VA-PRT: A Visualization Tool for Analyzing Post-translational Modification Retention Times

Brenner, Riley
Biostatistics Division
School of Public Health
University of Nevada Reno
Reno, Nevada
rileybrenner@unr.edu

Bertauche, Kurtis
Department of Computer Science
and Engineering
University of Nevada Reno
Reno, Nevada
kbertauche@unr.edu

Choi, Alexander
Biostatistics Division
School of Public Health
University of Nevada Reno
Reno, Nevada
alexanderchoi@unr.edu

Ryu, So Young*
Biostatistics Division
School of Public Health
University of Nevada Reno
Reno, Nevada
soyoungr@unr.edu

***Abstract*—Post-translational modifications (PTMs) play a critical role in many aspects of cell biology including cell growth, differentiation, and survival. Accurate identifications of PTM peptides and their PTM locations are very important in proteomics. The retention time information of PTM peptides can be used to improve the accuracy of PTM identification and experimental proteomic strategies. Therefore, understanding PTM effects on retention times will be beneficial for future research.**

Here, we present VA-PRT (Visualization tool for Analyzing Post-translational modification Retention Times) that systematically investigates the PTM effects on retention times using visualization techniques and advanced statistical methods. The usefulness of VA-PRT is demonstrated using a large synthetic proteomic and phosphoproteomic dataset. VA-PRT is implemented in R and will be freely available for public use.

***Index Terms*—post-translational modifications, mass spectrometry, retention time, visualization, hypothesis tests**

I. Introduction

Post-translational modification (PTM) is a substantial field of study in proteomics research [1]. PTMs have been linked to various cellular processes required for normal function. Study of specific PTM site effects is an ongoing research topic. It has previously been observed that the retention times of PTM peptides differ from their unmodified counterparts [2], [3]. Thus, retention time prediction may improve PTM identifications by reducing their false negatives and false positives. Further study of the retention time of PTM peptides can increase our knowledge about chromatographic behavior of PTM peptides, and may lead to better experimental proteomic strategies.

Here, we introduce a visualization tool for analyzing post-translational modification retention times (VA-PRT). This tool can be used to investigate the various aspects of retention time differences between unmodified peptides and their counterpart PTM peptides using various visualization techniques (e.g., scatterplots, boxplots, non-parametric regressions). It also provides hypothesis test results related to whether the effect of PTM on retention times is statistically significant. In this paper, we demonstrate the usefulness of VA-PRT to detect an effect of phosphorylation, one of the most important PTMs, on peptide retention times.

II. VA-PRT

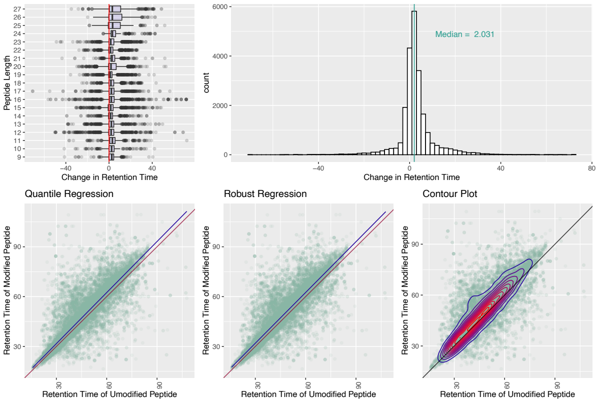


Fig. 1: An example output of VA-PRT. The following visualizations are shown: boxplots of changes in retention times in respect to peptide lengths (top left), a histogram of changes in retention times with a vertical line that represents the median value (top right), and scatterplots between retention times of unmodified peptides vs. their counterpart modified peptides with a quantile regression (bottom left), a robust regression (bottom middle), and a contour plot (bottom right).

