Application of the Cox Proportional Hazards Model for the Quantitative Analysis of LC-MS Proteomics Data

Ivan Arreola and David Han

Department of Management Science and Statistics, University of Texas at San Antonio, TX

ABSTRACT

Along with quantitative, analytical genomics, proteomics continues to be a growing field for determining the gene and cellular functions at the protein level. As the liquid chromatography mass spectrometry (LC-MS) experiments produce protein peak intensities data, statistical and computational techniques are required to conduct quantitative, analytical proteomics. The LC-MS proteomics data often have large quantities of missing peak intensities due to censoring of the low-abundance spectral features. Because of this, the observed peak intensities from the LC-MS method are all positive, skewed, and often left-censored. The classical survival analysis methods are ideal to detect differentially expressed proteins among different groups. These methods include the non-parametric rank sum (RS) tests such as the Kolmogorov-Smirnov (KS) and Wilcoxon-Mann-Whitney (WMW) tests, parametric survival models such as the accelerated failure time (AFT) model with popular lifetime distributions; log-normal (LN), log-logistic (LL), and Weibull (W) for modeling the peak intensity data. As an alternative approach, here we propose the Cox proportional hazards (PH) method, a popular semi-parametric model for modeling survival data. The proposed regression-based method allows for leniency on the hazard function by alleviating the requirements of distribution-specific hazard functions. With the hopes of gaining more insightful biological information for cellular functions at the protein level, the statistical properties of each method are investigated through a simulation study and an application to the Type I diabetes dataset.

***Keywords*:** Accelerated Failure Time Model, Cox Proportional Hazards Model, Liquid Chromatography Mass Spectrometry Proteomics, Survival Analysis, Type I Diabetes Mellitus

INRODUCTION

In the growing field of biomedical research, the focus has been expanding from DNA sequencing and related problems towards proteomics analyses. Proteomics is a growing area whose purpose is unraveling gene and cellular function at the protein level. Methods used in proteomic related problems aim to assess how encoded gene expressions co-operate in the cell to sustain biological life at the protein level, and how these pathways are disturbed in reaction to infectious diseases and cancers [1].

In the usual case, it is of interest to quantify the myriad of proteins in a given biological sample. There are a number of methods used to quantify the peptides within proteins, including liquid chromatography mass spectrometry (LC-MS), LC-MS/MS and the tandem mass spectrometry approach [2]. In the LC-MS approach, proteins are extracted from the given biological sample, digested into peptides and ionized. After the ionization of proteins, they are then

