Predicting clinical outcomes in neuroblastoma with genomic data integration

Ilyes Baali,[1†] Alp Emre Acar[1†], Tunde Aderinwale[2], Saber HafezQorani[3], Hilal Kazan[4*]

[1]Department of Electric-Electronics Engineering, Antalya International University, Antalya, Turkey

[2]Institute of Applied Sciences, Antalya International University, Antalya, Turkey

[3]Graduate School of Informatics, Department of Health Informatics, Middle East Technical University, Ankara, Turkey

[4]Department of Computer Engineering, Antalya International University, Antalya, Turkey

[†]These authors have contributed equally to this work.

[*]To whom correspondence should be addressed; E-mail: hilal.kazan@antalya.edu.tr

Neuroblastoma is a heterogeneous disease with diverse clinical outcomes. Recently collected genome-wide datasets provide opportunities to infer neuroblastoma subtypes more accurately than existing classification of risk groups. To this end, we used machine learning techniques to predict overall survival and event-free survival profiles of patients. Using the model that we trained on SEQC cohort, we can predict patient survival in an independent cohort with high accuracy (AUROC: 0.96) indicating the applicability of the model to different datasets. Additionally, we used unsupervised learning techniques that can effectively integrate multiple high-dimensional datasets to identify subgroups of patients with distinct survival profiles after stratification based on MYCN expression. These subgroups can improve treatment stratification of neuroblastoma patients.

Introduction

Neuroblastoma is the second most common solid tumor in childhood. The disease can have a large variety of clinical outcomes ranging from spontaneous regression to relentless progression despite extensive therapies. Chromosomal amplification of the MYCN locus occurs in 25 of all neuroblastomas and is associated with poor prognosis (*1*). Apart from MYCN amplification, a limited set of additional variables such as age at diagnosis, stage of disease etc. are used to stratify patients into distinct risk groups. Recent progress on high-throughput technologies enables the collection of genome-wide measurements across large set of patients in cohorts. We utilized the diverse data types provided by the SEQC cohort (i.e., neuroblastoma challenge in CAMDA 2017) to develop statistical

models that can predict clinical outcomes in neuroblastoma. Also, we employed an unsupervised learning strategy to identify subgroups that have significantly diverse survival profiles.

Results

Validation of the SEQC model on an independent cohort

We first performed supervised learning using support vector machines (SVM) within the SEQC dataset. The mean cross-validation accuracy of this model is close to the best accuracy reported for the same dataset (2). We then used our model trained on SEQC data to predict overall survival (i.e., occurrence of death from disease) and event-free survival (i.e., occurrence of progression, relapse or death) profiles of patients in an independent cohort that is called `Versteeg dataset` hereafter. This dataset includes the gene-expression measurements and clinical data for 88 patients. Figure 1a and 1b shows the ROC curve for predicting overall survival and event-free survival profiles, respectively.

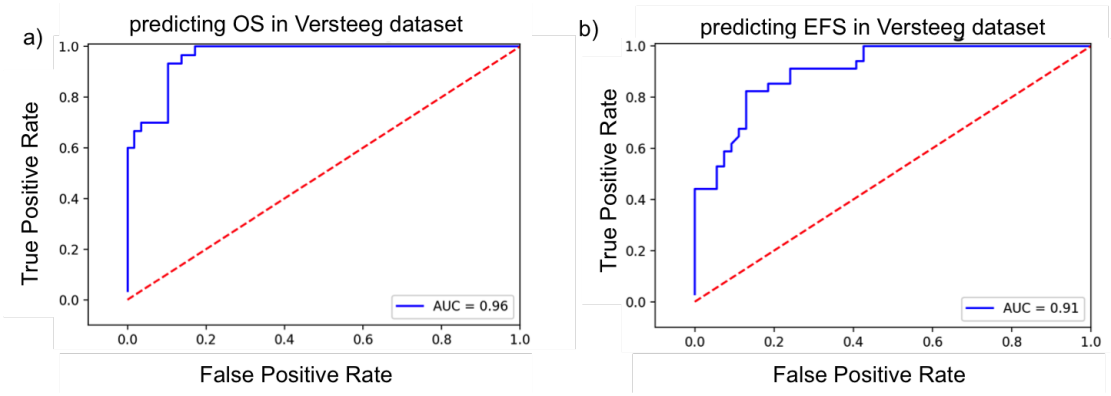

Figure 1: a) ROC curve for predicting overall survival profiles of patients in Versteeg dataset. b) ROC curve for predicting event-free survival profiles of patients in Versteeg dataset.

