Journal of Chromatography B xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Chromatography B

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

The use of LC predicted retention times to extend metabolites identification with SWATH data acquisition

Tobias Bruderer[a], Emmanuel Varesio[b], Gérard Hopfgartner[a,*]

[a] Life Sciences Mass Spectrometry, Department of Inorganic and Analytical Chemistry, University of Geneva, Quai Ernest-Ansermet 30, CH-1211 Geneva 4, Switzerland
[b] School of Pharmaceutical Sciences, University of Lausanne, University of Geneva, Rue Michel Servet 1, CH-1211 Geneva 4, Switzerland

ARTICLE INFO

Keywords:
Liquid chromatography
High resolution mass spectrometry
SWATH
LC retention time prediction
QSRR
Metabolomics

ABSTRACT

The application of predicted LC retention time to support metabolite identification was evaluated for a metabolomics MS/MS database containing 532 compounds representative for the major human metabolite classes. LC retention times could be measured for two C18 type columns using a mobile phase of pH = 3.0 for positive ESI mode (n = 337, 228) and pH = 8.0 for negative ESI mode (n = 410, 233). A QSRR modelling was applied with a small set of model compound selected based on the Kennard-Stone algorithm. The models were implemented in the R environment and can be applied to any library. The prediction model was built with two molecular descriptors, LogD2 and the molecular volume. A limited set of model compounds (LC CalMix, n = 16) could be validated on two different C18 reversed phase LC columns and with comparable prediction accuracy. The CalMix can be used to compensate for different LC systems. In addition, LC retention prediction was found, in combination with SWATH-MS, to be attractive to eliminate false positive identification as well as for ranking purpose different metabolite isomeric forms.

1. Introduction

The identification of low molecular weight compounds from liquid chromatography mass spectrometry analyses, to support metabolomics investigations, is mainly based on accurate mass measurements and liquid chromatographic retention comparison with authentic standards. Due to the large chemical space of the metabolome, the assignment of peak features (m/z and retention time) to molecules remains a challenging task [1–3]. Four major parameters can be considered including: elemental formula, product ion spectrum, ionization polarity and chromatographic retention time. With modern high resolution mass spectrometry elemental formula calculation based on accurate mass and isotopic match is straight forward. Product spectra assigned with libraries is still limited and does not differentiate isomers. As in many cases standards are not available, chromatographic prediction plays a major role as shown for analyte screening or identification [4–10]. Quantitative structure retention relationships (QSSR) is well suited for retention times prediction, based on molecular physicochemical properties [11]. The major advantage of QSRR compared to other retention prediction approaches is that it can be applied for any given compound based on its molecular descriptors calculated from its structure. LogD and LogP have been extensively applied as suitable molecular descriptor in particular for reverse phase chromatography. QSSR was applied for global LC–MS based metabolomics analysis by Creek et al. [4]. They developed a model based on a multiple linear regression between the logD(pH), five additional molecular descriptors and the measured retention times for 120 metabolite standards on a HILIC column. They reported a reduction in false positive annotations by 40% for putatively annotated metabolites from cell extracts. Recently Cao et al. [12] proposed a QSRR model for non-targeted lipid analysis with multiple linear regression (MLR) modelling and the random forests (RF) approach with significantly improved retention prediction.

These types of QSRR models can be used for retention prediction but they are limited by the need for large sets of model compounds (i.e. typically more than 100 reference standards) [4,12,13]. Furthermore, their precisions are somewhat limited which is regularly attributed to the lack of accuracies from the used physicochemical descriptors like the logP, pK_A and logD(pH) [6] which are calculated from open source software, often based on two dimensional molecular descriptors and more rarely from three dimensional descriptors [13].

A strategy has been presented by Andries et al. [14] for the reduction of the number of model compounds in QSRR models. A simple logP linear regression model, Abraham's solvation equation [15–17] and the Quantum Indices model [11] were compared for a small set of model compounds across 76 different LC conditions obtained from the literature. The use of small number of model compounds i. e. 7–15 did not

* Corresponding author.
E-mail address: gerard.hopfgartner@unige.ch (G. Hopfgartner).

http://dx.doi.org/10.1016/j.jchromb.2017.07.016
Received 6 November 2016; Received in revised form 30 June 2017; Accepted 7 July 2017

