Tissue Transcriptome - A Potential Feature for Postmortem Interval Predictions

Noah Herrick1, Logan Whitehouse1, Abdul Rehman Basharat3, Alexander Krohannon3, George E Sandusky2, Micheal Yard1, Susan Walsh1

1Department of Biology, Indiana University-Purdue University, Indianapolis

2Department of Pathology and Laboratory Medicine, Indiana University School of Medicine

3Department of BioHealth Informatics, Indiana University-Purdue University, Indianapolis

ABSTRACT

Accurate prediction of the postmortem interval (PMI) is one of the most important and complex tasks performed by forensic coroners. Transcriptomics offers a better alternative for PMI estimation as differential transcript expression within various tissues can provide precise indicators and ultimately make better predictions, accounting for the tissue types, forensic contexts, and other confounding factors. The Genotype-Tissue Expression (GTEx) Project provides RNA sequencing data. Here in this study, we have utilized a dataset from the GTEx project (dbGaP accession number: phs000424.v7.p2) which was generated using 9,777 tissue samples from 550 subjects for developing PMI prediction models. We are mining this data to further enhance key predictors associated with PMI using blood and salivary gland tissue types. Several machine-learning techniques have been trained and tested using the dataset to find the most useful and practical model to obtain the optimum prediction accuracy. Although the precision of data is slightly compromised, a categorical approach offers flexibility in prediction by widening the PMI period. This novel study will immensely aid forensic investigators in PMI prediction and can contribute to a broader understanding of postmortem biological processes. Furthermore, validation of the model with biological assays has been performed on other cohorts to evaluate prediction performance.

KEYWORDS

Postmortem interval, time since death, machine learning, prediction modeling, deep learning, differential gene expression, bioinformatics, computational biology, computational forensics

INTRODUCTION

Death triggers biological responses that can be investigated on a molecular level by tapping into the transcriptome to reveal site-specific changes undergone by a deceased body. The discovery of specific genes playing major roles at certain time points presents the opportunity to increase prediction accuracy. This can help forensic investigators accurately estimate time since death, referred to as the postmortem interval (PMI). Recent research suggests that gene expression across tissues is not conserved; therefore, some tissues are better reporters of PMI than others [1]. Although one hopes this insightful information from the tissue samples would be accessible at a crime scene to estimate PMI, it is not always the case. Therefore, finding the set of genes capable of predicting PMI within a tissue type readily available at crime scenes is critical. Considering this, we investigate blood and salivary gland tissue types [1, 2].

Investigators have attempted to determine PMI, the elapsed time between death and discovery of a body, with numerous biological, chemical, biochemical, and physical indicators for civil and criminal investigations [3]. Maintaining and developing ample methods to determine PMI is advantageous as death is entirely circumstantial and different approaches are necessary according to evidence type availability.

A few recent publications suggested the potential robustness in predicting PMI with machine learning techniques. In 2017, Colby Hunter et al. applied a dilution series approach, termed the “Gene Meter” approach, to predict PMI of zebrafish and mice using linear regression analyses in the examination of thousands of postmortem gene transcription profiles [4]. Belk et al. proposed predicting PMI analyzing microbiome data with random forest regression models in 2018 [5].

In this study, several machine-learning techniques have been implemented to find the best prediction model of PMI using gene transcripts per million (TPMs). Here, we have selected classification models to deal with the complexity of the data. Four classes were chosen as classifiers: Category I = premortem, Category II = 0-6 hours, Category III = 6-12 hours, and Category IV = 12-24 hours. Multinomial regression, decision tree classifiers, and k-nearest neighbor model performances have been evaluated for their accuracy at predicting PMI.

Because of recent developments in deep learning techniques and exhibition of its utility, neural network models have steadily made their impact on biological research as well [6]. Through these approaches, it is possible to gain greater insight into the complex interactions found in large biological datasets [7]. In one study, a deep learning approach has been utilized to estimate genes which are expressed postmortem and their expression levels over time using large, multi-sample omics datasets [8]. Moreover, deep neural network models provide a novel approach for predicting phenotype from genotype data and vice versa [9]. It is with this in mind that we developed a deep neural network model to predict PMI using large-scale gene expression data.

METHODS

1. **Neural Network Models**

Different types of artificial neural networks were created in order to attempt to predict PMI. A feed forward neural network model was created attempting to predict the scalar minute value of time since death. This model was created and optimized using the Keras in combination with the Talos package (<https://autonomio.github.io/docs_talos/#introduction>) a specialized toolset that allows for the testing of many hyperparameters and reporting of the best combination for a model. The model consisted of two hidden layers, one with 5000 neurons and the next with 3000, and the Talos results stated that the ReLu activation function worked best with both of those. The model was vastly underwhelming, only produced a top accuracy of 35% with the top performing hyperparameter combinations, and led us to pursue alternative methods.