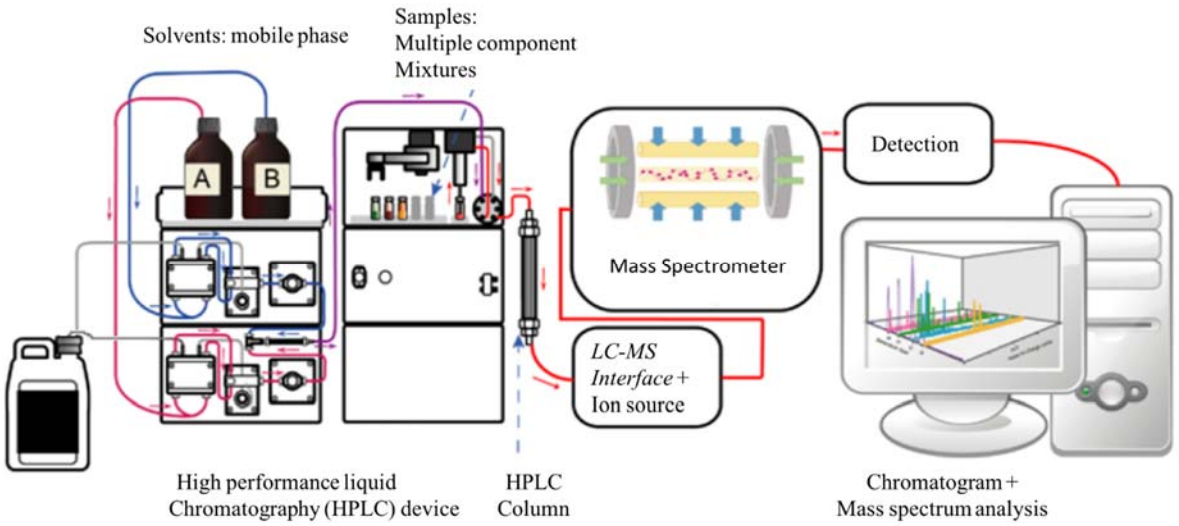

Figure 1. Schematic illustration of the LC-MS procedure [5]

introduced to the mass spectrometry for scanning. This is where mass charge and observed peak intensities are obtained. Then the peak intensities are matched for peptide identification and finally the peptide information is rolled up to the protein level. Then, one measures the amplitude of proteins from the identification step in the sample. Figure 1 above schematically illustrates the procedure of the LC-MS method.

In the quantification step of the analysis, the estimation of the abundance of proteins is carried out in one of the following ways: spectral counts, label-free methods, or isotopic labeling experiments [2]. Spectral counts method utilizes the number of peak intensities for a given peptide/protein while label-free methods such as intensity-based quantitation use peak intensities to assess peptide or protein abundance and other label-free methods are constructed from unlabeled peak intensities corresponding with the mass spectrum of extracted ions. After the quantification of proteins is completed, it is of interest to evaluate differentially expressed proteins. The overall objective of differential expression analysis is to distinguish across groups for biomarker discovery or to provide information for studying the pathways of certain diseases or cancers.

With the LC-MS proteomic data in the context of survival analysis, one often comes across censored observations which are characterized by an asymmetric distribution [3,4]. In order to evaluate differentially expressed proteins, the censored observations are usually transformed, normalized and/or imputed. To compensate for censored observations, various imputation techniques are used including the row mean imputation (RM), *k* nearest neighbor (KNN), and Bayesian principal components analysis [2]. Moreover, the probabilistic principal components analysis (PPCA) is a ubiquitous technique in data analysis and is based on the combination of the expectation maximization (EM) algorithm with a probability model.

Once the data are imputed and transformed, standard statistical techniques such as the two-sample *t*-test or linear regression methods are used to identify differentially expressed proteins. Other techniques to analyze the transformed data include the AFT models with LN, LL and Weibull distributions as well as the WMW and KS tests. It is worth noting that along with the AFT model, the Cox proportional hazards (PH) model is another popular technique to study the time-to-event data.

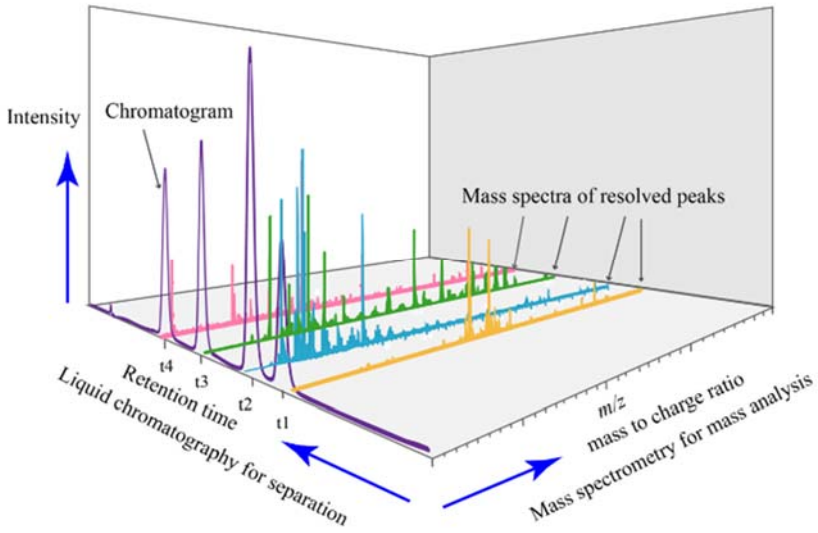

Figure 2. The LC-MS spectrum of each resolved peak [5]

The Cox PH model has been extensively applied in the medical and biostatistics fields while the AFT model has been applied in the industrial problems.

