ICIC Express Letters
Volume **6**, Number **5**, May **2012**

ICIC International ©2012 ISSN 1881-803X
pp. 1253–1257

A PATHWAY-BASED APPROACH FOR ANALYZING MICROARRAY DATA USING RANDOM FORESTS

Chin Hui Shi[1], Mohd Saberi Mohamad[1], Safaai Deris[1]
and Zuwairie Ibrahim[2]

[1]Artificial Intelligence and Bioinformatics Research Group
Faculty of Computer Science and Information Systems
[2]Department of Mechatronics and Robotics
Center for Artificial Intelligence and Robotics
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81310 UTM Skudai, Johor, Malaysia
wyce86@gmail.com; { saberi; safaai }@utm.my; zuwairiee@fke.utm.my

Received August 2011; accepted October 2011

Abstract. *Although machine learning methods, such as random forests, have been developed to correlate survival outcomes with a set of genes, less study has assessed the abilities of these methods in incorporating pathway information for analyzing microarray data. In general, genes that are identified without incorporating biological knowledge are more difficult to interpret. Thus, the pathway-based survival analysis using machine learning methods represents a promising approach for generating new biological hypothesis from microarray studies. The two popular variants of random forests used in this research for survival data are random survival forests and bivariate node-splitting random survival forests. There are three types of datasets used for this research and each dataset with a three-level outcome. This research which compared the four splitting rules available in random survival forests to identify log-rank test is the most accurate in terms of prediction error. To evaluate the accuracy of pathway based survival approach, this research considered employing area under the receiver operating characteristic curve for censored data. The use of random survival forests for survival outcomes in analyzing microarray data allows researchers to obtain results that are more closely tied with the biological mechanism of diseases.*

Keywords: Pathway, Survival outcomes, Microarray data, Random forests, Random survival forests, Bivariate node-splitting random survival forests

1. Introduction.

Although numerous methods have been developed to analyze microarray data based on single genes or the whole set of genes, they do not make use of pathway information. In the past several years, microarray data have been used for survival analysis through several methods. Correlating gene expression data with survival outcomes on the pathway level may lead to biologically more meaningful information for prognosis biomarkers. The two popular variants of random forests for the analysis of survival data are random survival forests [1] and bivariate node-splitting random survival forests [2].

This research describes a pathway-based method using random survival forests to analyze gene expression data with survival outcomes on the pathway level. The proposed method represents a promising approach for researchers to identify important pathways for predicting patient prognosis and discover important genes within those pathways. This research which also compared random survival forests with different split criteria to assess the performance of the proposed approach in identifying random survival forests with log rank test is the most accurate in terms of prediction error. The proposed method was applied to three different datasets where one dataset with a three-level outcome.

2. **Materials and Methods.** The two versions of random forests used for this research are *random survival and bivariate node-splitting random survival forests*. There is a difference between random forests classification and regression and the random survival forests. The key difference is the outcome of interest is a set of survival times with the corresponding censoring indicator for the survival counterpart. Figure 1 shows the flow chart of pathway analysis for survival outcomes using random survival forests.

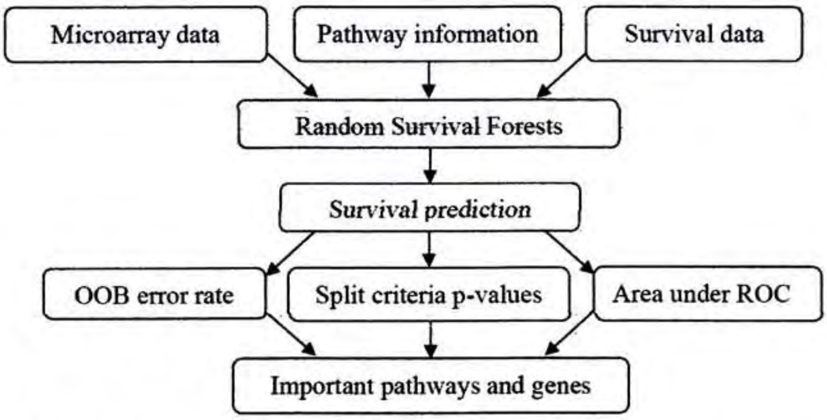

FIGURE 1. The flow chart of pathway analysis for survival outcomes using random survival forests

2.1. **Random survival forests.** Random Survival Forests (RSF) is an extension of Breiman's Random Forests to survival analysis settings. Algorithm uses a binary recursive tree growing procedure with different splitting rules for growing an ensemble cumulative hazard function. An "out-of-bag" estimate of Harrell's concordance index [3] is provided for assessing prediction. The algorithm used by random survival forests is broadly described as follows:

Step 1 Draw ntree bootstrap samples from the original data. Note that each bootstrap samples excludes approximately one-third of the sample data called out-of-bag (OOB).

Step 2 A survival tree is grown for each of the bootstrap sample.

Step 3 At each node of the tree, select $\sqrt{m}$ predictors at random for splitting.

Step 4 Split on a predictor using a survival splitting criterion. A node is split on that predictor which maximizes survival differences across daughter nodes.

