Talanta 185 (2018) 592–601

Contents lists available at ScienceDirect

Talanta

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

A novel strategy for retention prediction of nucleic acids with their sequence information in ion-pair reversed phase liquid chromatography

Chao Liang, Jun-Qin Qiao, Hong-Zhen Lian*

State Key Laboratory of Analytical Chemistry for Life Science, Collaborative Innovation Center of Chemistry for Life Sciences, School of Chemistry & Chemical Engineering and Center of Materials Analysis, Nanjing University, 163 Xianlin Avenue, Nanjing 210023, China

ARTICLE INFO

Keywords:
Nucleic acids
Ion-pair reversed-phase liquid chromatography
Retention prediction
Chromatographic hydrophobic and electrostatic interaction index

ABSTRACT

In this work, retention behaviors of oligonucleotides and double-stranded deoxyribonucleic acids (dsDNAs) have been investigated in ion-pair reversed-phase liquid chromatography (IP-RPLC). We demonstrated that classic linear solvent strength (LSS) model is applicable for describing isocratic retention of oligonucleotides and dsDNAs, which indicated that nucleic acids share the similar retention mechanism as other common small molecules in IP-RPLC. The separation of nucleic acids in IP-RPLC is driven by both hydrophobic and electrostatic interactions. We defined the parameter $\ln k_w/S$ obtained from LSS model in IP-RPLC as chromatographic hydrophobic and electrostatic interaction index (CHEI). CHEI of a nucleic acid has been revealed to correlate well with its gradient retention time. Notably, we proposed a strategy for retention prediction based on CHEI and base sequence information of nucleic acids. Corresponding to base locus, each base sequence is converted to a featured locus vector consisting of zeros and ones. CHEI prediction models were established by support vector regression (SVR) algorithm with locus vectors. Predicted CHEI values have been applied to predict retention times under desired gradient elution runs. This protocol is easy to grasp and worth pursuing further development for more precise retention prediction performance of nucleic acids in IP-RPLC.

1. Introduction

It is hard to pick up a newspaper or a news magazine without hearing some new applications of DNA technology. DNA related technologies, such as DNA sequencing [1] and genotyping techniques [2], boost the development of life sciences. Synthesized and chemically modified oligonucleotides, the so-called antisense oligonucleotides, have shown promise as the therapeutic agents in antisense therapy [3]. Thus, development of fast and reliable methods for nucleic acid analysis is highly desired. Ion-pair reversed-phase liquid chromatography (IP-RPLC) is one of the most suitable key technologies for separation and analysis of a large size range of both single (oligonucleotides) and double-stranded deoxyribonucleic acids (dsDNAs) [4]. The most noteworthy parameter in nucleic acids separation by IP-RPLC is the retention time, which is highly related to the structure of nucleic acids. Precise retention prediction is of great significance not only for rapid optimization of IP-RPLC separation conditions of nucleic acids but also for facilitating the elucidation of their existing status in physiological environment.

There have been a few investigations into the relationship between structure and retention behavior regarding to oligonucleotides. It has been found that the retention of oligonucleotides in IP-RPLC is influenced by both base composition and base sequence. Gilar et al. proposed an empirical formula for predicting oligonucleotide retention based on summation of retention times determined for individual nucleotides [5]. However, this formula was derived at high column temperatures, and it has been confirmed not applicable for oligonucleotides forming partial double-helical structure and hairpin structure at lower temperatures [6]. Sturm et al. [6] and Kohlbacher et al. [7] constructed the models with support vector regression (SVR). These models included temperature dependent secondary structural features of oligonucleotides, making it possible to predict retention times of oligonucleotides at fixed column temperatures. Since the secondary structures of oligonucleotides were simulated by software, this strategy was highly dependent on the reliability of secondary structure prediction methods. Lei et al. [8] proposed a different approach with the same dataset of oligonucleotides reported by Sturm et al. They used base sequence autocorrelation (BSA) descriptors to develop quantitative structure-retention relationship (QSRR) models. Their models allowed retention estimation at a fixed column temperature or at an arbitrary column temperature between 30 and 80 °C. Although satisfactory for retention prediction, the generation of BSA features for oligonucleotides was