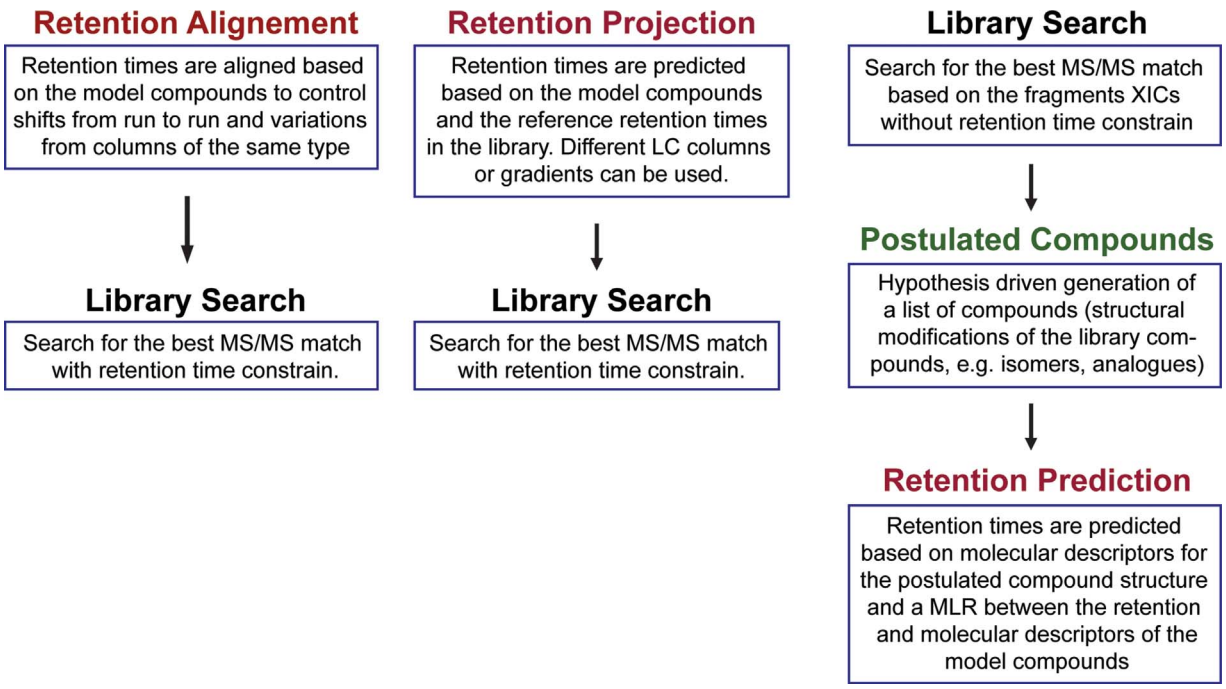

Fig. 1. Strategies to apply in-silico LC retention time prediction to support metabolites indentification.

Table 1
Recommended Model Compound Mixture for both LC columns Waters T3 and C18 (n = 3-20) selected with the Kennard-Stone algorithm with Euclidian distance for the two molecular descriptors logD2 (consensus logP and classic pKa, calculated with ACD Labs 2012, Build 2076, July 25, 2012) and MV (Molecular Volume calculated with ACD Labs 2010).

Nr.	Model Compounds Selection (pH = 3.0)			Model Compounds Selection (pH = 8.0)		
	Compound	logD (pH = 3.0)	MV	Compound	logD (pH = 8.0)	MV
1	Chenodeoxycholic acid	3.75	347.8	Purine	−0.48	81.5
2	2-Pyrrolidinone	−0.80	81.2	Hyodeoxycholic acid	0.72	347.8
3	Chlorogenic acid	−0.49	214.5	Chlorogenic acid	−4.06	214.5
4	3-Chlorotyrosine	−1.42	147.8	Cortisol	1.66	281.3
5	Cortisol	1.66	281.3	3-Chlorotyrosine	−1.61	147.8
6	N-Methyl-a-aminoisobutyric acid	−2.57	114.5	Indole-3-carboxylic acid	−0.91	114.4
7	Epicatechin	0.57	182.1	Octadecanedioic acid	0.77	314.9
8	Diphenhydramine	0.61	249.2	Tetradecanedioic acid	−1.29	248.8
9	Sphinganine	2.57	325.0	3,5-Diiodo-L-tyrosine	−1.20	180.0
10	Loratadine	4.12	303.5	Melatonin	1.74	197.6
11	1,11-Undecanedicarboxylic acid	3.07	232.3	Estrone	3.38	232.1
12	Ranitidine	−2.82	265.4	Pyridoxamine	−0.96	131.1
13	Tryptophanol	1.70	132.1	Ranitidine	−0.15	265.4
14	Ribothymidine	−1.49	163.7	Ribothymidine	−1.51	163.7
15	2-Piperidinone	−0.50	98.9	Salicylic acid	−0.78	100.3
16	Melatonin	1.74	197.6	Biocytin	−2.72	301.7
17	Luteolin	2.35	172.9	L-Aspartyl-L-phenylalanine	−3.64	204.4
18	2-Hydroxyphenethylamine	−2.57	124.1	1-Methylhistidine	−3.07	122.6
19	Liothyronine	1.58	272.6	N-Acetylserotonin	1.13	172.0
20	5-Methoxytryptophol	1.67	156.1	5-Methoxytryptophol	1.67	156.1

result in any significant loss of prediction accuracy. They applied the Kennard-Stone [18] algorithm to obtain an equidistant distribution of the molecular descriptors over the investigated variable space. Falchi et al. [10] have shown that Kernel-Based, partial least squares quantitative structure-retention relationship model can be a useful tool for metabolite identification. They build a model based on 1383 compounds and considering different chemical classes and demonstrated that their model succeeded in the RT prediction of drug metabolites.

In addition, for unknowns of interest, product ion spectra are acquired and identification is further enabled with the use of libraries or *denovo* spectra interpretation [19].

Data Independent Acquisition (DIA) such as MS^E or SWATH have gained interest as they enable the recording of full scan and product ion spectra in a single LC–MS run. While SWATH/MS has mostly been used for proteomics [20] these features are valuable for metabolomics studies [21,22] and provide qualitative and quantitative analysis in the same run (QUAL/QUAN) with enhanced selectivity of the precursor ions compared to MS^E as the size of the windows can be selected. Compared to DDA, SWATH spectra enable no only to perform library search on any precursor but also perform structural analog search based on product fragment by using the chromatographic profile.

In the present work, to allow a better assignment of metabolites in complex samples and in particular of isomeric metabolites, a QSRR modelling was applied with a small set of model compound ($n < 20$) selected based on the Kennard-Stone algorithm for a larger set of 532 compounds present in metabolomics datasets and referenced in the Human Metabolome Database [23]. SWATH MS/MS spectra were available in positive and negative mode for 532 compounds [24]. The