rfTSP: A Non-parametric Predictive Model with Order-based Feature Selection for Transcriptomic Data

by

Kelly Cahill

BS Mathematics, University of Pittsburgh, 2016

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Kelly Cahill

It was defended on

September 27, 2019

and approved by

George Tseng, ScD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Silvia Liu, PhD, Department of Pathology,School of Medicine, University of Pittsburgh

Jenna Carlson, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Thesis Advisor: George Tseng, ScD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

rfTSP: A Non-parametric Predictive Model with Order-based Feature Selection for Transcriptomic Data

Kelly Cahill, MS

University of Pittsburgh, 2019

Abstract

Genomic data has strong potential to predict biologic classifications using gene expression data. For example, tumor subtype can be determined using machine learning models and gene expression profiles. We propose the use of Top Scoring Pairs in combination with machine learning to improve inter-study prediction of genomic profiles. Inter-study prediction refers to two studies that are completely independent either in terms of platform or tissue. Top Scoring Pairs (TSPs) rank pairs of genes according to how well they are expressed between different groups of subjects. For example, gene A will be lowly expressed in cases, and gene B will be highly expressed in controls, while gene A will be highly expressed in controls, and gene B will be lowly expressed in cases. The pairs demonstrate an inverse relationship with respect to one and another. Using TSPs act not only as a feature selection step, but also allows for a non parametric method that transforms the continuous expression data to 0,1, which is based on the rank of the pairs. Due to the robust nature of the transformed data, our methods demonstrate that the use of TSP binary data is much more effective in prediction than continuous data, particularly in cross study prediction. Furthermore, we extend the use of TSPs to not only binary and multi-class label prediction, but also continuous classification. The objective of this paper is to demonstrate how using dichotomized data from TSPs as the feature space for machine learning methods, particularly random forest, returns stronger prediction accuracy across independent studies than traditional machine learning techniques with log2 and quantile normalization of data. This work has significant public health impact as accurate genomic prediction is crucial for early detection of many serious illnesses such as cancer.

Table of Contents

1.0 Introduction 1

2.0 Methods 4

2.1 TSP and KTSP for binary classification 4

2.2 rfTSP For binary classification 5

2.3 Multi-class prediction 7

2.4 Continuous prediction 8

3.0 Results 11

3.1 Binary prediction 11

3.2 Multi-class prediction 13

3.3 Continuous prediction 14

4.0 Discussion 17

5.0 Supplementary Tables 19

Bibliography 21

List of Tables

1 Glioblastoma tumor prediction 12

2 Confusion matrix for breast cancer tumor sub-type using rfTSP (true x predicted) 13

3 Sensitivity and specificity analysis for breast cancer tumor sub-type using rfTSP 14

4 Confusion matrix for breast cancer tumor sub-type using RF with log2 normalization 14

5 Sensitivity and specificity analysis for breast cancer tumor sub-type using RF with log2 normalization 15

6 Data descriptions 19

7 Training and testing assignments for IPF data 20

List of Figures

1 TSP Example, (A) binary, (B) multi-class, (C) continuous 3

2 Binary prediction on breast and lung tissue 12

3 Age prediction from brain tissue BA47. (A) rfTSP (B) RF with log2 normalization 16

1.0 Introduction

With the recent advancements of high-throughput technologies such as microarray and sequencing, omics data has become widely available to researchers interested in further understanding disease associations and gene expression. As more recent research supports the relationship between genes and the development of diseases, such as tumor genesis of cancers, the data holds potential answers to the most pertinent biologic questions in medical research. The availability of the data, and the relevance of the biologic significance that the data presents, has encouraged new statistical and bioinformatic methods. Particularly, supervised machine learning methods have been the most relevant tools for clinical use in disease status classication. Treatment and prevention for various cancers implement expression based-biomarkers for disease prediction and subtyping, drug response, and survival. For example, MammaPrint, Breast Cancer Index BCI and PAM50 are technologies frequently used in breast cancer patients. MSK-IMPACT is also used to screen for mutations in genes such as *EGFR*, *KRAS*, and *ALK* that are commonly seen in lung cancer tissues. Despite these advances, many challenges in disease classication still exist. In the generation of transcriptomics data, many aspects such as the genomic platform, tissue, and cohort vary greatly between studies. Because of this, reproducibility of disease classication studies using expression data has been difficult to achieve but is essential for clinical application. To ensure that these classication methods are solid enough for clinical use, inter-study prediction (a prediction model that is built in one study and tested in a completely independent test study) is necessary for validation. The characteristics of appropriate independent training and testing sets include expression profiles that are generated on different platforms (sequencing vs microarray), from different subjects, or from different (but related) brain tissues. For example, a training cohort may have been generated on an affymetrix platform, however, if a testing cohort is generated using Illumina sequencing technology, machine learning methods alone will generate a batch effect and return systematic biases, due to the noise generated between these platforms. Typical genomic models such as those used in Ogutu et al. (2004) and You-Jia et al. (2008) rely on single study prediction, which presents biased study specific

results and is not clinically applicable, or, commonly used machine learning methods with raw expression data that returns poor prediction accuracy. In addition to the bias presented in inter-study prediction, the noise generated requires a strong model that is robust enough to minimize the effects from varying factors (platform, cohort, and tissue type) of the studies. Literature has presented many applications of continuous expression data to machine learning methods, such as support vector machines and other linearly based models (Madan, Babu M. 2004), such as lasso and elastic net, as well as simple normalization methods (Youjia, Hua et al. 2008). While these methods work well for within study predictions and do provide some feature selection techniques to reduce data size applied to the models, they lack the robustness required for inter-study analysis.

