Chapter 7
Applications of Peptide Retention Time in Proteomic Data Analysis

Chen Shao

Abstract In proteomic studies, liquid chromatography is commonly used to separate peptide mixtures prior to mass spectrometry (MS) detection. As an independent dimension of information from the information provided by the MS, peptide retention time information has been proven to be able to aid proteomic data analysis in many aspects. So far, some popular software has offered options for this information for MS data acquisition and analysis. This chapter is a brief review of current methodologies of retention time prediction and application in proteomic analysis.

Keywords Retention time · Peptide identification · Quality control

7.1 Retention Time Prediction

A peptide's retention time (RT) is defined as the length of time elapsed from the injection of a sample into the chromatography system to the detection of peak maximum of a peptide. It depends on its chemical structures of peptides, along with the interaction between the environment (mobile and stationary phase, temperature, pH, etc.). Therefore, peptide RTs in a particular liquid chromatography (LC) condition can be predicted based on chemical structure-related properties of peptides, such as amino acid composition, sequence, hydrophobicity, and other physicochemical properties [1].

The task of RT prediction is to calculate a retention scale for each peptide in the given LC condition, e.g., to calculate the hydrophobicity scale in reverse-phase LC. A simple idea is to measure or predict retention coefficients for individual amino

C. Shao (✉)
National Key Laboratory of Medical Molecular Biology, Department of Pathophysiology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, 5 Dong Dan San Tiao, Beijing, China
e-mail: scshaochen@126.com

© Springer Science+Business Media Dordrecht 2015

Y. Gao (ed.), *Urine Proteomics in Kidney Disease Biomarker Discovery*, Advances in Experimental Medicine and Biology 845,
DOI 10.1007/978-94-017-9523-4_7

acids, and then, the retention scale of a peptide is predicted as the sum of retention coefficients of its constituent amino acids. The amino acid retention coefficients can be predicted either by a set of synthetic peptides with residues substituted by each of the twenty amino acids [9] or linear regression models based on peptides with various amino acid compositions [2, 21, 22, 31].

In the recent years, prediction models were refined by employing peptide sequence information and more intelligent computational algorithms, as well as large size of datasets that could prevent the problem of overfitting in data training [16, 27]. N-terminal residues were found to be influence factors to peptides' retention behavior due to the ion-pairing retention mechanism [19]. Taking into account of this effect, Krokhin et al. developed a widely used prediction model, sequence-specific retention calculator (SSRCalc) [16]. This model added a series of sequence-related correction factors to the previous model that predict peptide retention scales by the summation of individual amino acid retention coefficients [9]. Besides three of the N-terminal residues, these correction factors included C-terminal residues, nearest-neighbor effect of charged side chains (Lys, Arg, and His), peptide length, isoelectric point, hydrophobicity, propensity to form helical structures, etc. Another comprehensive model was built by Petritis et al. [27] based on artificial neural network. Similar to SSRCalc, their model embodied peptide properties such as length, sequence, nearest-neighbor amino acids, hydrophobicity, and hydrophobic moment, as well as predicted secondary structures as the input nodes of the neural network. Some other prediction models were developed in similar idea, but with different choices of peptide properties and statistical models [15, 29, 23].

The refined modes improved the prediction accuracy (R^2) significantly from approximately 0.91–0.92 to 0.96–0.98 [17]. However, these conclusions were based on limited size of datasets and reported by the authors themselves. A blind comparison of the most updated versions of prediction models would help greatly in the selection of proper prediction model for practical use. Besides, considering that models based on sequence information and intelligent computational algorithms often require a lot of computational time and large size of training datasets, the simpler and linear prediction models that provide less, but also sufficient prediction accuracy may be selected in some cases, such as on-the-fly RT prediction and calibration [10].

7.2 Application of RT Information in Proteomic Analysis

7.2.1 Peptide Identification Based on LC-MS Data

Accurate mass and time tag (AMT tag) is a well-known strategy to identify peptide sequences based on LC-MS data, which was firstly invented to identify the *Deinococcus radiodurans* proteome [34, 38]. Given the fact that many possible peptide species are unlikely to be detected in a particular biological system, this strategy assumes that peptides that are detectable in a biological system can be separated by a two-dimensional mass and RT vector [44]. Two main steps are included in this

strategy. In the first step, an AMT database for a particular organism or type of biological sample is constructed based on high-confident peptide identifications from previous replicate LC-MS/MS analysis. Secondly, peptides are identified from LC-MS experiments by matching measured mass and normalized elution time (NET) features to the existing database.