The computational and statistical analyses of the proteomics data are still not mature yet and here we propose the Cox PH model as an alternative way to analyze the proteomics data in the context of survival analysis for obtaining biological insight in the human disease conditions such as diabetes. The performance of the proposed method is compared to that of the AFT model with LN, LL, and Weibull distributions in addition to the rank sum tests with no imputation (NI), RM, KNN, PPCA and lastly the *t*-test with NI, RM, KNN, and PPCA. By comparing these survival analysis methods through a simulation study and an application to a real dataset, we aim to identify the consistent method that can be used not only in the proteomics analyses but also in the wide array of other survival studies.

STATISTICAL METHODS

t-Test A standard statistical method for differential expression analysis is the two-sample *t*-test. Under the assumption of the equal variances and the normally distributed populations, the log_2 transformed peak intensity data $\boldsymbol{X}$ have the following test statistic

$$TS_i = \frac{\bar{X}_{iD} - \bar{X}_{iC}}{S_p\sqrt{2/n}} \quad \text{for } i = 1,2,\dots,M$$

where S_p is the pooled standard deviation, C and D represent the control and diseased groups, respectively. M is the total number of proteins and n is the total number of samples associated with each protein. Under the null hypothesis, TS_i has a *t*-distribution with $2n-2$ degrees of freedom. If the number of comparisons exceeds two, then the *t*-test can be generalized to an *F*-test. However, the assumption of normality could be violated for the transformed data, which results in loss of power in detecting differential expressions among different groups.

Rank Sum Tests Two non-parametric methods such as the Wilcoxon-Mann-Whitney (WMW) and the Kolmogorov Smirnov (KS) rank sum test are considered

for testing the null hypothesis that the distributions of the comparison groups are identical. The non-parametric tests are robust by relaxing the normality assumption and thus enjoy distribution-free characteristics. However, when the normality assumption indeed holds, the non-parametric methods can lose efficiency but they are still more robust when the outliers are present in the dataset.

To elaborate, the KS test evaluates the null hypothesis that the distributions of the comparison groups are identical against the alternative hypothesis that they differ. The hypotheses under consideration are specifically formulated as

$$H_0: F_i(t) = G_i(t)$$

$$H_A: F_i(t) \neq G_i(t) \text{ for at least one } t$$

where t is the observed peak intensity. The goal is to distinguish if there are any differences between the two comparison groups. Let $F_i(\cdot)$ and $G_i(\cdot)$ be continuous distributions for the populations being compared for the i-th protein so that $C_{i1}, C_{i2}, \dots, C_{in} \overset{iid}{\sim} F_i(\cdot)$ and $D_{i1}, D_{i2}, \dots, D_{in} \overset{iid}{\sim} G_i(\cdot)$. Then, we calculate the empirical distributions of $F_{in}(t)$ and $G_{in}(t)$ as

$$F_{in}(t) = \frac{\sum_{j=1}^{n} C_{ij} \leq t}{n}$$

$$G_{in}(t) = \frac{\sum_{j=1}^{n} D_{ij} \leq t}{n}$$

Both $F_{in}(t)$ and $G_{in}(t)$ represent the sum of all peak intensities in the comparison groups (*viz.*, $\sum_{j=1}^{n} C_{ij}$ or $\sum_{j=1}^{n} D_{ij}$) are less than the observed peak intensities t over the number of samples (n) associated with each protein. The KS test statistics J_i for the i-th protein is

$$J_i = \frac{n^2}{d} \max_{-\infty<t<\infty} |F_{in} - G_{in}|$$

where d is the greatest common divisor of n. If we want to test at level α, we compare J_i to j_α, therefore rejecting the null hypothesis if $J_i \geq j_\alpha$. The above test statistic is formulated under the assumption that there are equal samples in both comparison groups but the test can also be generalized to an unequal sample size case.