Top scoring pairs (TSPs), as proposed by Geman et al. (2004), perform classifcation of case and control subjects using rank based gene pairs (see section methods 2.1 for details). Because of the ranking nature of TSPs, the model becomes non parametric. Figure 1 illustrates the definition of a strong top scoring gene pair. Figure 1A presents a binary TSP of breast cancer subtypes. The triangle symbol indicates one gene while the dot symbol indicates another gene. The left of figure A includes gene expression levels for subjects who have Her2 tumors and on the Basal on the right. The TSP pattern is evident as we see the level of expression changes in each gene across the subtypes. Figure 1B extends the binary TSP to a multi-class TSP across an additional breast cancer subtype. Lastly, Figure 1C is an example of a continuous TSP using expression data from brain region Broadman Area 47 and subject age. We use the rank based pattern to dichotomize the subjects according to their direction of gene expression for each TSP. The details are presented in the methods section. The use of binary data, proves to be very helpful in inter-study prediction, because this reduces noise due to the batch effect between studies and noise within the studies that can occur during poor quality sequencing. Previous papers have applied TSPs as the feature space to simple prediction models, such as majority vote (Bahman, Asfari et al. 2015, Pitts, Todd M et al. 2010). In our paper, we present the use of TSPs as a feature pre-screening method in tree based machine learning methods. The result is a non parametric model that is robust enough to compensate for the heterogeneity among studies. In this article, not only do we propose methods for binary prediction using TSPs in conjunction with tree methods,

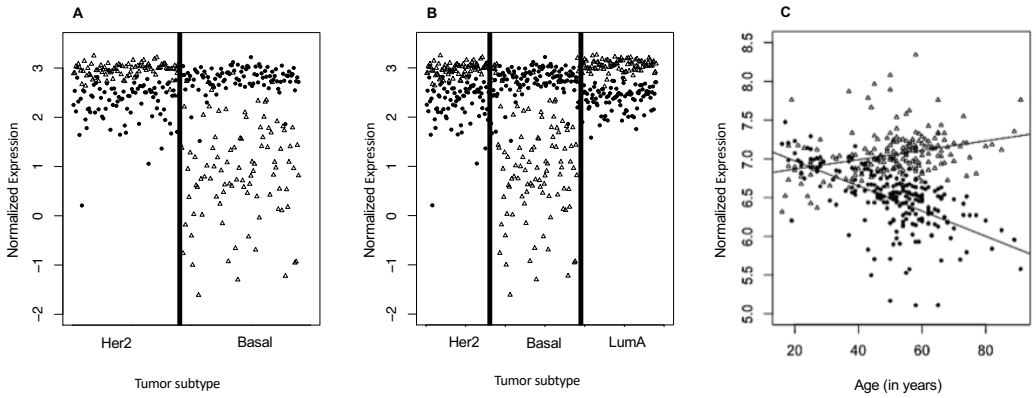

Figure 1: TSP Example, (A) binary, (B) multi-class, (C) continuous

but also an algorithm for selection TSPs in continuous and multi-class prediction. Transcriptomics data sets from lung, breast, and brain tissue are used to evaluate the performance of our methods.

2.0 Methods

2.1 TSP and KTSP for binary classification

Top scoring pairs were originally introduced by Geman et al. (2004), for the use of binary classification. Define X = $\{x_{gn}\}$ be a gene expression intensity matrix for all gene g and sample n, such that $1 \leq g \leq G$ and $1 \leq n \leq N$. Define y_n as the class label for sample n and $y_n \in \{0, 1\}$ for binary outcomes. A TSP is conceptually defined as a gene pair, i',j' $\in [1, G]$ such that when samples in the learning set are classified as $y = 0$, $x_{i'} > x_{j'}$ generally holds,and when samples in the learning set are classified as $y = 1$, the opposite order $x_{i'} < x_{j'}$ is more frequent. The reverse relationship also holds and can be extended to multi-class Top scoring pairs are selected by the probability, $T_{ij}(C)$, indicating the probability that expression levels of gene i are less than gene j, given a subject is classified to label C. Mathematically, $T_{ij}(C) = \Pr(x_{in} < x_{jn} | \text{y} = C)$. $T_{ij}(C)$ can be estimated empirically by the number of subjects that $x_{in} < x_{jn}$ over the total number of subjects with class label, C.

$$\hat{T_{ij}}(C) = \frac{\sum_{n=1}^{N} I(x_{in} < x_{jn}) I(y_n = C)}{\sum_{n=1}^{N} I(y_n = C)}$$

