Integrated Survival Analysis of mRNA and microRNA Signature of Patients with Breast Cancer Based on Cox Model

WEINA ZHANG¹ and YILUN ZHANG²

ABSTRACT

Precision medicine has become a trend in the field of biomedical research. Analysis of mRNA and microRNA (miRNA) is a basis of it. Although the analyses of mRNA and miRNA in different cancers have been studied widely, only a few integrated studies concerning their interactions on large patient cohorts have been done. In this article, we study the integrated significance of mRNA, miRNA, and their interactions in the survival analysis of breast cancer (BRCA). We detect differential expression genes, then we analyze significant genes of BRCA by the Cox model. The model described in this article has better performance than the model without predictors of interactions of miRNA and mRNA. The average C index considering miRNA–target interactions is 0.7, which achieves 84% improvement compared with the model without interactions, which is 0.38.

Keywords: Cox model, microRNA, miRNA–target interactions, mRNA, precision medicine, survival analysis.

1. INTRODUCTION

PRECISION MEDICINE IS A NEW APPROACH to treating and preventing diseases that take into account individual genes, climate, and lifestyle variation for each person. Genetic analysis is the foundation of precision medicine. It is key to precision medicine that figures out significant genes of the disease by genes and clinical big data. mRNA, microRNA (miRNA), and miRNA–target interactions all influence the relapse and therapy of diseases.

This article studies significant genes of breast cancer (BRCA). Although much is known about mRNA and miRNA in BRCA, a few studies concerned miRNA–target interactions. Volinia et al. (2013) used a univariate Cox regression and Kaplan–Meier (KM) test to analyze mRNA and miRNA, but they did not consider miRNA–target interactions. miRNA is a noncoding RNA molecule of ~22 nucleotides. It is involved in the regulation of post-transcriptional gene expression in plants and animals (Bartel, 2004). miRNAs play significant roles in neoplasia, metastasis, and therapeutic effect in different cancers (Friedman et al., 2009; Garzon et al., 2009; Ambros, 2011; Starega-Roslan et al., 2011; Iuliano et al., 2013). miRNAs cannot influence disease independently, they influence disease by regulating mRNAs.

¹Department of Data Engineering, School of Computer Science and Cybersecurity, Communication University of China, Beijing, China.

²Department of Electrical and Computer Engineering, University of Miami, Miami, Florida.

So, we should not consider mRNAs and miRNAs independently, we should consider interactions between miRNAs and mRNAs.

In this study, we make a global expression analysis of mRNA and miRNA and their interactions on 626 BRCA cases from The Cancer Genome Atlas (TCGA) by penalized Cox model. Experiments show that the Cox model in this article has better performance than the model without miRNA–target interactions. This article includes the following studies. (1) Properly preprocessing: removing batch effect genes and non-coding mRNAs. (2) Finding differential expression genes of BRCA. (3) Finding miRNA–target interactions. (4) Building an interpretable and stable Cox model by comprehensive analysis of mRNAs, miRNAs, and miRNA–target interactions of BRCA.

The remainder of this article is organized as follows. In Section 2, we introduce the Cox proportional hazards model. In Section 3, we show the methods and tests of the penalized Cox model on real data. In Section 4, we report our results and make discussions.

2. COX PROPORTIONAL HAZARDS MODEL

Survival analysis is used to model time-to-event data. It reveals relationships between the given covariates and time distributions. Cox proportional hazards model (Cox, 1972) is one of the most common approaches to survival analysis. Cox proportional hazards model is a famous semiparametric model having advantages of parametric and nonparametric models. It is similar to the parametric model in expression form, but it is not influenced by the distribution of parameters of model. Considering the usual survival analysis framework, we have data for the i th observation with the form (y_i, X_i, δ_i) , where y_i is the observed time and δ_i denotes the status of observed event, such as dead, relapse, or recovery. If δ_i is 1, it means the event has occurred at time $T = y_i$. If δ_i is 0, it means y_i is the right-censoring time. As in regression, X_i is a vector of p potential predictors $(x_{i1}, x_{i2}, \dots, x_{ip})$. We further let $t_1 < t_2 < \dots < t_m$ be the ascending list of unique failure times. The Cox model assumes a semiparametric form for the hazard.

