Semester Project Report

Predicting Cell Type and State using only Gene Expressions

Submitted by:
Peshal Agarwal
Student ID: 18-949-537
Course ID: 401-3630-04L

Supervisors:
Dr. Natalie R. Davidson
Dr. Gunnar Rätsch
Dr. Lukas Meier

1 Introduction

One of the popular ways to understand genetic functions is to directly map a genetic change (genotype) to a biological characteristic (phenotype) and study the connections. However, we know that phenotypic response results from genetic change through a long chain of complex relationships. Some have also attempted to model pairwise interaction between the genes, however, even this does not cover the complete picture. Recently, Yu et al. [2016] have presented a strategy based on hierarchical models to map the effect of genotype to phenotype. This allows us to combine the effect of genetic variations at various scales by utilizing hierarchical information provided by gene ontology (GO) databases. The authors introduced a concept of "ontotype" representing the biological processes between the genotype and the phenotype. They create a graph from genotype to phenotype via ontotype based on GO and study interactions between various genes.

It has been an open research question for decades to identify which mutations are causative of cancer. Thus, we must not just look for mutations and instead check for pathway activation and gene expressions since, it allow us to make predictions at an individual level rather than for a cohort of people. It also helps us understand the efficacy of a particular drug. So far, in the literature, people have looked at pathways independent of one another but in this project we aim to study the interaction between multiple pathways and learn the interaction between them. It is important to note that two different mutations can lead to same pathway activation while a single mutation can lead to multiple pathway activations.

Tumors comprise mostly of epithelial cells and mutations transform them into cancerous cells. These cancers grow by themselves multiplying in large numbers spreading into local tissues through metastasis. "The epithelial–mesenchymal transition (EMT) [1] is a process by which epithelial cells lose their cell polarity and cell-cell adhesion, and gain migratory and invasive properties to become mesenchymal stem cell." EMT is one such process that occurs during metastasis (Yeo et al. [2017]).

"Hypoxia [1] is a medical condition in which the body or a region of the body is deprived of adequate oxygen supply at the tissue level". When tumor cells develop such a condition, it is referred to as tumor hypoxia. Rapid growth of tumor cells leads to lower concentration of oxygen in certain regions of tissues. Cancer cells are known to change their metabolism in such difficult micro-environments so that they continue to grow. Due to this change, standard therapies do not work efficiently leading to cancer progression. Thus, it is important to study as hypoxia along with a pathways it induces.

According to studies (Muz et al. [2015], Joseph et al. [2018]), hypoxia microenvironment induces EMT in different types of cancers such as ovarian cancer, prostrate cancer and breast cancer. This could be because they share some of the pathways. People have studied these in isolation but in this project we model them together, trying to understand about their interplay. We use gene expression data publicly accessible databases and apply machine learning models to differentiate between hypoxic and non-hypoxic tissues as well as between epithelial and mesenchymal.

2 Modeling

The Cancer Genome Atlas (TCGA) is an ongoing collective effort funded by the US government to systematically record the mutations in genes that cause cancer. The Genotype-Tissue Expression (GTEx) is a project that provides access to tissue-gene expression data apart from other services for a comprehensive study of gene expressions and regulation. We deal with cancer genomic data related to hypoxia and EMT pathways from both of these databases. We start with simple classification

[1] Defintion taken from www.wikipedia.org

tasks and then slowly move to complex analysis. Our goal is to predict the hypoxic state and tissue type simultaneously, but we initially classify each of them separately and then analyse multi-output prediction.

Note that since there are no separate train and test files, we always split the data into training (two-third) and testing (one-third), stratifying on the labels since the data is imbalanced almost always. We also perform 5-fold cross-validation to search for hyper-parameters using grid search, wherever applicable, all throughout our analysis. We also mean center and scale our data. Whenever we refer to accuracy, it is always on the "unseen" test data.

2.1 Hypoxia

We first consider the hypoxia data containing 34 gene (known to be associated with hypoxia) expression for each of the 11,234 samples from real patients from both the databases. We aim to predict tumor vs normal based on the assumption that normal samples are more normoxic than tumor samples. We intially take samples only from TCGA database, dropping to 9461 samples. We implement the classic logistic regression, basic decision tree algorithm, and XGBoost classifier to predict hypoxic samples. Results shown in Table 1 clearly show that XGBoost outperforms the other two on the test data not just in terms of accuracy but also the AUROC and AUPRC scores. Thus, we always implement XGBoost classifier for our further analysis.