Each pair of genes is then given a score, S_{ij}, representing the difference of the probabilities between the two class labels: $S_{ij} = T_{ij}(1) - T_{ij}(0)$. This score can also be empirically estimated as $\hat{S}_{ij} = \hat{T}_{ij}(1) - \hat{T}_{ij}(0)$. Since the probability $T_{ij}(C)$ ranges from 0 to 1, the gene pair score by definition ranges from -1 to 1. TSP will be selected based on this score value. A strong TSP pattern will have score close to 1 or -1, meaning that the probability of seeing $x_i < x_j$ will be very high in one of the classes, C_1, but very low in another class C_2. In terms of cases and controls, for example, $\hat{S}_{ij} = -1$ implies $x_i > x_j$ in cases (1) while $x_i < x_j$ in controls (0), and the vice versa. The high absolute value of the score $\mid \hat{S}_{ij} \mid$ can be used as selection criteria for top pairs. For binary classification, Geman et al. (2004), and other

papers (Pitts, Todd M et al. 2010) used only one top scoring pair to classify subjects. Given a new testing sample $\vec{x} = \{x_i ... x_G\}$ and the TSP i', j', the class label is determined by:

$$\begin{cases} \hat{C}(\vec{x}) = & 1 \quad \text{if } x_{i'}^{(test)} - x_{j'}^{(test)} \leq 0 \\ & 0 \quad \text{if } x_{i'}^{(test)} - x_{j'}^{(test)} > 0 \end{cases}$$

The TSP classifier $\hat{C}_{i'j'}(\vec{x}^{(test)})$ is based on only one TSP, hence, the method is not very robust as it is sensitive to noises in the data. Bahman et al. (2015), proposed the kTSP algorithm to combine multiple TSPs for a more robust classification. Instead of choosing just one TSP, the top K TSPs, $\{(i'_1, j'_1) ... (j'_K, j'_K)\}$, are selected with highest absolute value score, $\mid \hat{S}_{ij} \mid$. Given, the new test sample, $\vec{x}$ is now classified by $\hat{C}(\vec{x}^{(test)}) = \underset{C}{\text{argmax}} \sum_{k=1}^{K} I(\hat{C}_K(\vec{x}^{(test)}) = C)$. The kTSP is an ensemble classifier that aggregates multiple weak classifiers by majority vote. To avoid the potential for ties, the top number K is typically selected to be odd in binary classification. Previous papers including Kim et al. (2016) have used cross validation and optimization to decide the optimal number, K, for number of TSPs. While the kTSP algorithm is beneficial from its non-parametric property, as discussed in the introduction section, simple majority vote does not create enough model complexity for efficient learning, where many samples will be classified right on the decision boundary. Furthermore, kTSP algorithm assigns equal weight to all top scoring pairs but ignores their inter-relationships/correlations.

2.2 rfTSP For binary classification

To overcome the above limitations, we develop a Robust tree-based model with order-based feature selection method (rfTSP). Specifically, we propose a two-stage approach which selects top scoring pairs as the feature space (stage 1) and dichotomizes these gene features as input for tree-based prediction models (stage 2), including random forest (RF), boosting and bagging with classification and regression tree (CART). The first step of our method is the selection and dichotomization of the top scoring pairs. We calculate the score, $\hat{S}_{ij}$, for each gene pair using the same method as introduced in section 2.1. The absolute value of

$\hat{S}_{ij}$ indicates the strength of rank perturbation of gene pairs between two class labels, where $|\hat{S}_{ij}|$ closing to 1 indicates high differential pattern and 0 means almost no association with outcomes. Top TSPs will be selected by this score. Then their expression continuous data will be transformed into binary format according to the ranks of gene expression of selected pairs for each sample. To be specific, for a top scoring pair with gene i and j for sample n, their expression value x_{in} and x_{jn} will be binarized to $B_{(i,j),n}$ as follows,

$$\begin{cases} 1 \text{ if } \hat{S}_{ij}(x_{im} - x_{jm}) \leq 0 \\ 0 \text{ if } \hat{S}_{ij}(x_{im} - x_{jm}) > 0 \end{cases}$$

$$\begin{cases} B_{(i,j),n} = & 1 \quad \text{if } x_{in} \leq x_{jn} \\ & 0 \quad \text{if } x_{in} > x_{jn} \end{cases}$$

By the above feature selection and dichotimization procedure, the continuous data matrix $\mathrm{X} = \{x_{gn}\}$ will be transformed into a binary matrix $B = \{b_{kn}\}$, where $1 \leq k \leq K$ and $1 \leq n \leq N$ indicates gene pair k of sample n. In the scenario of classification, top TSP will be selected from training data $X^{(train)}$ and binarized as $B^{(train)}$. These top features will be applied to new testing data $X^{(test)}$ and binarized as $B^{(test)}$.

