ANALYZING TUMOR GENE EXPRESSION DATA WITH DEEP LEARNING

Soumya Ram and Prachi Sinha

ABSTRACT

Current methods use probabilistic differential expression analysis software to find relevant pathways for cancerous tissue types. However, one can harness deep learning's ability to handle complexity to improve existing approaches. Here, we train a deep learning classifier on lung tissue. We perform guided backpropagation to find the relevant genes in the input, for the prediction. We benchmark this method against DESeq using gene ontology analysis. Our classifier is very accurate, with 99% accuracy. Through the gene ontology analysis, our deep learning methods uncovers many additional pathways that are supported by the literature.

INTRODUCTION

Advances in RNA-seq capabilities along with reduced costs have lead to rapid developments and applications of such methodologies to profiling gene expression levels, which can be used to improve understanding and identification of disease biomarkers, and further down the line, even improve drug development[6][10][21]. There are many steps in the pipeline for RNA-seq analysis, and one of the most important ones is differential gene expression analysis. Typically, such analysis is used to identify genes which are differentially expressed in two or more conditions. Currently, most expression analysis tools rely on statistical methods to identify differentially expressed genes (DEGs), because they use read counts that are probabilistically assigned to transcripts[1]. Commonly used expression analysis tools such as EdgeR[20] and DESeq[2] use discrete probability distributions, such as the Poisson or negative binomial[3][4]. Recently, some new tools have been developed to analyze gene expression data in order to differentiate cancer types using deep-learning models[7][9]. Unlike most differential analyses used to identify markers for specific phenotypes which compare normal to disease expression levels, these deep learning models for gene expression analysis focus on comparing and classifying tumor types.

In this paper, we seek to develop a novel tool that uses deep learning methods to predict biomarkers for cancer types by identifying DEGs for lung cancer by comparing normal and cancerous tissues. To enable better biomedical analyses, UCSC compiled the UCSC Toil RNAseq Recompute Compendium[8], which aligned and normalized RNA samples from multiple datasets, two of which we are interested in: The Cancer Genome Atlas (TCGA) and

Genotype-Tissue Expression (GTEx). As described above, we aim to develop a tool that can identify biomarkers for cancer types through differential expression analysis of normal and cancerous tissue samples. For the purposes of testing and developing our proposed methodology, we decided to focus on lung cancer. The Toil dataset contains over 1000 samples for lung tissue and RNA-seq data for over 60,000 genes. It would be very difficult to use standard differential analysis software tools on this scale of data, and even many deep learning methods such as KNNs would be difficult to use because of the dimensions of the data. We are able to overcome these problems by first pre-processing the data to only consider genes with high variance across normal and cancerous tissue, and then representing the data as images which can be classified as cancerous or non cancerous using a convolutional neural network (CNN) [8]. Simply being able to classify samples as cancerous or non-cancerous using gene expression data doesn't have much clinical significance, but we can take our deep learning model and apply backpropagation to identify which factors (genes) contribute most strongly to the classification. Making the assumption that the genes which contribute most strongly to the classification process are potential biomarkers for disease, we can use the results of back propagation to find the top 250 genes which may be related to lung cancer.

Because lung cancer is fairly well researched and many of its biomarkers and related pathways are well documented, we plan to validate our results using gene enrichment pathway analysis[12], which identifies statistically significant biological pathways that are enriched in a gene list, and by comparing our list of genes to documented lung cancer biomarkers in current literature[8][13][14]. In addition, in order to compare the effectiveness of our method with existing statistical expression analysis tools, we plan to compare our results to the DEGs identified by DESeq on a reduced number of samples. We chose DESeq because it has been reviewed as one of the most balanced software in regards to precision, accuracy and sensitivity[5].

SPECIFIC AIMS

Aim 1: Classify cancerous and normal tissue samples

Acquire aligned and normalized RNA-seq gene expression data samples for normal and cancerous lung tissue, and preprocess the data to reduce the number of relevant genes and represent the data as images with reshaped (reduced) dimensions. Develop a convolutional neural network to classify images of gene expression data as cancerous or non-cancerous.

Aim 2: Identify and validate relevant genes

Modify and apply Grad-Cam, a network interpretability tool that highlights the inputs that are relevant to classification, to our classifier. Take the top 250 genes identified, and validate the results in three different ways. First use gene enrichment pathway analysis to see if the identified genes are actually relevant to lung cancer. Second, compare the generated list of genes to

documented lung cancer biomarkers in current literature. Lastly, take a random subset of samples and use DESeq to generate a list of DEGs which can be compared to our list to determine the effectiveness of our method.