* Corresponding author.
E-mail address: hzlian@nju.edu.cn (H.-Z. Lian).

https://doi.org/10.1016/j.talanta.2018.04.030
Received 14 December 2017; Received in revised form 2 April 2018; Accepted 7 April 2018
Available online 10 April 2018

relatively complex. Studzińska and Buszewski [9] established several QSRR models for oligonucleotides with molecular modeling-descriptors calculated using Hyperchem software. However, the correlations of these models were often low, indicating they are insufficient for retention prediction of oligonucleotides. Different with rigid small molecules, oligonucleotides are usually folded, contorted and coiled, and then might form distinct 3D structures in chromatographic mobile phase systems. As for dsDNAs, their retention behaviors were found mainly depended on the size. Huber et al. [10] proposed a size-dependent model for rapid sizing of dsDNAs under a given mobile phase. Zhang et al. [11] developed another size-dependent model, which could predict retention of dsDNAs under arbitrary linear gradient elution runs. However, the suitability of size-dependent models for analyzing dsDNAs with minor difference in size has not been validated. In our previous work, we found that retention of dsDNAs shorter than 40 bp is influenced by the base composition and base sequence other than size [12]. Thus, both base composition and base sequence would influence retention behaviors of oligonucleotides and dsDNAs.

The retention characteristic of nucleic acid in IP-RPLC is that a small change in the mobile phase strength would cause a sharp change in retention factor, and therefore gradient elution is beneficial for nucleic acid separation [4]. An important thing should be noticed is that some of the above models were applicable for predicting retention time under a fixed chromatographic condition. Models proposed by Sturm et al. [6], Kohlbacher et al. [7], and Lei et al. [8] could predict retention at different column temperatures. In optimization of the chromatographic condition, the starting concentration of organic modifier in mobile phase and the gradient slope are often more important than the column temperature. The size-dependent model by Zhang et al. [11] could predict retention times of dsDNAs under arbitrary linear gradient runs. However, this model has not been used for oligonucleotides or dsDNAs with minor difference in size, which inspired us to re-investigate the retention behaviors of nucleic acids in IP-RPLC.

Zhang et al.'s model was derived from a retention equation established by Shoenmakers et al. [13] for solute under a linear gradient elution run. This retention equation was retrospectively based on linear solvent strength (LSS) model [14]:

$$\ln k = \ln k_w - S\varphi \quad (1)$$

where φ is the volume fraction of the organic modifier in the mobile phase, $\ln k_w$ is the extrapolated retention factor corresponding to pure water as mobile phase and S is a constant calculated via linear regression analysis. LSS model is commonly established in RPLC with an adequate range of organic modifier concentration [14]. In some references, LSS model was also found applicable in IP-RPLC for small molecules [15,16]. Early, Gilar et al. reported a linear dependency of ln$k$ on the acetonitrile (ACN) concentration φ for oligodeoxythymidines $(dT)_2$ to $(dT)_{30}$ at 60 °C [5]. In Zhang et al.'s work [11], although used in establishing retention predicting models, the LSS model had not been demonstrated applicable for dsDNAs with isocratic retention data. If retention behaviors of oligonucleotides and dsDNAs could be described by the LSS model like small molecules, should other retention equations established for small molecules also be suitable for nucleic acids?