To decide the number of TSPs for the training model, we propose a hypothesis test procedure which utilizes a binominal proportion test. Recall from the TSP selection algorithm, that $\mid S_{ij} \mid = \mid T_{ij}(1) - T_{ij}(0) \mid$, where $T_{ij} = \Pr(x_i < x_j | y = C) = P_{ij}(C)$. A score of S_{ij} close to 1 implies that $P_{ij}(C = 1) >> P_{ij}(C = 0)$, and a score, S_{ij}, close to -1 implies the reverse: $P_{ij}(C = 1) << P_{ij}(C = 0)$. For these scores, pair (i, j) would be categorized as strong top scoring pair. When $S_{ij} = 0$, this implies that $P_{ij}(C = 1) = P_{ij}(C = 0)$, meaning (i, j) is not likely to be a top scoring pair. Under the null assumption, for any given pair (ij), there is a 0.5 probability that $P_{ij}(C = 1) = P_{ij}(C = 0)$, meaning the possibility of any pair following a top scoring relationship with respect to case and control subjects is random. The alternative distribution is that $P_{ij}(C = 1) \neq P_{ij}(C = 0)$. If a pair of genes is truly top scoring, we would expect to see the alternative probability not equal to .5, meaning the top scoring relationship is not random. The two sample proportion test can generate a p value distribution that calculates the proportion of observed scores that are more extreme than

the null scores. If (i, j) is any pair; T_{ij} will follow a binomial distribution $T_{ij} \sim$ BIN(n, p = .5) where n is the number of subjects in the training set. The hypotheses are as follows:

$$\begin{cases} H_o : P_{ij}(C = 1) = P_{ij}(C = 0) \\ H_a : P_{ij}(C = 1) \neq P_{ij}(C = 0) \end{cases}$$

Different p value cut-offs give us different number of TSPs. While a p value cut off of .05 may seem to return many false positives, machine learning methods, such as random forest, only require the very top pairs as the feature space and have imbedded feature selection methods as well. Because of this, we do not need to worry about having too many false positives as the input to our feature space. Alternatively, cross validation can also be used.

For the second stage of our method, binary classifier will learn from the training binarized matrix $B^{(train)}$ and perform prediction on the new testing binarized matrix $B^{(test)}$. We selected some well-developed classifiers, including random forest (RF), boosting and bagging with classification and regression tree (CART). These tree-based methods, on the one hand, will provide robust prediction results between training and testing binary data. On the other hand, gene-pair features will be ranked and selected at each split node of the decision tree. These top gene-pairs and associated gene features are straightforward for biological interpretation. To select the top number of gene pairs for classifier input, we will apply 5-fold cross validation (CV) method to tune the parameters. That is, training samples will be equally divided into 5 exclusive partitions, and for each CV, 4-folds will serve as training samples and the remaining 1 fold will serve as testing. The overall performance will be evaluated by the Youden index (sensitivity + specificity -1) of the 5-fold prediction result. Top number K will be tuned from 5 to 50 and the best K value will be the one with the highest CV Youden index.

2.3 Multi-class prediction

We directly used the methodology outlined in the binary setting to address the problem of multiple classes. Instead of having two class labels, we now have m class labels. Let

$\mathbf{c} = (c_1, c_2, c_3, ...c_m, ...c_M)$ be a vector of class labels. Similar to the previous binary method, we can calculate a score $S_{ij} = T_{ij}(c_m) - T_{ij}(c'_m)$ using one-versus-others strategy. We can view $T_{ij}(c_m)$ as we would in the binary setting: the probability that gene expression, x_i is less than gene expression x_j given subject class label is c_m. $T_{ij}(c'_m)$ is the probability that gene expression, x_i is less than gene expression x_j given subject class label is not equal to c_m. Mathematically, we can write this as: $T_{ij}(c'_m) = Pr(x_i < x_j | y \neq c_m) = Pr(x_i < x_j | y = \bigcup_{1 \leq i \leq M; I \neq m} c_i)$. $\hat{T}_{ij}(c_m)$ and $\hat{T}_{ij}(c'_m)$ are calculated empirically as:

$$\hat{T_{ij}}(c_m) = \frac{\sum_{n=1}^{N} I(x_{in} < x_{jn}) I(y_n = c_m)}{\sum_{n=1}^{N} I(y_n = c_m)}$$

$$\hat{T_{ij}}(c'_m) = \frac{\sum_{n=1}^{N} I(x_{in} < x_{jn}) I(y_n \neq c_m)}{\sum_{n=1}^{N} I(y_n \neq c_m)}$$

so that we have an empirical score, $\mid \hat{S_{ijm}} \mid = \mid \hat{T_{ij}}(c_m) - \hat{T_{ij}}(c'_m) \mid$. We repeat this process for each class label, $1 \leq m \leq M$. Due to biological variation, some class labels are better separated from others but some are less. Because of this, rather than using the same TSP score threshold for each class label, we choose equal number of TSPs in each class label. We then use cross validation techniques to choose an optimal number of pairs from each group. This way, each class label has a fair chance of being predicted, and we don't risk adding bias to the model by only having gene pairs that have strong prediction power for a subset of well-separated class labels.

