Predicting Gene Ontology Enrichment of Cancer Microarray Datasets Using Machine Learning

Ethan Jacob Moyer [1], Ifeanyi Osuchukwu [1], Tony Kabilan Okeke [1]

[1] School of Biomedical Engineering, Science and Health Systems, Drexel University, Philadelphia, Pennsylvania, USA

Course : BMES 483/543
Instructor : Ahmet Sacan
Date : 2022-06-07
Dataset(s) : CuMiDa Datasets

Abstract

By analyzing 34,524 probes in 10 cancer types across 20 data sets (tumor versus normal) from the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array (GPL570) microarray platform, we built a multioutput logistic regression classifier to predict gene ontology (GO) enrichment for a small subset of important terms. By using a leave-one-out cross validation classification accuracy, we displayed that such a prediction is possible with accuracies ranging from as 45% - 80% for 22 gene ontology terms. After implementing a forward feature selection on our gene feature set accuracies improved, ranging from 80% - 100%. Our findings potentially highlight that specific gene expression levels play a vital role in mapping to gene ontology terms. This would also indicate that selected genes are most responsible for the functional properties of a given gene product.

1 Introduction

Cancer is characterized by an overt proliferation of cells originating from an initial mutation in the body. Different cancer types may attack the body via different mechanisms depending on the malignancy of the cancer, its origin, and its interactions with the body's immune system [1]. Identifying these mechanisms may allow for novel cancer therapeutics and improve outcomes for cancer patients.

Extracting biologically significant meaning from cancer-related genomic data sets is a difficult task. Cancer studies leverage many tools for determining biologically relevant mechanisms related to the disease. Once such tool is microarrays [2-5]. Microarray technology serves as an informative tool for acquiring gene-level expression data. They can be used for a variety of purposes in research and clinical studies, such as measuring gene expression and detecting specific deoxyribonucleic acid (DNA) sequences (e.g., single-nucleotide polymorphisms, or SNPs) [6].

How can we highlight the genes that are contributing most to protein function in disease such as cancer? The large number of genes on microarrays unrelated to protein function makes answering this question difficult; however, machine learning serves as a possible solution for determining which genes from a microarray platform are most important for different independent pathways. Building accurate prediction models that are easy to interpret is further complicated by the fact that these interactions are often multifaceted and non-linear, whereas most common machine learning approaches rely on a linear combination of terms. This work serves as an initial step for exploring how machine learning may aid in exploring this topic.

We selected a total of 21 Datasets, all of which contain two groups - a normal (healthy) group and a cancer group; all the datasets were generated on the GPL570 microarray. The differentially expressed genes were identified in each dataset, and a hypergeometric test was used to identify enriched Gene Ontology (GO terms [7]. One of the selected datasets was excluded because no significantly differentially expressed genes were found, so a total of 20 datasets were included in further analyses. We

rely on the annotated GO terms present in the GEO Platform File (GPL).

2 Datasets

We leveraged the Curated Microarray Database (CuMiDa). CuMiDa was created for the purpose of benchmarking machine learning methods, so it is an appropriate resource for exploring our methods in this paper [8]. We limited our analysis to cancer versus normal data sets that were run on the GPL570 platform. GPL570 was selected since the GPL files available through the Gene Expression Omnibus are annotated with Gene Ontology terms. We identified 21 such data sets. One data set did not fit our inclusion criteria of having at least one significantly differentially expressed gene. The GPL570 microarray contains 54,675 probes; probes that were not associated with any Gene Ontology terms were excluded, leaving only 34,524 probes.

Cancer	# Data sets
Prostate	3
Lung	3
Colorectal	3
Renal	2
Head/Neck	2
Gastric	2
Breast	2
Pancreatic	1
Liver	1
Leukemia	1

Table 1. Data set cancer types. The cancer types for CuMiDa data sets are tabulated in this table. Only those that satisfy the inclusion criteria for this work are included. Each data set is used as a single sample in the logistic regression model with expression values as features.

ID	Type	Samples
GSE16515	Pancreatic	51
GSE26910	Breast	12
GSE42568	Breast	116
GSE62232	Liver	91
GSE42743	Head/Neck	103
GSE12452	Head/Neck	40
GSE71935	Leukemia	46
GSE46602	Prostate	49
GSE26910	Prostate	12
GSE55945	Prostate	17
GSE27262	Lung	48
GSE19804	Lung	114
GSE18842	Lung	90
GSE66270	Renal	28
GSE53757	Renal	143
GSE79973	Gastric	20
GSE19826	Gastric	24
GSE41328	Colorectal	18
GSE8671	Colorectal	63
GSE32323	Colorectal	33

Table 2. Data set size. The CuMiDa data set sizes are tabulated in this table. Only those that satisfy the inclusion criteria for this work are included. Each data set is used as a single sample in the logistic regression model with expression values as features.

