Developing a Comprehensive R/Bioconductor Package for DNA Methylation-based Human Cancer Classification and Phenotyping

Jacob Fanale, SEAS 2026

PI: Zhou, Wanding, PhD, Assistant Professor of Pathology and Laboratory Medicine

Abstract

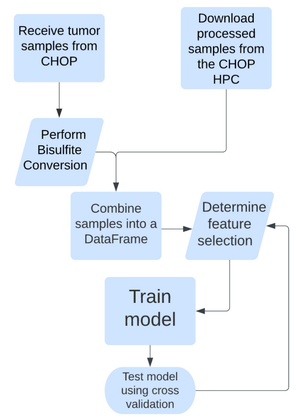
Cancer DNA methylome encodes rich and detailed molecular footprints of a tumor's cell of origin and oncogenic mechanism. Machine learning models have proven successful in predicting tumors of the central nervous system. We developed an R/Bioconductor package, CytoMethIC (Cytosine Methylation Intelligence) to provide an open-source solution for comprehensive human cancer phenotyping, encompassing automated determination of cancer type, subtype, and clinical oncology attributes such as tumor stage, cell of origin, aneuploidy, sample purity, race and sex. Our package encapsulates six different machine learning frameworks: random forest, support vector machine, multilayer perceptron, extreme gradient boosting, k-nearest neighbor, and Naïve-Bayes. Each framework utilizes optimal selection techniques and is tested to predict 33 human cancer types (91 subtypes) in The Cancer Genome Atlas cohort and 66 brain cancer types (82 subtypes) in the Children's Brain Tumor Network cohort. These datasets are profiled using different generations of Infinium BeadChip platforms. We evaluated the models for accuracy, confidence, interpretability, algorithm runtime and model storage. Our user-friendly, standard-compliant informatics facilitate the use of machine learning models for DNA methylation-based cancer classification in clinical diagnosis.

Objectives

- Examine the efficacy of 6 different machine learning classifiers for determining tumor class using the DNA Methylome.
- Test the predictability of different cancer-related attributes using Random Forest Classifier models. These attributes include tumor subtype, cell of origin, tumor stage, aneuploidy, sample purity, race, sex, and determining if a tumor is benign or malignant.
- Develop an open-source package to facilitate the simple use and distribution of the classifiers we create.

Methods and Workflow

We sourced our tumor and cancer samples from The Cancer Genome Atlas (TCGA) and from the Molecular Bioinformatics lab at the Berlin School of Integrative Oncology. We used the CHOP Respublica high-performance computing cluster to store tens of thousands of tumor samples, as well as to run our classification scripts. To train these models, we used the following R packages: randomForest, e1071, xgboost, and tensorflow. In addition, we used R's ggplot2 library to generate figures and visualize our models. We used GitHub to store all scripts that were used to ensure effective version control.



Above: Workflow for generating classification models

Model Comparison Results

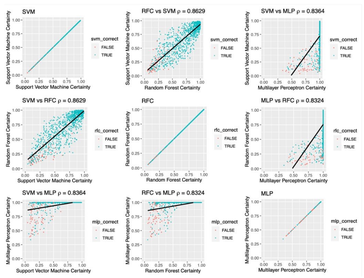
We tested our 6 different models on two different datasets each: a CNS tumor dataset and a pan-cancer dataset. We found that the multilayer perceptron, random forest, and support vector models were the most effective at solving both classification problems, in both overall accuracy and average class wise accuracy.

	Classifier	Model Train Runtime (s)	Model Test Runtime (s)	Overall Accuracy	Average Class Accuracy	Interpretable certainty score?
Capper Prospective	RFC	272.03	0.25	0.9427	0.8860	Y
	SVM	142.21	11.168	0.9409	0.9119	Y
	MLP	21.79	0.1422	0.9382	0.8773	Y
	XGB	746.01	0.021	0.9127	0.8127	Y
	KNN	38.56	NA*	0.8918	0.8489	Y
	GNB	5.65	39.784	0.8864	0.7830	NA**
TCGA	MLP	72.33	0.344	0.9580	0.9522	Y
	SVM	357.83	38.084	0.9565	0.9462	Y
	XGB	2990.87	0.062	0.9410	0.9106	Y
	RFC	1353.44	0.805	0.9153	0.8608	Y
	KNN	398.30	NA*	0.8799	0.8303	Y
	GNB	8.35	115.170	0.8136	0.8196	NA**

Model training is done with 2146 samples and 6636 features.
Model testing is done with 1104 samples and 6636 features
* The K-Nearest-Neighbors package used does not allow for separated model training and testing.
** The Naïve-Bayes package used does not allow for computation of certainty for each prediction.

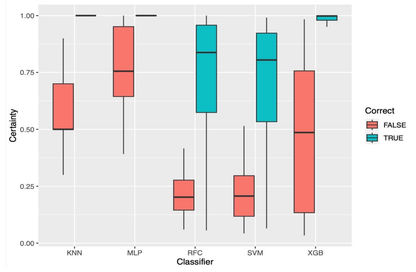
Above: Table of accuracies and runtimes for different models

These three models produce an interpretable certainty score, so we further examined this by comparing the certainty scores in a pairwise fashion by model, as seen in the scatterplots at right. Based on Spearman's Rho correlation values, we found that all certainty scores from these 3 models are comparable to each other.