2.4 Continuous prediction

In this section, we extend the use of TSPs to the continuous data regression setting. The definition of a TSP in the continuous case is similar to the binary and multi-class case, but pairs of genes are ranked with respect to a monotone continuous outcome. Let y_i $(i = 1...N)$ be the age for subject i, and $Y = \{y_1....y_N\}$ is the vector of length N for age variable. We first select the genes that undergo age-related changes, by fitting robust linear regression to each of genes. Only top outcome related genes are used to construct TSPs for our model.

A pair of genes (x_i, x_j) is a TSP if the age trajectories of two genes intersect at a certain age (we call this as "crossing point", y^*, where $y_1 < y^* < y_N$). For $y < y^*$, $x_i > x_j$ and for $y > y^*$, $x_i < x_j$, and vice versa. For the first stage of our algorithm, we propose a pipeline to select TSPs, which incorporate the usage of crossing monotone continuous outcomes. The idea is to quantify if the ranks of a pair of genes differ before and after the crossing point. The detailed steps are as follows:

(i) Let $z_n = x_{in} - x_{jn}$ be the difference in expression levels for gene pair (i, j) and subject n, then we have $z = \{z_1 ... z_N\}$. To find the crossing age, we fit a regression to z with respect to y as $\hat{z} = \hat{a} + \hat{b}y$, where we can derive the crossing point $y^* = \frac{-a}{b}$. (ii) The crossing point and the regression line at $\hat{z} = 0$ define four mutually exclusive sets of subjects:

$\mid A \mid = \{n : y_n \leq y^* \cap z_n \geq 0\}$. $\mid A \mid$ is the number of subjects with outcome (y_n) less than the crossing point and z_n greater than 0.

$\mid B \mid = \{n : y_n \geq y^* \cap z_n \geq 0\}$. $\mid B \mid$ is the number of subjects with age (y_n) greater than the crossing point and z_n greater than 0.

$\mid C \mid = \{n : y_n \leq y^* \cap z_n \leq 0\}$. $\mid C \mid$ is the number of subjects with age (y_n) less than the crossing age and z_n less than 0.

$\mid D \mid = \{n : y_n \geq y^* \cap z_n \leq 0\}$. $\mid D \mid$ is the number of subjects with outcome (y_n) greater than the crossing point and z_n less than 0.

(iii) Under the null assumption, where the number of subjects before and after the crossing point will follow binomial distributions with m_1 as the number of subjects in regions $\mid A \mid + \mid C \mid$ and m_2 as the number of subjects in regions $\mid B \mid + \mid D \mid$, respectively. $m_1 \sim \text{BIN}(\mid A \mid + \mid C \mid, p_1)$ and $m_2 \sim \text{BIN}(\mid B \mid + \mid D \mid, p_2)$, where p_1 and p_2 are the proportions of subjects that fall into regions $\mid A \mid$ and $\mid D \mid$, respectively. We can estimate these values as sensitivity and specificity by $\hat{p_1} = \frac{A}{A+C}$ and $\hat{p_2} = \frac{D}{D+B}$. (iv) We define a Youden index $J = p_1 + p_2 - 1$ to quantify the strength of a gene pair, which can be estimated as $\hat{J} = \hat{p}_1 + \hat{p}_2 - 1$. Indices close to 1 or -1 indicate pairs with strong top scoring patterns. The corresponding 95% confidence interval for $\hat{J}$ is $\hat{J} \pm 1.96 * \sqrt{var(\hat{J})}$ and can be used to determine suitable cut offs for the Youden index. Once the TSPs are selected, the original expression data is transformed from continuous to binary, in a similar fashion as the binary case. For each subject n, given a TSP (i', j'), if $x_{in} > x_{jn}$ then a binary coding 1 will be assigned, otherwise 0 will be given

exactly as described in section 2.1. This results in an k x n matrix of indicators 0 and 1, where k is the number of selected top scoring pairs. For the second stage of our method, similarly to the binary case in section 2.1, we treat the transformed binary data as an input to tree-based machine learning methods.

While the number of top scoring pairs used in the model can be arbitrarily selected as pairs that have a Youden index magnitude close to one, we used cross validation to select an optimal Youden index cut off. We repeated 10-fold cross validation for a lower bound confidence interval of .6,.7,.8, and .9, for positive Youden indices, and an upper confidence limit of -.6,-.7,-.8, and -.9 for negative Youden indices. We selected the cut offs that returned the lowest average mean square error across all cross validated iterations.