3 Methods

Differentially expressed genes between the cancer and normal groups were identified for each of the data sets using a q-value cut-off of 0.05 and a fold change cut-off of 1.5. We performed false discovery rate correction on the results of a two-sided two-sample t-test assuming independence and equal variances. We used the calculated fold-change values for the 34,524 genes as our features, yielding a 20 x 34,524 data matrix for analysis. The results of the differential expression analysis were used in a hypergeometric test to identify enriched GO terms for each dataset. GO terms that were identical across multiple datasets were excluded. This resulted in 720 GO terms across the 20 data sets.

We preprocessed these target GO terms first by removing those that are absent in less than 45 percent of samples and absent in more than 55 percent of samples. This results in 23 GO term targets. We then performed a correlation assessment to determine whether any of these GO terms were

strongly correlated with each other; highly correlated GO terms were excluded leaving us with 22 GO terms across the 20 data sets.

Logistic regression is a categorical classifier that relies on the logistic function shown in Equation 1 [9].

$$Equation\ 1\text{: } f(x) = \frac{1}{1+e^{x}}$$

Although it is typically used for binary classification to model the probability of one event occurring using the log-odds of a linear combination of predictor variables, it can be applied to multi-output classification as well. We are able to adapt this binary classification method to multiple outputs by simply fitting one model for each independent output variable.

This method was implemented in Python 3.7.13 through Google Collaboratory. The model was accessed using SciKit Learn. Before running the machine learning model, our feature set was standardized by removing the mean and scaling to unit variance. We used leave-one-out cross-validation to assess the fitness of our model. In leave-one-out cross-validation, each sample is used once as a test set while the remaining samples from the training set. We performed our feature selection method in two steps. First, the relative importance of each attribute was computed using an Extra Tree classifier with the number of tree estimators set at 50. Consequently, the attribute dataset was reduced to a 20 by 938 matrix. This helped to reduce the computational intensity of performing a forward feature selection method on 34,524 features for each GO term. Instead, feature selection was applied to our dataset resulting from the extra tree classifier. Selected genes were recorded for each GO term.

4 Experiments and Results

Our first experiment was simply to classify each of the 22 GO terms using the raw 34,524 by 20 data matrix. As expected, the resulting accuracies varied tremendously . Figure 1 displays the leave-one-out average accuracy for each GO term.

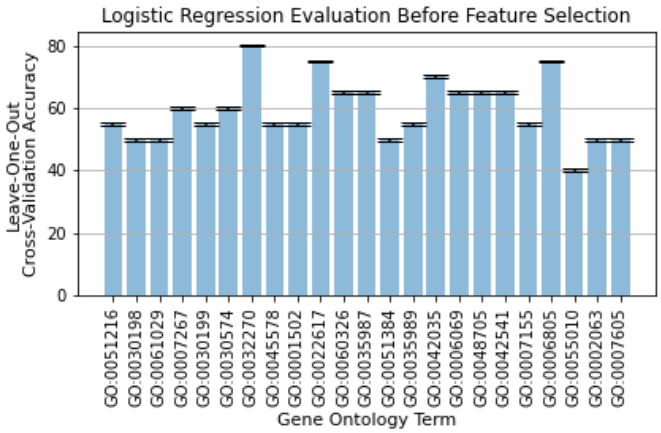

Figure 1. Logistic Regression Evaluation Before Feature Selection. This figure shows the leave-one-out cross-validation accuracy of the logistic regression classifier for each GO term.

Our second experiment was to perform forward feature selection. Figure 2 displays the leave-one-out average accuracy for each GO term.

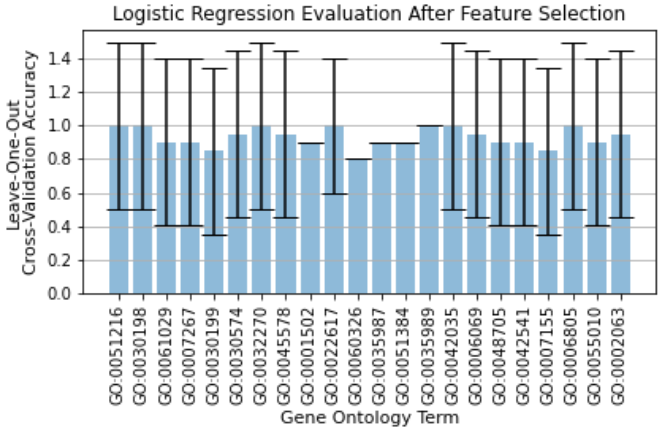

Figure 2. Logistic Regression Evaluation After Feature Selection. This figure shows the leave-one-out cross-validation accuracy of the logistic regression classifier on the different GO terms tested.

Our Logistic regression model improved .The amount of features needed to accurately predict each GO term varied from 1-4. We drastically reduced the noise in our data, and identified key genes that potentially play the most vital role in determining the presence of GO term, for a given dataset. Table 2. displays GO terms and their selected genes from our forward selection model.

