Predicting Gene Ontology Enrichment of Cancer Microarray Datasets Using Machine Learning

Ethan Jacob Moyer, Ifeanyi Osuchukwu, Tony Kabilan Okeke

Introduction

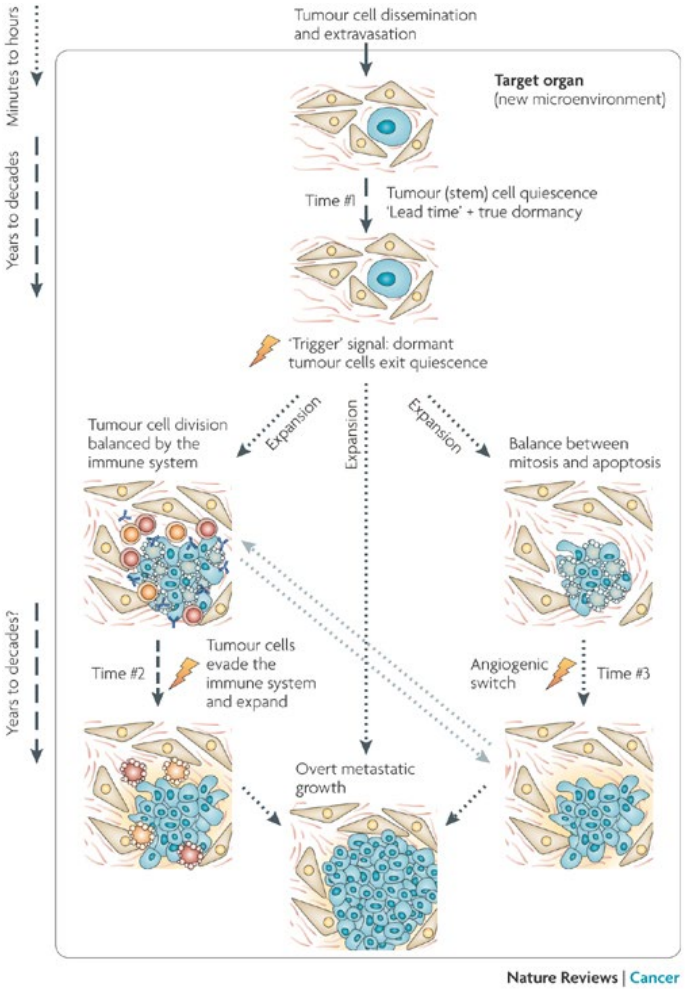
Problem Description

- Extracting biologically significant information from genomic data is a difficult task.
- Microarrays
    - Microarray technology can be used for a variety of purposes in research and clinical studies, such as measuring gene expression and detecting specific DNA sequences (e.g., single-nucleotide polymorphisms, or SNPs)
- 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 building accurate prediction models that are easy to interpret difficult.

Introduction

Background

- Cancer is characterized by an overt proliferation of cells
- Different cancer types may attack the body via different mechanisms
- Identifying these mechanisms may allow for novel cancer therapeutics