There are similar methods that are also identify peptides based on the accurate measurements of mass and RT [11, 24, 41]. These methods do not need to construct a reference database prior to peptide identification. Instead, features are matched by measured mass and RT between different LC-MS/MS runs. Then, peptide identifications from MS/MS spectra can be transferred from one single run to the others. In a study of urinary proteome [25], using "match between runs" option implemented in MaxQuant software [3], the authors were able to increase number of protein identifications from an average of 462 to 633 in a single run.

Saving the effort from MS/MS analysis, AMT tag and similar methods can improve the efficiency and coverage of proteomic analysis. The success of these methods depends on the complexity of biological system as well as the resolution of both MS instruments and LC systems. False discovery rate (FDR) or confidence of peptide identification can be estimated by decoy database searching (shifting masses of all peptides in the AMT database by a certain value) [28] or statistical models [20, 37, 43]. Study of computational simulation showed that for organisms with relative small proteomes, such as *Deinococcus radiodurans*, modest mass and RT accuracies were sufficient for confident peptide identifications by the AMT tag strategy. For more complex proteome, such as human proteome, more strict criteria should be used. The majority of proteins could be uniquely identified within the tolerances of 1 ppm for mass and 0.01 for NET [26].

7.2.2 Peptide Identification from MS/MS Spectra

RT information has been used to improve peptide identification from MS/MS spectra in several ways. One strategy is to incorporate RT information into a discriminant function along with other peptide-spectrum matching parameters, such as SEQUEST scores [39]. This discriminate function was trained based on data from a known protein mixture. When applying to human plasma proteome analysis, it achieved a 16 % increase of positive peptide identifications.

Predicted RT information can serve as a validation parameter for peptide identification results generated by database searching programs. Kawakami et al. [12] validated peptide identifications by the correlation between measured and predicted RTs. Peptide identifications within a certain correlation tolerance were accepted as high-confident identifications. Several studies reported that number of true positive peptides increased significantly by the combination use of RT filter and lower threshold of database searching score [15, 29, 33].

Besides the application of predicted RT information, Sun et al. built up an empirical RT database based on high-confident peptide identifications from

repeated LC-MS/MS runs of a urine sample [40]. This database was used to validate MS/MS identifications for new urine samples. The bottleneck of the empirical database method is that it can only be applied to peptides that were previously detected in a particular proteome, whereas every peptide sequence can have a predicted RT value. However, this method still has its value because it avoids the problem of incorrect RT prediction, which is evitable due to the complex nature of peptide retention behavior.

7.2.3 Post-translational Modification Identification

PTM on a peptide alters not only its molecular mass, but also its physicochemical property (e.g., hydrophobicity), resulting in RT shifts. The RT difference between modified and unmodified peptide (ΔRT) provides a new dimension of information in additional to mass shift (ΔM) in PTM identification.

Previous studies reported lots of instances that peptides with different modification types or different modification sites elute in different RTs [4, 13, 32, 42]. Zybailov et al. [45] depicted the ΔRT distributions of dozens of modification forms detected in a plant proteome. They found that the direction of RT shifts correlated well with the hydrophobicity shifts of the modified peptides for the majority of modifications. Combination of ΔRT and ΔM constrains can efficiently reduce the FDR in PTM identification [32], especially for studies on low-resolution mass spectrometers. For example, deamidation of a peptide results in a mass shift of only 0.984 Da, which could not be accurately distinguished from its unmodified form by a low-resolution LCQ mass analyzer. A study [4] based on synthetic peptide pairs observed that deamidated peptides elute about 3 min later than the corresponding unmodified forms in RPLC. Deamidation detection accuracy was improved from 42 to over 93 % by filtering original SEQUEST identifications by both ΔRT and ΔM constrains.

ΔRT information was also used to improve the algorithms for fast search of unrestricted modifications. The Delta Accurate Mass and Time (DeltAMT) algorithm [7] calculates a two-dimensional delta vector (ΔM, ΔRT) for each pair of spectra obtained in a LC-MS/MS run. The whole set of spectrum pairs are composed of two classes, those from modified and unmodified forms of the same peptide and those from two unrelated peptides. Thus, there are two classes of delta vectors, modification-induced ones and random-induced ones. Bivariate Gaussian mixture models are employed to discriminate modification-induced distributions from random ones. Then, putative modifications could be identified and reported with (ΔM, ΔRT) information as well as the putative modified and unmodified spectrum pairs. Since this algorithm does not use any fragment ion information from MS/MS spectra, it is able to find out high-confident modifications in a very fast speed. However, this algorithm is limited to high abundant modifications, since vector distributions of low abundant modifications are not usually distinguishable from random ones.