GO Term	Selected Genes
Cartilage development (GO:0051216)	ATP8B3, DEFB106A /// DEFB106B, PLA2G12A, COL6A5
Extracellular matrix organization (GO:0030198)	TAF3, USP22, CACNB2
Eyelid development in camera-type eye (GO:0061029)	KLHL12, APLP2
Cell-cell signaling (GO:0007267)	GNMT
Collagen fibril organization (GO:0030199)	RAPGEF5
Collagen catabolic process (GO:0030574)	RAPGEF5
Positive regulation of cellular protein metabolic process (GO:0032270)	WNT6, PSMD6, CTNNA1
Negative regulation of b cell differentiation (GO:0045578)	OLFM2, ACVR1B
Cartilage condensation (GO:0001502)	CDHR3, FASTKD2
Extracellular matrix disassembly (GO:0022617)	MYB, COL6A5, ADAMTSL1
Cell chemotaxis (GO:0060326)	SPAG11A /// SPAG11B, PSMD6
Endodermal cell differentiation (GO:0035987)	IL17RA, ASB14
Response to glucocorticoid (GO:0051384)	WDR60, MYO18B
Tendon development (GO:0035989)	FOXA2, HPS6, LSG1, IPMK
Regulation of cytokine production (GO:0042035)	TBX6, BARD1
Ethanol oxidation (GO:0006069)	PPP2R5E, TBC1D1, SOX5
Skeletal system morphogenesis (GO:0048705)	EGF
Hemoglobin biosynthetic process (GO:0042541)	RRM2
Cell adhesion (GO:0007155)	PLA2G12A
Xenobiotic metabolic process (GO:0006805)	HSPA9, DNTTIP2, PLXNA3, POT1
Ventricular cardiac muscle tissue morphogenesis (GO:0055010)	CDHR3, DIP2A, EIF3A
Chondrocyte development (GO:0002063)	DDX25, LSG1

Table 2. Selected Genes for Different GO Terms. The features selected for each of the GO terms are tabulated above.

5 Discussion

The results of this work have a significant research impact in two ways. First, the methods explored in this work serve as a basis for classifier exploration of GO terms from microarray data. This serves as a potential future work as many other classifiers may be well-suited for this classification problem. Second, the feature selection technique explored in this paper allows us to subset the genes under study to those that are the most important for each GO term. Determining whether these selected features correspond to the predicted GO term is also left as future work.

Our results are likely influenced by the fact that we performed single output label classification. The target GO terms were unrelated and chosen based on a naive present/absence heuristic. This is a major shortcoming of our study as many gene products and gene pathways involve several GO terms. Another future work includes performing this same workflow but for categories of related GO terms. These need to be selected based on certain biological mechanisms, which can be done using a subpopulation analysis of the occurrence of GO annotations on the array platform. This would allow for greater biological validation as the targets would correspond to something much more biologically meaningful.

6 References

1. Dunn, G. P., Old, L. J., & Schreiber, R. D. (2004). The immunobiology of cancer immunosurveillance and immunoediting. *Immunity*, *21*(2), 137-148.

2. Kumar, R., Sharma, A., & Tiwari, R. K. (2012). Application of microarray in breast cancer: An overview. *Journal of pharmacy & bioallied sciences*, *4*(1), 21.
3. Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., ... & Chinnaiyan, A. M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, *6*(1), 1-6.
4. Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., ... & Chinnaiyan, A. M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences*, *101*(25), 9309-9314.
5. Glinsky, G. V., Berezovska, O., & Glinskii, A. B. (2005). Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *The Journal of clinical investigation*, *115*(6), 1503-1521.
6. Cutler, D. J., Zwick, M. E., Carrasquillo, M. M., Yohn, C. T., Tobin, K. P., Kashuk, C., ... & Chakravarti, A. (2001). High-throughput variation detection and genotyping using microarrays. *Genome research*, *11*(11), 1913-1925.
7. Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic acids research, 32(suppl_1), D258-D261.
8. Feltes, B. C., Chandelier, E. B., Grisci, B. I., & Dorn, M. (2019). Cumida: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. Journal of Computational Biology, 26(4), 376-386.
9. Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). Logistic regression (p. 536). New York: Springer-Verlag.
10. Asif, M., Martiniano, H. F., Vicente, A. M., & Couto, F. M. (2018). Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. PloS one, 13(12), e0208626.
11. Eilers, P. H., Boer, J. M., van Ommen, G. J., & van Houwelingen, H. C. (2001, June). Classification of microarray data with penalized logistic regression. In Microarrays: optical technologies and informatics (Vol. 4266, pp. 187-198). International Society for Optics and Photonics.
12. Sartor, M. A., Leikauf, G. D., & Medvedovic, M. (2009). LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics, 25(2), 211-217.
13. Shen, L., & Tan, E. C. (2005). Dimension reduction-based penalized logistic regression for cancer classification using microarray data. *IEEE/ACM Transactions on computational biology and bioinformatics*, *2*(2), 166-175.