The WMW rank sum test is a distribution-free two-sample test under the assumption that only the locations of the populations differ. The null hypothesis states that the two distributions do not differ by Δ_i for the i-th protein. In essence, $D_i = C_i + \Delta_i$ where D_i and C_i differ by Δ_i, the location-shift parameter or treatment effect. The alternative hypothesis represents that the two distributions do differ by Δ_i for the i-th protein. Therefore, we have the following hypothesis for the WMW test

$$H_0: \Delta_i = 0$$

$$H_A: \Delta_i \neq 0$$

The WMW rank-sum test statistics W_i is computed by combining the comparison groups, each consisting of size n. Then the combined $2n$ observations is ranked and the test statistic for the i-th protein is derived from the sum of the ranks assigned to the observations from a comparison group in the ordered combined group. Therefore, W_i is

$$W_i = \sum_{j=1}^{n} S_{ij}$$

where S_{ij} is the rank associated with the j-th sample from a selected treatment comparison group of the i-th protein. For this two-sided test, if $W_i \geq \omega_{\alpha/2}$ or if $W_i \leq n(2n+1)\omega_{\alpha/2}$ then W_i is rejected where $\omega_{\alpha/2}$ is a nominal value found from statistical software. Both the KS test and the WMW test have a similar limitation to the two-sample t-test. That is,

they are restricted to two samples and prevent the adjustment of covariates.

AFT Model Under the assumption that the missing data are censored, the AFT model can be applied to compare protein-level expressions across the groups of interests. Let t_{ij} be the observed peak intensity for the *i*-th protein in sample *j* with Z_{ij} being an indicator variable assigning the group membership; 1 if sample *j* is in the treatment group and 0 if it is in the control group. Additionally, $S_i(t|Z_{ij})$ is the survival function defined as the probability that the *i*-th protein peak intensity is greater than some value $t|Z_{ij}$. Then, the AFT model is formulated as

$$S_i\left(t\middle|Z_{ij}\right) = S_{i0}\{\exp(\theta_i Z_{ij})t\}$$

defining the relation between the survival function $S_i\left(t\middle|Z_{ij}\right)$ and the acceleration factor $\exp(\theta_i Z_{ij})$ for the *i*-th protein. Here, S_{i0} is the survival function at the baseline levels of all covariates included in the model. More formally, in the current application, the baseline survival function is the survival function for the control group. The acceleration factor $\exp(\theta_i Z_{ij})$ shows how the survivor function for the *i*-th protein changes from the baseline survival function as the covariate changes. θ_i represents the effect of the *i*-th peak intensity on its predicted survival peak $S_i(t)$. The AFT model can be rewritten in terms of a linear relationship between the log intensities and the group indicator variable because applying the model to peak intensity data, it is assumed that the effects of covariates are multiplicative to the predicted survival function of the peak intensity. Hence, we now have

$$Y_{ij} = \log\left(t_{ij}\right) = \mu_i + \gamma_i Z_{ij} + \sigma_i W_{ij}$$

where μ_i and σ_i are mean and scale parameters in relation to the *i*-th protein while W_{ij} is the error term in the model for the *i*-th protein and the regression coefficient γ_i is the effect of treatment compared to the control group of the log-transformed peak intensity data. The most common parametric distributions for the AFT model are exponential, gamma, Weibull, log-normal, and log-logistic.

