Journal of Pharmaceutical and Biomedical Analysis 127 (2016) 94–100

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

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

Exploiting non-linear relationships between retention time and molecular structure of peptides originating from proteomes and comparing three multivariate approaches

Petar Žuvela[a], Katarzyna Macur[b], J. Jay Liu[a], Tomasz Bączek[c,*]

[a] Department of Chemical Engineering, Pukyong National University, 365 Sinseon-ro, 608-739 Busan, Republic of Korea
[b] Laboratory of Mass Spectrometry, Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk, Kładki 24, 80-822 Gdańsk, Poland
[c] Department of Pharmaceutical Chemistry, Medical University of Gdańsk, Hallera 107, 80-416 Gdańsk, Poland

ARTICLE INFO

Article history:
Received 30 October 2015
Received in revised form 11 January 2016
Accepted 23 January 2016
Available online 27 January 2016

Keywords:
Quantitative structure-retention relationships (QSRR)
LC–MS/MS
Genetic Algorithms
Non-linear relationships
Proteomics

ABSTRACT

Peptides' retention time prediction is gaining increasing popularity in liquid chromatography–tandem mass spectrometry (LC–MS/MS)-based proteomics. This is a promising approach for improving successful proteome mapping, useful both in identification and quantification workflows. In this work, a quantitative structure-retention relationships (QSRR) model for its direct prediction from the molecular structure of 185 peptides originating from 8 well-characterized proteins and two *Bacillus subtilis* proteomes has been developed. Genetic Algorithm (GA) was used for selection of a subset of molecular descriptors coupled with three machine learning methods: Support Vector Regression (SVR), Artificial Neural Networks (ANN), and kernel Partial Least Squares (kPLS) for regression. Final GA-SVR, GA-ANN, and GA-kPLS models were validated through an external validation set of 95 peptides originating from the human epithelial HeLa cells proteomes. Robustness and stability was ensured by defining their applicability domain. The descriptors of the developed models were interpreted confirming a causal relationship between parameters of molecular structure and retention time. GA-SVR model has shown to be superior over the others in terms of both predictive ability, and interpretation of the selected descriptors.

1. Introduction

The discoveries in the field of systems biology highlight the fact that many different types of molecules (e.g., genes, mRNAs, proteins and metabolites) co-exist in the living organisms in contextual relationships [1]. Widely investigated biomolecules—proteins, are important components of these interactive networks. With the rising development of new analytical techniques, experimental approaches and bioinformatics tools, especially in the field of mass spectrometry (MS)-based proteomics, large-scale studies on structure and function of entire protein families has become possible. This led to birth of a novel interdisciplinary scientific domain: proteomics [2].

In an area referred to as shotgun proteomics [3], in which proteins are broken down into peptides, reversed-phase high performance liquid chromatography coupled with tandem mass spectrometry (RP-LC–MS/MS) allows for their fast, accurate, and direct analysis [4]. RP-LC is powerful in separating them, while MS/MS allows for their accurate identification by making use of database search algorithms such as Sequest [5], which was utilized in this work.

Peptides' retention time is a parameter that can be easily extracted from LC–MS data, and can support their identification. As there is usually little a priori knowledge about peptides composing the investigated proteomic samples in the discovery studies, the prediction of their retention times becomes more and more popular [6]. Recently, peptide retention time prediction has been gaining a lot of attention also in data-independent-acquisition MS proteomics workflows, such as SWATH, allowing for simultaneous qualitative and quantitative proteome profiling [7]. Among different approaches for peptides' retention time prediction, quantitative structure-retention relationships (QSRR) [8–11] models are steadily being integrated as an important segment of shotgun proteomics. Their applications include: proteome-wide retention time prediction [6], improvement of peptide identification [13], and prediction of their elution order [12]. Scarcity of peptide QSRR

* Corresponding author. Fax: +48 58 3491635.
E-mail address: tbaczek@gumed.edu.pl (T. Bączek).

http://dx.doi.org/10.1016/j.jpba.2016.01.055
0731-7085/

applications and publications is in part due to complex modeling of peptide structures, and high computational cost of their optimization. Apart from molecular structure optimization, variable selection is another crucial aspect of QSRR model development. Our previous work [14] has shown that a properly optimized Genetic Algorithm (GA) [15] exhibits exceptional performance in tackling this issue, where it was coupled with Partial Least Squares [16] for regression. However, since the relationship between molecular descriptors and retention time is non-linear for peptides with long sequences [12] (i.e., with a mass of over 5 kDa), use of machine learning methods is encouraged. There are a few studies detailing development of machine learning-based QSRR models.

Thereby, Tian et al. [17] employed Support Vector Regression (RF) [18,19]. Random Forests (RF) [20], and Gaussian Process (GP) [21] for modeling a set of 2042 peptides originating from the *Drosophila melanogaster* proteome. The authors report satisfactory model performance and confirm the usefulness of the mentioned methods for systematic QSRR modeling of a proteome-wide peptide dataset.

Shinoda et al. [6] employed Artificial Neural Networks (ANN) [22] for proteome-wide predictions of retention times of peptides originating from the *Escherichia coli*. Petritis et al. also employed ANNs for development of a QSRR model based on a training set of peptides originating from *Deinococcus radioduranans* proteome validated with a set of peptides originating from *Shewanella oneidensis* proteome [23]; as well as a model for peptides originating from *Saccharomyces cerevisiae* [24]. In both instances the authors used experimental descriptors to model LC retention time, requiring even further experiments. This limitation was overcome by Golmohammadi et al. [25] in which the authors used theoretical descriptors to develop SVR and ANN QSRR models for a set of 93 peptides with a known amino acid composition.

