Journal of Chromatography A, 1403 (2015) 54–62

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Least absolute shrinkage and selection operator and dimensionality reduction techniques in quantitative structure retention relationship modeling of retention in hydrophilic interaction liquid chromatography

Emilia Daghir-Wojtkowiak[a,1], Paweł Wiczling[a,1], Szymon Bocian[b,3], Łukasz Kubik[a,1], Piotr Kośliński[c,2], Bogusław Buszewski[b,3], Roman Kaliszan[a,*], Michał Jan Markuszewski[a,*]

[a] Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Al. Gen. Hallera 107, 80-416 Gdańsk, Poland
[b] Department of Environmental Chemistry and Bioanalytics, Faculty of Chemistry, Nicolaus Copernicus University, Gagarin Street 7, 87-100, Toruń, Poland
[c] Department of Toxicology, Ludwik Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Torun, M. Curie-Skłodowskiej 9, 85-094 Bydgoszcz, Poland

ARTICLE INFO

Article history:
Received 25 February 2015
Received in revised form 7 May 2015
Accepted 11 May 2015
Available online 19 May 2015

Keywords:
Hydrophilic interaction liquid chromatography (HILIC)
QSRR
Nucleosides
Pterins
Least absolute shrinkage and selection operator (*LASSO*)
Stationary phases

ABSTRACT

The objective of this study was to model the retention of nucleosides and pterins in hydrophilic interaction liquid chromatography (HILIC) via QSRR-based approach.

Two home-made (Amino-P-C18, Amino-P-C10) and one commercial (IAM.PC.DD2) HILIC stationary phases were considered. Logarithm of retention factor at 5% of acetonitrile ($\log k_{ACN}$) along with descriptors obtained for 16 nucleosides and 11 pterins were used to develop QSRR models. We used and compared the predictive performance of three regression techniques: partial least square (PLS), the least absolute shrinkage and selection operator (*LASSO*), and the *LASSO* followed by stepwise multiple linear regression.

The highest predictive squared correlation coefficient (Q^2_{LOOCV}) in PLS analysis was found for Amino-P-C10 ($Q^2_{LOOCV} = 0.687$) and IAM.PC.DD2 ($Q^2_{LOOCV} = 0.506$) and the lowest for IAM.PC.DD2 ($Q^2_{LOOCV} = -0.01$). Much higher values were obtained for the LASSO model. The Q^2_{LOOCV} equaled 0.9 for Amino-P-C10, 0.66 for IAM.PC.DD2 and 0.59 for Amino-P-C18. The combination of *LASSO* with stepwise regression provided models with comparable predictive performance as the *LASSO*, however with possibility of calculating the standard error of estimates.

The use of *LASSO* itself and in combination with classical stepwise regression may offer greater stability of the developed models thanks to more smooth change of coefficients and reduced susceptibility towards chance correlation. Application of QSRR-based approach, along with the computational methods proposed in this work, may offer a useful approach in the modeling of retention of nucleoside and pterin compounds in HILIC.

1. Introduction

Hydrophilic interaction liquid chromatography (HILIC) is a variant of normal phase liquid chromatography (NP-LC). It is characterized by a complex separation mechanism mostly driven by polar forces, such as hydrogen bonds and dipole–dipole interactions. HILIC is mainly used to analyze polar compounds with poor solubility in NP-LC or insufficient retention in reversed-phase liquid chromatography [1]. The usefulness of HILIC has been confirmed in the analysis of amino acids [2], lipids and phospholipids [3,4] or endogenous metabolites of clinical interest [5,6].

* Corresponding authors at: Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gen. J. Hallera 107, 80-416 Gdańsk, Poland. Tel.: +48 58 349 14 93; fax: +48 58 349 14 93/19 62.
E-mail addresses: roman.kaliszan@gumed.edu.pl (R. Kaliszan), markusz@gumed.edu.pl (M.J. Markuszewski).
[1] Tel.: +48 58 349 14 93; fax: +48 58 349 19 62.
[2] Tel.: +48 52 585 38 96; fax: +48 52 585 35 29.
[3] Tel.: +48 56 611 43 08; fax: +48 56 611 48 37.

http://dx.doi.org/10.1016/j.chroma.2015.05.025
0021-9673/

Modified nucleosides and pterins have been proposed as possible markers of pathophysiological status [7–10]. The nucleosides originate from post-transcriptional modifications of the genome, whereas pterins are produced during immune system activation. Nucleosides and pterins are polar compounds, which are difficult to analyze in RP-HPLC mode. Thus, HILIC mode may constitute an alternative [11,12]. Rodríguez-Gonzalo et al. [13] described the method of *on-line* extraction of nucleosides and deoxynucleosides using HILIC coupled with MS/MS detection. Similar approach was applied by Tuytten et al. [14], who also used *on-line* automated extraction, coupled with hydrophilic interaction chromatography ESI-MS for targeted metabolomics analysis of nucleosides. In the literature, quantitative determination of eight pterin compounds has already been described using HILIC mode [15].

The retention mechanism is often investigated by Quantitative Structure Retention Relationships (QSRR). QSRR relates structural parameters of investigated compounds with their retention data and serves to predict retention of compounds or classify stationary phases [16,17]. For the proper development of QSRR equations, chromatographic retention data should be collected for a representative number of compounds, followed by further validation of the obtained models.