Clustering patients into subgroups with multi-view kernel k-means

MYCN expression is one of the best predictors of survival in neuroblastoma. However, there is still some variability in survival profiles after stratification based on MYCN expression. We aimed to explain this remaining variability with RNA-seq data. As such, we further clustered the patients that have low and high MYCN expression (N= 90 and N=408 respectively). The threshold to split the patients based on MYCN expression was chosen to minimize the log-rank test p-value from Kaplan-Meier analysis. As expected, there is a high overlap between the patients with high MYCN expression and patients with MYCN amplification. We used multi-view kernel k-means (MKKM) to integrate two versions of the same RNA-seq dataset that are processed in different ways (See Methods).

The number of clusters was selected as 2 based on mean silhouette score. The weights of the views for the two versions of the RNA-seq data were 0.43 and 0.57 (i.e., MAV and RPM processing of RNA-seq data) for patients with low MYCN expression. Similarly, the view weights were 0.46 and 0.54 when clustering the patients with high MYCN expression. These weights give better silhouette scores over uniform weights indicating the advantage of using multi-view kernel k-means. Also, we used all the available features rather than going through an initial feature selection procedure.

Figure 2a and 2b plot the survival curves of identified clusters from patients with low and high MYCN expression, respectively. Log-rank test gives p-values of 0.06 and 0, respectively. We also tried including the microarray data in addition to RNA-seq datasets; however this did not improve the silhouette score and the log-rank test p-value. We re-

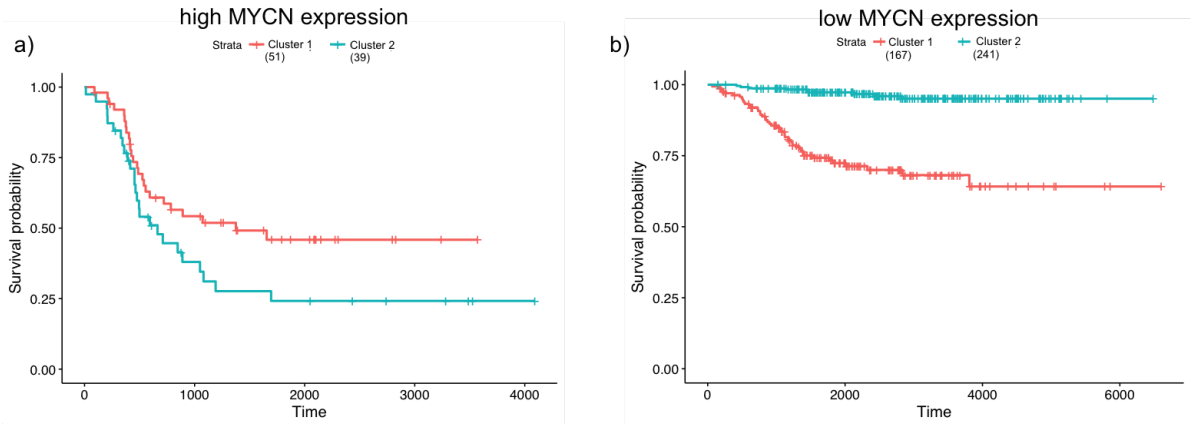

Figure 2: a) Survival curves of MKKM-inferred subgroups identified within the patients with high MYCN expression. Log-rank test p-value is 0.06. b) Survival curves of MKKM-inferred subgroups identified within the patients with low MYCN expression. Log-rank test p-value is 0.

peated the same analysis with the subset of the patients (145 patients) for which aCGH data is available. We again split the patients into two groups based on MYCN expression. Because the group of patients with high MYCN expression is small (N=24), we only cluster the patients with low MYCN expression (N = 121). Figure 3 shows the survival curves of identified clusters for patients with low MYCN expression. The difference between Figure3a and 3b is due to the inclusion of aCGH data. Namely, including aCGH data in addition to RNA-seq datasets results in a lower p-value (0.07 vs 0.0008).