In order to know if the information from TCGA is sufficient for GTEx, we train on the former and predict on the latter. While GTEx contains only non-hypoxic samples but the model ends up predicting less than one-fifth correctly, leading to rejection of our hypothesis. We also try to include metadata such as type of study and gender but it hardly helps to improve model performance.

We now take the complete dataset (TCGA and GTEx both) and train the XGBoost classifier. The confusion matrix (Table 2) shows that it wrongly predicts only 44 hypoxic samples as normal. We also look at top five "important" features (genes) that help in classification, using $gain$[2] as the metric of importance. The results conform with the findings in the literature. To analyze the difference in expression, we make boxplots (Figure 1) for each of these genes. The plots clearly show that expression values of the hypoxic samples are more spread and contains most outliers indicating abnormality, which indicates their role in activating hypoxia pathways.

2.2 Tissue Type

Now, we aim to train a model that can distinguish between epithelial and mesenchymal tissue just based on gene expression values. Once again, we initially consider samples only from TCGA database. It contains 7753 samples out of which only 237 are mesenchymal. Figure 2 shows the count of hypoxic samples for each of the tissue types in the TCGA dataset. Moreover, we have 249 genes (predictors) in contrast to just 34 genes (hypoxia) for each sample. Despite such large dimensions, Logistic Regression is able to classify with 99% accuracy on the test data. Table 3 shows the confusion matrix, with just 7 false negatives. However, when we try to predict on GTEx data while training on TCGA the accuracy slips down to 55%. This could be because the tissues studied in the two databases are different and all the samples in GTEx data are non-hypoxic but not so in TCGA.

This prompts us to separately train and test on GTEx data. We remove samples for which the tissue type is unknown leaving us with 1504 samples out of which 972 are mesenchymal. The accuracy now jumps back to over 99% over 497 test samples. Table 5 shows the confusion matrix,

[2]Gain implies the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model.

Model	AUPRC	AUROC	Accuracy
Logistic Regression	0.9930	0.9318	0.9395
Decision Tree	0.9895	0.8790	0.9232
XGBoost	0.9994	0.9938	0.9824

Table 1: Comparison of models for classifying hypoxic and normal samples. XGBoost clearly beats all others in terms of all the three metrics.

	Predicted Hypoxic	Predicted Normal
Hypoxic	751	44
Normal	32	2881

Table 2: Confusion matrix of XGBoost prediction on data combined from TCGA and GTEx. It achieves an accuracy of 0.9795 with AUPRC of 0.9988 and AUROC of 0.9958

with just 3 false negatives. However, when we combine the two databases and re-train our logistic regression model, we achieve about 98% accuracy with 15 false negatives (refer to Table 6).

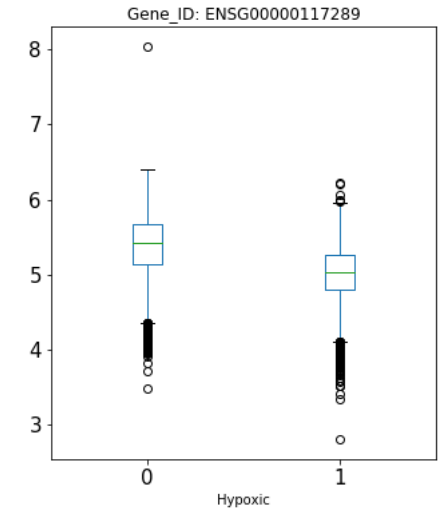
2.3 Hypoxia and EMT

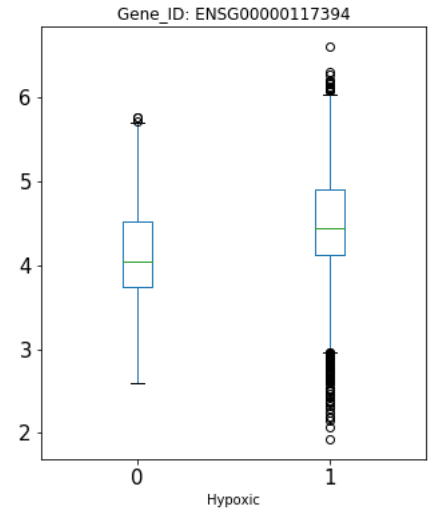
Now that we are able to classify hypoxic samples and tissue types one at a time, we aim to classify them together initially casting it as multi-class problem. Out of the 7753 samples from TCGA, the pie chart (Figure 3) tell us that most of them are non-hypoxic epithelial tissues while only one is hypoxic-mesenchymal. We implement the linear SVM algorithm to separate the data in four classes but it ends up putting everything in the majority class. We then moved to XGBoost which led to an accuracy of 97.85%. The confusion matrix (refer to Table 4) shows no prediction of mesenchymal non-hypoxic since we just had only one sample.

