control subjects and worked out the preparation of urine samples for ^{1}H NMR spectra acquisition. M.V.L. coordinated the collection of urine samples of patients with T2DM. M.V.L. and M.P.H. provided a metabolic interpretation of pure components found to be expressed in ^{1}H NMR spectra of urine of 62 patients with T2DM. All of the co-authors read the paper and contributed to the paragraphs that correspond with their areas of expertise.

Funding

The work performed has been supported through grant IP-2016-06−5235 "Structured decompositions of empirical data for computationally assisted diagnosis of disease" funded by the Croatian Science Foundation.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Nicholson, J. K.; Lindon, J. C. *Nature* **2008**, *455*, 1054−1056.
(2) Robinette, S. L.; Brüschweiler, R.; Schroeder, F. C.; Edison, A. S. *Acc. Chem. Res.* **2012**, *45*, 288−297.
(3) Schicho, R.; Shaykhutdinov, R.; Ngo, J.; Nazyrova, A.; Schneider, C.; Panaccione, R.; Kaplan, G. G.; Vogel, H. J.; Storr, M. *J. Proteome Res.* **2012**, *11*, 3344−3357.
(4) Shaykhutdinov, R. A.; MacInnis, G. D.; Dowlatabadi, R.; Weljie, A. M.; Vogel, H. J. *Metabolomics* **2009**, *5*, 307−317.
(5) Jukarainen, N. M.; Korhonen, S.-P.; Laakso, M. P.; Korolainen, M. A.; Niemitz, M.; Soininen, P. P.; Tuppurainen, K.; Vepsäläinen, J.; Pirttilä, T.; Laatikainen, R. *Metabolomics* **2008**, *4*, 150−160.
(6) Emwas, A.-H. M.; Salek, R. M.; Griffin, J. L.; Merzaban, J. *Metabolomics* **2013**, *9*, 1048−1072.
(7) Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M. *Anal. Chem.* **2006**, *78*, 4430−4442.
(8) Soininen, P.; Haarala, J.; Vepsäläinen, J.; Niemitz, M.; Laatikainen, R. *Anal. Chim. Acta* **2005**, *542*, 178−185.
(9) Kopriva, I.; Jerić, I.; Hadžija, M. P.; Hadžija, M.; Lovrenčić, M. V.; Brkljačić, L. *Anal. Chim. Acta* **2019**, *1080*, 55−65.
(10) Merkx, D. W. H.; Westphal, Y.; van Velzen, E. J. J.; Thakoer, K. V.; de Roo, N.; van Duynhoven, J. P. M. *Carbohydr. Polym.* **2018**, *179*, 379−385.
(11) Allen, G. I.; Peterson, C.; Vanucci, M.; Maletić-Savatić, M. *Stat. Anal. Data Min.* **2013**, *6*, 302−314.
(12) Cherni, A.; Piersanti, E.; Anthoine, S.; Chaux, C.; Shintu, L.; Yemloul, M.; Torrésani, B. *Faraday Discuss.* **2019**, *218*, 459−480.
(13) Kopriva, I.; Jerić, I. *Chemom. Intell. Lab. Syst.* **2014**, *137*, 47−56.
(14) Kopriva, I.; Jerić, I. *Anal. Chem.* **2010**, *82*, 1911−1920.
(15) Kopriva, I.; Jerić, I.; Smrečki, V. *Anal. Chim. Acta* **2009**, *653*, 143−153.
(16) Lawson, R.; Hanson, C. *Solving Least Squares Problems*; SIAM: Philadelphia, US, 1995.
(17) Duarte, N. C.; Becker, S. A.; Jamshidi, N.; Thiele, I.; Mo, M. L.; Vo, T. D.; Srivas, R.; Palsson, B. O. *Proc. Acad. Nat. Sci.* **2007**, *104*, 1777−1782.
(18) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; et al. *Nucleic Acids Res.* **2007**, *35*, D521−D526.
(19) Tibshirani, R. *J. Roy. Stat. Soc. B* **1996**, *58*, 267−288.
