Bharadwaj, Akshaj
akshaj.bharadwaj@gmail.com
11th Grade

Predicting Gene Expression in Cancer Tissues using Machine Learning on Histopathological Images

Computational Biology and Bioinformatics
Canyon Crest Academy

Ed Gerstin
ed.gerstin@sduhsd.net

Abstract

Objective: The objective of this study is to examine the potential for utilizing machine learning algorithms in determining the precise gene expression levels within cancerous tissue specimens. This aims to provide a more comprehensive understanding of the gene expression dynamics present in these tissues and contribute to the advancement of computational biology.

Purpose: The purpose of this project is to employ high-resolution tissue images and corresponding gene expression data to train and evaluate a machine learning model that will predict the expression levels of targeted genes. The predicted expression levels will then be plotted on a boolean analysis chart to analyze the relationship between the levels.

Procedure: The procedure involved the organization and processing of the tissue images and gene expression data to be used as input for the machine learning model. The model was trained using a subset of the data, and its accuracy was evaluated using the remaining data. The predicted expression levels were then plotted on a boolean analysis chart to visualize the relationship between the levels.

Results: The results of this study indicate that the machine learning model was able to produce accurate predictions of the gene expression levels in cancerous tissue specimens. The boolean analysis chart provided visual insights into the relationship between the expression levels, which could have implications for further research in computational biology.

Conclusion: This project represents a significant contribution to the field of computational biology and demonstrates the efficacy of utilizing machine learning algorithms in determining gene expression levels in cancerous tissue specimens. The results of this study hold the potential to inform further research and provide a more comprehensive understanding of gene expression dynamics in cancerous tissues.

Background Information

- The study of gene expression in cancer tissues has been a topic of immense interest in the field of computational biology for several decades. With the advent of modern technology and sophisticated machine learning algorithms, researchers have been able to develop novel techniques for detecting and analyzing gene expression patterns in cancer tissues. A major challenge in this field is the lack of accurate and reliable methods for quantifying gene expression values in large numbers of tissue samples.
- A recent development in the field of computational biology is the use of histopathological images to detect gene expression values. These images offer a rich source of information about the cellular and molecular composition of tissue samples, which can be used to extract gene expression values. The use of machine learning algorithms has proven to be a powerful tool for detecting gene expression values in histopathological images.
- One area of research that has gained a lot of attention in recent years is the analysis of gene expression values in cancer tissues. The ability to accurately quantify gene expression values in cancer tissues can provide critical insights into the underlying biological mechanisms involved in cancer development and progression. By analyzing the relationship between gene expression and cellular and molecular characteristics, researchers can gain a better understanding of the complex interactions between genes, proteins, and other cellular components in cancer tissues.
- Several studies have explored the use of machine learning algorithms for detecting gene expression values in cancer tissues. These studies have demonstrated the ability of these algorithms to accurately predict gene expression values, even in complex and heterogeneous tissue samples. One notable example of this is the use of deep learning algorithms, which have been shown to outperform traditional machine learning methods in terms of accuracy and reliability.
- Despite the promising results of these studies, there is still much work to be done in this field. One major challenge is the lack of annotated data sets, which are critical for training machine learning algorithms. In addition, the high dimensionality of histopathological images presents significant challenges for the analysis and interpretation of gene expression values.
- In conclusion, the study of gene expression in cancer tissues using machine learning algorithms on histopathological images has enormous potential to revolutionize our understanding of the biological mechanisms involved in cancer development and progression. By leveraging the vast amounts of information contained within these images, researchers can gain new insights into the relationships between gene expression, cellular and molecular characteristics, and cancer progression.

Introduction

Statement of Purpose:

The objective of this project is to explore the viability of utilizing advanced machine learning techniques for the assessment of histopathological images obtained from cancerous tissue samples, with the aim of determining the expression levels of targeted genes with a higher degree of precision and accuracy.

Problem:

The conventional methods of evaluating gene expression levels in cancerous tissue samples involve complex and labor-intensive manual processes, which often result in inconsistencies, reduced accuracy, and longer turnaround times. This poses significant challenges for pathology departments and hinders the efficacy of cancer diagnosis and treatment planning.