Step 5 Repeat Steps 3 and 4 until each terminal node contains no more than 0.632 times the number of events.

Step 6 Calculate an ensemble cumulative hazard estimate by combining information from the ntree trees.

Step 7 Compute an out-of-bag (OBB) error rate for the ensemble CHF.

2.2. **Bivariate combination split.** By considering the much-reduced dimension in a pathway-based setting, a bivariate splitting criterion is feasible. From the four splitting rules, LR split criterion performed the best and is chosen to implement three approaches to use bivariate splitting strategies in random survival forests by modifying the C embedded code in the R program [2]. The strategy is to split on the best pair of covariates at every node split by changing Step 4 of the above algorithm:

Step 4 Using the LR splitting criterion, a node is split using the predictor pair from Step 3 that maximizes the survival differences between daughter nodes by finding best split of the form $x_i + x_j \leq c$ for $i \neq j$.

This approach is called bRSF LR for bivariate random survival forests with LR splitting criterion. *This strategy helps take into account the correlations among genes in the* pathway.

3. Results and Discussion.

3.1. Datasets. A total of 1308 pathways were used for the analysis. These pathways are wired diagrams of genes and molecules from KEGG and BioCarta. There are also a few signal processing pathways and others related to human diseases. A summary for the real data used in this analysis is given in Table 1. The Breast dataset [4] is classified into luminal, basal and apocrine classes. Both Lung_a [5] and Lung_b [6] datasets are lung cancer datasets and provide outcomes as 'good' or 'poor' ones.

TABLE 1. The datasets used in this research

Dataset	Number of patients	Genes	Response type
Breast	49	22215	Three tumor types
Lung_a	86	7129	Normal/tumor
Lung_b	62	12600	Normal/tumor

3.2. Experimental results. This research paper applied RSF to assess their abilities in giving biological insights based on three different microarray datasets. To assess the stability of the concordance error rate, this research paper first calculated the 10-fold cross validation for RSF with different spilt criteria using 1308 pathways from the three different datasets. From Table 2, the observed random survival forests with log rank test split rule (RSF LR) gave the lowest mean of error rates among the three different datasets used. From the finding, the LR splitting criterion is chosen for this analysis because the mean of error rates are the lowest and it performs best in terms of prediction error.

TABLE 2. Mean of error rate for three different datasets using RSF

Datasets	logrank	conserve	logrankscore	random
Breast	0.5393	0.5624	0.5548	0.5646
Lung_a	0.5420	0.5730	0.5562	0.5813
Lung_b	0.5511	0.5784	0.5850	0.5835

Table 3 shows the best five pathways of RSF on breast cancer dataset among 435 pathways. It has been found that the highest AUC value was achieved by sulfur metabolism pathway, as it has been suggested that defective sulfur metabolism might be related to carcinogenesis [7]. Besides, the lowest p-value was achieved by BC-multi-step pathway. The new multi-step pathways of breast cancer progression have been delineated through genotypic phenotypic correlations in the past few years [8]. The lowest mean of error rate was achieved by BC-IL-3 signaling pathway. Small mean of error rate based on gene in a given pathway would indicate the pathway as potentially interesting.

Table 4 shows the best five pathways of breast cancer dataset using bRSF. This dataset seemed to do slightly better than the results in Table 3. BC-Ras-independent pathway was shown to have highest AUC value and BC-deregulation of c-myc-induced pathway was shown to have lowest p-value and lowest mean of error rate. The ras mutation is infrequent in breast cancer although aberrant function of ras signal transduction pathway is thought to be common in human breast cancer [9]. Myc deregulation contributes to breast cancer development and progression and is associated with poor outcomes [10]. Application of RSF and bRSF to two other datasets had done as well.

TABLE 3. The best five pathways of RSF on breast cancer dataset

Pathways	Number of genes	RSF LR p-value	RSF LR AUC	Mean of error rate
BC-IL 3 signaling pathway	15	0.0100	0.6539	0.3817
BC-Multi-step pathway	10	0.0042	0.7459	0.3931
BC-Mechanisms of transportation	19	0.1685	0.6955	0.4044
BC-Role of PPAR-gamma	12	0.1325	0.6923	0.4006
Sulfur metabolism	9	0.0231	0.7784	0.4108

TABLE 4. The best five pathways of bRSF on breast cancer dataset

Pathways	Number of genes	bRSF LR p-value	bRSF LR AUC	Mean of error rate
Alkaloid biosynthesis	6	0.0019	0.7905	0.3503
BC-Deregulation of C-myc-induced pathway	15	0.0017	0.8710	0.3172
BC-Erk and PI-3 Kinase	33	0.0435	0.7952	0.4141
BC-Presenilin action	10	0.0519	0.7762	0.4722
BC-Ras-Independent pathway	20	0.0125	0.8883	0.3424

Time-dependent ROC analysis is an extension of the concept of ROC curves for time-dependent binary disease variables in censored data. Figure 2 shows the time-dependent ROC curve from censored survival data for breast cancer dataset using RSF LR. Each pathway is represented by different color in the ROC curve. From the finding, the AUC value for Sulfur metabolism is the highest among others. The area under the curve for this pathway is the largest compared with others and for the ROC curve, it yield a point in ROC space which is nearer to the y-axis which means the curve will be at the upper left corner of the ROC space.