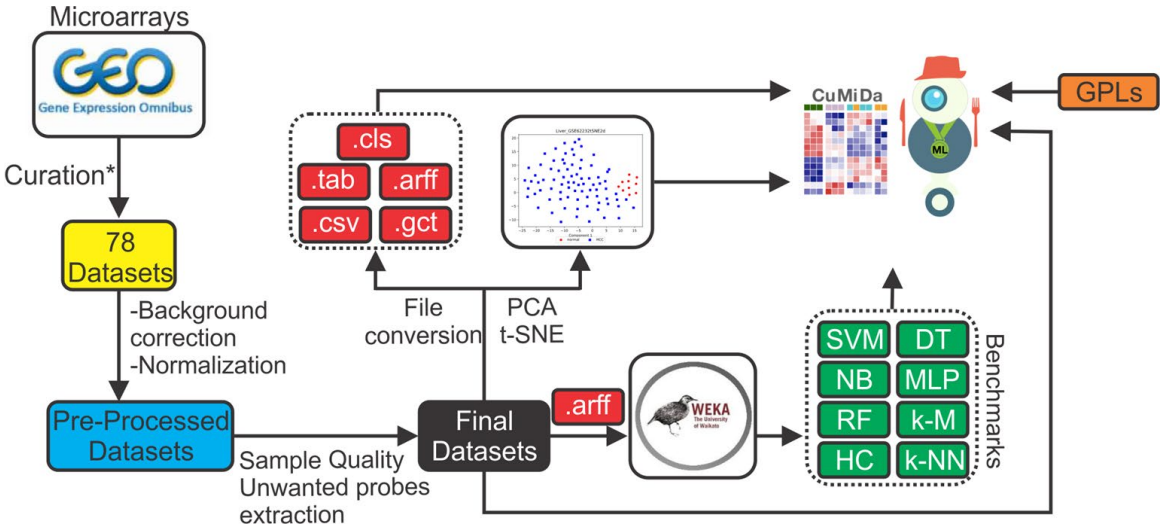
Introduction

Goals

- Determine if it possible to predict Gene Ontology Terms (GO term) using Logistic regression as our machine learning model and gene expression data as our featureset.
- Highlight which probset genes contribute to predicting the presences of a GO term the most.

The Datasets

- We leveraged the Curated Microarray Database (CuMiDa) which was created to benchmark machine learning methods on gene expression data.
  - CuMiDa data sets have undergone background correction and normalization, and low quality probes were excluded.



The Datasets

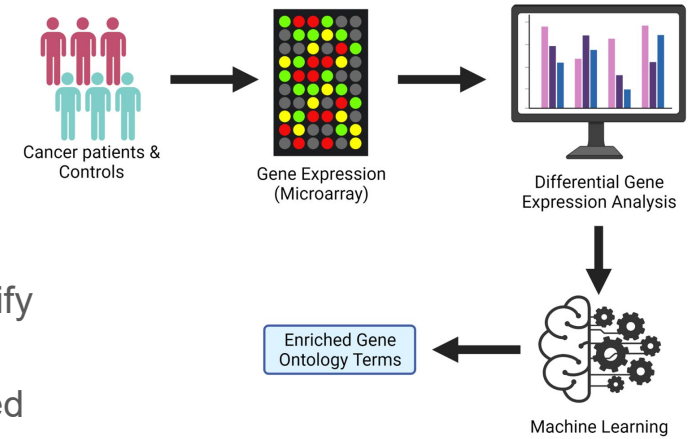
- We limited our analysis to datasets containing only cancer and normal samples that were run on the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array).
  - We included 20 datasets (across 10 cancer types) in our analysis.
  - The GPL570 microarray contains 54,675 probes. Probes without corresponding GO terms were excluded, leaving us with 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

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

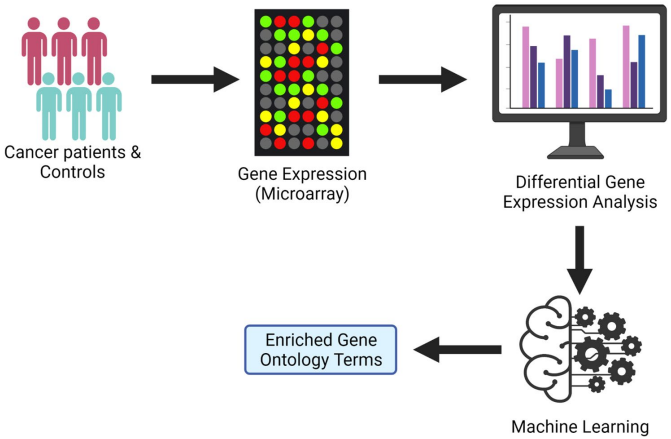
Methods

- The differentially expressed genes between the cancer and normal groups were identified for each of the data sets.
    - We used a q-value cut-off of 0.05 and a fold-change cut-off of 1.5 after performing FDR correction to identify DEGs.
- We then performed hypergeometric tests to identify enriched GO terms in each dataset.
    - GO terms that were either identical or highly correlated were excluded from further analysis.
    - We also excluded GO terms absent in less than 45% or more than 55% of datasets.
- This gave us a 20 x 34,524 feature matrix and a 20 x 22 matrix of GO labels.