METHODS

Our first step was to train a deep learning classifier on cancerous vs non-cancerous lung tissue. For our data, we chose the normalized read-counts for lung tissue from the GTEx and TCGA datasets. We acquired our data from the UCSC RNA-Seq Recompute, which preprocessed both datasets such that they were consistent.

We then did further preprocessing for our deep learning classifier. Because we wanted to feed in an "image", we organized genes such that those from the same chromosome were next to each other and reshaped the array of values into a 105x105 pixel image. In addition, we filtered out genes with a variance across samples less than 1.5. Lastly, we normalized our image so all values were between 0 and 1. We had 1300 samples in total, where 300 were from normal tissue and 1000 were from cancerous tissue.

We designed a small neural network in PyTorch to act as the classifier. We chose to keep it small due to our limited computing power and small sample size. The details are found below.

Figure 1

First Part:	Second Part:
nn.Conv2d(1, 32, kernel_size=5, padding=1) *nn.BatchNorm2d(32)* *nn.Conv2d(32, 64, kernel_size=5, padding=1)* *nn.BatchNorm2d(64)* *nn.Conv2d(64, 100, kernel_size=3, padding=1)* *nn.BatchNorm2d(100)*	*nn.Dropout2d(p=0.25, inplace=False)* *nn.MaxPool2d(kernel_size=2, padding=0, stride=2)* *nn.Linear(100*12*12, 512)* *nn.Linear(512,1)*

Then, we applied guided-backpropagation to the neural network in order to predict the locations in the input that were most relevant to the final result. After this, we extracted the ~500 most relevant genes and analyzed their correlations using gene ontology software. We benchmarked these results against gene ontology software applied to the results of the DESeq analysis. We chose the gene ontology software DAVID, as it was a long-standing software that had been cited over 16,000 times.

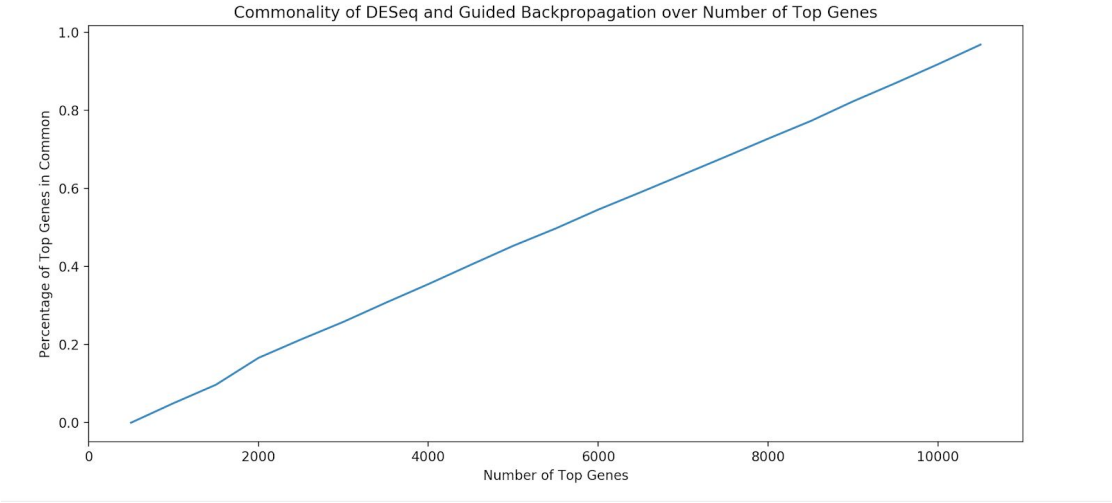
RESULTS

Aim 1: Classify cancerous and normal tissue samples

We were initially concerned about our small sample size and disparity in sample percentages, as our data was 23% normal tissue and 77% cancerous tissue. However, our classifier performed very well. Our accuracy was 99%. Cancerous tissue was classified correctly 100% of the time, and normal tissue was classified correctly 97% of the time.

Aim 2: Identify and validate relevant genes

Our two methods for acquiring the most relevant genes, DESeq and Guided-Backpropagation on the input, yielded very different results. Below is a graph displaying the percentage commonality in their top genes, as a function of the total number of genes.



Subsequently, when analyzing the different top gene lists with gene ontology software, we also saw different results.

For DESeq, the Disease Classes and Kegg Pathway results are below, respectively.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GAD_DISEASE_CLASS	CARDIOVASCULAR	RT		56	27.6	1.3E-2	2.0E-1
	GAD_DISEASE_CLASS	UNKNOWN	RT		25	12.3	2.7E-2	2.2E-1
	GAD_DISEASE_CLASS	REPRODUCTION	RT		17	8.4	2.9E-2	1.6E-1

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Neuroactive ligand-receptor interaction	RT		9	4.4	2.6E-3	2.1E-1

For Guided-Backpropagation, the Disease Classes Kegg Pathway results are below, respectively.

