Journal of Chromatography B, 877 (2009) 433–440

Contents lists available at ScienceDirect

Journal of Chromatography B

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

Standardization of retention time data for AMT tag proteomics database generation

I.A. Tarasova[a], V. Guryča[b,c], M.L. Pridatchenko[a], A.V. Gorshkov[d], S. Kieffer-Jaquinod[b,c], V.V. Evreinov[d], C.D. Masselon[b,c,*], M.V. Gorshkov[a]

[a] Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Moscow, Russia
[b] CEA, iRTSV, Laboratoire d'Etude de la Dynamique des Protéomes, Grenoble F-38054, France
[c] INSERM, U880, Grenoble, F-38054, France
[d] N.N. Semenov's Institute of Chemical Physics, Russian Academy of Sciences, Moscow, Russia

ARTICLE INFO

Article history:
Received 4 September 2008
Accepted 17 December 2008
Available online 25 December 2008

Keywords:
Reversed phase liquid chromatography
Retention time correlation
Proteomics
AMT tag
Mass spectrometry

ABSTRACT

The combination of liquid chromatography (LC) with mass spectrometry (MS) has become a mainstream proteome analysis strategy. In LC–MS, measured masses possess their "universal" scale derived from atomic mass tables. In contrast, the observed LC retention times (RT) are not tied to a conventional time scale, and depend on experimental conditions. However, RT data, being explicitly orthogonal to MS, offer relevant information for proteome characterization. We present here a strategy for peptides RT data standardization, based on the generation of a standard scale using retention prediction models, which enables sharing of identification databases in the context of multi-laboratory research.

1. Introduction

Modern proteomics strategies often rely on the so-called "Shotgun" approach, based on a combination of liquid chromatography (LC) and tandem mass spectrometry (MS/MS) for identifying peptides in complex mixtures of digested proteins. Most often, the chromatographic separation is merely used as a mean of reducing complexity of the mixture delivered to the mass spectrometer. Nonetheless, the possibility to use both MS/MS and LC data for peptide identification and sequencing has attracted considerable interest [1–7], given that chromatography provides information about the primary structure which is complementary to the MS data.

Because it can provide high-quality separation for a great variety of chemical species, reverse-phase high performance liquid chromatography (RP-HPLC) is a preferred method for the separation of complex mixtures according to the analyte hydrophobicity and size. In proteomics, RP-HPLC using linear solvent gradients with aqueous/organic mobile phases is by far the most frequent, and provides superior results for proteins and peptides separation prior to mass spectrometry (MS). In these applications, the need to balance LC separation efficiency and MS detection requirements restricts the use of mobile phases. For the same reasons, columns for proteomics application have to conform to strict quality criteria since the weakly buffered mobile phases used can contribute to poor peak shape if free silanols or residual metals are present. It is therefore not surprising to see similar chromatographic methods in most reports on proteomics, with linear gradients from water to acetonitrile, using formic or acetic acid as ion pairing reagent, and separations at room temperature. Nevertheless, differences in RP-HPLC protocols published in the proteomics literature include changes in gradient steepness, flow rate, and column parameters such as length, diameter, particles diameter and pores size. This results in different observed retention times (RT) measured for the same species in different research laboratories.

A noted trend in proteome analyses is the increase in processing of data content during LC–MS/MS experiments, which are compiled into continuously updated databases. In particular, high throughput methodologies relying on Accurate Mass and retention Time (AMT) measurements are increasingly gaining momentum [8–11]. It is worth mentioning that identification database compilation is a labor-, sample-, and time-consuming task, which has to be repeated in each laboratory working on a given proteome due to the specificity of the measured RT. It would undoubtedly be very beneficial to translate these databases across laboratories working on the same biological material. In addition, there is growing awareness, in the proteomics community, of the need to provide means to fairly

* Corresponding author.
E-mail address: christophe.masselon@cea.fr (C.D. Masselon).

1570-0232/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2008.12.047

compare data obtained across laboratories working on different instrumental platforms and using slightly different analytical protocols.

In "Shotgun" proteomics, the collected mass spectrometric data possess their absolute "universal" value, derived from atomic mass tables; but LC data, in the form of RT, are not tied to a conventional time scale, and may vary depending on the separation protocols used (gradient profile, flow rate, mobile phase composition), types of LC columns (column size, pore and particle sizes, adsorbent type, manufacturer), as well as the HPLC instrument. This can make the translation of identification databases problematic. Such a translation requires that a simple relationship between RT in different conditions be sought. In other words, there is a need for standardization of the RT obtained under particular experimental conditions, i.e. the introduction of a relative RT scale independent of the LC protocols, systems, or conditions used.