3.0 Results

3.1 Binary prediction

We tested our methods on lung, breast, and brain tissue. For binary prediction, we used data on lung disease from the six different sources, breast cancer data from Wang et al. (2005) and glioblastoma brain cancer from The Cancer Genome Atlast (TCGA). Table 6 in the appendix shows subject and platform information for all 6 lung disease studies as well as the breast cancer study. The outcomes of interest were case versus control in lung disease, estrogen receptor positive versus negative in breast cancer subjects, and malignant glioblastoma versus benign tumor in brain tissue. Within each disease, we have multiple independent studies from different subjects and platforms. Figure 2 shows the results of our application. In Figure 2A, we use 6 different independent lung disease studies across different cohorts and platforms to train and classify disease status using random forest with log2 normalization, rfTSP, and kTSP. The 6 lung studies were assigned to testing or training, such that only the 3 largest sets are used as training. Table 7 in the appendix lists all of the training and testing study pairs. prediction accuracy is calculated as the number of correctly classified subjects over the number of total subjects. Overall, we can see that the non-parametric rfTSP methods outperform random forest with normalized data. In 2B, we train our model using RNAseq data from breast and lung tissue and test our model using microarray data from lung and breast. Not only do we apply our method to random forest, but we also evaluated other commonly used machine learning methods as well, such as bagging, boosting, and CART. We also directly compare our method to the kTSP method used in Bahman et al. (2015). For both diseases, the methods using dichotomized data from top scoring pairs outperform the classical machine learning methods. Due to the inherent noise between the studies, the robustness of dichotomized data proves to be imperative for inter-study prediction. Because machine learning methods have imbedded feature selection methods, people commonly pour many features into the methods, requiring more computing power and time. Because our method is two stage and selecting TSPs acts as a feature

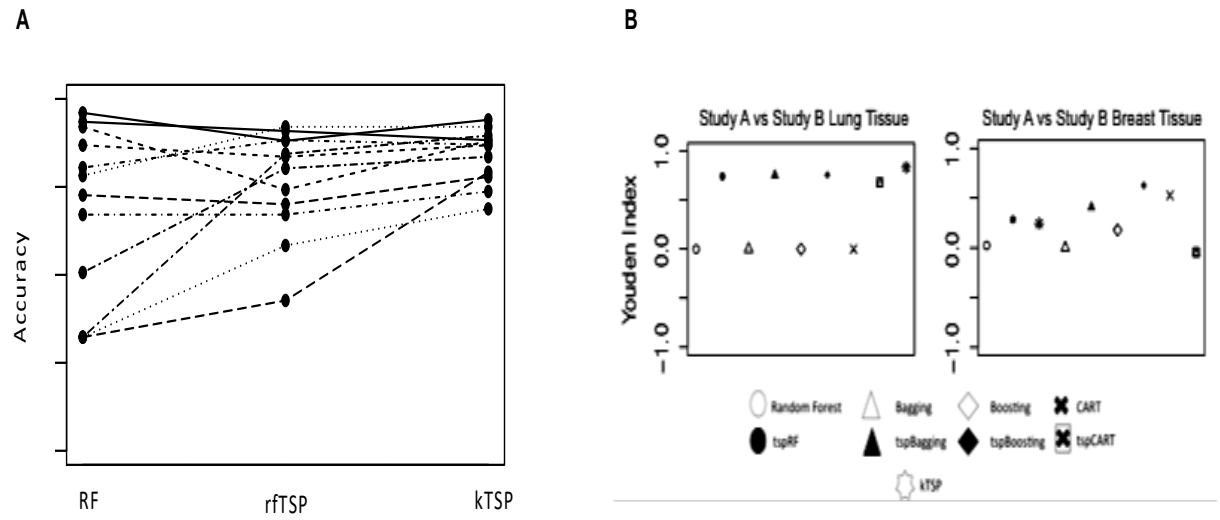

Figure 2: Binary prediction on breast and lung tissue

selection step, we are able to reduce the required computing time.

Table 1 below also explores how the algorithms perform under different normalization methods. We apply data from brain tissue to predict glioblastoma status using TCGA data. We build the model in an RNAseq set and test the model in microarray, using random forest alone with quantile normalization as well as rfTSP and kTSP. rfTSP has the highest accuracy, while random forest with quantile normalization fails completely. It's also interesting to note that rfTSP has the highest sensitivity and specificity at .8 and .84 respectively. While kTSP's performance is comparable, it's specificity drops to .65.

Table 1: Glioblastoma tumor prediction

	RF+QN	rfTSP	kTSP
Accuracy	.4301	.8172	.7204
CI	(.33,.54)	(.72,.89)	(.62,.81)

3.2 Multi-class prediction

We next tested our multi-class prediction method using breast cancer data from TCGA and MetaBric. We have three PAM50 tumor subtype classifications present in the data, HER2, Basal-like, and Luminal A. We build our model using RNAseq data from TCGA and test our method in microarray data from MetaBric. As mentioned in the methods section 2.2, we used 10 fold cross validation within the training set to select the optimal number of top scoring pairs. Based on having optimal accuracy, 40 TSPs from each class label group were selected, so a total of 120 pairs were used as input features to the model. Again, this greatly reduces computational time that would be required to run random forest if without an effective feature selection step. Our method, rfTSP was able to generate 78 percent accuracy in METABRIC data, while RF alone produced only 40 percent accuracy. Table 2 and Table 3 show the confusion matrix as well as sensitivity and specificity breakdown of the rfTSP prediction result, while Tables 4 and 5 show the breakdown of the random forest used only with log2 normalization. Our method is able to classify each luminal A subject correctly and 70 percent of basal subjects correctly. Furthermore, the specificity across all three classes remains high. Random forest alone classifies each subject as HER2. Because HER2 appeared genetically more similar to both basal-like and luminal A cancers in the MetaBric and TCGA, when there is noise present between the studies, random forest selects the class label that is seemingly an average of the other two, allowing for a "safer" classification. The dichotomized and rank based data in rfTSP, however, allows the three individual class labels to remain unique even in the presence of noice during inter-study prediction.