In this work, we employed three state-of-the-art machine learning methods: Support Vector Regression (SVR) [18,19]. Artificial Neural Networks (ANN) [22], and kernel Partial Least Squares (kPLS) [26].

The developed models were validated through an independent external set of 95 peptides originating from the human epithelial HeLa cells proteome. Their applicability domains were defined, and they were thoroughly interpreted confirming a causal relationship between the molecular descriptors and retention time.

2. Materials and methods

2.1. RP-LC–MS/MS analysis

2.1.1. Sample preparation

Aqueous solutions of bovine milk β-casein, human serum albumin, bovine serum albumin, chicken egg ovalbumin, ribonuclease B, bovine milk lactoglobulin, bovine myoglobin, insulin-like growth factor-binding protein 1 (purified from human amniotic fluid using a previously reported procedure [27]) were prepared in a concentration of 3 mg/mL.

Bacillus subtilis proteins, at the concentration of 1.2–1.5 mg/mL, were obtained after extraction from endospores of the strains 168 and Δ*prpE* as reported in [28].

Human epithelial HeLa cells (2×10^6, seeded in 6 cm plates in 1 mL growth medium per well) were cultured as described by Doszczak et al. [29]. Subsequently, the cells were incubated for 4 h with: (1) 100 U/mL recombinant human interleukin-1 alpha (IL-1α; eBioscience, Vienna, Austria)—HeLaIL-1α; (2) 25 μg/mL cycloheximide (CHX)—HeLaCHX; (3) both 100 U/mL IL-1α and 25 μg/mL CHX—HeLaIL-1α/CHX; or (4) alone as a control—HeLaK. After stimulation, the media were discarded. The cells were washed on ice in phosphate buffered saline (Oxoid, Basingstoke, UK) and harvested from the plate. Subsequently, they were centrifuged ($1000 \times g$) and the cell precipitates were lysed with lysis buffer: 10 mM Tris–HCl pH 7, 1 mM EDTA, 250 mM saccharose; with freeze-thawing and sonication to obtain total lysis. The protein concentration in the collected supernatants was about 0.2 mg/mL.

The protein samples were reduced by incubating with 100 mM dithiotreitol (DTT, in freshly prepared 100 mM ammonium bicarbonate), at 60 °C for 30 min. After that, trypsin was added to the samples (enzyme to substrate ratio 1:50). The proteolytic digestion was performed for 12 h at 37 °C and trifluorocaetic acid (TFA) added to quench the reaction. The concentrations of the obtained peptide mixtures were about 0.1 μg/μL.

The protein standards, chemicals (DTT, TFA, ammonium bicarbonate, CHX, lysis buffers components) and MS-grade trypsin used in this study were purchased at Sigma-Aldrich (Steinheim, Germany), if not otherwise stated. Water used in the presented experiments was deionized by passing through a Direct-Q™ system (Millipore, Bedford, MA, USA).

2.1.2. RP-LC–MS/MS conditions

Samples were analyzed on the Finnigan LC-UV-MS/MS LTQ linear ion trap MS system with ESI ion source (Thermo Finnigan, San Jose, CA, USA). The instrument was controlled by Thermo Xcalibur software 1.4. The chromatographic separation of the peptides mixtures was achieved using the XTerra MS C18 (2.1 × 100 mm, 3.5 μm) column (Waters, Milford, MA, USA) at the flow rate of 200 μL/min and the linear 90 min gradient time, from 0% to 60% of solvent B,. The mobile phases: solvent A – 0.1% aqueous solution of TFA, and solvent B – 0.1% TFA in MS-grade acetonitrile (Sigma–Aldrich, Steinheim, Germany), were mixed on-line. The mass spectrometer was operated in the positive ion mode using the following constant instrumental conditions: source voltage 4.62 kV, capillary voltage 40.97 V and capillary temperature 219.96 °C. The collision-induced dissociation was used to generate MS/MS spectra in the linear ion trap. It was performed with an isolation width of 3 Da (m/z) and the activation amplitude of 35% of ejection RF amplitude, which corresponds to 1.58 V.

2.1.3. Protein identification

Upon acquiring the MS/MS spectra, they were automatically searched against the protein database (*fasta, downloaded from UniProtKB) with the use of the Sequest Algorithm, included in Bioworks 3.0 (Thermo Finningan, San Jose, CA, USA). Washburn et al. [30] filtering criteria: X_{corr} [5] values of at least 1.9, 2.2 and 3.75, for +1, +2 and +3 charged tryptic peptides, respectively, and ΔC_n [5] values above 0.08 were applied in peptide identification. Experimental retention times ($t_{R,exp}$) of the identified peptides were defined at peak intensity maximum. Therefore, 185 peptides originating from (1) eight model proteins, (2) *B. subtilis* proteome, and 95 peptides originating from (3) human epithelial HeLa cells proteomes were used for QSRR model development and validation.

2.2. Model development

In order to calculate a set of descriptors for QSRR modeling, molecular structures of 280 peptides were modeled using the powerful sequence editor in HyperChem Professional 8 (Hypercube Inc., Gainesville, Florida, USA) software. Modeled structures were solvated in a water box of a defined length. Subsequently, they were optimized using the Molecular Mechanics [31] method with CHARMM (BIO+) force field [32]. and Polak–Ribière conjugate gradient algorithm [33] employed until the root mean square gradient (RMS) value of 0.1 kcal mol^{-1} Å^{-1} was reached. Final structures were used as input into Dragon 6.0 (Talete, Milano, Italy) software, and 4885 descriptors were calculated. Preliminary variable selection was performed by removing descriptors with a relative