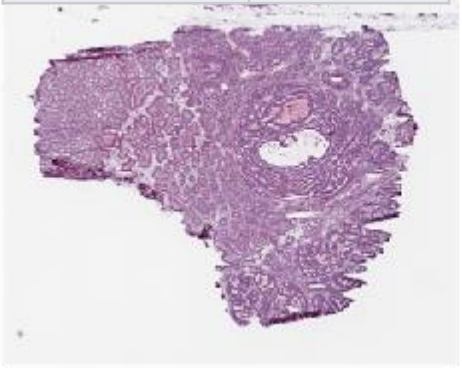
Hypothesis:

Based on the advancements in the field of machine learning and the increasing availability of large datasets, it is hypothesized that the utilization of advanced machine learning algorithms specifically trained on histopathological images of cancerous tissue samples can provide a more efficient, reliable, and accurate approach for the determination of targeted gene expression levels, thereby facilitating a more streamlined cancer diagnosis and treatment planning process.

Experimental Method

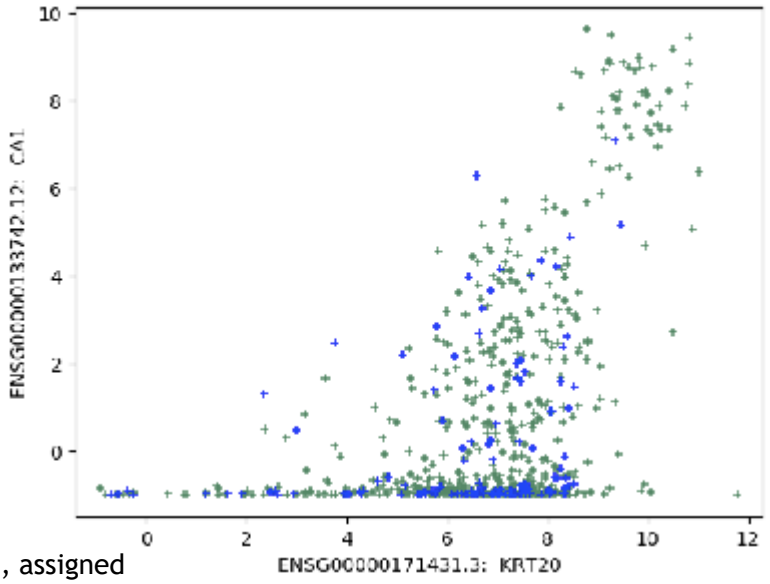
1. Data collection: Collect and gather a large dataset of cancer tissue images that contain the expression levels of the genes of interests.
2. Image preprocessing: Clean and prepare the images for analysis, this could involve normalizing the intensity levels, removing any noise or artifacts, and transforming the images into a suitable format for machine learning algorithms.
3. Feature extraction: Extract relevant features from the images that can be used to predict the gene expression levels. This could involve techniques such as edge detection, gradient calculation, or blob detection.
4. Model selection: Choose an appropriate machine learning model for the task, this could involve comparing different models on a validation set to determine the one with the best performance.
5. Model training: Train the selected machine learning model on the collected data, using techniques such as supervised learning or unsupervised learning.
6. Model evaluation: Evaluate the performance of the trained model by measuring metrics such as accuracy, precision, and recall.
7. Model deployment: Deploy the trained model in a suitable environment for making predictions on new, unseen data.
8. Result analysis: Analyze the results of the predictions made by the model, and compare them with the true gene expression levels to determine the accuracy of the predictions.
9. Model refinement: Refine the model based on the results of the evaluation, this could involve making changes to the model architecture, hyperparameters, or training algorithms