Table 2: Confusion matrix for breast cancer tumor sub-type using rfTSP (true x predicted)

	Basal	Her2	LumA
Basal	33	11	6
Her2	1	34	15
LumA	0	0	50

Table 3: Sensitivity and specificity analysis for breast cancer tumor sub-type using rfTSP

	Basal	Her2	LumA
Sensitivity	.7	.5	1
Specificity	.84	1	.79

Table 4: Confusion matrix for breast cancer tumor sub-type using RF with log2 normalization

	Basal	Her2	LumA
Basal	20	20	10
Her2	15	20	15
LumA	19	12	19

3.3 Continuous prediction

Lastly, we test our continuous method to predict subject's molecular age using brain tissues from Broadmann areas 47 and 11 from the University of Pittsburgh Medical Center. While the two brain areas are similar and are both generated using RNA-seq technology, there is always inherent noise associated with cross tissue prediction. In Figure 4, we use our method, rfTSP, and RF alone to build the model in BA47 and test in BA11. We select the number of top scoring pairs based on 10-fold cross validation within the BA47 data at various confidence interval cut offs. We find that a confidence interval cut off of .7 and -.7 is optimal and it estimates about 150 pairs. Likely, due to the more complicated nature of continuous prediction, more features are needed for random forest to succeed. Our method returns a mean squared error of 6.54, while random forest alone has a mean square error of 7.93. As seen in the multi-class example, random forest has a tendency to predict towards the middle of a class label spectrum. Because of this, random forest is often criticized for its poor prediction along the tail. In our scenario, we see that random forest does struggle

Table 5: Sensitivity and specificity analysis for breast cancer tumor sub-type using RF with log2 normalization

	Basal	Her2	LumA
Sensitivity	.37	.39	.43
Specificity	.69	.70	.71

to accurately predict those over age 60. While rfTSP still shows this pattern, looking at figure 3, we can see that the subjects over age 60 our model predicted a higher age than subjects over age 60 using RF alone. rfTSP in the continuou setting is also robust enough to maintain the information provided from the tail end subjects in our training set.

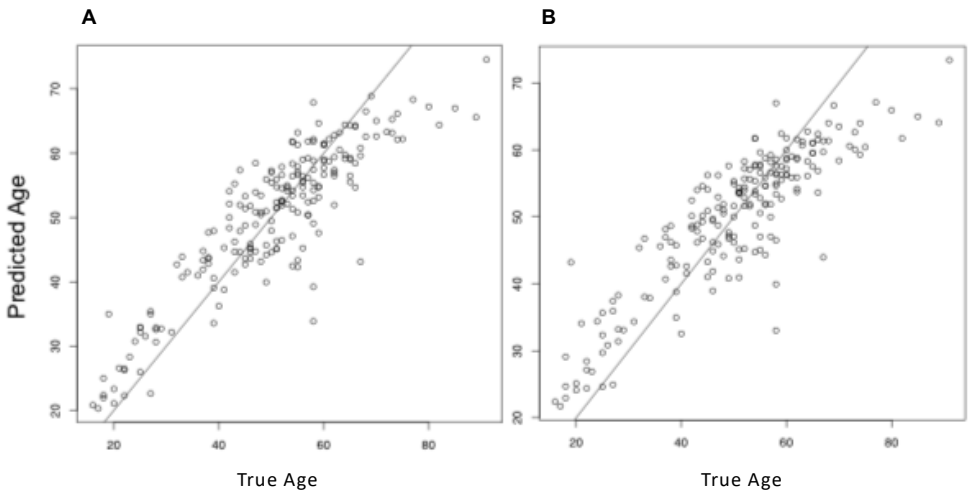

Figure 3: Age prediction from brain tissue BA47. (A) rfTSP (B) RF with log2 normalization