Above: Pairwise correlations of certainty scores in different models

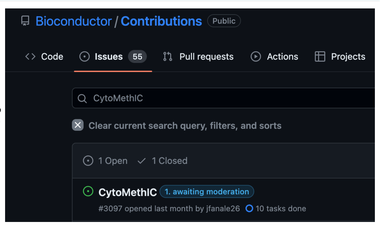
Additionally, we wanted to determine if a threshold for a "trustable" classification existed. As can be observed at right, for the three models mentioned above, there is a clear separation between correctly and incorrectly classified samples. This threshold is around 0.93 for MLP, and around 0.50 for RFC and SVM.



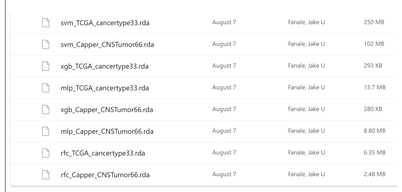
Above: Boxplot of certainty scores, separated by correct classification status

Bioconductor Package Development

Throughout the testing of these models, we developed an open-source R/Bioconductor package that is currently being reviewed by the Bioconductor staff for publication. This software, named CytoMethIC, greatly simplifies the use of our cancer type and phenotype models. The software abstracts the use of these models to a single function, allowing a user with little to no experience coding in R to classify samples.



Above: Opened Bioconductor Issue for submitted package



Above: Machine Learning Models uploaded for CNS tumor classification and pan-cancer classification

Additionally, due to the large file size of our machine learning classifier models, we created a supplementary Experiment Hub package to publish our developed models. When these publications are approved, we will add functionality to connect the Experiment Hub and Bioconductor packages such that the user can use our classifiers without needing to manually download them.

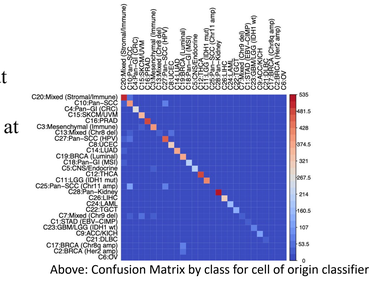
Predictability Test Results

	Model	Overall Accuracy	Average Class Accuracy
COO28	RFC	0.8000	0.7068
	SVM	0.7986	0.7049
	RFC	0.8910	0.8730
CNSTumor82	SVM	0.8773	0.8800
	RFC	0.9826	0.9698
Sex	SVM	0.9830	0.9708
	RFC	0.3713	0.2224
TumorStage21	SVM	0.3696	0.2188
	RFC	0.6816	0.4684
TCGASubtype91	SVM	0.6830	0.4656
	RFC	0.8277	0.2214
Race	SVM	0.8271	0.2191
	RFC	0.5372	0.1396
Aneuploidy	SVM	0.5372	0.1397
Purity	Regression	0.9839*	NA
		* (1 - MSE)	

Above: Table of overall and class accuracies for predictability cross-validation experiment

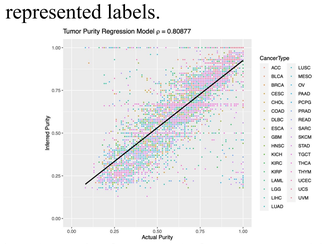
We trained a random forest and support vector model for each classification label shown at left to ascertain the predictability of different cancer subtypes and phenotypes. We held model parameters and features constant for this experiment. As is shown at left, all labels are predictable, apart from tumor type and aneuploidy, which may be due to feature selection.

Further examining the predictions, we found that the TCGA dataset is very imbalanced for some label sets, as is apparent in the confusion matrix at right. Despite this, our models generally performed very well when faced with under-represented labels.



Above: Confusion Matrix by class for cell of origin classifier

Additionally, our tumor purity regression model had a high spearman's correlation constant. As can be seen at right, the model was able to infer sample purity to a high degree of accuracy. This accuracy is maintained across carcinomas, sarcomas, leukemias, and lymphomas.



Above: Scatter plot of tumor purities from regression model, colored by cancer type

Conclusions

- Random forest, support vector, multilayer perceptron, and extreme gradient boosting models are the most accurate when presented with a multiclass DNA methylome classification problem. Additionally, these models all produce interpretable certainty scores.
- Most tumor subtypes and phenotypes are predictable, and those which are not are likely predictable using different, manually selected feature inputs.
- There is major class imbalance in both datasets that we used for training models, and this must be considered in the analysis of results.
- The software and algorithms that our lab developed show incredible promise for improving the accuracy and efficiency of oncological diagnoses.

Citations and Acknowledgements

- Capper D, Jones DTW, Sill M, et al. DNA methylation-based classification of central nervous system tumours. Nature. 2018;555(7697):469-474. doi:10.1038/nature26000
- A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.
- Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2023). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-13, <https://CRAN.R-project.org/package=e1071>.
- Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y, Yuan J (2023). xgboost: Extreme Gradient Boosting. R package version 1.7.5.1, <https://CRAN.R-project.org/package=xgboost>.
- Allaire J, Tang Y (2022). tensorflow: R Interface to 'TensorFlow'. R package version 2.11.0, <https://CRAN.R-project.org/package=tensorflow>.
- Other Contributors: Zhou, Wanding, PhD, Principal Investigator. Goldberg, David, PhD Candidate. Li, Marilyn, MD, MS. Xu, Weixuan, PhD. Xu, Feng, PhD. Kraya, Adam, PhD.
- The results here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga