Partially Ordered Expression Features Improves Survival Prediction in Cancer

Mustafa Buyukozkan [*]
Institute of Computational Biology
Helmholtz Zentrum Muenchen
mustafa.buyukozkan@helmholtz-muenchen.de

Halil İbrahim Kuru [*]
Department of Computer Engineering
Bilkent University
Ankara Turkey, 06800
ibrahim.kuru@bilkent.edu.tr

Oznur Tastan
Department of Computer Engineering
Bilkent University
Ankara, Turkey, 06800
oznur.tastan@cs.bilkent.edu.tr

1 Introduction

Predicting the survival of cancer patients is critical for choosing patient specific treatment strategies. Survival prediction has been traditionally based on clinical or pathological factors such as patient age and tumor stage. With the availability of high-throughput data expression quantities are also incorporated in the models [6, 1, 8]. The survival models that are built with molecular expression profiles rely on the individual expression quantities of the molecules in the tumors. However, in the cell molecules interact with each other and in cancer these interactions are dysregulated in various ways. A better representation of the molecular abundance that accounts for these dysregulations has potential to increase the predictive performance of survival models and help reach biomarkers that are readily interpretable.

To reach results that are biologically relevant, we suggest using partial ordering of the expression quantities in lieu of individual expression values. In this work, we focused on protein expression data as it is more stable; however, the same framework is applicable to other molecular types as well. We built random forest survival (RSF) models [5] with partial order features of protein expression data and compare them with the models trained with individual protein expression features in 8 different cancer types. The results demonstrate that partial order features have better predictive performance in the majority of the cancers. Accounting order dysregulation of proteins unveil predictive features with direct relevance to the biological mechanism of cancer. Below, we first describe the methodology and next results obtained.

2 Methods

2.1 Datasets

The Cancer Genoem Atlas protein expression data and patient survival data are obtained from USCS Cancer Browser (https://genome-cancer.ucsc.edu) (April 11, 2017). The protein expression was quantified by reverse phase protein array (RPPA). We worked with eight different cancer types, which include: ovarian adenocarcinoma (OV), breast invasive carcinoma (BRCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), bladder

[*] Co-first authors of the paper. M. Büyüközkan conducted this research at Bilkent University.

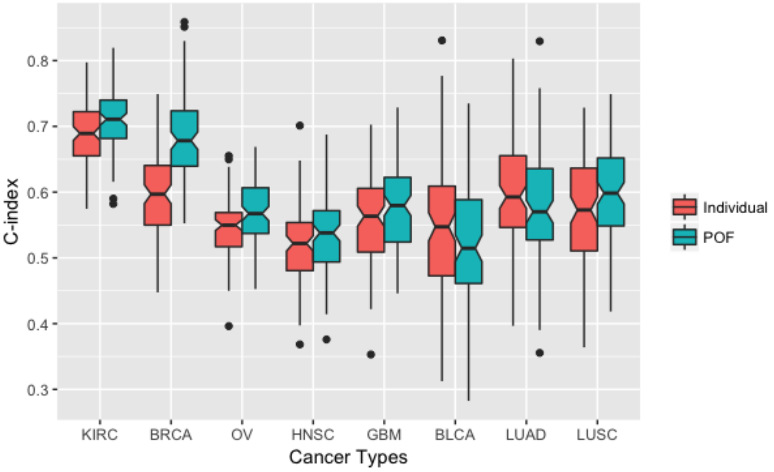

Figure 1: Comparison of RSF model performances that are trained with individual features and partial order features(POF) for different cancer types.

urothelial carcinoma (BLCA). For each cancer type, the data is of the form, $D = \{\mathbf{X}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{n}$; n is the number of patients. For each patient $\mathbf{X}$ is the derived features from protein expression data and $\mathbf{y} = \{S, \delta\}$ is the survival data, where S is survival time and δ denotes censoring; if $\delta = 1$ the data is right-censored and 0 it is not censored.

2.2 Partial Ordered Features

Let $X_i^{(k)}$ and $X_j^{(k)}$ denote the expression values for protein i and j in patient k. The partial ordered features (POF) for this patient is defined as:

$$X_{i,j}^{(k)} = \begin{cases} 1 & \textit{if } X_i^{(k)} > X_j^{(k)} \\ -1 & \textit{otherwise} \end{cases} \tag{1}$$

$X_{i,j}^{(k)} = 1$ indicates that the molecule i is more abundant with respect to molecule j for this patient, whereas $X_{i,j}^{(k)} = -1$ indicates otherwise. For every pair of proteins, we derive a partial ordered feature. This nonlinear representation aims at capturing expression dysregulation among proteins.

2.3 Model Training and Performance Comparison

For each cancer type, we randomly split the samples into two 100 times: 80% as the training set and 20% as the test set. For each split, two types of RSF models are built using the training data where only the feature representations differed. In the first type of models, the individual expression values were input as features and in the second one we use the suggested POF representation. In both cases, we performed a feature selection step; the association of individual features with survival was decided based on the hazard ratio from the univariate Cox model [7]. Likelihood ratio test p-values were used to assess the significance of hazard ratio; features with p-value ≤ 0.05 are retained for model training. For each cancer type and feature representations, 100 models were trained. Finally, the models were evaluated by the Concordance-Index (C-index) [3] on the test data.

2.4 Assessing Feature Importance

We quantified the importance of a feature by the difference between the performances of the original ensemble and the ensemble where this feature's assignments are randomized in the patients [4]. Large difference between the two implies the feature contributes to the model. Average of feature importance for a POF is calculated over the ensemble models in which it is selected.