$$h_i(t|X) = h_0(t) \exp(X_i^T \beta), \quad (1)$$

where $h_i(t|X)$ is the hazard for observation i at time t , $h_0(t)$ is a baseline hazard, and β is a p -dimensional regression coefficient vector. Then we infer the i th observation partial likelihood is

$$L_i = \frac{\exp(X_i \beta)}{\sum_{l \in R(t_i)} \exp(X_l \beta)}, \quad (2)$$

where $R(t_i)$ denotes indices of the risk set at time t_i with $y_i \geq t_i$. If all observations are independent, their partial likelihood can be multiplied with each other, that is,

$$L = \prod_{i=1}^n L_i = \prod_{i=1}^n \left\{ \frac{\exp(X_i \beta)}{\sum_{l \in R(t_i)} \exp(X_l \beta)} \right\}^{\delta_i}. \quad (3)$$

Scaled log partial likelihood (m is the number of failure times),

$$l = \frac{2}{n} \left(\sum_{i=1}^m \left\{ X_i \beta - \log \left[\sum_{l \in R(t_i)} \exp(X_l \beta) \right] \right\} \right). \quad (4)$$

A key feature of genomic data is that dimension p can be much larger than the sample size n , so that traditional methodology cannot be directly applied. To make inferences for a high-dimensional Cox model, a variety of regularization approaches have been proposed. Tibshirani (1997) proposed the Lasso method. Zou and Hastie (2005) proposed an elastic network in linear regression. Park and Hastie (2007) applied elastic network in the Cox model. After that, there are many methods to solve the Cox model (Eq. 1). Simon et al. (2011) solved the Cox model with elastic network by the coordinate descent method. We estimate it by maximizing the partial likelihood as follows.

$$\hat{\beta} = \arg \max_{\beta} \left\{ \frac{2}{n} \left[\sum_{i=1}^m \left(X_i \beta - \log \left(\sum_{l \in R(t_i)} \exp(X_l \beta) \right) \right) \right] - \lambda P_{\alpha}(\beta) \right\}, \quad (5)$$

where

$$\lambda P_{\alpha}(\beta) = \lambda \left(\alpha \sum_{i=1}^p |\beta_i| + \frac{1}{2} (1 - \alpha) \sum_{i=1}^p \beta_i^2 \right).$$

3. METHODS

The algorithm proposed in this article is implemented with *R*. The RNA-Seq, microRNA-Seq, and clinic data were downloaded from TCGA (<https://tcga-data.nci.nih.gov/tcga/>) in January 2018. We selected 626 BRCA samples with mRNA, miRNA, and clinical data among which 560 samples are censored data, others are uncensored data.

3.1. Data preprocessing

3.1.1. Differential expression gene analysis. Leek et al. (2010) mentioned batch effects may be correlated with an outcome of interest and lead to incorrect conclusions. The batch effect occurs when tests are influenced by laboratory conditions, reagent lots, and variations in staff. So we need to remove batch effects before building the model. Noncoding RNA refers to an RNA sequence that does not encode protein sequences or that can only produce noncoding RNA molecules. Finding differential expression (DE) genes between healthy people and patients is the foundation of precision medicine. DE analysis refers to analyzing the normalized read count data of different experimental groups by statistical methods to discover quantitative changes in expression levels. There are different methods for DE analysis such as edgeR (Robinson et al., 2010) and DESeq (Anders and Huber, 2010) based on negative binomial distributions or baySeq (Hardcastle and Kelly, 2010) and EBSeq (Leng et al., 2013), which are Bayesian approaches based on a negative binomial model. In this article, we use Deseq to analyze DE. First, we remove batch effects (Ritchie et al., 2015) and noncoding mRNAs of the original 20531 mRNAs and 1870 miRNAs. Then we perform differential gene analysis on the processed data. Finally, we get 807 mRNAs and 608 miRNAs.

3.1.2. miRNA–target interactions. miRNAs are a type of noncoding single-stranded RNA molecules of ~22 nucleotides. They are involved in the regulation of post-transcriptional gene expression in plants and animals. They play important regulatory roles in many biological processes, including differentiation, metabolism, development, and cellular signaling. Thus, identification of miRNA–target mRNAs is important for studying biological processes of disease. There are some miRNA–target tools, such as TargetScan (Agarwal et al., 2015), miRanda (Enright et al., 2003; John et al., 2004; Betel et al., 2008, 2010), TarpmiR (Ding et al., 2016), and miRWalk (Sticht et al., 2018). miRWalk can provide comprehensive prediction and experimentally verified miRNA–mRNA interactions. miRWalk not only records miRNA binding sites within the whole sequence of a gene, but also combines this information with a comparison of binding sites resulting from 12 existing miRNA–target prediction programs to build new comparative platforms of binding sites for the promoter (4 prediction data sets), cds (5 prediction data sets), 5′- (5 prediction data sets), and 3′-UTR (13 prediction data sets) regions. It also documents experimentally verified miRNA–target interaction information collected through an automated text-mining search and data from existing resources (miRTarBase, PhenomiR, miR2Disease, and HMDD) offering such information. We get 77,697 pairs of miRNA–mRNA interactions by miRWalk2.0.