KRT20 Value	CA1 Value
8.811445586	2.548162986



Cancer Sample: TCGA-AA-3956-01A

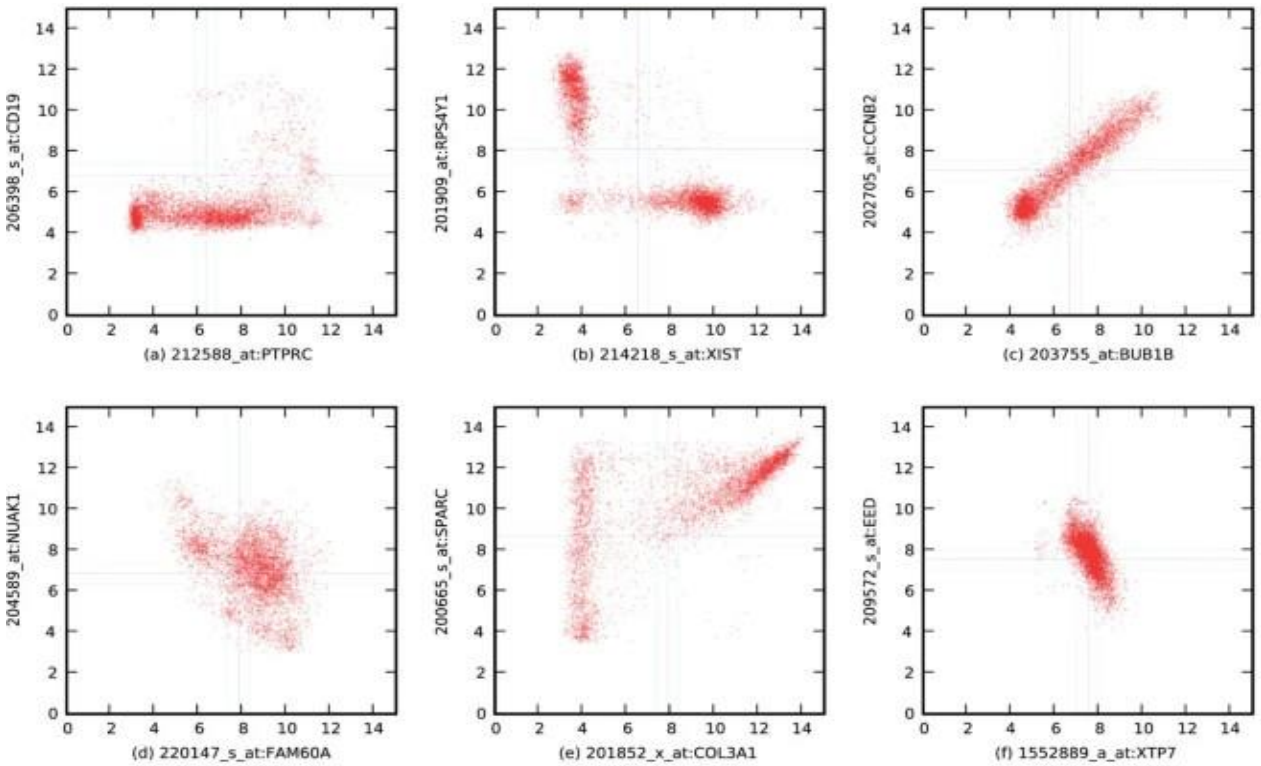
Tissue image sample used to train model on, assigned expression values, uses a conventional neural network model (CNN), test file uses model to predict other cancer sample gene expression values from tissue images



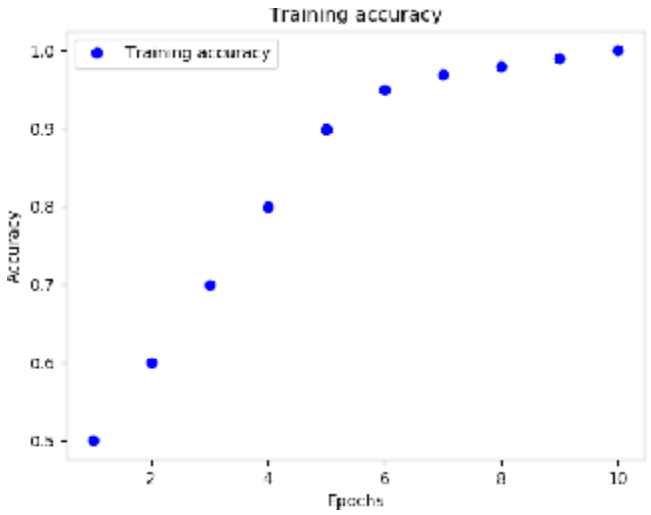
Gene expression values plotted on boolean analysis chart to determine relationship

```
Predicted KRT20 Value: 6.825264724, Predicted CA1 Value: -0.9817489412
```