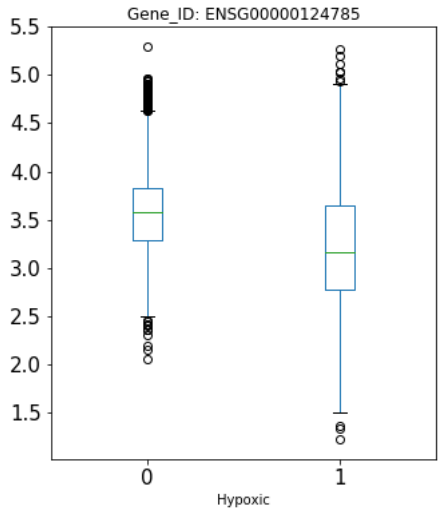
We also modeled it as a multi-output classification instead of multi-class, with multi-output classifier wrapped around the XGBoost model. Since, we have two-dimensional vector as the output, we calculate the F1 score (0.91) instead of accuracy. Also, we observe that false positive and false negatives numbers have reduced (comparing Table 8 and 9 with 4) in the confusion matrix indicating the interplay between hypoxia and EMT pathways.

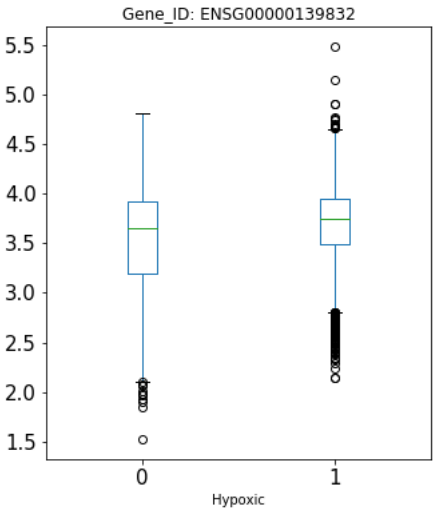
	Predicted Mesenchymal	Predicted Epithelial
Mesenchymal	79	7
Epithelial	17	2456

Table 3: Confusion matrix of prediction of tissue type on TCGA data. It achieves an accuracy of 0.9906 with AUPRC of 0.9993 and AUROC of 0.9888 using Logistic regression.









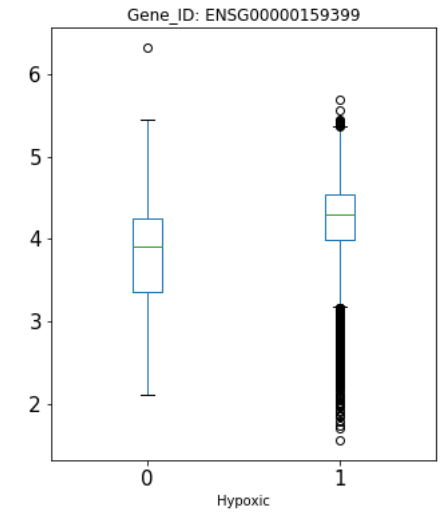


Figure 1: Box plots of log-scaled gene expression values of the five "most important" genes for hypoxic and non-hypoxic samples. The above plots indicate that the gene expression levels of hypoxic samples behave more erratically compared to the non-hypoxic samples. We observe higher variance and more outliers in hypoxic samples conforming with correlation of genes with cell status.