3 Results

3.1 Performance of Partially Ordered Features

Figure 1 compares the distribution of C-indices for 100 models trained with two feature representation: individual expression values and POF representation. In 6 out of 8 cancer types, POF representation achieves better results, for 5 of them the difference was statistically significant (Wilcoxon signed-rank test, $\alpha = 0.05$). Among all cancer types, most improvement was observed in BRCA (p-value $\leq 1.5e - 16$) and in KIRC (p-value $\leq 2.1e - 14$).

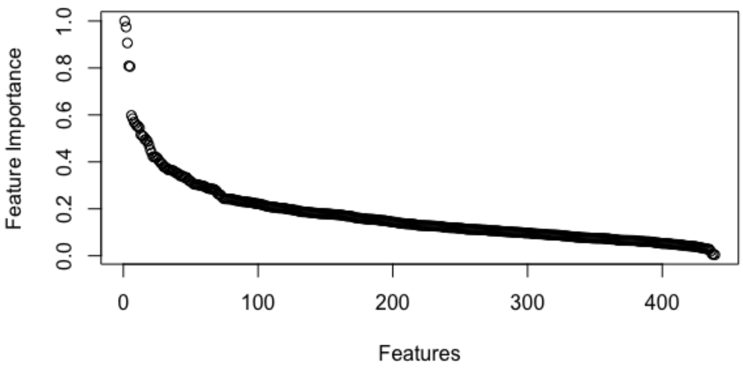

Figure 2: Distribution of average feature importance values in BRCA. Feature importances are scaled in the [0 -1] range.

3.2 Feature Importance in BRCA

Figure 2 displays the normalized importance of features that have been selected in $>= 95\%$ of the models trained in BRCA. Of these, we select the top performing 50 POF and displayed them as a network (Figure 3). Interestingly, most of these pairwise features were also related to each other through common proteins; and they participate in common pathways. Dishevelled homolog 3 (DVL3), which is part of the WNT signaling pathway, have many interactions whose dysregulation is predictive of survival. There are recent studies that points to importance of DVL3 in resistance to cancer drugs [2].

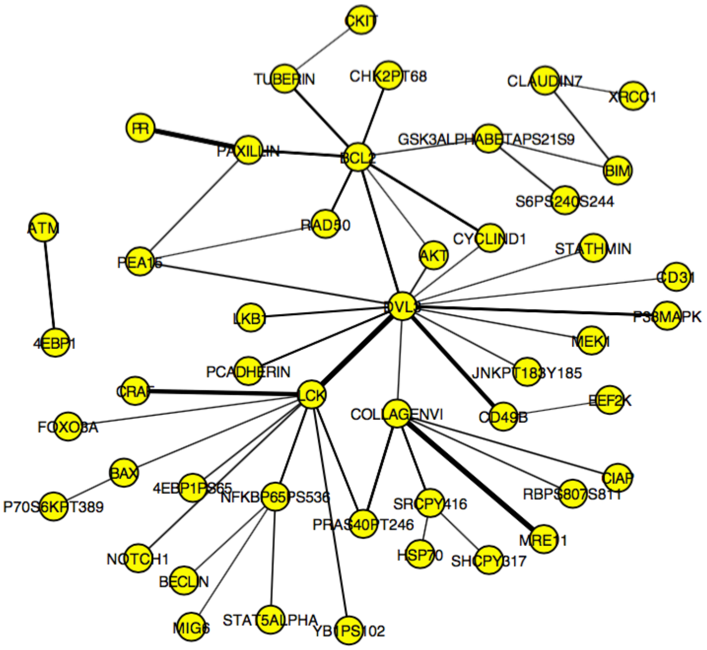

Figure 3: 50 top partially ordered features in BRCA. Nodes represent proteins; edges represent the partially ordered features. Edge thickness is proportionally average feature importance.

3.3 Conclusion

The representation of data in terms of pairwise comparison of expression values provides a way of incorporating nonlinear interactions between the individual molecules. Inputting in a nonlinear survival model such as RSF further allow representing data in higher nonlinear interactions, leading to better survival models. The representation brings insight into the cellular interactions that are disrupted in cancer and are associated with survival.

References

[1] A. Fernandez-Teijeiro, R. A. Betensky, L. M. Sturla, J. Y. Kim, P. Tamayo, and S. L. Pomeroy. Combining gene expression profiles and clinical parameters for risk stratification in medulloblastomas. *Journal of clinical oncology*, 22(6):994–998, 2004.

[2] S. Gao, I. Bajrami, C. Verrill, A. Kigozi, D. Ouaret, T. Aleksic, R. Asher, C. Han, P. Allen, D. Bailey, et al. Dsh homolog dvl3 mediates resistance to igfir inhibition by regulating igf-ras signaling. *Cancer research*, 74(20):5866–5877, 2014.

[3] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[4] H. Ishwaran and U. Kogalur. Package ?randomforestsrc.? 2015. ht tp. *cran. r-project. org/web/packages/randomForestSRC/randomForestSRC. pdf. Accessed March*, 23, 2015.

[5] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

[6] K. Shedden, J. M. Taylor, S. A. Enkemann, M.-S. Tsao, T. J. Yeatman, W. L. Gerald, S. Eschrich, I. Jurisica, T. J. Giordano, D. E. Misek, et al. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature medicine*, 14(8):822–827, 2008.

[7] T. Therneau. A package for survival analysis in s. r package version 2.37-4. *URL http://CRAN. R-project. org/package= survival. Box*, 980032:23298–0032, 2013.

[8] Y. Yuan, E. M. Van Allen, L. Omberg, N. Wagle, A. Amin-Mansour, A. Sokolov, L. A. Byers, Y. Xu, K. R. Hess, L. Diao, et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nature biotechnology*, 32(7):644–652, 2014.