7.2.4 Time-scheduled Targeted Proteomic Analysis

Multiple reaction monitoring (MRM) is the method of choice in targeted proteomics. It is a highly sensitive method for accurate quantitation of low abundance proteins in complex protein mixtures. This method needs a sufficient dwell time for each transition to maintain sensitivity and a reasonable cycle time to ensure accurate quantitation. Thus, only a limited number of transitions can be measured in each cycle, limiting its throughput [30]. Time-scheduled transition acquisition (tMRM) offers a solution that can remarkably increase the throughput of traditional MRM experiment without compromising its performance. In this method, the whole gradient time is split into small time windows, and transitions are monitored only in selected windows centered around the expected RT of peptides. Thus, with the same dwell time setting and number of transitions monitored in each duty cycle, tMRM is able to measure many times of transitions in the whole gradient time [36].

A key point to the success of tMRM is to define proper RT window that can capture the entire peptide elution profile from baseline to baseline. This depends on accurate prediction of peptide RTs for each injection. In spite of strict control of the LC system, RT shifts between injections are inevitable, especially when experiments lasting for days to weeks to analysis large amounts of samples. To fit in with the RT shifts, predefined RT windows need to be regularly corrected or repredicted, reducing the efficiency and robustness of tMRM experiment. To aid this situation, on-the-fly RT calibration methods have been developed and integrated in the instrument operating software [8, 14].

This method makes use of a set of well-characterized landmark peptides to calibrate RTs of targeted peptides. Landmark peptides could be either spiked-in synthetic peptides [6, 8] or endogenous peptides that distribute in a broad range of the whole gradient. At any time point, RT windows of subsequent targeted peptides are adjusted based on a local linear regression model generated by the last two eluted landmark peptides. RT windows of peptides elute between the first and second landmark peptides can be simply adjusted by RT shift of the first landmark peptide to calibrate the difference in dead volume. Broad RT windows are set for all landmark peptides as well as peptides elute before the third landmark peptide to ensure that they can be captured without or with minimal calibration.

This method achieved over 90 % success rates on analyses of 180 targeted peptides in a gradient from 0.5 to 2 % solvent B per minute, as well as a nonlinear gradient [8]. It could also precisely correct RT shifts caused by other factors such as change of loading amounts of samples [6] and different LC columns [14]. This method significantly increases the robustness of the entire tMRM workflow by compensating for several commonly occurred changes in experimental conditions, reducing the requirement of LC reproducibility in analysis. Researchers can be rescued from offline RT calibration of LC system and refinement of RT prediction models, saving experimental time, and importantly, precious biological samples.

7.3 Discussion and Perspective

It has been well proven that using RT information could benefit proteomic data analysis. However, its application in practical proteomic analysis has so far been restricted because RT information is of lower resolution compared to mass information, and importantly, peptide RT alters in different LC conditions. Krokhin and colleagues addressed this issue by optimizing their SSRCalc prediction model by four popularly used LC conditions in proteomics. These LC conditions are 300 Å-TFA, 100 Å-TFA, 100 Å-formic acid, and 100 Å-pH 10 [5, 16, 18]. However, since there are hundreds of choices of mobile and stationary phases and other LC parameters in practice, it is an impossible task to pretention retention scales for all LC conditions. A more flexible solution is to train and test the prediction model in the same LC run [15]. Theoretically, this solution is able to adapt all LC conditions. The limitation of this solution is that it needs a sufficient set of high-confident peptide identifications for model training, which is not always available in a single LC run. Another prediction model, ELUDE, is the combination of the above two solutions [23]. When sufficient data are available, ELUDE derives a new RT index for the condition at hand; otherwise, it selects and calibrates a pretrained model from a library of predictors. Model selection and calibration processes are performed automatically by robust statistical methods in ELUDE, facilitating its practical use. However, it should be noted that the accuracy and efficiency of all prediction models are still needed to be tested blindly by datasets covering a great variety of LC conditions.

LC alignment is another important technology in this field. Slight changes of LC conditions and inevitable RT shifts between LC runs can be adjusted by this technology [8]. A recent review of LC alignment methods can be found at [35]. A good idea is to employ a set of spiked-in synthetic peptides as landmarks for LC alignment or to correlate predicted retention scales and measured RTs for each run. These peptides are designed to span a wide range of hydrophobicity, allowing accurate alignment for the entire LC profile. For example, six synthetic peptides were employed to optimize the SSRCalc model in different LC conditions (2009); the eleven iRT standard peptides were used for on-the-fly RT calibration in tMRM analysis [6].