In this present work, LSS model was first demonstrated applicable for nucleic acids with isocratic retention data of some oligonucleotides and short dsDNAs in IP-RPLC. And the parameters $\ln k_w$ and S in the LSS model were discussed in detail for oligonucleotides and dsDNAs respectively. Then retention times of oligonucleotides and short dsDNAs under different linear gradient runs were tried to be predicted with the parameter $\ln k_w/S$ calculated from LSS model. The parameter $\ln k_w/S$ was first described by Valko et al. as the chromatographic hydrophobicity index (CHI) in RPLC for estimating octanol/water partition coefficients (log$P$) of small molecules [17]. CHI is the volume fraction of organic modifier required to achieve an equal distribution of an analyte between mobile and stationary phases ($k = 1$, $\ln k = 0$). Then, Krokhin and Spicer [18] used a similar parameter named as hydrophobicity index (HI) for peptides retention study in RPLC. HI is the concentration of organic modifier that yields a retention factor of 10 under isocratic elution conditions for any peptide. Good linear correlations have been found between gradient retention times and CHI or HI values, which enlightened us to predict retention of nucleic acids with this parameter. Peptides are weak ionizable amphoteric compounds, while nucleic acids are strong ionizable acids. Besides, peptides are more hydrophobic than nucleic acids. In nucleic acids separation, ion-pair reagents are often necessary. Except for the hydrophobic interactions, ion-ion interactions play an important role in IP-RPLC separation of nucleic acids. So we defined $\ln k_w/S$ in IP-RPLC as the chromatographic hydrophobic and electrostatic interaction index (CHEI) although it has the same expression as CHI. Obviously, CHEI is dependent on the chromatographic condition (column, mobile phase and column temperature) used. Considering the difficulty to establish LSS model for nucleic acids with isocratic retention data, we calculated $\ln k_w/S$ from gradient retention data. CHEI was then demonstrated sufficient for retention prediction of nucleic acids under different linear gradient elution runs. Many strategies for peptides retention prediction from their sequence information have been reported. SSRCalc [19] and ELUDE [20] are two well-trained predictors available at websites based on peptide sequence widely used in proteome analysis. Recently, Lu et al. [21] proposed locus-specific retention predictor (LsRP) based on base locus information and SVR for predicting peptide retention time. In LsRP, a series of vectors were designed as the idea of translating amino acid order into vectors. Like peptides whose sequences consist of 20 kinds of amino acids, the sequences of nucleic acids consist of 4 kinds of nucleotides. However, there are no well-established retention predictors for nucleic acids. Herein, we elucidated that modified LsRP could be applied for predicting CHEI or retention time of nucleic acids. Finally, an easily prehensible protocol was proposed for predicting retention times of nucleic acids under desired linear gradient elution runs. In order to demonstrate the universal applicability of the LsRP based strategy for nucleic acid retention prediction, two datasets were used in this work. One is from references, which was used by Sturm et al. [6] and Lei et al. [8] for retention prediction model establishment and evaluation. The other was from our experiments performed on a common commercial silica-based C_{18} column with mostly used mobile phase composed of ACN and triethylammonium acetate (TEAA) at column temperature of 30 °C.

2. Experimental section

2.1. Materials

HPLC grade ACN and triethylamine (TEA) were purchased from Honeywell (Ulsan, Korea) and TEDIA (Fairfield, OH, USA), respectively. Acetic acid (≥ 99.5%, analytical reagent) was purchased from Sinopharm Chemical Reagent Co., Ltd. (Shanghai, China). Ultrapure water was used throughout the experiment.

59 oligonucleotides ranging from 5 to 25 bases were synthesized by Invitrogen Biotechnology Co., Ltd. (Shanghai, China). 12 short dsDNAs (20 bp, 25 bp, or 30 bp) were synthesized by GenScript Co., Ltd. (Nanjing, China). All the nucleic acids were provided in lyophilized form and then dissolved in proper amount of water to get the final concentration of 10 μmol/L before use. The sequences of oligonucleotides and dsDNAs are listed in Table S1 of the Supplementary material.

2.2. Apparatus

A Waters 2695 Alliance separation module (Milford, MA, USA) was employed consisting of a vacuum degasser, a quaternary pump and an auto-sampler, and a Waters 996 photodiode-array (PDA) detector set at 260 nm. Data acquisition and processing were performed on a Waters Empower chromatography manager system. The pH values of mobile phases were measured after mixing buffer with ACN using a FiveEasy