Using gene-expression analysis to predict complete remission in patients with acute myeloid leukemia

Ophir Gal[1], Noam Auslander[2,3], Yu Fan[4*], Daoud Meerzaman[4*]

[1]Department of Computer Science, University of Maryland, College Park, MD 20742, United States of America

[2]Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892, United States of America

[3]Center for Bioinformatics and Computational Biology, Department of Computer Science, University of Maryland, College Park, MD 20742, United States of America

[4]The Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, MD 20850, United States of America

*Corresponding authors: meerzamd@mail.nih.gov; yu.fan@nih.gov

ABSTRACT

Machine learning is a useful tool for advancing our understanding of the patterns and significance of biomedical data. Given the growing trend on the application of Machine learning (ML) techniques in precision medicine, here we present a ML technique which predicts the likelihood of complete remission (CR) in patients diagnosed with acute myeloid leukemia (AML). In this study, we explored the question of whether machine-learning algorithms designed to analyze gene-expression patterns obtained through RNA sequencing (RNA-seq) can be used to accurately predict the likelihood of complete remission (CR) in pediatric AML patients who have received induction therapy.

We employed tests of statistical significance to determine which genes were differentially expressed in the samples derived from patients who achieved CR after two courses of treatment and the samples taken from patients who did not benefit. We tuned classifier hyperparameters to optimize performance and used multiple methods to guide our feature selection as well as our assessment of algorithm performance.

To identify the model which performed best within the context of this study, we plotted receiver operating characteristic (ROC) curves. Using the top 75 genes from the k-nearest neighbors algorithm (K-NN) model (K=27) yielded the best area-under-the-curve (AUC) score that we obtained: 0.84. When we finally tested the previously unseen test data set, the top 50 genes yielded the best AUC=0.81. Pathway enrichment analysis for these 50 genes showed that the GDP-fucose biosynthesis pathway is the most significant with an adjusted P-value = 0.0092, which may suggest the vital role of N-glycosylation in AML.

INTRODUCTION

RNA sequencing (RNA-seq) and other high-throughput next-generation sequencing platforms have emerged as powerful approaches for discovering pathogenic pathways and potential targets for clinical intervention in patients with acute myeloid leukemia (AML) [1]. Using whole-transcriptome sequencing, our previous work compared the profiles of core-binding factor acute myeloid leukemia (CBF-AML) cases to those characterized by normal karyotypes (NK), illuminating similarities and differences with respect to gene-expression signatures and splicing events as well as RNA fusions that typify the inv(16) versus the t(8;21) AML subtypes [2].

In concert with the rise of large-scale omics-oriented sequencing, machine-learning algorithms have increasingly been applied to gene expression analysis aimed at classifying tumors, predicting survival, identifying therapeutic targets, and classifying genes according to function [3-7]. Significant results have been shown for predicting outcomes of large B-cell lymphoma [8], and hepatitis B virus–positive metastatic hepatocellular carcinomas [9] as well as documenting diverse pathologic responses to chemotherapy in patients with breast cancer [10]. Using gene-expression profiling of data generated by microarrays in conjunction with both supervised and unsupervised learning, Bullinger et al. identified prognostic subclasses in adult AML; the research group also constructed an optimal 133-gene predictor of overall survival [11]. Yeoh et al. performed classification, subtype discovery, and outcome prediction in patients with pediatric acute lymphoblastic leukemia (ALL) [12]. However, no previous study has specifically addressed expression differences among large cohorts of pediatric and young-adult AML patients with regard to CR. In this study, we compare pre-treatment gene-expression profiles using three supervised learning algorithms to discover predictors of complete remission.