5 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GAD_DISEASE_CLASS	NORMALVARIATION	RT		9	6.2	5.5E-3	9.4E-2
	GAD_DISEASE_CLASS	METABOLIC	RT		39	26.7	7.5E-3	6.6E-2
	GAD_DISEASE_CLASS	VISION	RT		9	6.2	2.7E-2	1.5E-1
	GAD_DISEASE_CLASS	OTHER	RT		15	10.3	3.5E-2	1.5E-1
	GAD_DISEASE_CLASS	REPRODUCTION	RT		10	6.8	6.2E-2	2.1E-1

6 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Fat digestion and absorption	RT		4	2.7	1.1E-3	1.1E-1
	KEGG_PATHWAY	Steroid hormone biosynthesis	RT		3	2.1	3.9E-2	8.5E-1
	KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450	RT		3	2.1	6.0E-2	8.6E-1
	KEGG_PATHWAY	Neuroactive ligand-receptor interaction	RT		5	3.4	6.0E-2	7.8E-1
	KEGG_PATHWAY	Chemical carcinogenesis	RT		3	2.1	6.8E-2	7.5E-1
	KEGG_PATHWAY	Prostate cancer	RT		3	2.1	8.1E-2	7.4E-1

Unfortunately, neither of the Disease classes was accurate. The overlap across both methods was the Disease class of reproduction, which does not have any relevance to lung cancer.
The overlapping pathway, Neuroactive ligand-receptor interaction, was found to be relevant in lung cancer by other sources[16].

Four of the additional KEGG pathways found using the deep learning method, fat digestion and absorption[17], steroid hormone biosynthesis[18], metabolism of xenobiotics by cytochrome P450[19], and chemical carcinogenesis were found to be relevant to lung cancer through the literature.

However, the additional pathway of prostate cancer was not found to be relevant.

DISCUSSION

We were surprised by the accuracy of our classifier, as 99% is a very high accuracy rate. This points to the existence of a clear biomarker of cancerous lung tissue, showing that our original goal is at least possible to achieve.
In addition, we were also surprised by the sheer difference in the top genes for guided backpropagation and DESeq. An explanation for this could be that guided backpropagation takes

into account pixel interactions and dependencies, as opposed to DESeq which models genes as being independent. This could also explain the diversity of pathways discovered by guided backpropagation versus DESeq. From the greater number of discovered pathways supported by the literature, one would conclude that guided backpropagation is a better tool than DESeq.However, it is also more prone to giving false positives, such as with the prostate cancer pathway. Therefore, it provides a good starting point for possible pathways to check. However,the fact that the gene ontology analysis gives the wrong overall disease classifications leads us to doubt the validity of the results. Different gene ontology packages need to be explored before making a definite conclusion.

FUTURE GOALS

For our future goals, we'd like to experiment with different gene ontology software. It was strange that DAVID did not output lung cancer as the disease, so it is unclear if this error is due to the top genes data or the gene ontology software. Trying different software can resolve this question.

In addition, we'd like to benchmark DESeq and guided backpropagation on different cancer. This would establish a strong overall relationship.

Lastly, to get the most relevant genes from guided backpropagation, the gradients for all of the images were averaged, and the most significant pixels were taken. However, if the cancerous images had high levels for certain genes and cancerous genes had low levels, this averaging would cancel out important genes. Therefore, it would be useful to repeat the same experiments averaging just the cancerous images.

ORIGINAL PROPOSAL

Our original proposal was very different from our project. Our initial idea was to use adversarial perturbations to "learn" the functional requirements for a cure. However, based on class feedback we realized that this was not feasible and changed our project accordingly.

EXPERIENCE

Overall, we had a positive experience with our project. Initially, we had a lot of challenges trying to acquire the RNASeq data. The xenapython package we were using kept giving us 500 Errors, preventing us from downloading the data. We spent approximately ~8 hours trying to figure out the reason for this before migrating to another package. In retrospect, we should try out different packages if one is giving us many errors.

PEER REVIEW

We found the peer review to be quite helpful. Our initial idea of adversarially perturbing a cancer vs normal tissue classifier to learn a functional cure was not feasible, and we learned this through the review. In addition, the peer reviews provide valuable feedback to narrow our scope. Taking these two main points into account, we completely revised our project.

DIVISION OF LABOR

Soumya worked on the deep learning classifier and the guided backpropagation. Prachi worked on the differential expression analysis and pathway analysis. Both worked on the data preprocessing together.