Cox PH Model The Cox PH model is a widely used model in the field of biostatistics because of its ability to handle censored time-to-event data. Censored data arise when some units of observations are investigated for variable lengths of time but do not experience the event under study [4]. The PH model is often used to study the effects of the predictor variables such as treatment, age, gender, height weight, education, income, and blood pressure for predicting the survival outcomes. Exponentiating these predictors gives the relative risk for the covariate in question [4]. The Cox PH model is a very well-known semiparametric model with the leniency of not modeling the distribution-specific hazard function $\lambda(t)$. Let T be the failure time of interest and let Z be the set of time-dependent covariates. Now, let $Z(t)$ denote the value of Z at time t and let $\bar{Z}(t) = \{Z(s): 0 \leq s \leq t\}$ denote the history of the covariate up to time t. For convenience, let us formulate the effects of the covariates on failure time through the hazard function $\lambda_0(t)$ and we obtain the conditional hazard function of T given $\bar{Z}$ as

$$\lambda(t|\bar{Z}) = P\left(T \in [t, t+dt) \mid T \geq t, \bar{Z}(t)\right)$$

where $[t, t+dt)$ is an extremely small interval [4]. The Cox PH model is then given as

$$\lambda(t|\bar{Z}) = \lambda_0(t)\, e^{\beta' Z(t)}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function and it illustrates the hazard function for a subject with covariate values all equal to zero. Also, β is a set of unknown regression parameters in the model [3,4]. Since the Cox PH model allows for estimating and making inferences for the regression coefficients without assuming the population distribution for the baseline hazard, which is often unknown, this model is frequently used to analyze the prognostic factors in clinical and medical research.

SIMULATION STUDY

We simulated B = 30 datasets from LN, Weibull, and LL distributions. Each dataset was comprised of 10 samples in each of the two comparison groups, control and treatment, for total n = 20 with m = 1600 proteins, 40% of which were differentially expressed. Differentially expressed proteins were generated by introducing the differences in log means between the two comparison groups. The differences were allowed to vary from 1.05 to 1.50. Additionally, we varied the percentage of missing observations over the values of 0%, 20%, and 40% in order to examine its effect on the model performances. Five approaches were used to handle the simulated censored data. These approaches are NI, RM, KNN, and PPCA imputations as described in the previous sections.

Table 1. The results of the simulation study with the number of differentially expressed proteins at 5% FDR

Censoring Rate	0			20			40		
	LN	W	LL	LN	W	LL	LN	W	LL
t-test (NI)	413	435	404	309	451	333	188	415	212
t-test (RM)	413	435	404	147	347	170	19	105	22
t-test (KNN)	413	435	404	341	469	346	31	54	62
t-test (PPCA)	413	435	404	255	428	275	36	101	37
KS-test (NI)	286	379	305	244	413	272	120	336	161
KS-test (RM)	286	379	305	48	148	45	21	44	29
KS-test (KNN)	286	379	305	180	377	190	16	38	29
KS-test (PPCA)	286	379	305	89	204	124	30	61	31
RS-test (NI)	383	434	393	298	424	317	173	352	193
RS-test (RM)	383	434	393	82	292	100	8	47	13
RS-test (KNN)	383	434	393	317	423	327	54	57	66
RS-test (PPCA)	383	434	393	191	324	228	46	68	55
AFT-LN (LC)	413	433	404	419	463	417	422	490	427
AFT-W (LC)	348	395	320	361	527	331	368	516	353
AFT-LL (LC)	403	454	406	405	453	427	419	477	439
Cox PH (LC)	66	111	66	63	109	66	69	109	58

The missing observations were treated as the left-censored (LC) data for implementing the survival models using the survival package in R. Since the conventional survival package does not have the left-censored data analysis option for the Cox PH model, we used the icenReg package in R instead. For the KNN imputation, the function knn.impute in the impute package was used with $k = 3$ nearest neighbors, and the pca function of the pcaMethods package was used to impute the data for the PPCA approach.

Table 1 above provides the number of differentially expressed proteins at a true FDR of 5%. It is observed that when there is no missing data present, the *t*-test detects the same number of differentially expressed proteins across all imputation methods and performs as well as the AFT models. The rank-sum test also performed relatively well in detecting differentially expressed proteins. Unfortunately, the Cox PH model did not perform well for detecting differentially expressed proteins. As the proportion of missing data rises, the AFT model constantly outperformed all other standard methods in detecting differentially expressed proteins. The rank-sum tests, the Cox PH model, and in some cases the *t*-test had the least power in detecting differentially expressed proteins. Examining the performance of the AFT model alone, it was found that the AFT model under Weibull distribution had the least power in detecting true differentially expressed proteins when compared to the log-normal and log-logistic distributions. Surprisingly, the *t*-test and rank-sum tests with 'NI' performed well across all percentages of missingness. This is due to the fact that this method is not taking into consideration of the censored nature of the data. Nonetheless, the simulation study indicates that the AFT model is a recommended method of analysis when detecting differentially expressed proteins.