VA-PRT is a visualization tool that allows researchers to investigate the effect of PTMs on retention time of peptides (Figure 1). VA-PRT contains several visualization components including but not limited to (i) histograms of changes in retention times between unmodified and PTM peptides at both overall-level and a specific PTM type level, (ii) scatterplots of retention times of unmodified vs. PTM peptides, which can reveal a linear or non-linear trend of PTM effects on retention times at both the overall-level and at specific PTM type level, and (iii) peptide-specific retention time distribution boxplots. VA-PRT also performs PTM specific hypothesis tests about the effect of a specific PTM on retention times using Wilcoxon signed-rank tests.

VA-PRT is implemented in R. The R package will be freely available for public use in the future. Based on user-defined parameters and input files, VA-PRT automatically outputs

The work was supported by NSF CAREER DBI-2046122.

* Corresponding author.

various graphs and hypothesis test results that will be useful for investigating the effect of PTM on retention times. VA-PRT can produce several important visuals using one function (Figure 1), while additionally allowing the user to create individual scatterplots or histograms. The visualizations are saved as pdf files and the hypothesis test results are saved as text files.

VA-PRT contains several functions that allow for visual analysis based on paired modified and unmodified peptides with an option of a filtering step (e.g., setting upper and lower bounds on retention time). The package **ggplot2** is frequently utilized in VA-PRT. Details about major statistical methods used in the VA-PRT are listed as follows:

A. Hypothesis tests

VA-PRT performs non-parametric paired analyses to test the effects of PTMs on retention time. Specifically, it uses a Wilcoxon signed-rank test [4] to investigate whether (paired) retention times between modified and unmodified peptides are significantly different. Noting that the normality of data is often violated in retention times, a Wilcoxon signed-rank test is an appropriate choice for this purpose. VA-PRT reports both mean and median changes of retention times with their corresponding p-values. Small p-values (e.g., less than 0.05) imply that retention times between modified and unmodified peptides are significantly different. In other words, the effects of PTMs on retention time are significant.

B. Robust and Quantile Regressions

Considering outliers and high leverage data points that may be present in the dataset, VA-PRT uses robust regressions and quantile regressions instead of least squares regressions. In robust regression, an M-estimation procedure with a Huber weight function is employed [5], [6]. In the Huber weight function, weights of observations with small residuals are one, while observations with larger residuals have smaller weights. Thus, potential outliers have small weights. In VA-PRT, the function **rlm** from the **MASS** package is used for the robust regression [7].

VA-PRT employs the 50^{th} quantile regression, which is a median regression, to explore a linear trend of data. The regression computes the 50^{th} conditional quantile function of the response variable (e.g., retention times of modified peptides), given the covariate (e.g., retention times of unmodified peptides). In simple terms, a median regression aims to fit a line using the median values of the response variable rather than the mean values. Similar to robust regression, the line of fit is less influenced by outliers than a least square regression approach. The median regression is additionally considered since the distributions of retention times may not be normal, but highly skewed. VA-PRT uses the function **rq** from the **quantreg** package with the modified version of the Barrodale and Roberts algorithm for *ll*-regression [8], [9].

C. Contour Plots

VA-PRT visualizes the distribution of paired retention times of unmodified peptides vs. modified peptides using a 2-dimensional density contour plot. In this visualization, the 2D kernel density estimation is used. This contour plot is useful to explore scatterplots with many overlapped observations. The function **kde2d** from the **MASS** package is employed [7].

D. Local Polynomial Regressions

VA-PRT uses a local polynomial regression to explore a non-linear trend of data. It smooths the response variable (e.g., retention times of modified peptides) as a function of the covariate (e.g., retention times of unmodified peptides). The function **loess** from the **stats** package with default settings is used in VAR-PRT [10]. This approach displays a non-linear trend of data. However, it may not be robust in the presence of outliers.