REFERENCES

[1] Conesa, A., Madrigal, P., Tarazona, S. et al. A survey of best practices for RNA-seq data analysis. Genome Biol 17, 13 (2016) doi:10.1186/s13059-016-0881-8

[2] Anders, Simon, and Wolfgang Huber. “Differential expression analysis for sequence count data.” Genome biology vol. 11,10 (2010): R106. doi:10.1186/gb-2010-11-10-r106

[3] Bullard, James H et al. “Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.” BMC bioinformatics vol. 11 94. 18 Feb. 2010, doi:10.1186/1471-2105-11-94

[4] Mark D. Robinson, Gordon K. Smyth, Moderated statistical tests for assessing differences in tag abundance, Bioinformatics, Volume 23, Issue 21, 1 November 2007, Pages 2881–2887, https://doi.org/10.1093/bioinformatics/btm453

[5] Costa-Silva J, Domingues D, Lopes FM (2017) RNA-Seq differential expression analysis: An extended review and a software tool. PLoS ONE 12(12): e0190152. https://doi.org/10.1371/journal.pone.0190152

[6] Mortazavi, A., Williams, B., McCue, K. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621–628 (2008) doi:10.1038/nmeth.1226

[7] Boyu Lyu and Anamul Haque. 2018. Deep Learning Based Tumor Type Classification Using Gene Expression Data. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '18). ACM, New York, NY, USA, 89-96. DOI: https://doi.org/10.1145/3233547.3233588

[8] Toil enables reproducible, open source, big biomedical data analyses.
Nat Biotechnol. 2017 Apr 11;35(4):314-316. doi: 10.1038/nbt.3772.
John Vivian, Arjun Arkal Rao, Frank Austin Nothaft, Christopher Ketchum, Joel Armstrong, Adam Novak, Jacob Pfeil, Jake Narkizian, Alden D Deran, Audrey Musselman-Brown, Hannes Schmidt, Peter Amstutz, Brian Craft, Mary Goldman, Kate Rosenbloom, Melissa Cline, Brian

O'Connor, Megan Hanna, Chet Birger, W James Kent, David A Patterson, Anthony D Joseph, Jingchun Zhu, Sasha Zaranek, Gad Getz, David Haussler & Benedict Paten
[9] Gao, Feng et al. “DeepCC: a novel deep learning-based framework for cancer molecular subtype classification.” Oncogenesis vol. 8,9 44. 16 Aug. 2019, doi:10.1038/s41389-019-0157-8
[10] Rapaport, Franck et al. “Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.” Genome biology vol. 14,9 (2013): R95. doi:10.1186/gb-2013-14-9-r95
[11] Integrative analysis of the melanoma transcriptome.
Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, Johnson LA, Robinson J, Verhaak RG, Sougnez C, Onofrio RC, Ziaugra L, Cibulskis K, Laine E, Barretina J, Winckler W, Fisher DE, Getz G, Meyerson M, Jaffe DB, Gabriel SB, Lander ES, Dummer R, Gnirke A, Nusbaum C, Garraway LA
Genome Res. 2010 Apr; 20(4):413-27.
[12] Reimand, J., Isserlin, R., Voisin, V. et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc 14, 482–517 (2019) doi:10.1038/s41596-018-0103-9
[13] Biomarkers in the lung cancer diagnosis: a clinical perspective.
X. Li, T. Asmitananda, L. Gao, D. Gai, Z. Song, Y. Zhang, H. Ren, T. Yang, T. Chen, M. Chen Neoplasma. 2012; 59(5): 500–507. doi: 10.4149/neo_2012_064
[14] Molecular biomarkers for lung adenocarcinoma
Olivier Calvayrac, Anne Pradines, Elvire Pons, Julien Mazières, Nicolas Guibert
European Respiratory Journal Apr 2017, 49 (4) 1601734; DOI: 10.1183/13993003.01734-2016
[15] arXiv:1610.02391
[16]Shi, Ke, et al. "Identification of Key Genes and Pathways in Female Lung Cancer Patients Who Never Smoked by a Bioinformatics Analysis." *Journal of Cancer* 10.1 (2019): 51
[17]Yang, Yang, Meng Wang, and Bao Liu. "Exploring and comparing of the gene expression and methylation differences between lung adenocarcinoma and squamous cell carcinoma." *Journal of cellular physiology* 234.4 (2019): 4454-4459..
[18]Fan, Ziwei, et al. "Association between the CYP11 family and six cancer types." *Oncology letters* 12.1 (2016): 35-40.
[19]Zhang, Ji Y., Yue Fen Wang, and Chandra Prakash. "Xenobiotic-metabolizing enzymes in human lung." *Current drug metabolism* 7.8 (2006): 939-948.
[20] Robinson, Mark D et al. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” Bioinformatics (Oxford, England) vol. 26,1 (2010): 139-40. doi:10.1093/bioinformatics/btp616
[21] Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC bioinformatics. 2010;11(1):422. pmid:20698981