Talanta 192 (2019) 32–39

Contents lists available at ScienceDirect

Talanta

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

Positive matrix factorization: A data preprocessing strategy for direct mass spectrometry-based breath analysis

Xue Li[a,b,*,1], Dandan Huang[c,d,1], Jiafa Zeng[a,b], Chak Keung Chan[c], Zhen Zhou[a,b]

[a] Institute of Mass Spectrometry and Atmospheric Environment, Jinan University, No. 601 Huangpu Avenue West, Guangzhou 510632, China
[b] Atmospheric Pollution Online Source Analysis Engineering Research Center of Guangdong Province, Jinan University, Guangzhou 510632, China
[c] School of Energy and Environment, City University of Hong Kong, Hong Kong, China
[d] State Environmental Protection Key Laboratory of the Cause and Prevention of Urban Air Pollution Complex, Shanghai 200233, China

ARTICLE INFO

Keywords:
Breath analysis
Positive matrix factorization
Mass spectrometry
Exhalome
Volatile organic compounds

ABSTRACT

Interest in exhaled breath has grown considerably in recent years, as breath biosampling has shown promise for non-invasive disease diagnosis, therapeutic drug monitoring, and environmental exposure. Real time breath analysis can be accomplished *via* direct online mass spectrometry (MS)-based methods, which can provide more accurate and detailed data and an enhanced understanding of the temporal evolution of exhaled VOCs in the breath; however, the complicated chemical composition and large raw datasets involved in breath analysis have hindered the discovery of sources contributing to the exhaled VOCs. The positive matrix factorization (PMF) receptor model has been widely used for source apportionment in atmospheric studies. Since the exhaled VOCs contain compounds from various sources, such as alveolar air, mouth air and respiratory dead-space air, PMF may be also helpful for source apportionment of exhaled VOCs in the breath. Thus, this study explores the application of PMF in the pretreatment of direct breath measurement data. The results indicate that (i) endogenous compounds and background contaminants sources can be readily distinguished by PMF in data obtained from replicate measurements of human exhaled breath at single time points (~30 s/measurement), which may benefit both exhalome investigations and the identification of exposure biomarkers; (ii) sources resolved from online measurement data collected over longer periods (1.5 h) can be used to isolate the evolution of exhaled VOCs and investigate processes such as the pharmacokinetics of ketamine and its major metabolites. Therefore, PMF has shown promise for both data processing and subsequent data mining for the ambient MS-based breath analysis.

1. Introduction

When breath biosampling began to show promise for the non-invasive diagnosis of various diseases (*e.g.*, lung cancer, COPD) [1–8] and assessment of environmental exposure [9–11], interest in exhaled breath has grown considerably since the 1980s. For example, an average of 205 volatile organic compounds (VOCs) has been reported in breath samples from normal, healthy individuals [12]. The possible sources of exhaled VOCs include mouth air, respiratory dead-space air and alveolar air. It is noteworthy that the chemical composition of the alveolar air is different from that of the mouth air and respiratory dead-space air. In the alveolar air, there are more metabolic compounds, which are released from the blood-alveolar air exchange [13].

A number of direct mass spectrometry (MS)-based methods have been developed that excel in characterizing the real time chemical composition of exhaled breath; these techniques include the proton transfer reaction MS (PTR-MS) [14,15], selected-ion flow-tube MS (SIFT-MS) [16–18], trace atmospheric gas analyzer (TAGA) [19,20], and emerging technologies such as secondary electrospray ionization MS (SESI-MS) [21–23], extractive electrospray ionization MS (EESI-MS) [24,25], and plasma ionization MS [26]. Online MS techniques have the advantages of recording the variations of exhaled VOCs in high time resolution (*e.g.*, several seconds), providing the time series of the VOCs during the short exhalation period. However, because ambient ionization generally occurs in laboratory air, VOCs in the laboratory air may constitute a VOC "source" and thus should be distinguished from exhaled VOCs during data pretreatment [27].

The positive matrix factorization (PMF) receptor model was

* Corresponding author at: Institute of Mass Spectrometry and Atmospheric Environment, Jinan University, No. 601 Huangpu Avenue West, Guangzhou 510632, China.
E-mail address: tamylee@jnu.edu.cn (X. Li).
[1] There authors contributed equally to this work.

https://doi.org/10.1016/j.talanta.2018.09.020
Received 24 April 2018; Received in revised form 20 August 2018; Accepted 8 September 2018
Available online 10 September 2018
0039-9140/

Table 1
Contributions of each source and, the assigned elemental composition and compound name to individual *m/z*.