The most popular modeling technique in QSRR-oriented studies is multiple linear regression (MLR) with ordinary least square (OLS) estimation. MLR has distinct advantages in terms of inference and provides an interpretable model of compounds' behavior in a given chromatographic system. The OLS estimator works properly if descriptors are non-redundant (uncorrelated) and leads to the unbiased estimates. However, in QSRR modeling we almost always deal with high dimensional and redundant data. In such a case, the computation of the OLS estimation is no longer effective as it introduces a lot of variance in the estimates of regression coefficients, which often leads to overfitting and poor predictive performance of the developed model [18].

Therefore, considering high dimensional space, where the number of descriptors exceeds the number of observations, dimensionality reduction techniques should be used [19].

Supervised partial least square (PLS) regression is another method often used in QSRR studies. It evaluates the presence of linear relationships between variables in high dimensional data matrix. Conceptually, variable selection based on PLS is a discrete process, in which predictors are either kept in or dropped out of the model. PLS models the relationship among different variables by means of latent variables, which handle the problem of multicollinearity. Hence, it makes this technique useful and widely-adopted in the analysis of high dimensional data.

However, it should be noted, that continuous process of variable selection can also be used to fit the model. Regression methods that are based on regularization constrain (penalize) or (in other words) regularize the coefficient estimates by shrinking them towards zero. Regularization is used to prevent overfitting by penalizing complex models with large number of parameters.

There are limited studies on nucleosides and pterins in a view of QSRR modeling utilizing HILIC mode. Therefore, in this study, we:

(i) inspected the similarity and dissimilarity of nucleosides and pterins, based on the calculated descriptors,
(ii) built models based on PLS and the *LASSO* and compared their predictive ability, introduced the combination of *LASSO* and stepwise MLR, followed by full assessment of model performance,
(iii) developed and interpreted the obtained QSRR models.

2. Materials and methods

2.1. Materials

In the study, 16 nucleosides (including 2 deoxynucleosides): adenosine (A), guanosine (G), uridine (U), cytidine (C), inosine (I), 5-methyluridine (5mU), 1-methyladenosine (1mA), 2-methyladenosine (2mA), pseudouridine (Ps), 3-methyluridine (3mU), N4-acetylcytidine (N4Ac), 2-O-methylguanosine (2-O-mG), 5-methylcytidine (5mC), N2,N2-dimethylguanosine (N2N2dmG), 8-bromoguanosine (8-BrG), 2-O-methyladenosine (2-O-mA), and 11 pterins: 6,7-dimethylpterin (6,7-dmP), L-monapterin (L-monaP), xanthopterin (Xan), 6-methylpterin (6mP), isoxanthopterin (iXanP), 7-biopterin (7-bioP), 6-hydroxymethylpterin-monophosphate (6-OH-mP), biopterin (BioP), neopterin (NeoP), pterin (Pt) and pterin acid (Pt.acid), were subjected to the analysis. The standards of nucleosides and pterins were purchased from Sigma–Aldrich (St. Louis, MO, USA) and Dr. B. Schircks Lab-oratories (Jona, Switzerland), respectively.

Deionized water purified with Direct–QUV (Millipore, France) system, was used to prepare aqueous solutions of nucleosides and pterins. Stock concentrations of analyzed solutions were 0.25 mM. Acetonitrile of HPLC grade was purchased from Chempur (Piekary Slaskie, Poland).

Three stationary phases operating in HILIC mode were tested. Two of them were home-made N,O-dialkylphosphoramidate stationary phases: Amino-P-C18, Amino-P-C10 (Chair of Environmental Chemistry and Bioanalytics, Nicolaus Copernicus University, Toruń, Poland) and the third one was commercial Immobilized Artificial Membrane IAM.PC.DD2, 150 × 4.6, d_p = 12 μm (Regis Technologies, Morton Grove, IL, USA). Home-made stationary phases were packed into stainless steel 250 × 4.6 mm chromatographic columns. The structures of chemically bonded stationary phases used in the study are presented in Fig. 1. The physicochemical parameters of the investigated stationary phases are presented in Table 1 and in references [20,21].

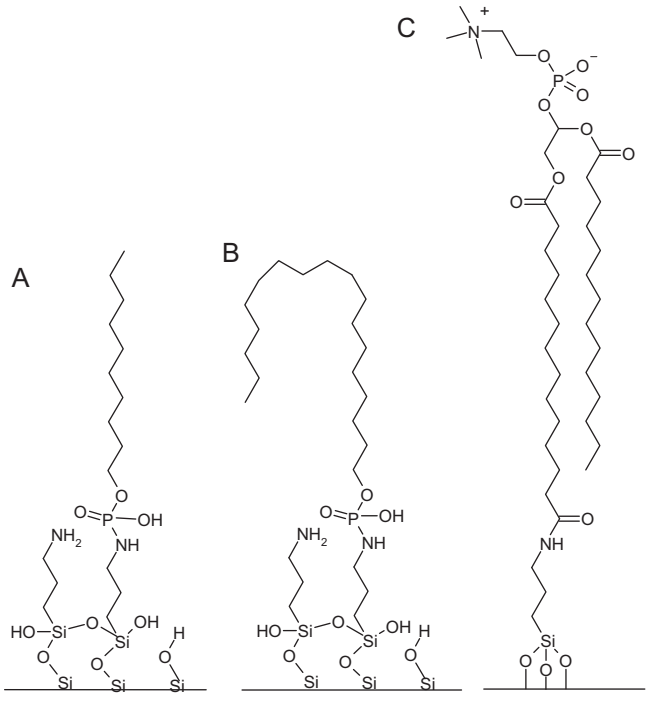

Fig. 1. Structures of new chemically bonded stationary phases: Amino-P-C10 (A) and Amino-P-C18 (B), and stationary phases used for comparison: IAM.PC.DD2 (C).