(20) Slawski, M.; Hein, M. *El J. Stat.* **2013**, *7*, 3004−3056.
(21) Yaghoobi, M.; Wu, D.; Davies, M. E. *IEEE Signal Process. Lett.* **2015**, *22*, 1229−1233.
(22) http://www.mehrdadya.com/.
(23) Hwang, T. L.; Shaka, A. J. *J. Magn. Reson., Ser. A* **1995**, *112*, 275−279.
(24) https://sites.google.com/site/slawskimartin/code (accessed on Sept 15, 2021.web
(25) Kim, D.; Sra, S.; Dhillon, I. S. *SIAM J. Sci. Comput.* **2010**, *32*, 3548−3563.
(26) Benjamini, Y.; Hochberg, Y. *J. Roy. Stat. Soc. B* **1995**, *57*, 289−300.
(27) Lefort, G.; Liaubet, L.; Canlet, C.; Tardivel, P.; Père, M.-C.; Quesnel, H.; Paris, A.; Iannuccelli, N.; Vialaneix, N.; Servien, R. *Bioinformatics* **2019**, *35*, 4356−4363.
(28) Maulidiani; Rudianto; Mediani, A.; Kathib, A.; Ismail, A.; Hamid, M.; Lajis, N. H.; Shaari, K.; Abas, F. *Metabolomics* **2017**, *13*, 131.
(29) Kai, M.; Kamada, T.; Baba, Y.; Shitomoto, M.; Setoyama, S.; Otsuji, S. *Clin. Chim. Acta* **1980**, *108*, 259−266.
(30) Sone, H.; Shimano, H.; Ebinuma, H.; Takahashi, A.; Yano, Y.; Iida, K. T.; Suzuki, H.; Toyoshima, H.; Kawakami, Y.; Okuda, Y.; et al. *Metabolism* **2003**, *52*, 1019−1027.
(31) Sullivan, M. A.; Forbes, J. M. *EBioMedicine* **2019**, *47*, 590−597.
(32) Exner, M.; Hermann, M.; Hofbauer, R.; Kapiotis, S.; Speiser, W.; Held, I.; Seelos, C.; Gmeiner, B. M. K. *FEBS Lett.* **2000**, *470*, 47−50.
(33) Friedrich, N. *J. Endocrinol.* **2012**, *215*, 29−42.
(34) Bohus, E.; Coen, M.; Keun, H. C.; Ebbels, T. M. D.; Beckonert, O.; Lindon, J. C.; Holmes, E.; Noszál, B.; Nicholson, J. K. *J. Proteome Res.* **2008**, *7*, 4435−4445.
(35) Yousri, N. A.; Mook-Kanamori, D. O.; El-Din Selim, M. M.; Takiddin, A. H.; Al-Homsi, H.; Al-Mahmoud, K. A. S.; Karoly, E. D.; Krumsiek, J.; Do, K. T.; Neumaier, U.; et al. *Diabetologia* **2015**, *58*, 1855−1867.
(36) Mook-Kanamori, L.; Pi, Z.; Zhou, Y.; Liu, Y.; Wei, M.; Song, F.; Liu, Z. *RSC Adv.* **2017**, *7*, 16494.
(37) Ogawa, S.; Shimizu, M.; Nako, K.; Okamura, M.; Ito, S. *J. Clin. Exp. Neuropsychol.* **2018**, *3*, 04.
(38) Shi, G.-J.; Li, Y.; Cao, Q.-H.; Wu, H.-X.; Tang, X.-Y.; Gao, X.-H.; Yu, J.-Q.; Chen, Z.; Yang, Y. *Biomed. Pharmacother.* **2019**, *109*, 1085−1099.
(39) Goldberg, D. M.; Yan, J.; Soleas, G. J. *Clin. Biochem.* **2003**, *36*, 79−87.
(40) Messana, I.; Forni, F.; Ferrari, F.; Rossi, C.; Giardina, B.; Zuppi, C. *Clin. Chem.* **1998**, *44*, 1529−1534.