MATERIALS AND METHODS

We obtained 473 bone marrow specimens from 473 patients, both children and young adults with ages ranging between 8 days to 28 years who had been diagnosed with de novo AML. For comparison, we acquired an additional 20 bone marrow specimens from normal, healthy individuals. All samples were obtained by written consent from the parents/guardians of minors from the Children's Oncology Group clinical trial AAML1031. The Institutional Review Board at Fred Hutchinson Cancer Research Center has reviewed and approved this study. It is filed under IR File #9950 (Biology of the Alterations of the Signal Transduction Pathway in Pediatric Cancer. The number of samples with clinical information regarding complete remission used in this study was 414. RNA-seq was performed on all 493 samples using the Illumina platform

HiSeq2000 (https://www.illumina.com). Reads were mapped to Ensembl Gene IDs (http://useast.ensembl.org/), which belong to 31 biotypes including protein-coding sequences, non-coding sequences, and pseudogenes, among others. RPKM (Reads Per Kilobase per Million mapped reads) values were calculated for each gene. Genes that had a count of at least one per million (CPM) in at least three samples were retained. Quantile normalization was applied among all samples. Python library sci-kit learn (http://scikit-learn.org/stable/) modules of commonly used statistical models and algorithms were directly implemented in the scripts. Gene set enrichment analysis (GSEA) was performed using the online tool Enrichr (http://amp.pharm.mssm.edu/Enrichr/), as well as our in-house OmicPath (v 0.1) R package.

Feature Selection

PCA (principal component analysis) was performed to examine the general pattern of the data, remove outliers, and select algorithms appropriate for our data.

RNA-seq expression data of m samples by n genes were used as inputs, and learn the mapping using $F: x \to \{CR, Not\ in\ CR\}$

$$X \in R^{m \times n}, y \in \{0,1\}^{m \times 1}$$

Samples were divided into a training set (N=331) and a test set (N=83). Three classifiers: k-nearest neighbors algorithm (K-NN), Support Vector Machine (SVM), and Random Forest were applied to select features for the training set via 5-fold cross-validation. With the features selected, the classifier was tested on the same training set (N=331). The classifier with the best performance was then tested on the remaining test set (N=83).

KNN Classifier

We performed 100 iterations of a 5-fold cross-validation in which we carried out a t-test for initial feature selection in each fold in order to identify the 100 most statistically significant genes that were the most differentially expressed between the CR (positive class) and NCR (negative class). We found that using more than 100 genes did not improve performance. For the second feature selection, from the genes identified by t-testing, we used two algorithms: Hill Climbing [13] (sequential feature addition) and Randomized Lasso [14] (using the model's feature weights as ranks and selecting the highest ranking feature). At each fold, an AUC was computed from the corresponding validation set. Following the 100 iterations, we generated a list of the features (genes) that were ranked by the average of AUC computed using those specific genes. Essentially, the genes that helped yield the best AUCs were ranked highest.

SVM Classifier

Due to the requirement for balanced data (approximately the same number for each of the two classes being predicted), a smaller subset of 114 samples ($N_{(CR)}$=57, $N_{(NCR)}$=57; 91 for the training set and 23 for the test set) was used for the SVM. Processes similar to those described above for K-NN were applied to SVM classifiers with one exception: we used a third method, Recursive Feature Elimination for the second feature selection in addition to Hill Climbing and Randomized Lasso.

Random Forest Classifier

We used two methods of feature selection, one simple and one complex. In the first method, we trained the classifier once on 4/5 of the training set and used the built-in "feature importance" attribute to rank the features to be used a second time on the same 4/5 dataset. We then tested the classifier on the remaining 1/5 of the training set (validation set). In the more complex feature selection method, we carried out 100 iterations of 5-fold cross-validation while aggregating the feature importance computed at each fold, and then used these values to compute a Spearman correlation between each gene and the AUC computed in all folds. We used the genes with the highest correlation scores to train on the same 4/5 training set, and then tested the method on the 1/5 validation set.

Following Feature Selection

At the end of each feature selection, the same cross-validation procedure was employed to generate the AUC results when testing the validation set. The final AUC result of the (chosen) K-NN classifier was a simple, one-episode period of training on the training set with the selected genes followed by testing on the unseen test set with the same selected genes.

RESULTS

According to PCA (principal component analysis), these 414 AML samples with clinical information regarding complete remission did not cluster by CR or NCR status, nor by age/year of diagnosis. There is no obvious outliers, so all of them were included in this study.

AUC results from different K values were used to estimate optimal K for the K-NN classifier. Figure 1 shows that statistically significant genes identified from the t-test can help improve the AUC results and that K=27 yielded the best average AUC. With the optimal K=27, ROC curves were produced using two feature-selection methods: Hill Climbing and Randomized Lasso (Figure 2). Overall, the Hill Climbing resulted in better results with the best AUC=0.84.

