Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot

RT-SVR+q: A strategy for post-mascot analysis using retention time and q value metric to improve peptide and protein identifications

Weifeng Cao[a], *Di Ma*[b], *Arvinder Kapur*[c], *Manish S. Patankar*[c], *Yadi Ma*[d], *Lingjun Li*[a,b,*]

[a]*Department of Chemistry, University of Wisconsin-Madison, 777 Highland Ave., Madison, WI 53705, United States*
[b]*School of Pharmacy, University of Wisconsin-Madison, 777 Highland Ave., Madison, WI 53705, United States*
[c]*Department of Obstetrics and Gynecology, University of Wisconsin-Madison, 777 Highland Ave., Madison, WI 53705, United States*
[d]*Department of Computer Sciences, University of Wisconsin-Madison, 777 Highland Ave., Madison, WI 53705, United States*

ARTICLE INFO

Article history:
Received 11 April 2011
Accepted 17 August 2011
Available online 24 August 2011

Keywords:
Shotgun proteomics
Database search
Support vector regressor
Retention time
q Value
Peptide identification

ABSTRACT

Shotgun proteomics commonly utilizes database search like Mascot to identify proteins from tandem MS/MS spectra. False discovery rate (FDR) is often used to assess the confidence of peptide identifications. However, a widely accepted FDR of 1% sacrifices the sensitivity of peptide identification while improving the accuracy. This article details a machine learning approach combining retention time based support vector regressor (RT-SVR) with q value based statistical analysis to improve peptide and protein identifications with high sensitivity and accuracy. The use of confident peptide identifications as training examples and careful feature selection ensures high R values (>0.900) for all models. The application of RT-SVR model on Mascot results (p=0.10) increases the sensitivity of peptide identifications. q Value, as a function of deviation between predicted and experimental RTs (ΔRT), is used to assess the significance of peptide identifications. We demonstrate that the peptide and protein identifications increase by up to 89.4% and 83.5%, respectively, for a specified q value of 0.01 when applying the method to proteomic analysis of the natural killer leukemia cell line (NKL). This study establishes an effective methodology and provides a platform for profiling confident proteomes in more relevant species as well as a future investigation of accurate protein quantification.

1. Introduction

With the advent of high-throughput mass spectrometry (MS), shotgun proteomics has been employed as a major tool to analyze biological samples and produce thousands of MS or tandem MS spectra [1–3]. The methods of choice for annotating these spectra with sequences are currently database search engines such as Mascot, SEQUEST, OMSSA and X! Tandem [4–7]. The algorithms generally score the similarities between the experimental and theoretical spectra and rank the best match with the highest score as the predicted peptide spectrum match (PSM). However, the top PSM is not necessarily correct due to the scoring scheme and quality of a spectrum. Target/decoy search strategy and the resulting false discovery rate (FDR) calculation are used to assess the confidence of reported PSMs [8]. However, there is a tradeoff

Abbreviations: NKL, natural killer leukemia cell line; RPLC, reversed phase liquid chromatography; MS/MS, tandem mass spectrometry; PSM, peptide spectrum match; RT, retention time; ΔRT, deviation between predicted and experimental RT; SVR, support vector regressor; FDR, false discovery rate; MIT, Mascot identity threshold; MHT, Mascot homology threshold

* *Corresponding author at:* School of Pharmacy, University of Wisconsin-Madison, 777 Highland Ave., Madison, WI 53705, United States.
E-mail addresses: wcao2@wisc.edu (W. Cao), lli@pharmacy.wisc.edu (L. Li).

1874-3919/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jprot.2011.08.013

between sensitivity and accuracy of peptide or protein identifications that FDR has to manage [9,10]. In order to increase sensitivity while maintaining accuracy one can incorporate retention time predictors [11–22] as a post-Mascot analysis tool to increase confidence for peptide identifications.