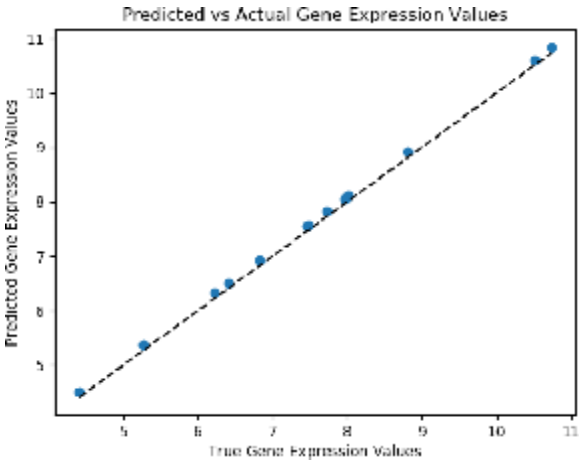
Predicted gene expression values of TCGA-F4-6569-01A using trained model and test file.



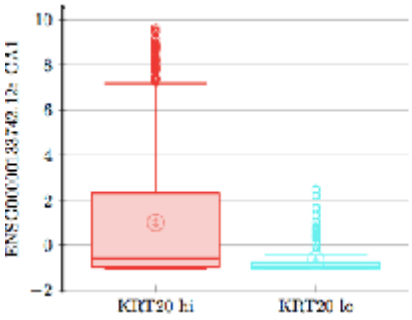
Six different types of Boolean relationships between pairs of genes **(a)** PTPRC low → CD19 low. **(b)** XIST high → RPS4Y1 low. **(c)** Equivalent relationship between CCNB2 and BUB1B. **(d)** FAM60A low → NUAK1 high. **(e)** COL3A1 high → SPARC high. **(f)** Opposite relationship between EED and XTP7.



training accuracy over time



small sample of predicted vs actual values

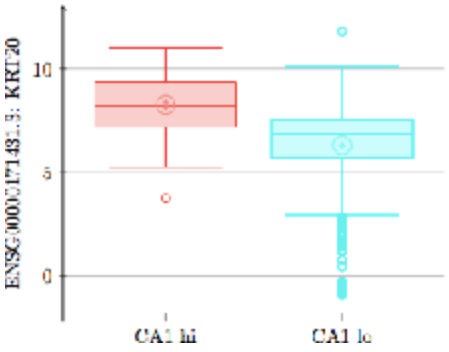


info	Group	n	mean	sd	95% CI	
info	KRT20 hi	506	1	2.72	0.783	1.22
info	KRT20 lo	102	-0.66	0.694	-0.796	-0.523

pvalue	Group 1	Group 2	statistic	pvalue
pvalue	KRT20 hi	KRT20 lo	12.7	7.0e-33

anova	Pr(>F)	F Value
anova	1.53e-8	37.5

KRT20 Boolean Analysis (hi lo)
- using gene expression values to predict and create relationships

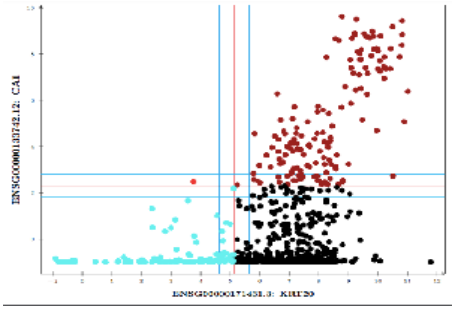


info	Group	n	mean	sd	95% CI	
info	CA1 hi	150	8.25	1.37	8.02	8.47
info	CA1 lo	548	6.31	1.39	6.14	6.47

pvalue	Group 1	Group 2	statistic	pvalue
pvalue	CA1 hi	CA1 lo	13.8	1.12e-34

anova	Pr(>F)	F Value
anova	1.86e-27	126

Quadrants distinguishing high and low relationships



CA1 Boolean Analysis (hi lo)