Figure 1. AUCs from different Ks used to estimate an optimal K value for K-NN classifier

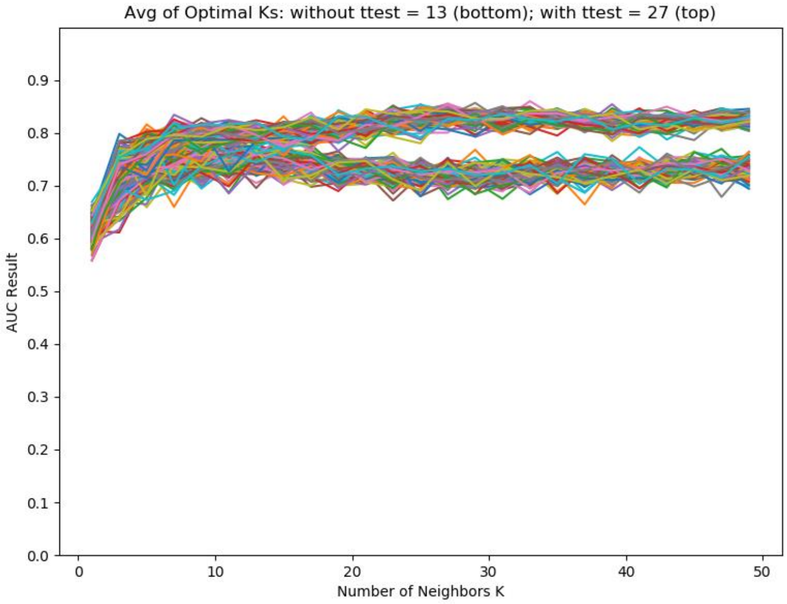
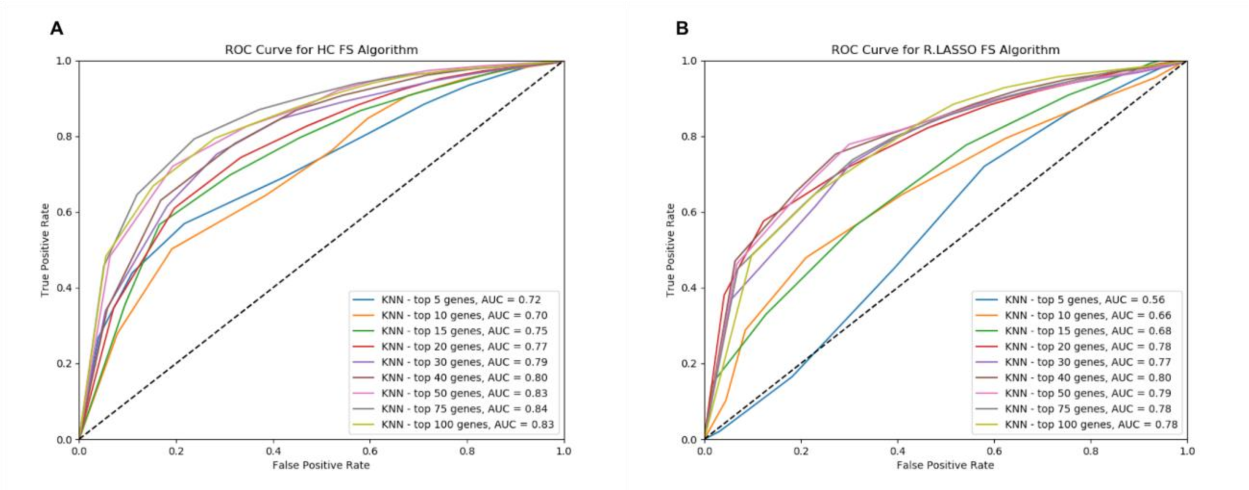


Figure 2. ROC curves of K-NN using two-feature selection methods: (A) Hill Climbing and (B) Randomized Lasso