The retention time (RT) of a peptide is defined as the elapsed time between the time of injection and the time of elution of the peak maximum. Previous studies demonstrated that the retention time of a peptide is the function of various peptide parameters, including amino acid composition [23], N-terminal or C-terminal residues [14], location of amino acids within the primary structure [16], peptide length or mass [12], and hydrophobicity [20]. Many sophisticated models have been constructed to predict retention time and used predicted RT to improve peptide identification. For example, Krokhin et al. [14] trained a linear model (SSRC) by linearly correlating RT with a comprehensive hydrophobicity which integrates residue's hydrophobicity and structural and positional effects. Strittmatter et al. [21] proposed an artificial neural network (ANN) peptide RT prediction model by using positional amino acid information to yield a 16% increase in peptide identification for a complex sample (human plasma) [16]. Klammer et al. [22] adopted a support vector regressor dynamically trained for each chromatographic run, with which 50% more positive peptide identifications were obtained at a false positive rate of 3%. Although a great deal of effort has been made in this field to improve protein identifications in shotgun proteomics, there are still some challenging issues to be addressed. The comparison of the previously published models indicates that the static linear model depends on the chromatographic condition and thus prediction bias would occur when the model is used for a different condition. ANN model needs an extremely large dataset (~345,000 training examples) [21] that is often impractical for application. Dynamic SVR model is suitable for relatively small dataset and avoids the RT variation between different chromatographic runs. However, its performance is modest compared to the other two models. In addition, the deviation between predicted and experimental RTs (ΔRT) is favorably used to filter out false positives when applying the trained RT predictor to real data. However, there are limitations in previous approaches to determine a suitable ΔRT threshold. With SSRC [24], the ΔRT threshold was determined by tentatively checking recovery of peptide predictions with varying arbitrary ΔRT values like ±4, ±2, and ±1 min. In [22], optimal ΔRT threshold was selected from a range of ΔRT values (0–240 min) at which the highest numbers of true positives were obtained across the largest number of FDR values (in a range of 0.5–10%). These approaches could lead to under or overestimation of peptide identifications by unsuitable ΔRT threshold. Hence, it is necessary to develop a state-of-the-art RT predictor which can increase the sensitivity to maximize the number of predictions while ensure the accuracy of peptide identifications at the same time.

Given that a dynamic SVR model is more universal and practical for real application, we developed a SVR based RT predictor (RT-SVR) to be used in conjunction with Mascot search results obtained from 2D LC–MS/MS experiments. Our proposed RT-SVR model was constructed with multiple peptide spectral matches (PSM) which were obtained from Mascot search results (at FDR ~ 1%) for each run. When applying the trained RT-SVR model to real data for examining peptide identifications, instead of choosing a ΔRT threshold arbitrarily or trained with a set of FDR, we introduced a method called q value assessment to define a dynamic ΔRT threshold that improves the confidence of evaluation for peptide identifications. By using this statistical method, q value rather than ΔRT is employed as the cut-off criteria to filter out the false positives. q Value metric was first proposed by Storey et al. [25] to analyze genomic data and later on it was revised and applied to MS-based proteomics by Kail et al. [26,27]. The q value can be understood as the minimal FDR at which a peptide spectra match (PSM) can be accepted. In practice, q value can be associated with any PSM in a dataset. Since q value is calculated from all PSMs in a dataset, it is considered as a statistical result for the whole dataset like FDR. Previous studies have shown that q value is equivalent to FDR estimation and no bias will be introduced toward under or overestimation [28,29]. Thus, q value assessment is viewed as a more accurate and reliable estimation of error rate. In our study, a modified q value was calculated based on target and decoy PSMs and then was assigned to a peptide prediction. Finally, we can unambiguously filter out the false positives for a given q value threshold such as 0.01 (Fig. 1). By applying our strategy to proteomic analysis of the natural killer leukemia cell line (NKL), the trained RT-SVR models for all datasets obtained from sample fractions show high performance with R value above 0.900. The peptide and protein identifications increase by up to 89.4% and 83.5% respectively in comparison with Mascot search results (at FDR 1%) with a q value of 0.01. Our results thus demonstrate the utility of the RT-SVR with q value assessment as a robust and reliable method for post-Mascot analysis in proteomic applications.

In addition, we combined RT-SVR and Mascot score screening (Mascot Identity Threshold) to rescue those peptide identifications missed by RT-SVR. This combined RT-SVR method yields more peptide and protein identifications.

In order to evaluate the general applicability of our RT-SVR strategy we applied the model to an independent set of large-scale yeast proteomic data acquired using a Thermo LTQ mass spectrometer (downloaded from PeptideAtlas (PAe001337) and processed by the combined RT-SVR). As a comparison, 566 unique proteins were predicted at a q value of 0.01 in contrast with 470 with Mascot (MIH, FDR 1%) and 499 unique proteins reported by Trans-Proteomic Pipeline (TPP, probability filter 0.010). This result suggests that the RT-SVR model is independent of instruments used for shotgun proteomics and is generally applicable to proteomic data analysis acquired on multiple mass spectrometric platforms.

RT-SVR was written in Java. The windows-based graphical user interface can be freely downloaded from http://pages.cs.wisc.edu/~yadi/bioinfo/rtsvr/rtsvr.html.

2. Materials and methods

2.1. *Sample preparation for proteomic analysis*

10^7 NKL cells were harvested and washed three times with ice-cold PBS. Cells were lysed with 100 μL RIPA lysis buffer (Formulation: 50 mM Tris–HCl (pH7.4), 150 mM NaCl, 0.1%