3.2. Analysis based on regularized Cox model

We analyze data modeled by the Cox model with *R* package glmnet (Hastie and Qian, 2014). Predictor variables are differential mRNAs, miRNAs, and miRNA–target mRNA pairs. The whole data set is randomly split into one training data set and one test data set. The ratio of the training set to the validation set is 3. We use C index to assess the prediction performance of the model, log-rank test, and area under the

FIG. 1. ROC curves of model A. (a) The ROC curve has an AUC 0.6 at the longest observation time. (b) The ROC curve has an AUC 0.62 at the 95% observation time. (c) The ROC curve has an AUC 0.57 at the 75% observation time. The observation time means time of all samples. The longest observation time is the maximum value of time. The 95% observation time is 95th percentiles of observation time. The 75% observation time is 75th percentiles of observation time. AUC, area under the ROC curve; ROC, receiver operating characteristic.

receiver operating characteristic (ROC) curve (AUC) to assess the discrimination of the model. Harrell et al. (1996) and Pencina and D'Agostino (2004) developed the overall C index as a metric describing the probability of conditional agreement between survival time and predict time. Concordance is defined as $\Pr(\text{agreement})$ for any two randomly chosen observations, in this case, agreement means that the observation with shorter survival time also has a larger risk. We compute C index using `SurvConcordance` function from `survival` package (Therneau and Lumley, 2014). Log-rank test (Mantel, 1966) compares the differences between KM survival curves.

We build a Cox model using a training set and compute PIs (prognostic indices) of every patient in test set with $PI = X^T \hat{\beta}$, where $\hat{\beta}$ is the coefficient of survival model. Patients are arranged in descending order of PI. In one experiment, the top 40% are high-risk groups and the last 40% are low-risk groups. In the other experiment, the top 50% are high-risk groups and the last 50% are low-risk groups. A log-rank test is only used to assess whether the observation is classified into the correct group. As for the results of the division, the AUC is needed to evaluate the performance. ROC curve is a probability curve. It is a performance

FIG. 2. KM curve of model A. **(a)** The KM curve is for model A that uses the 40% PI as the threshold for the high- and low-risk groups. **(b)** The KM curve is for model A that uses the 50% PI as the threshold for the high- and low-risk groups. KM, Kaplan–Meier; PIs, prognostic indices.

FIG. 3. ROC curves of model B. **(a)** The ROC curve has an AUC 0.95 at the longest observation time. **(b)** The ROC curve has an AUC 0.83 at the 95% observation time. **(c)** The ROC curve has an AUC 0.76 at the 75% observation time. The observation time means time of all samples. The longest observation time is the maximum value of time. The 95% observation time is 95th percentiles of observation time. The 75% observation time is 75th percentiles of observation time.

measurement for classification problems at various threshold settings. AUC is the area under the ROC curve that indicates the discriminative ability of the model. The higher the AUC, the better the model is in distinguishing between classes. We compute the AUC by survival ROC package (Saha and Heagerty, 2013).

4. RESULTS AND DISSCUSSIONS

mRNAs influence overall survival of BRCA, miRNAs are modulators of the cellular processes responsible for cancer that are encoded by mRNAs. Because of these relations, we performed integrated survival analysis on a large BRCA cohort of 626 patients, using genome-wide data for mRNA, miRNA expression, and miRNA–target interactions. We assumed model A is a model that only considers predictors without miRNA–target interactions, and model B is a model that takes miRNA–target interactions into considerations. Because of random subsampling during building the model, the same model has slightly

FIG. 4. KM curves of model B. **(a)** The KM curve is for model B that uses the 40% PI as the threshold for the high- and low-risk groups. **(b)** The KM curve is for model B that uses the 50% PI as the threshold for the high- and low-risk groups.

different C index on test in different runs. The average C index is 0.38 in model A and 0.7 in model B when we run models A and B 10 times, respectively. Experiments show model B has better performance than model A.

The following figures show the ROC and KM curves of models A and B in one run. Figure 1 shows ROC curves of model A at different observation times in one run. Figure 2 shows KM curves of model A at different thresholds in one run. Figure 3 shows ROC curves of model B at different observation times in one run. Figure 4 shows KM curves of model B at different thresholds in one run.

Experiments show miRNA–target interactions influence the concordance and discrimination of the model prediction. When we build a gene model for precision medicine, we should consider not only mRNAs and miRNAs, but also miRNA–target interactions. The approaches in this article are not only applicable to BRCA data but also to other disease data.