Previous efforts to implement standardization procedures for LC data mostly focused on ways to improve reproducibility of RT measurement on a particular instrumental setup using an established protocol [3,12,16,27]. The main idea behind these approaches was that fixing the LC protocol does not prevent retention time scattering between different HPLC runs for identical samples, because of column aging or variations in mobile phase preparation, etc.... Therefore, standardization becomes essential for data comparison especially for complex samples, as encountered in proteomics. Most standardization methods to date were based on the usage of an internal or external reference, or standard. One of the assumptions is that LC data scale linearly in day-to-day runs on a given instrument and for a given LC protocol (gradient profile, flow rate, etc.). Within this assumption, a standard sample can be used to obtain "relative retention time" using the following equation: $t_i = \mathrm{RT}_{i,\mathrm{exp}}/\mathrm{RT}_{st}$, where $\mathrm{RT}_{i,\mathrm{exp}}$ is the retention time of the sample compound, and RT_{st} the retention time of the internal standard. Note that this simple approach is limited to data obtained using the same LC protocol; and a change in gradient slope, for instance, may result in different relative retention times for the same compound. A more sophisticated standardization procedure using external standards was suggested by Sapirstein et al. [12] who proposed to use as standards several selected peaks from a specific protein sample which was analyzed before and after the sample of interest. The RTs of these peaks were then used as anchor points in a piecewise calibration algorithm to normalize the chromatograms of samples run in the interval between two of the standards. The proposed normalization algorithm demonstrated a fivefold improvement in the precision of chromatographic data over a period of several months of data collection. In another work, Petritis et al. [3], have proposed to use a Genetic Algorithm for normalization, which was set to optimize two variables of a linear equation, $y = ax + b$. The variable a normalized the gradient slope, and the variable b normalized the LC run start time (dead volumes, delay time, etc.). The optimization of these variables was performed for each separation and the normalization of RT into a 0–1 range and was based on 6 peptides chosen as calibration standard which were specific for the proteomes under study. Over the course of many experiments, the RT normalized using this procedure deviated from the mean by about 1% for the identified peptides. It is assumed that LC conditions in these experiments were the same or at least similar. In summary, previous efforts dealing with peptide RT standardization ranged from very simple to highly sophisticated. It is of particular significance that all these attempts were limited in scope to the effect of an unwanted change in LC separation on RT and offered time scale tied up to specific calibration standards separated under fixed LC conditions. In addition, most authors referred, at least implicitly, to a linear relationship between the measured RT [3,8,12,16,27].

When expanding the scope of standardization methods to deliberate changes in separation conditions, the first problem one is faced with is the question of the reference: could one measured retention time constitute a reliable reference for all further measurements, calibrations and alignments? When comparing two or more runs obtained under identical or similar LC conditions, the choice is not so crucial. However, when multiple datasets acquired under variable conditions are to be brought to the same scale, it becomes important to carefully choose what to align with. It is clear that simply choosing an experimental dataset as a reference is not only arbitrary, but risky, since this particular dataset can be prone to errors in RT estimations. One way to deal with this issue has been independently proposed by McIntosh and co-workers [8] and by us [13]: it consists in the conversion of experimental RT values into a scale corresponding to an intrinsic property of the peptide sequence.

McIntosh et al. suggestion was based on linking peptide LC data with their predicted hydrophobicity values. Using peptides identified with high confidence, they estimated the parameters of a linear equation relating hydrophobicities with RT for a particular experiment. The RT normalization was performed using the Sequence Specific Retention Calculator (SSRCalc) [4], an RT prediction algorithm. In the underlying model of SSRCalc, peptides relative hydrophobicities are assumed to be proportional to RT. These authors claimed the independence of normalized RT on the separation conditions (e.g. the gradient slopes) to combine data from multiple different LC configurations into a single AMT database.

In the present work, we assess the feasibility of LC data standardization using a normalized RT scale tied up with aminoacid interaction energies, using a model introduced by Gorshkov et al [14]. This model of peptide separation is based on the Liquid Chromatography at Critical Conditions applied to biomolecules (BioLCCC) [6–7,15]. It takes into account exclusion effects during peptide separation and the corresponding normalized RT scale is considered sequence specific and generally independent of the LC protocols. Due to the fact that only a few phenomenological parameters are used in the model (determined from the number of aminoacid residues and C- and N- terminal groups) it can be easily adapted for a large variety of solid and mobile phases.

The key issue of RT standardization using predicted properties of peptide sequences is the assumption of linear correlation between experimental retention times acquired under different separation conditions. In the present work, following previous evidence by Casal et al [16], we tested this assumption for a range of experimental parameters such as columns parameters, mobile phase compositions and gradient slope typically used in proteomics experiments [17,18].

Finally, we demonstrate an approach for standardization of peptide RT by conversion of measured values to a standard scale, independent of the instrument or method used. While any of the sequence-dependent RT prediction algorithms [3–4,6–7,19–21] can be used for the purpose of this work, we have selected the additive model pioneered by Meek [22] and recently refined by Krokhin et al. [4], and the BioLCCC model proposed by Gorshkov et al [6]. Both models performed equally well.

2. Experimental

Cytochrome c digest and 6 protein digest were purchased from Dionex/LCPacking (Dionex, Amsterdam, Netherlands) and used as recommended. After a careful analysis of MS/MS data, we found that the molecular structures of two peptides differed from the sequences specified in the Dionex data sheet: IFVQKCAQCHTVEK should be designated correctly as CAQCHTVERL + heme, and KGEREDLIAYLK as GEREDLIAYLKK. The *Cytochrome c* peptides used as retention time calibrants are recapitulated in Table 3. The 6 protein digest standard includes *Cytochrome c*, lysozyme,