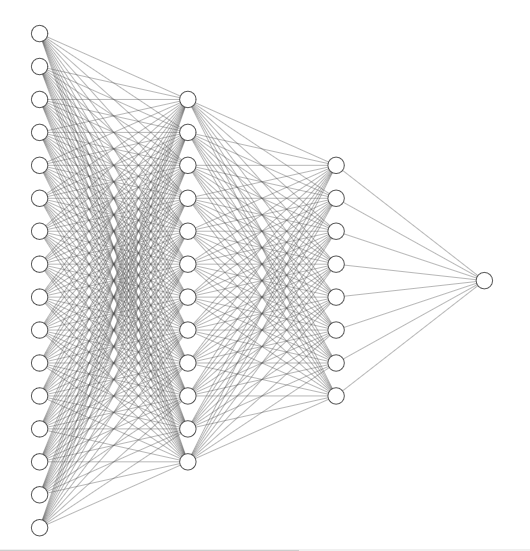


Figure 1. Scaled model structure of feed forward neural network.

1. **A multinomial regression model using penalized regularization**

Each of the 56,202 genes with TPM counts for 167 whole blood and 66 minor salivary gland samples were tested with partial correlation respective to age and sex covariates; Bonferroni correction threshold was applied to account for the false discovery rate of running thousands of partial correlations. There were 4767 genes in whole blood and 286 genes in minor salivary gland tissue types with p < 0.05.

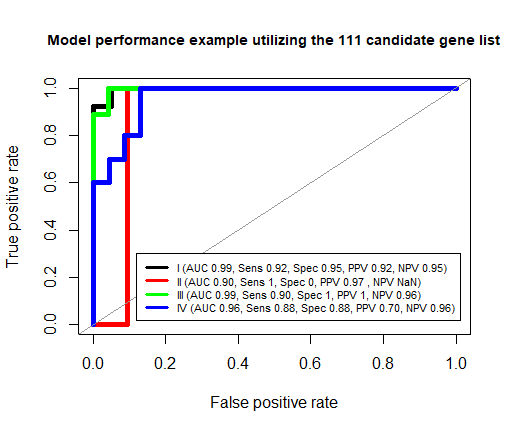
The optimal regression penalty narrowed down the list of genes required for expression profiling and PMI prediction. Alpha values zero (Ridge), 0.5 (Elastic Net), and one (Lasso) were cross-validated 100 times (80:20) with a multinomial logistic regression model (Supplementary Table 1) using an internal leave one out cross-validation (LOO-CV) approach for optimal λ parameter.

Each iteration consisted of a randomized 80:20 training and testing set of the total cohort (individuals n=167). Gene lists used to build each model were compiled and inspected for feasibility in real-world applications with regards to the number of genes needing expression profiling for prediction model input. The total gene count as shown in Supplementary Tables 1 and 2 accounts for overlap of genes per category.

The optimal alpha found was 0.5 (Elastic Net) due to higher mean AUC compared to Lasso, and a lower gene list count compared to Ridge. A final gene list for model building & testing was selected (n= 111 genes) over 100 iterations using 100% of individuals, internal LOO-CV, and λ of one standard error of the minimal MSE (Figure 2).

1. **Other Methods**

We are also exploring the utility of few other types of machine learning models including decision trees, k-nearest neighbors, and random forest. The decision tree may help us in removing the irrelevant features from the data, thus improving the model design for more complex deep learning model. We also want to explore the performance of the unsupervised k-nearest neighbor model and random forest. We are hoping that these models will provide us with significant insight into the data.



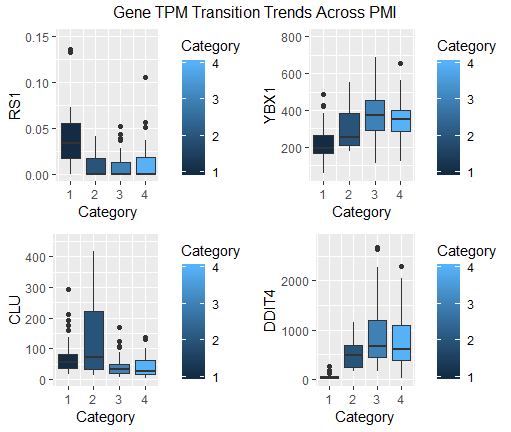
**B**

**A**

Model performance example utilizing the 111-candidate gene list

**Figure 2 A)** Categorical distribution of genes from the final 111 candidate gene list. This set was derived from 100% of the Whole Blood dataset and is intended for use on external data sets to further evaluate performance. **B)** ROC curve depicting an example of each categories’ performance metrics from model training using 80% and model testing using 20% of the total cohort (n=167) of the top 111 candidate gene list. **C)** Top predictor gene TPMs, based on coefficients used in the 100% built model, across the four categories offering insights to the biological changes occurring postmortem. \*Category I = premortem, Category II = 0-6 hours, Category III = 6-12 hours, Category IV = 12-24 hours

Gene TPM expression profiles across PMI



**C**

DATASET

RNA sequencing data was retrieved from the GTEx project release V7 (dbGaP Accession phs000424.v7.p2). The V7 release, in its entirety, includes 11,688 samples from 53 different tissue sites and 714 donors. Gene TPM values, derived from “Whole Blood” and “Minor Salivary Gland” tissue samples processed by Illumina TruSeq.v1, were extracted for model building with corresponding sample attributes e.g. age, sex, death classification, sample ischemic times, etc.