ACKNOWLEDGMENTS

Thanks are due to XiaoDong Cai (University of Miami) for conceiving the idea and Xin Zhou (University of Miami) for valuable discussion.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

This study is supported by the PhD research promotion project of Communication University of China (No. CUC200D030).

REFERENCES

- Agarwal, V., Bell, G.W., Nam, J.-W., et al. 2015. Predicting effective microrna target sites in mammalian mrnas. *elife* 4, e05005.
- Ambros, V. 2011. Micrnas and developmental timing. *Curr. Opin. Genet. Dev.* 21, 511–517.
- Anders, S., and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Biol.* 11, R106.
- Bartel, D.P. 2004. Micrnas: Genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297.
- Betel, D., Koppal, A., Agius, P., et al. 2010. Comprehensive modeling of microrna targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* 11, R90.
- Betel, D., Wilson, M., Gabow, A., et al. 2008. The microrna. org resource: Targets and expression. *Nucleic Acids Res.* 36(suppl_1), D149–D153.
- Cox, D.R. 1972. Regression models and life-tables. *J. R. Stat. Soc. Ser. B (Methodol.)* 34, 187–202.
- Ding, J., Li, X., and Hu, H. 2016. Tarpmir: A new approach for microrna target site prediction. *Bioinformatics.* 32, 2768–2775.
- Enright, A.J., John, B., Gaul, U., et al. 2003. Microrna targets in drosophila. *Genome Biol.* 5, R1.
- Friedman, R.C., Farh, K.K.-H., Burge, C.B., et al. 2009. Most mammalian mrnas are conserved targets of micrnas. *Genome Res.* 19, 92–105.
- Garzon, R., Calin, G.A., and Croce, C.M. 2009. Micrnas in cancer. *Annu. Rev. Med.* 60, 167–179.
- Hardcastle, T.J., and Kelly, K.A. 2010. bayseq: Empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinform.* 11, 422.
- Harrell, F.E., Lee, K.L., and Mark, D.B. 1996. Tutorial in biostatistics multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* 15, 361–387.
- Hastie, T., and Qian, J. 2014. Glmnet vignette. Available at: www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf. Accessed September 20, 2016.
- Iuliano, R., Vismara, M.F.M., Dattilo, V., et al. 2013. The role of micrnas in cancer susceptibility. *BioMed Res. Int.* 2013, 591931.

- John, B., Enright, A.J., Aravin, A., et al. 2004. Human microRNA targets. *PLoS Biol.* 2, e363.
- Leek, J.T., Scharpf, R.B., Bravo, H.C., et al. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733.
- Leng, N., Dawson, J.A., Thomson, J.A., et al. 2013. Ebsseq: An empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics.* 29, 1035–1043.
- Mantel, N. 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* 50, 163–170.
- Park, M.Y., and Hastie, T. 2007. L1-regularization path algorithm for generalized linear models. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 69, 659–677.
- Pencina, M.J., and D'Agostino, R.B. 2004. Overall c as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Stat. Med.* 23, 2109–2123.
- Ritchie, M.E., Phipson, B., Wu, D., et al. 2015. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47–e47.
- Robinson, M.D., McCarthy, D.J., and Smyth, G.K. 2010. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 26, 139–140.
- Saha, P., and Heagerty, P. 2013. Introduction to survivalROC: An r package for survival-roc method. Available at: <https://CRAN.R-project.org/package=survivalROC> Accessed October 8, 2019.
- Simon, N., Friedman, J., Hastie, T., et al. 2011. Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* 39, 1.
- Starega-Roslan, J., Koscińska, E., Kozłowski, P., et al. 2011. The role of the precursor structure in the biogenesis of microRNA. *Cell. Mol. Life Sci.* 68, 2859.
- Sticht, C., De La Torre, C., Parveen, A., et al. 2018. mirwalk: An online resource for prediction of microRNA binding sites. *PLoS One* 13, e0206239.
- Therneau, T.M., and Lumley, T. 2014. Package 'survival'. *Survival analysis Published on CRAN.* Available at <https://CRAN.R-project.org/package=survival> Accessed October 8, 2019.
- Tibshirani, R. 1997. The lasso method for variable selection in the cox model. *Stat. Med.* 16, 385–395.
- Volinia, S., and Croce, C.M. 2013. Prognostic microRNA/mrna signature from the integrated analysis of patients with invasive breast cancer. *Proc. Natl. Acad. Sci.* 110, 7413–7417.
- Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 67, 301–320.

Address correspondence to:

Dr. Weina Zhang
School of Computer Science and Cybersecurity
Communication University of China
No.1 Dingfuzhuang East Street Chaoyang District
Beijing 100024
China

E-mail: zhangweina@cuc.edu.cn