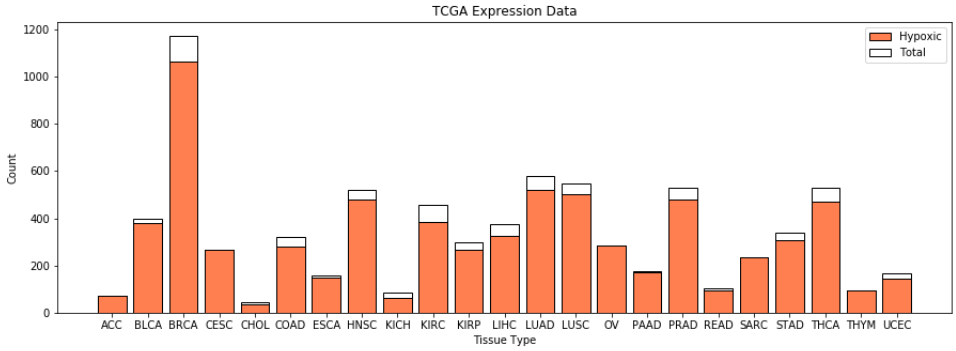


Figure 2: Distribution of hypoxic samples in TCGA data for each of the tissue type.

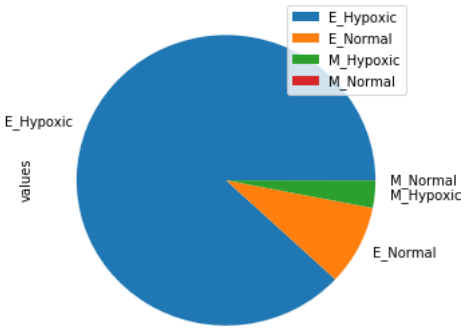


Figure 3: Distribution of tissue state and type for samples in TCGA data. E and M stands for Epithelial and Mesenchymal.

	Predicted E, False	**Predicted E, True**	**Predicted M, False**	**Predicted M, True**
E, False	2225	11	3	0
E, True	36	198	0	0
M, False	4	0	81	0
M, True	0	1	0	0

Table 4: Confusion matrix of XGBoost prediction of tissue type and normality on TCGA data. It achieves an accuracy of 0.9785

	Predicted Mesenchymal	**Predicted Epithelial**
Mesenchymal	314	3
Epithelial	1	179

Table 5: Confusion matrix of prediction of tissue type on GTEx data. It achieves an accuracy of 0.9920 with AUPRC of 0.9997 and AUROC of 0.9998 using logistic regression

	Predicted Mesenchymal	**Predicted Epithelial**
Mesenchymal	369	15
Epithelial	20	2651

Table 6: Confusion matrix of prediction of tissue type on GTEx and TCGA data combined using 249 genes known to be associated with EMT. It achieves an accuracy of 0.9885 with AUPRC of 0.9995 and AUROC of 0.9970 using logistic regression

	Predicted Hypoxic	**Predicted Normal**
Hypoxic	587	29
Normal	36	2308

Table 7: Confusion matrix of prediction of tissue state on GTEx and TCGA data combined using all 278 genes known to be associated with EMT and Hypoxia. It achieves an accuracy of 0.9780 with AUPRC of 0.9979 and AUROC of 0.9932 using logistic regression

	Predicted Mesenchymal	**Predicted Epithelial**
Mesenchymal	79	7
Epithelial	4	2469

Table 8: Confusion matrix of prediction of tissue type on TCGA data using 249 genes known to be associated with EMT.

	Predicted Hypoxic	**Predicted Normal**
Hypoxic	2309	15
Normal	30	205

Table 9: Confusion matrix of prediction of hypoxia on TCGA data using 249 genes known to be associated with EMT

References

Michael Ku Yu, Michael Kramer, Janusz Dutkowski, Rohith Srivas, Katherine Licon, Jason F Kreisberg, Cherie T Ng, Nevan Krogan, Roded Sharan, and Trey Ideker. Translation of genotype to phenotype by a hierarchy of cell subsystems. *Cell systems*, 2(2):77–88, 2016.

Chang Dong Yeo, Nahyeon Kang, Su Yeon Choi, Bit Na Kim, Chan Kwon Park, Jin Woo Kim, Young Kyoon Kim, and Seung Joon Kim. The role of hypoxia on the acquisition of epithelial-mesenchymal transition and cancer stemness: a possible link to epigenetic regulation. *The Korean journal of internal medicine*, 32(4):589, 2017.

Barbara Muz, Pilar de la Puente, Feda Azab, and Abdel Kareem Azab. The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy. *Hypoxia*, 3:83, 2015.

Joel P Joseph, MK Harishankar, Aruthra Arumugam Pillai, and Arikketh Devi. Hypoxia induced EMT: A review on the mechanism of tumor progression and metastasis in OSCC. *Oral oncology*, 80:23–32, 2018.