CONCLUSION

This paper aims to find the model that can most accurately predict PMI with gene TPMs. Initially, the model will be trained and tested using the GTEx project release V7 data; however, testing our models on external data is critical for out of sample error evaluations. Collection of external data is currently in the process with collaborators at the Indiana University School of Medicine Department of Pathology. If the model proves accurate on external data, this will be a novel forensic tool to the benefit of forensic investigators.

REFERENCES

1. Pokupcic, K., *Blood as an Important Tool in Criminal Investigation.* Journal of Forensic Science and Criminal Investigations, 2017.

2. Chatterjee, S., *Saliva as a Forensic Tool.* Journal of Dental Problems and Solutions, 2018.

3. Ferreira, P.G., et al., *The effects of death and post-mortem cold ischemia on human tissue transcriptomes.* Nature communications, 2018. **9**(1): p. 490.

4. Hunter, M.C., A.E. Pozhitkov, and P.A. Noble, *Accurate predictions of postmortem interval using linear regression analyses of gene meter expression data.* Forensic science international, 2017. **275**: p. 90-101.

5. Belk, A., et al., *Microbiome data accurately predicts the postmortem interval using random forest regression models.* Genes, 2018. **9**(2): p. 104.

6. Min, S., B. Lee, and S. Yoon, *Deep learning in bioinformatics.* Briefings in bioinformatics, 2017. **18**(5): p. 851-869.

7. Ching, T., et al., *Opportunities and obstacles for deep learning in biology and medicine.* Journal of The Royal Society Interface, 2018. **15**(141): p. 20170387.

8. Chen, Y., et al., *Gene expression inference with deep learning.* 2016. **32**(12): p. 1832-1839.

9. Md. Mohaiminul Islam, Y.W.a.P.H., *Deep Learning Models for Predicting Phenotypic Traits and Diseases from Omics Data*, in *Artificial Intelligence - Emerging Trends and Applications*. 2017, Intechopen.

SUPPLEMENTARY TABLES

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **100 Randomized Cross Validated Models**  (80% training: 20% testing) | | | | |
| **Alpha** | **Mean AUC (Blood)** | **PMI** | **λ = minimal MSE** | **λ = 1 SE of min. MSE** |
| **α = 0.0** | **λ = min. MSE** | **Category** | **Genes** | **Genes** |
| **Ridge** | 86.12 ± 5.22 |  | (Blood) | (Blood) |
| **λ = 1 SE of min. MSE** | **I** | 4767 | 4767 |
| 85.81 ± 5.09 | **II** | 4767 | 4767 |
|  | **III** | 4767 | 4767 |
|  | **IV** | 4767 | 4767 |
|  | **Total** | 4767 | 4767 |
| **α = 0.5** | **λ = min. MSE** | **I** | 400 | 303 |
| **Elastic Net** | 81.82 ± 4.87 | **II** | 124 | 43 |
| **λ = 1 SE of min. MSE** | **III** | 396 | 225 |
| 82.22 ± 4.45 | **IV** | 352 | 137 |
|  | **Total** | 997 | 637 |
| **α = 1.0** | **λ = min. MSE** | **I** | 208 | 117 |
| **Lasso** | 81.12 ± 5.04 | **II** | 57 | 22 |
| **λ = 1 SE of min. MSE** | **III** | 213 | 116 |
| 81.96 ± 4.55 | **IV** | 194 | 72 |
|  | **Total** | 575 | 311 |

**Supplementary Table 1:** Balancing Alpha regularization penalty with the categorical gene list

|  |  |  |  |
| --- | --- | --- | --- |
| **100 Randomized Cross Validations**  (80% training: 20% testing) | | | |
| **Alpha** | **PMI** | **λ = minimal MSE** | **λ = 1 SE of min. MSE** |
| **α = 0.5** | **Category** | **Mean AUC** | **Mean AUC** |
| **Elastic Net** | **I** | 0.993 ± 0.011 | 0.987 ± 0.017 |
| **II** | 0.892 ± 0.102 | 0.868 ± 0.116 |
| **III** | 0.946 ± 0.038 | 0.944 ± 0.032 |
| **IV** | 0.917 ± 0.046 | 0.909 ± 0.050 |

**Supplementary Table 2:** Optimal Model performance on 111 selected gene list