Methods

Datasets

We downloaded the RNA-seq, microarray and aCGH datasets for the SEQC cohort from CAMDA website. We used two versions of the RNA-seq data: `SEQC_NB_MAV_G_log2.txt`

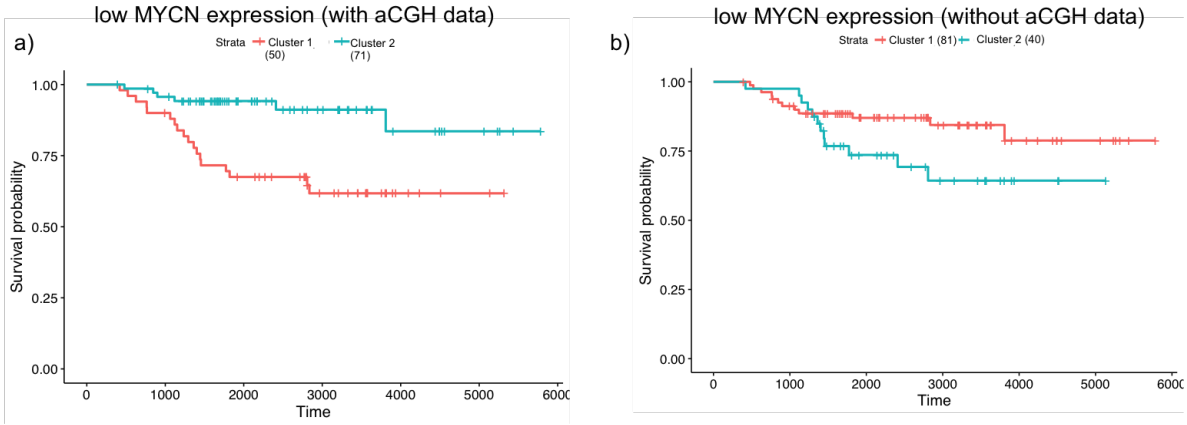

Figure 3: a) Survival curves of MKKM-inferred subgroups identified within the patients with low MYCN expression. RNA-seq and aCGH dataseta are combined. Log-rank test p-value is 0.0008. View weights are 0.45 (MAV) and 0.55 (RPM). b) Survival curves of MKKM-inferred subgroups identified within the patients with low MYCN expression. Only RNA-seq datasets are used. Log-rank test p-value is 0.07. View weights are 0.28 (MAV), 0.34 (RPM) and 0.37 (aCGH).

downloaded from CAMDA website and `GSE62564_SEQC_NB_RNASeq_log2RPM.txt` downloaded from GEO website for entry GSE62564. We downloaded the Versteeg dataset from R2 database (http:// r2.amc.nl). The GEO accession number for Versteeg microarray data is GSE16476 (*3*). We used the survival package in R to perform the Kaplan-Meier analysis (*4*).

Machine Learning Methods

We used the SVC function available in Python's scikit-learn library for prediction on `Versteeg dataset`. Linear and RNF kernels were used to train the OS and EFS models respectively. The best gamma parameter for the RBF kernel and C parameter was set with cross-validation. For multi-view kernel k-means we used the code provided in the paper by Gonen et al (*5*).

Discussion

The availability of genome-wide datasets for cancer patients have increased rapidly in recent years. Methods that can effectively integrate these datasets can improve our understanding of cancer development and progression. To this end, we used supervised and unsupervised learning strategies to predict patient survival in neuroblastoma. Our supervised model can accurately predict overall survival and event-free survival profiles of neuroblastoma patients in an independent cohort. We also inferred subgroups that have distinct survival rates by combining multiple multi-dimensional datasets. Including

microarray data in addition to RNA-seq datasets provided no improvement in silhouette-score or log-rank test p-value indicating redundancy between the two datasets. However, including aCGH data in addition to RNA-seq datasets have improved the silhouette-score and resulted in a lower log-rank test p-value. Overall, our results suggest that integration of multi-modal datasets can improve subtype definition in cancer.

References

1. RC Seeger, GM Brodeur, H Sather, A Dalton, SE Siegel, KY Wong, and D Hammond. Association of multiple copies of the n-myc oncogene with rapid progression of neuroblastomas. *N. Engl. J. Med.*, 483:589, 2012.

2. F Hertwig J Thierry-Mieg W Zhang et. al. W Zhang, Y Yu. Comparison of rna-seq and microarray-based models for clinical endpoint prediction. *Genome Biology*, 16:133, 2015.

3. JJ Molenaar, J Kosterand, DA Zwijnenburg, and P van Sluis et. al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis. *Nature*, 483:589, 2012.

4. TM Therneau. *A Package for Survival Analysis in S*, 2015. version 2.38.

5. M Gonen and AA Margolin. Localized data fusion for kernel k-means clustering with application to cancer biology. *Advances in Neural Information Processing Systems*, 2014.