m/z	Contribution[a]			Elemental composition ($[M-H]^-$)	$\Delta m/z$[d] (mDa)	Compound	Ref.[e]
	S-1[a]	S-2[b]	S-3[c]				
72.992	45	26	29	C_2HO_3	− 1.1	Glyoxylic acid	[35]
87.008	42	11	46	$C_3H_3O_3$	− 1.1	Pyruvic acid	[35,36]
88.003	28	55	18	$C_2H_2NO_3$	− 1.1	/	/
89.023	39	0	61	$C_3H_5O_3$	− 1.4	Lactic acid	[35,36]
93.034	0	95	5	C_6H_5O	− 1.4	Phenol	[37–39]
94.980	0	100	0	H_3O_4Si	− 0.9	Silicate	/
96.959	29	61	10	HSO_4	− 1.2	Sulfate	/
101.024	35	30	35	$C_4H_5O_3$	− 1.1	Acetoacetic acid	[40–43]
103.003	24	55	21	$C_3H_3O_4$	− 1.1	Malonic acid	[43,44]
103.039	41	39	19	$C_4H_7O_3$	− 1.1	Hydroxybutyric acid	[40–43]
108.021	0	99	1	$C_6H_4O_2$	− 1.1	/	/
109.029	13	87	0	$C_6H_5O_2$	− 1.1	Hydroquinone	/
111.008	38	28	34	$C_5H_3O_3$	− 1.2	/	/
111.946	10	62	28	SO_5	− 1.1	/	/
112.985	37	58	5	$C_2O_2F_3$	− 1.1	Trifluoroacetic acid	[45]
113.023	36	31	33	$C_5H_5O_3$	− 1.1	/	[43]
115.038	99	0	1	$C_5H_7O_3$	− 1.1	/	[43]
117.019	25	64	11	$C_4H_5O_4$	− 1.0	/	[43]
123.008	97	3	0	$C_6H_3O_3$	− 1.1	/	/
125.024	28	60	12	$C_6H_5O_3$	− 1.1	/	/
127.039	35	27	38	$C_6H_7O_3$	− 1.1	/	[43]
129.018	33	50	17	$C_5H_5O_4$	− 1.0	/	[43]
129.054	35	42	24	$C_6H_9O_3$	− 0.9	/	[43]
141.019	39	46	15	$C_6H_5O_4$	− 0.9	/	/
143.034	32	49	19	$C_6H_7O_4$	− 0.8	/	[43]
143.070	26	48	26	$C_7H_{11}O_3$	− 0.8	/	[43]
144.962	100	0	0	$C_6H_3Cl_2$	− 0.8	Dichlorobenzene	[46,47]
146.959	95	5	0	/	/	/	/
155.035	39	43	19	$C_7H_7O_4$	− 0.7	/	/
157.050	31	49	21	$C_7H_9O_4$	− 0.7	/	[43]
157.087	19	52	29	$C_8H_{13}O_3$	− 0.6	/	[43]
159.065	29	60	11	$C_7H_{11}O_4$	− 0.7	/	[42]
165.040	0	79	21	$C_5H_9O_6$	− 0.6	/	/
171.066	28	47	25	$C_8H_{11}O_4$	− 0.6	/	[43]
171.103	18	48	34	$C_9H_{15}O_3$	− 0.5	/	/
173.082	24	60	15	$C_8H_{13}O_4$	− 0.6	/	/
179.056	8	92	0	$C_6H_{11}O_6$	− 0.4	Glucose	[48]
188.951	30	67	3	/	/	/	/
227.202	1	59	40	$C_{14}H_{27}O_2$	0.1	C14:0	[36]
234.957	8	92	0	/	/	/	/
239.060	28	19	52	/	/	/	/
255.234	4	66	31	$C_{16}H_{31}O_2$	0.6	Palmitic acid (C16:0)	[12,26,36,49]
278.984	63	37	0	/	/	/	/
283.264	0	63	37	$C_{18}H_{35}O_2$	0.8	Stearic acid (C18:0)	[26,36,49]
286.216	5	69	26	$C_{16}H_{30}O_4$	0.7	/	
301.240	12	53	35	$C_{17}H_{33}O_4$	0.8	/	/
380.909	57	43	0	/	/	/	/

[a] S-1: mouth/respiratory dead-space air.
[b] S-2: alveolar air.
[c] S-3: ambient air.
[d] $\Delta m/z$ is obtained by comparing the measured value and theoretical mass of the compounds.
[e] Ref., references cited.

developed by *Paatero and Tapper* [28,29] and can be used to quantify the contributions of various sources to a given set of samples by iteratively deriving the compositions, or "fingerprints," of the sources [30]. The model seeks to reconstruct the measured species concentrations and achieve chemical mass balance by varying the source profiles and temporal contributions. PMF uses both the sample concentration and user-provided uncertainty associated with the sample data to weight individual points, and then performs multivariate factor (or source) analysis to decompose the speciated data matrix into a matrix of source profiles and a matrix of source contributions. These source profiles can be compared to measured source profile information and/or emission inventories to aid in the identification of sources that may contribute to the given samples.

Because exhaled VOC analysis using direct MS methods involves both multiple VOC sources and high time-resolution measurements, PMF data preprocessing [31] should be useful in revealing the origins of VOCs in the breath. Thus, this study aims at investigating the application of PMF in the pretreatment of direct breath measurement data and exploring the possible sources of exhaled VOCs.

2. Methods

The data were analyzed using the PMF2 algorithm [32] with two input matrices, namely the sample concentration matrix and the error matrix. The sample concentration matrix was used as collected by the SESI-MS (see the Experimental Section in the Supporting Information). To obtain the error matrix, the detection limit (DL) of each *m/z* must be determined for the specific instrument. The DL of a specific *m/z* is the lowest signal intensity that the instrument can detect. The corresponding error matrix can thus be calculated using Eqs. (1) and (2).