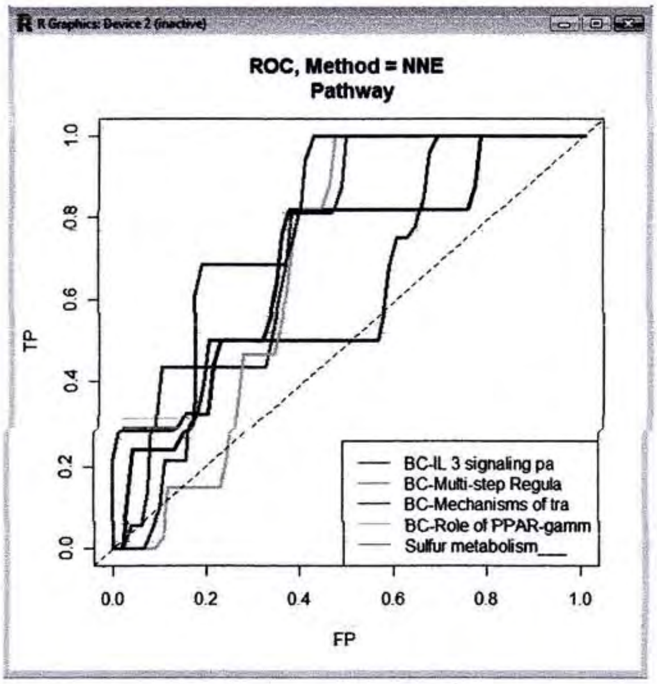

FIGURE 2. The time-dependent ROC curve of breast cancer dataset using RSF LR

Last but not least, the splitting criterion used in growing a tree have involved survival time and censoring information. The three different datasets have their own survival time and censoring status. Performance of random survival forests also depended on the censoring rate. If most of the cases are deaths, then the performance of error rate was good or vice versa. Even though the sample size for a dataset is larger than others, if most of the censoring in the dataset is higher and thus performance of error rate is poor.

4. **Conclusions.** This research described a pathway-based approach for analyzing microarray data with survival outcome using random survival forests with univariate and bivariate node splits. The LR test approach helps to identify pathways that are good at predicting patient's prognosis. AUC helps to identify pathways that are good at correctly predicting patients who progress or survive past a certain time. Lastly, the concordance error rate helps to identify the pathways that are biologically meaningful. The future work would certainly motivate and draw the interest of other researchers to develop other novel pathway-based methods based on multivariate variable selection for survival outcomes.

Acknowledgments. This work is financed by Institutional Scholarship MyPhD provided by the Ministry of Higher Education of Malaysia. We also would like to thank Universiti Teknologi Malaysia for supporting this research by UTM GUP research grants (vot number: Q.J130000.7123.00H67 and Q.J130000.7107.01H29).

REFERENCES

[1] H. Ishwaran, U. B. Kogalur, E. H. Blackstone and M. S. Lauer, Random survival forests, *Annals of Applied Statistic*, vol.2, no.3, pp.841-860, 2008.

[2] H. Pang, D. Datta and H. Zhao, Pathway analysis using random forests with bivariate node-split for survival outcomes, *Bioinformatics*, vol.26, no.2, pp.250-258, 2009.

[3] F. E. Jr. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee and R. A. Rosati, Evaluating the yield of medical tests, *J. Amer. Med. Assoc.*, vol.247, no.18, pp.2543-2546, 1982.

[4] P. Farmer, H. Bonnefoi, V. Becette, M. Tubiana-Hulin, P. Fumoleau, D. Larsimont, G. Macgrogan, J. Bergh, D. Cameron, D. Goldstein, S. Duss, A. L. Nicoulaz, C. Brisken, M. Fiche, M. Delorenzi and R. Iggo, Identification of molecular apocrine breast tumors by microarray analysis, *Oncogene*, vol.24, no.29, pp.4660-4671, 2005.

[5] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker and M. Meyerson, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, *Proc. of the National Academy Sciences of the United States of America*, vol.98, no.24, pp.13790-13795, 2001.

[6] D. G. Beer, S. L. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer and S. Hanash, Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nat. Med.*, vol.8, no.8, pp.816-824, 2002.

[7] A. Jamshidzadeh, M. Aminlari and H. R. Rasekh, Rhodanese and arginase activity in normal and cancerous tissues of human breast, esophagus, stomach and lung, *Arch. Iran. Med.*, vol.4, pp.88-92, 2001.

[8] P. T. Simpson, J. S. Reis-Filho, T. Gale and S. R. Lakhani, Molecular evolution of breast cancer, *J. Pathol.*, vol.205, no.2, pp.248-254, 2005.

[9] G. J. Clark and C. J. Der, Abberant function of the Ras signal transduction pathway in human breast cancer, *Breast Cancer Research and Treatment*, vol.35, no.1, pp.133-144, 1995.

[10] J. Xu, Y. Chen and O. I. Olopade, MYC and breast cancer, *Genes & Cancer*, vol.1, pp.629-640, 2010.