To use RT information as a parameter in data analysis, a proper tolerance value or window size should be set up firstly. This depends on the experimental reproducibility heavily. The wider the RT window is, the more false positives would be achieved. Therefore, there is also an urgent need to set up standards and quality control methods for LC experiments. With the joint effort of bioinformaticists and experimental biologists, RT information would be widely used in practical proteomic analysis in the near future.

References

1. Baczek T, Kaliszan R (2009) Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics. Proteomics 9:835–847
2. Browne CA, Bennett HPJ, Solomon S (1982) The isolation of peptides by high-performance liquid chromatography using predicted elution positions. Anal Biochem 124:201–208
3. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p. b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26:1367–1372
4. Dasari S, Wilmarth PA, Rustvold DL, Riviere MA, Nagalla SR, David LL (2007) Reliable detection of deamidated peptides from lens crystallin proteins using changes in reversed-phase elution times and parent ion masses. J Proteome Res 6:3819–3826
5. Dwivedi RC, Spicer V, Harder M, Antonovici M, Ens W, Standing KG, Wilkins JA, Krokhin OV (2008) Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. Anal Chem 80:7036–7042
6. Escher C, Reiter L, MacLean B, Ossola R, Herzog F, Chilton J, MacCoss MJ, Rinner O (2012) Using iRT, a normalized retention time for more targeted measurement of peptides. Proteomics 12:1111–1121
7. Fu Y, Xiu LY, Jia W, Ye D, Sun RX, Qian XH, He SM (2011) DeltAMT: a statistical algorithm for fast detection of protein modifications from LC-MS/MS data. Mol Cell Proteomics 10(5):M110–000455
8. Gallien S, Peterman S, Kiyonami R, Souady J, Duriez E, Schoen A, Domon B (2012) Highly multiplexed targeted proteomics using precise control of peptide retention time. Proteomics 12:1122–1133
9. Guo D, Mant CT, Taneja AK, Parker JMR, Hodges RS (1986) Prediction of peptide retention times in reversed-phase highperformance liquid chromatography I. Determination of retention coefficients of amino acid residues of model synthetic peptides. J Chromatogr 359:499–517
10. Henneman AA, Palmblad M (2013) Retention time prediction and protein identification. Methods Mol Biol 1007:101–118
11. Jaffe JD, Mani DR, Leptos KC, Church GM, Gillette MA, Carr SA (2006) PEPPeR, a platform for experimental proteomic pattern recognition. Mol Cell Proteomics 5:1927–1941
12. Kawakami T, Tateishi K, Yamano Y, Ishikawa T, Kuroki K, Nishimura T (2005) Protein identification from product ion spectra of peptides validated by correlation between measured and predicted elution times in liquid chromatography/mass spectrometry. Proteomics 5:856–864
13. Kim J, Petritis K, Shen Y, Camp DG 2nd, Moore RJ, Smith RD (2007) Phosphopeptide elution times in reversed-phase liquid chromatography. J Chromatogr A 1172:9–18
14. Kiyonami R, Schoen A, Zabrouskov V (2010) On-the-Fly retention time shift correction for multiple targeted peptide quantification by LC-MS/MS. Thermo Fisher Scientific Application note: 503
15. Klammer AA, Yi X, MacCoss MJ, Noble WS (2007) Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. Anal Chem 79:6111–6118
16. Krokhin OV (2006) Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. Anal Chem 78:7785–7795
17. Krokhin OV (2012) Peptide retention prediction in reversed-phase chromatography: proteomic applications. Expert Rev Proteomics 9:1–4
18. Krokhin OV, Spicer V (2009) Peptide retention standards and hydrophobicity indexes in reversed-phase high-performance liquid chromatography of peptides. Anal Chem 81:9522–9530