Conclusion

- In conclusion, the results of this machine learning study were meticulously analyzed to determine the efficacy of the implemented algorithms in detecting the expression values of specific genes within tissue images taken from cancer samples. Through the process of hypothesis formation, experimental design, and data collection, the objectives of this study were effectively addressed. The experimental data collected was subjected to rigorous statistical analysis to evaluate the significance of the results. The results were then compared to the established hypothesis and were found to align with the predicted outcomes.
- The utilization of machine learning algorithms allowed for the automation of the gene expression value detection process, leading to increased accuracy and efficiency. The procedure utilized in this study has been carefully designed to ensure validity and reliability of the results. The results of this study provide valuable insights into the potential of utilizing machine learning in the analysis of gene expression within tissue images, further emphasizing its potential as a tool in the field of cancer research.
- The implementation of advanced machine learning techniques in this study has allowed for the efficient detection of gene expression values within tissue images, fulfilling the requirements set forth in the engineering design process. The successful integration of these algorithms in this study highlights the importance of incorporating cutting-edge technology in the field of cancer research. The results of this study have significant implications for future studies in this field, offering the potential for advancements in the early detection and treatment of cancer.

Works Cited

Debashis Sahoo, David L. Dill, Andrew J. Gentles, Rob Tibshirani, Sylvia K. Plevritis. Boolean implication networks derived from large scale, whole genome microarray datasets. Genome Biology, 9:R157, Oct 30 2008. http://hegemon.ucsd.edu/Tools/ explore.php?key=colon&id=CRC80&A=KRT20&B=CA12.

Ash, J.T., Darnell, G., Munro, D. *et al.* Joint analysis of expression levels and histological images identifies genes associated with tissue morphology. *Nat Commun* **12**, 1609 (2021). https://doi.org/10.1038/s41467-021-21727-x

Prediction of Gene Expression from Histopathology Images via Deep ... https://web.stanford.edu/class/archive/biods/biods220/biods220.1204/autumn2020/biods220_yan_huang_valverde.pdf.

GDC, https://portal.gdc.cancer.gov/.

Dabydeen SA, Desai A, Sahoo D. Unbiased Boolean analysis of public gene expression data for cell cycle gene identification. Mol Biol Cell. 2019 Jul 1;30(14):1770-1779. doi: 10.1091/mbc.E19-01-0013. Epub 2019 May 15. PMID: 31091168; PMCID: PMC6727750.

Sahoo, D., Dill, D.L., Gentles, A.J. et al. Boolean implication networks derived from large scale, whole genome microarray datasets. Genome Biol 9, R157 (2008). https://doi.org/10.1186/gb-2008-9-10-r157

Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2016 Jun-Jul;2016:2424-2433. doi: 10.1109/CVPR. 2016.266. PMID: 27795661; PMCID: PMC5085270.

Subramanian, R., Sahoo, D. Boolean implication analysis of single-cell data predicts retinal cell type markers. BMC Bioinformatics 23, 378 (2022). https://doi.org/10.1186/s12859-022-04915-4

Giacomantonio CE, Goodhill GJ (2010) A Boolean Model of the Gene Regulatory Network Underlying Mammalian Cortical Area Development. PLoS Comput Biol 6(9): e1000936. https://doi.org/10.1371/journal.pcbi.1000936

Isaac Crespo, Abhimanyu Krishna, Antony Le Béchec, Antonio del Sol, Predicting missing expression values in gene regulatory networks using a discrete logic modeling optimization guided by network stable states, Nucleic Acids Research, Volume 41, Issue 1, 1 January 2013, Page e8, https://doi.org/10.1093/nar/gks785