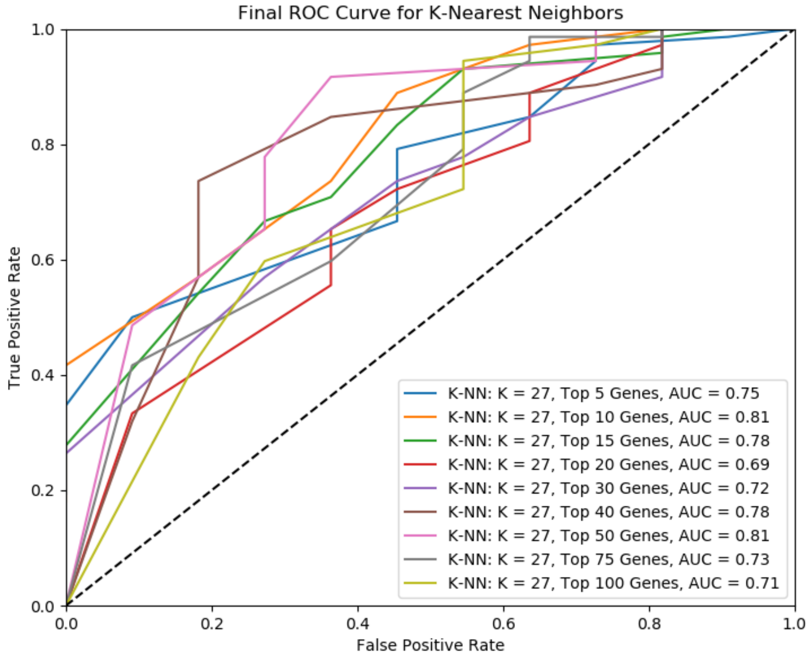


To compare the performance of K-NN and SVM classifier, the balanced data set with $N_{(CR)}=57$, and $N_{(NCR)}=57$ was split into training set (N=91) and the test set (N=23). Using a 5-fold cross-validation performed on the training set, ROC curves of K-NN and SVM algorithms were calculated using three feature-election methods: Hill Climbing, Recursive Feature Elimination, and Randomized Lasso. K-NN outperformed SVM, and Hill Climbing still resulted in better AUC results for K-NN (Figure S1).

Hyperparameter tuning for Random Forest suggested using 100 trees to have the best performance (AUC=0.74). The simple method resulted in better results with the best (training set) AUC=0.73 compared with the more complex approach (Figure S2).

Based on above observations, K-NN with Hill Climbing performed the best on the training data (N=331), yielding an AUC score of 0.84. When we tested this model on the remaining 1/5 of the data (N=83) using the top 50 genes with the best AUC scores from the training set yielded an AUC score of 0.81 (Figure 3).

Figure 3. Final K-NN model performance on test data (N=83)



Based on using these top 50 genes, our GSEA analysis using OmicPath showed that BATF (Basic leucine zipper transcriptional factor ATF-like) and RAC2 (Ras-related C3 botulinum toxin substrate 2) are related to a decreased IgM level with FDR (False Discovery Rate)=0.0073.

TSTA3 (GDP-L-fucose synthase) and RAC2 are related to an increased neutrophil cell number (FDR=0.0073). Pathway enrichment analysis using Enrichr showed that TSTA3 and FPGT (Fucose-1-phosphate guanylyltransferase) were mapped to the GDP-fucose biosynthesis pathway (Reactome 2016; https://reactome.org) with an adjusted P-value of 0.0092. These two genes were also mapped to the pathway's parent terms "Synthesis of substrates in N-glycan biosynthesis" and "Biosynthesis of the N-glycan precursor (dolichol lipid-linked oligosaccharide, LLO) and transfer to a nascent protein." This indicates the vital role of N-glycosylation in AML pathology and patient prognosis.