III. Data

Considering that phosphorylation is one of the most important post-translational modifications, we used a synthetic proteomic and phosphoproteomic dataset which contains $>$100,000 peptide sequences [2] to demonstrate the performance of the proposed visualization tool. The experimental and analysis details are shown in [2]. In brief, mixtures of 96 seed tryptic peptides and their synthesized variants were analyzed by an Orbitrap Velos using a beam-type collision-induced dissociation (HCD) fragmentation method. The data was then analyzed by the Mascot search engine (2.3.1; http://www.matrixscience.com) against the human IPI v3.72 including the sequences of all synthesized peptides.

The data was filtered using an E-value threshold of 0.01. To increase the accuracy of our analysis, only identifications matched to the sequences of synthesized peptides were considered. For each sequence, a retention time associated with the maximum intensity of the corresponding precursor ion at the time it was selected for an MS/MS fragmentation was extracted. VA-PRT then performed retention time matching between modified peptides and the corresponding unmodified peptides and visualization analyses.

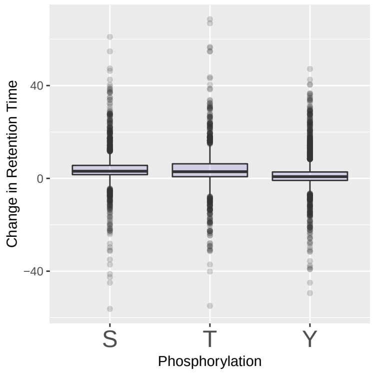
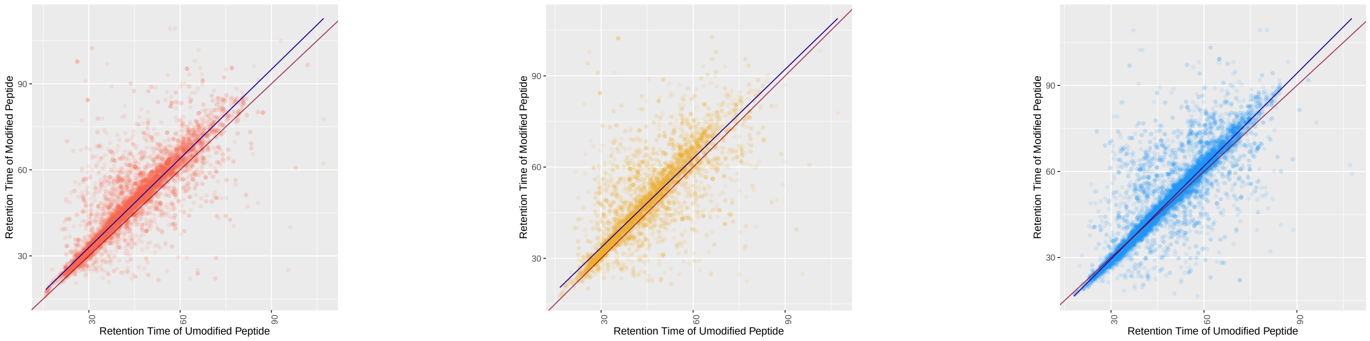


Fig. 2: Boxplots of retention time changes after serine (S), threonine (T), and tyrosine (Y) phosphorylations.



(a) Serine Phosphorylation (S)

(b) Threonine Phosphorylation (T)

(c) Tyrosine Phosphorylation (Y)

Fig. 3: Scatterplots between retention times of unmodified peptides vs. modified peptides with robust regression lines in blue. The red lines represent identify lines (e.g., y=x).

TABLE I: The effects of phosphorylations on retention times.