19. Krokhin OV, Craig R, Spicer V, Ens W, Standing KG, Beavis RC, Wilkins JA (2004) An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. Mol Cell Proteomics 3:908–919
20. May D, Liu Y, Law W, Fitzgibbon M, Wang H, Hanash S, McIntosh M (2008) Peptide sequence confidence in accurate mass and time analysis and its use in complex proteomics experiments. J Proteome Res 7:5148–5156
21. Meek JL (1980) Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. Proc Natl Acad Sci USA 77:1632–1636
22. Meek JL, Rossetti ZL (1981) Factors affecting retention and resolution of peptides in high-performance liquid chromatography. J Chromatogr 211:15–28
23. Moruz L, Tomazela D, Kall L (2010) Training, selection, and robust calibration of retention time models for targeted proteomics. J Proteome Res 9:5209–5216
24. Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, Brusniak MY, Vitek O, Aebersold R, Muller M (2007) SuperHirn—a novel tool for high resolution LC-MS-based peptide/protein profiling. Proteomics 7:3470–3480
25. Nagaraj N, Mann M (2011) Quantitative analysis of the intra- and inter-individual variability of the normal urinary proteome. J Proteome Res 10:637–645
26. Norbeck AD, Monroe ME, Adkins JN, Anderson KK, Daly DS, Smith RD (2005) The utility of accurate mass and LC elution time information in the analysis of complex proteomes. J Am Soc Mass Spectrom 16:1239–1249
27. Petritis K, Kangas LJ, Yan B, Monroe ME, Strittmatter EF, Qian WJ, Adkins JN, Moore RJ, Xu Y, Lipton MS, Camp DG 2nd, Smith RD (2006) Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. Anal Chem 78:5026–5039
28. Petyuk VA, Qian WJ, Chin MH, Wang H, Livesay EA, Monroe ME, Adkins JN, Jaitly N, Anderson DJ, Camp DG 2nd, Smith DJ, Smith RD (2007) Spatial mapping of protein abundances in the mouse brain by voxelation integrated with high-throughput liquid chromatography-mass spectrometry. Genome Res 17:328–336
29. Pfeifer N, Leinenbach A, Huber CG, Kohlbacher O (2009) Improving peptide identification in proteome analysis by a two-dimensional retention time filtering approach. J Proteome Res 8:4109–4115
30. Picotti P, Aebersold R (2012) Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods 9:555–566
31. Sakamoto Y, Kawakami N, Sasagawa T (1988) Prediction of peptide retention times. J Chromatogr 442:69–79
32. Savitski MM, Nielsen ML, Zubarev RA (2006) ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. Mol Cell Proteomics 5:935–948
33. Shen Y, Kim J, Strittmatter EF, Jacobs JM, Camp DG 2nd, Fang R, Tolie N, Moore RJ, Smith RD (2005) Characterization of the human blood plasma proteome. Proteomics 5:4034–4045
34. Smith RD, Anderson GA, Lipton MS, Pasa-Tolic L, Shen Y, Conrads TP, Veenstra TD, Udseth HR (2002) An accurate mass tag strategy proteome measurements. Proteomics 2:513–523
35. Smith R, Ventura D, Prince JT (2013) LC-MS alignment in theory and practice: a comprehensive algorithmic review. Brief Bioinform doi: 10.1093/bib/bbt080
36. Stahl-Zeng J, Lange V, Ossola R, Eckhardt K, Krek W, Aebersold R, Domon B (2007) High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. Mol Cell Proteomics 6:1809–1817
37. Stanley JR, Adkins JN, Slysz GW, Monroe ME, Purvine SO, Karpievitch YV, Anderson GA, Smith RD, Dabney AR (2011) A statistical method for assessing peptide identification confidence in accurate mass and time tag proteomics. Anal Chem 83:6135–6140

38. Strittmatter EF, Ferguson PL, Tang K, Smith RD (2003) Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. J Am Soc Mass Spectrom 14:980–991
39. Strittmatter EF, Kangas LJ, Petritis K, Mottaz HM, Anderson GA, Shen Y, Jacobs JM, Camp DG 2nd, Smith RD (2004) Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. J Proteome Res 3:760–769
40. Sun W, Zhang L, Yang R, Shao C, Zhang Z, Gao Y (2009) Improving peptide identification using an empirical peptide retention time database. Rapid Commun Mass Spectrom 23:109–118
41. Tolmachev AV, Monroe ME, Purvine SO, Moore RJ, Jaitly N, Adkins JN, Anderson GA, Smith RD (2008) Characterization of strategies for obtaining confident identifications in bottom-up proteomics measurements using hybrid FTMS instruments. Anal Chem 80:8514–8525
42. Xie H, Gilar M, Gebler JC (2009) Characterization of protein impurities and site-specific modifications using peptide mapping with liquid chromatography and data independent acquisition mass spectrometry. Anal Chem 81:5699–5708
43. Yanofsky CM, Kearney RE, Lesimple S, Bergeron JJ, Boismenu D, Carrillo B, Bell AW (2008) A Bayesian approach to peptide identification using accurate mass and time tags from LC-FTICR-MS proteomics experiments. Conf Proc IEEE Eng Med Biol Soc 2008:3775–3778
44. Zimmer JS, Monroe ME, Qian WJ, Smith RD (2006) Advances in proteomics data analysis and display using an accurate mass and time tag approach. Mass Spectrom Rev 25:450–482
45. Zybailov B, Sun Q, van Wijk KJ (2009) Workflow for large scale detection and validation of peptide modifications by RPLC-LTQ-Orbitrap: application to the Arabidopsis thaliana leaf proteome and an online modified peptide library. Anal Chem 81:8015–8024