4.0 Discussion

Supervised machine learning is a critical analytic tool for many applications in biologic research, including cancer, disease, and aging. As high-throughput genomic data becomes more publicly availabel, effective statistical and machine learning methods are required to reach the full potential that the data can offer. accurate inter-study prediction is necessary as single study classification often faces small sample size issues and study-specific bias. Furthermore, commonly used machine learning techniques often have poor prediction results in cross study application due to the between study noise. These typical genomic prediction models experience low reproducibility across independent studies because of a lack of robustness and expression measurement noises that is inevitable in varying tissue types and platform technology. Regardless, researchers and clinicians still wish to classify subjects across different cohorts, platforms, or tissue types; it remains a difficult problem in genomics that requires an accurage and robust model that will not falter under cross-study heterogeneity. To overcome these limitations, we propose the use of top scoring pairs as the nonparametric feature selection and data transformation engine for machine learning methods, particularly random forest. TSPs not only act as a feature selection step, but they also transform expression data from continuous measurement in gene features to binary status in gene pairs, which minimizes noise from outliers within and between studies. Because rfTSP is a rank based non-parametric algorithm, the method is particularly successful in cross platform prediction where we commonly see a baseline shift in expression between two studies, noisy studies due to poor sequencing quality, and biological variation present in cross-tissue prediction. We have implemented the use of top scoring pairs in not only binary classification but also multi-class and continuous classification. Our data applications demonstrate that top scoring pairs have tremendous potential in robust machine learning and can outperform typical genomic classification models. The use of top scoring pairs adds another feature selection stop, potentially reducing computational time while still providing crucial information to the training model. Despite the difficulties present in data, our examples prove that rfTSP can create a robust and accurate prediction. Our model can provide

patients and researchers accurate classification without restricting them to a single study prediction.

There are a few limitations and future directions to address. Currently, TSPs are selected using the observed patterns present in training set only. This creates some challenge as a gene pair that would be biologically relevant to the disease of interest and show a strong pattern in the testing data may not in the training set due to poor data quality or outliers. One way to overcome this would be to incorporate certain pathway analysis into the TSP selection process and include gene pairs in the model that may not show a strong TSP pattern in the training data but give us biological reason to believe they could happen in the testing set. Furthermore, there are some challenges in the random forest algorithm, particularly in predicting subjects whose class label is in the tail end. As mentioned in the multi-class and continuous applications, rfTSP does improve tail end prediction, but there is still room to improve the actual random forest algorithm itself. Lastly, our methods prove very well in application, but we lack theoretical proofs that rigorously demonstrate under which parameters our method will succeed. We plan to address these topics in future work as well as extend our continuous classification model to other outcomes than age, such as survival time. We will provide a publicly available R package "rfTSP" on github.

5.0 Supplementary Tables

Table 6: Data descriptions

Name	Study	Samples	Control	Case	Reference
Tedrow A	IPF	63	11	52	GSE47460
Tedrow B	IPF	96	21	75	GSE47460
Emblom	IPF	58	20	38	GSE17978
Konishi	IPF	38	15	23	GSE10667
Pardo	IPF	24	11	13	GSE2052
Larsson	IPF	12	6	6	GSE11196
Wang	BC	286	209	77	GSE2034

Table 7: Training and testing assignments for IPF data

	train	test
1	A	B
2	A	Emblom
3	A	Pardo
4	A	Larsson
5	A	Konishi
6	B	Emblom
7	B	Konishi
8	B	Pardo
9	B	Larsson
10	Emblom	Konishi
11	Emblom	Pardo
12	Emblom	Larsson

Bibliography

Afsari, Bahman, Elana J. Fertig, Donald Geman, and Luigi Marchionni. 2015. "SwitchBox: An R Package for k-Top Scoring Pairs Classifier Development." Bioinformatics. https://doi.org/10.1093/bioinformatics/btu622.

Breiman, Leo. 2001. "Random Forests." Machine Learning. https://doi.org/10.1023/A:1010933404324.

Dietterich, Thomas G. 2000. "Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization." Machine Learning. https://doi.org/10.1023/A:1007607513941.

Geman, Donald, Christian D'Avignon, Daniel Q. Naiman, and Raimond L. Winslow. 2004. "Classifying Gene Expression Profiles from Pairwise MRNA Comparisons." Statistical Applications in Genetics and Molecular Biology. https://doi.org/10.2202/1544-6115.1071.

Kim, Sung Hwan, Chien Wei Lin, and George C. Tseng. 2016. "MetaKTSP: A Meta-Analytic Top Scoring Pair Method for Robust Cross-Study Validation of Omics Prediction Analysis." Bioinformatics. https://doi.org/10.1093/bioinformatics/btw115.

Ogutu, Joseph O., Torben Schulz-Streeck, and Hans Peter Piepho. 2012. "Genomic Selection Using Regularized Linear Regression Models: Ridge Regression, Lasso, Elastic Net and Their Extensions." BMC Proceedings. https://doi.org/10.1186/1753-6561-6-S2-S10.

Pitts, Todd M., Aik Choon Tan, Gillian N. Kulikowski, John J. Tentler, Amy M. Brown, Sara A. Flanigan, Stephen Leong, et al. 2010. "Development of an Integrated Genomic Classifier for a Novel Agent in Colorectal Cancer: Approach to Individualized Therapy in Early Development." Clinical Cancer Research. https://doi.org/10.1158/1078-0432.CCR-09-3191.

Wang, Xingbin, Yan Lin, Chi Song, Etienne Sibille, and George C. Tseng. 2012. "Detecting Disease-Associated Genes with Confounding Variable Adjustment and the Impact on Genomic Meta-Analysis: With Application to Major Depressive Disorder." BMC Bioinformatics. https://doi.org/10.1186/1471-2105-13-52.