DISCUSSION

This study explored and evaluated different machine-learning algorithms for predicting complete remission in AML patients based on their pre-treatment gene expression signatures. It revealed a significant underlying genetic difference between patients with contrasting outcomes following treatment. GSEA results highlighted specific biological features that carry prognostic value for further exploration. For example, low IgM and leukocyte count > $50 \times 10^9/1$ have been demonstrated as two of the adverse predictors for the duration of complete continuous remission in childhood acute lymphoblastic leukemia (ALL) [15]. Fucose-containing glycans play important roles in selectin-mediated leukocyte-endothelial adhesion as well as various immunity and signaling processes. Alterations in expression or structure of fucosylated oligosaccharides have also been observed in cancer pathology. Conditional impairment in fucosylated glycan expression in mice exhibited altered myeloid development including aberrant proliferation of myeloid progenitors and an increased production of granulocytes which leads to neutrophilia. The loss of AB blood group antigen expression along with the increases in H and Lewis^y

expression are associated with poor prognosis. Increased expression of Lewis$^{x/a}$ structures, Tn/sialyl-Tn/T antigens, and β1,6 GlcNAc branching of N-linked core structures were observed in advanced cancers and related with poor prognosis. [16-19] This information may help physicians select more suitable courses of treatment, whether the treatment be more aggressive chemotherapy or an altogether novel alternative therapy.

Acknowledgements

The authors wish to acknowledge Dr. Meshinchi's group from Fred Hutchinson Cancer Research Center for data provision and Laura K. Fleming, Ph.D. for editorial critique.

Declaration of Conflicting Interests

The Authors declare that there is no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Tarlock, K. and S. Meshinchi, *Pediatric acute myeloid leukemia: biology and therapeutic implications of genomic variants.* Pediatr Clin North Am, 2015. **62**(1): p. 75-93.
2. Hsu, C.H., et al., *Transcriptome Profiling of Pediatric Core Binding Factor AML.* PLoS One, 2015. **10**(9): p. e0138782.
3. Tan, A.C. and D. Gilbert, *Ensemble machine learning on gene expression data for cancer classification.* Appl Bioinformatics, 2003. **2**(3 Suppl): p. S75-83.
4. Furey, T.S., et al., *Support vector machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 2000. **16**(10): p. 906-14.
5. Bair, E. and R. Tibshirani, *Semi-supervised methods to predict patient survival from gene expression data.* PLoS Biol, 2004. **2**(4): p. E108.

6. Lee, J.S. and S.S. Thorgeirsson, *Genome-scale profiling of gene expression in hepatocellular carcinoma: classification, survival prediction, and identification of therapeutic targets.* Gastroenterology, 2004. **127**(5 Suppl 1): p. S51-5.
7. Brown, M.P., et al., *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proc Natl Acad Sci U S A, 2000. **97**(1): p. 262-7.
8. Shipp, M.A., et al., *Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.* Nat Med, 2002. **8**(1): p. 68-74.
9. Ye, Q.H., et al., *Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning.* Nat Med, 2003. **9**(4): p. 416-23.
10. Ayers, M., et al., *Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer.* J Clin Oncol, 2004. **22**(12): p. 2284-93.
11. Bullinger, L., et al., *Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia.* N Engl J Med, 2004. **350**(16): p. 1605-16.
12. Yeoh, E.J., et al., *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.* Cancer Cell, 2002. **1**(2): p. 133-43.
13. Tsamardinos, I., Brown, L.E. & Aliferis, C.F. , *The max-min hill-climbing Bayesian network structure learning algorithm.* Machine Learning, 2006. **65**: p. 31.
14. Meinshausen, N.a.B., P., *Stability selection.* J. Roy. Statistical Society B, 2010. **72**(4): p. 417-473.
15. Miller, D.R., et al., *Prognostic factors and therapy in acute lymphoblastic leukemia of childhood: CCG-141. A report from childrens cancer study group.* Cancer, 1983. **51**(6): p. 1041-9.
16. Becker, D.J. and J.B. Lowe, *Fucose: biosynthesis and biological function in mammals.* Glycobiology, 2003. **13**(7): p. 41R-53R.
17. Smith, P.L., et al., *Conditional control of selectin ligand expression and global fucosylation events in mice with a targeted mutation at the FX locus.* J Cell Biol, 2002. **158**(4): p. 801-15.
18. Kim, Y.J. and A. Varki, *Perspectives on the significance of altered glycosylation of glycoproteins in cancer.* Glycoconj J, 1997. **14**(5): p. 569-76.
19. Orntoft, T.F. and E.M. Vestergaard, *Clinical aspects of altered glycosylation of glycoproteins in cancer.* Electrophoresis, 1999. **20**(2): p. 362-71.