Post-translational modification type	p-values	Mean change in retention time (min)	Median change in retention time (min)
All phosphorylations (STY)	$< 2.2 * 10^{-16}$	2.80	2.03
Serine Phosphorylation (S)	$< 2.2 * 10^{-16}$	3.90	3.02
Threonine Phosphorylation (T)	$< 2.2 * 10^{-16}$	3.56	2.86
Tyrosine Phosphorylation (Y)	$< 2.2 * 10^{-16}$	1.58	0.70

IV. Results and Discussions

We explored the effect of phosphorylation on retention times using VA-PRT. Table I displayed the mean and median effects of phosphorylation on retention times and Wilcoxon signed-rank test results that investigated the effects of phosphorylation on retention times. The overall effect of phosphorylation on retention times as well as individual effects of serine, threonine, and tyrosine phosphorylations on retention times were significant at a 99% confidence level (Table I). Phosphorylated peptides generally had longer retention times than their counterpart unmodified peptides. The median changes in retention times between phosphorylated and unmodified peptides were 2.03 minutes. Among serine, threonine, and tyrosine phosphorylations, serine phosphorylation had the most effects on retention time with 3.02 median changes in retention times (Figure 2). The peptide lengths were also significantly associated with changes in retention times between phosphorylated peptides and unmodified peptides with a p-value of $6.68*10^{-10}$ based on a linear regression analysis. Furthermore, we explored the changes in effects of phosphorylation on retention time with scatterplots and robust regression lines. As expected, when retention times of unmodified peptides increased, retention times of the corresponding phosphorylated peptides increased (Figure 3). However, robust regression lines were not parallel to identity lines, especially for serine- and tyrosine-phosphorylated peptides. The slopes of robust regression lines for serine- and tyrosine-phosphorylated peptides were larger than the identity lines. This implied that the effects of serine and tyrosine phosphorylations were not constant. The effects of serine and tyrosine phosphorylations on retention times became larger for peptides that eluted later. More investigations using other phosphoproteomic datasets would be beneficial to confirm the observed effects.

V. Conclusions

In this paper, we demonstrated the functions of VA-PRT and its usefulness using a synthetic proteomic and phosphoproteomic dataset. We anticipate that VA-PRT will aid researchers in comprehensively investigating the effect of various types of PTMs on retention times. With the help of VA-PRT, researchers will be capable of diagnosing how PTM affects retention time as well as identifying important factors (e.g., PTM type, peptide length, interaction effect) that may influence retention times of peptides.

References

[1] Liebler, DC. 2002. Introduction to proteomics: tools for the new biology. New Jersey, USA: Human Press Inc.

[2] Marx, H., Lemeer, S., Schliep, J.E., Matheron, L., Mohammed, S., Cox, J., Mann, M., Heck, A.J. and Kuster, B., 2013. A large synthetic peptide and phosphopeptide reference library for mass spectrometry–based proteomics. Nature biotechnology, 31(6), pp.557-564.

[3] Steen, H., Jebanathirajah, J.A., Rush, J., Morrice, N. and Kirschner, M.W., 2006. Phosphorylation analysis by mass spectrometry: myths, facts, and the consequences for qualitative and quantitative measurements. Molecular & Cellular Proteomics, 5(1), pp.172-181.

[4] Hollander, M., and Wolfe, D.A., 1973. Nonparametric statistical methods. John Wiley & Sons, pp.22-33, 68-75.

[5] Huber, P.J., 1981. Robust statistics. Wiley, 308p.

[6] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A., 2011. Robust statistics: the approach based on influence functions. Wiley.

[7] Venables, W.N. and Ripley, B.D., 2002. Modern applied statistics with S. Springer, New York, NY, pp. 271-300.

[8] Koenker, R. and d'Orey, V., 1994. Remark AS R92: A remark on algorithm AS 229: Computing dual regression quantiles and regression rank scores. Journal of the Royal Statistical Society. Series C (Applied Statistics), 43(2), pp.410-414.

[9] Koenker, R., 2005. Quantile regression Cambridge University Press New York.

[10] Cleveland, W.S., Grosse, E. and Shyu, W.M., 1992. Local regression models. Chapter 8 in Statistical models in S (JM Chambers and TJ Hastie eds.). Wadsworth & Brooks/Cole, Pacific Grove, CA.