Table 2. The results of various tests performed on the Type I diabetes mellitus data for detecting differentially expressed (DE) proteins at 5% FDR

Method	Assumptions	Total DE
***t*-test (RM)**	normality; parametric	79
KS (RM)	nonparametric	76
WMW RS (RM)	nonparametric	80
***t*-test (KNN)**	normality; parametric	70
KS (KNN)	nonparametric	66
WMW RS (KNN)	nonparametric	65
***t*-test (PPCA)**	normality; parametric	76
KS (PPCA)	nonparametric	69
WMW RS(PPCA)	nonparametric	69
AFT-LN (LC)	log-normal; parametric	126
AFT-W (LC)	Weibull; parametric	142
AFT-LL (LC)	log-logistic; parametric	131
COX PH (LC)	semi-parametric	86

APPLICATION TO TYPE I DIABETES MELLITUS

As a practical application, the statistical methods examined in the simulation study were applied to the Type I diabetes mellitus dataset discussed in [2]. The dataset is based on frozen human serum samples from the DASP between 2000 and 2009. The diabetes data consist of 10 healthy control subjects and 10 subjects with a recent diagnosis of Type I diabetes mellitus. The samples were analyzed using the accurate mass and tag method, and the final LC-FTICR MS datasets were processed using the PRISM Data Analysis system. Any observation below the lowest observable peak intensity within the given sample was considered missing (left-censored). Hence, the detection limit is sample specific for this dataset. This resulted in the dataset composed of 173 proteins with the missingness rate of 24%. Four approaches were implemented to handle the missing/censored observations in the dataset: RM, KNN, PPCA imputation, and LC.

Table 2 above provides the results of computation and compares the number of differentially expressed proteins at 5% FDR. It was found that the KS test with KNN imputation has the least power to detect differential expressions. The AFT model again outperformed the standard methods, and in particular, the AFT method with Weibull distribution outperformed all other tests under consideration. Interestingly, the Cox PH model (with the bootstrap sample size of 200) was the second best in detecting differentially expressed proteins as opposed to the rank sum tests and *t*-test with RM, KNN, PPCA imputation methods.

CONCLUSION

In this work, we implemented the statistical methods from classical parametric and non-parametric tests, and survival analysis to detect differentially expressed proteins based on LC-MS proteomics data. From the simulation study, it was found that the methods using 'NI' perform well across the board but the AFT model outperforms all standard methods under consideration. From the application to the diabetes dataset, it was found that along with the AFT model, the newly proposed semi-parametric Cox PH model for left censored data performs competitively better than other standard methods.

REFERENCES

[1] Listgarten, J. and Emili, A. (2005). "Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry," *Molecular and Cellular Proteomics*, **4:** 419–434.

[2] Tekwe, C.D., Carroll, R.J., and Dabney, A.R. (2012). "Application of survival analysis methodology to the quantitative analysis of LC-MS proteomics data," *Bioinformatics*, **28:** 1998–2003.

[3] Orbe, J., Ferreira, E., and Nunez-Anton, V. (2002). "Comparing proportional hazards and accelerated failure time models for survival analysis," *Statistics in Medicine*, **21:** 3493–3510.

[4] Fisher, L.D. and Lin, D.Y (1999). "Time-dependent covariates in the Cox proportional-hazards regression model," *Annual Review of Public Health*, **20:** 145–157.

[5] Liquid chromatography–mass spectrometry. *Wikipedia*. https://en.wikipedia.org/wiki/Liquid_chromatography%E2%80%93mass_spectrometry