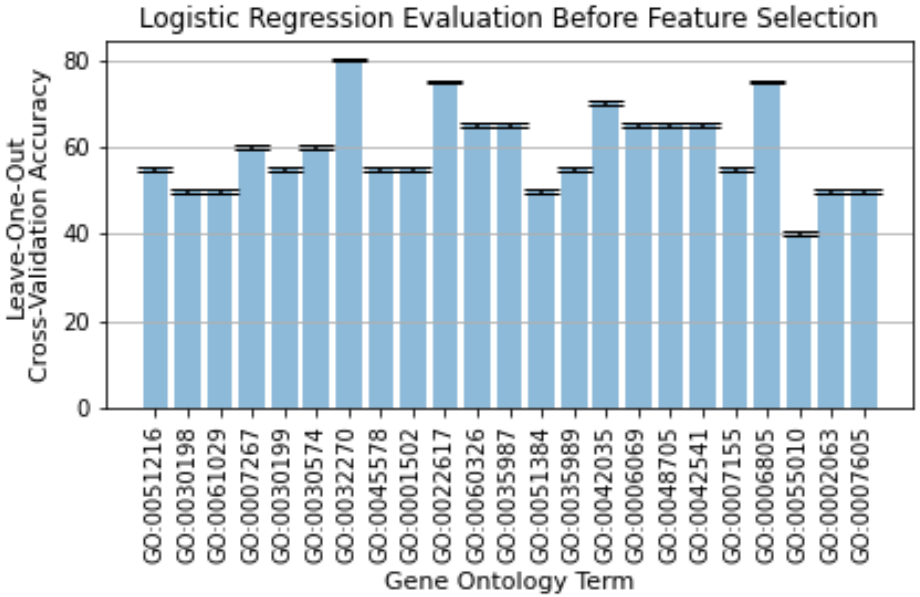
Methods

- Logistic regression
  - Single Label Output Classification
- Before running the machine learning model, our feature set was standardized using SciKit Learn's StandardScaler function. Which removes the mean and scaling to unit variance.
- Leave one out Cross Validation Score
- Feature Selection
  - The relative importance of each attribute was computed using an Extra Tree classifier with the number of tree estimators set at 50.
  - Forward Feature Selection



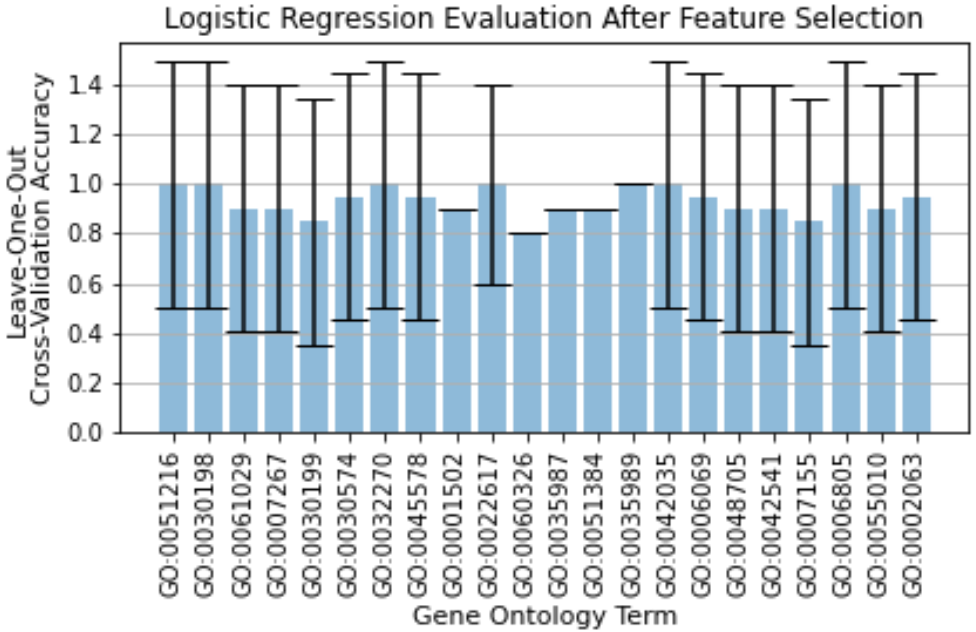
Results

- 22 GO terms
- 20 by 34,524 data matrix.



Results

- 22 GO terms
- Features selected from a 20 by 938 matrix



Results

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

GO Term	Selected Genes
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
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

Discussion

- 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.
- Feature Selection implemented in this study 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.
- Results are likely influenced by the fact that we performed single output label classification. 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.

References

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.

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.

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic acids research, 32(suppl_1), D258-D261.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). Logistic regression (p. 536). New York: Springer-Verlag.

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.

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.

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.