Comparing similarity measures over RNA-seq derived gene expression profiles

Kevin Liao[1]

[1]Biostatistics, UNC-Chapel Hill

Abstract

Over the past decade, RNA-seq data has emerged as a state-of-the-art assay for analyzing the transcriptome. Through sequencing short strands of RNA and mapping them to a reference genome or set of transcripts, the amount of gene activity, also known as gene expression, in a cell or tissue can be measured and used for predicting phenotypes. This process of predicting phenotypes from gene expression profiles typically utilizes machine learning algorithms to classify an unknown sample based on training data. Certain classification algorithms, such as K-Nearest Neighbors, rely on computing a similarity measure to make these predictions. Therefore, different similarity measures affect the performance of these classifiers. In this project, using data from the Sequence Read Archive (SRA), different similarity measures were implemented and compared to determine the optimal algorithm for predicting phenotypes using RNA-seq derived gene expression profiles. We labeled samples using 12 different tissue types as our phenotype, and a measure based on Pearson Correlation achieved the highest prediction accuracy of 84%.

1 Introduction

RNA-seq is a high-throughput DNA sequencing method that is used for mapping and quantifying the transcriptome. Following the central dogma of biology, genetic information encoded in DNA is transcribed into RNA, in which the total set of transcripts constitutes the transcriptome. Therefore, analyzing the transcriptome can reveal critical insight into how the genes within the genome function, which ultimately influence a number of factors such as a cells molecular components, responses to different conditions, progression of disease, and much more [1]. RNA-seq allows us to better understand gene functions by accurately measuring gene expression levels. It is important to note that gene expression can be calculated not only at the gene level, but also the isoform level due to how pre-mRNA is spliced to produce various isoforms. To measure gene expression, the general process is described in Li et al.[2]:

1. Isolate RNA from a sample
2. Convert RNA to cDNA fragments through reverse-transcription and fragmentation
3. Generate millions of reads from cDNA fragments using a high-throughput sequencer
4. Map reads to a reference genome or transcript set
5. Estimate expression levels using counts of reads mapped to each gene

In eukaryotic organisms such as humans, almost every cell in the body contains the same genome and thus the same set of genes. However, it is the varying gene activity between any two cells that results in a range of physical, biochemical, and functional differences that, for example, distinguish a brain cell from a skin cell [3]. This variation in gene activity, in the form of gene expression profiles, can then be used to predict a phenotype of the cell.

One common approach to making phenotype predictions involves the use of machine learning in order to perform classification. In this work, we used machine learning to predict an unknown sample's tissue type among 12 possible choices: brain, lung, liver, skeletal+muscle, colon, skin, heart, kidney, prostate, testis, ovary, and pancreas. This was done using a supervised algorithm called the K-Nearest Neighbors algorithm. This algorithm does so by computing a similarity measure between the unknown sample and each sample in the training dataset. The most similar sample and its tissue type are then returned, and the unknown sample is classified using the same label as the returned sample.

Due to the reliance of the algorithm on the similarity measure, different similarity measures will affect the performance of the classifier. This project seeks to address this issue by identifying the similarity measure that optimizes phenotype prediction using RNA-seq derived gene expression. Five measures of similarity were chosen and compared in this project: 1) Euclidean distance 2) Pearson correlation 3) Spearman Rank correlation 4) Median Absolute Deviation and 5) Poisson Model-based similarity.

2 Methods

2.1 Dataset Creation

Our dataset consisted of 1488 samples compiled across a number of studies from the Sequence Read Archive (SRA). Each sample in the dataset is associated with a gene expression profile in the form of a vector and its tissue label. In our gene expression vectors, the read counts aligned to each transcript were converted into relative expression levels using a measure called transcripts per million (TPM). This metric normalizes for sequencing depth and gene length, and is advantageous for comparing samples directly because the sum of all TPMs in each sample are the same.

2.2 K-Nearest Neighbors Algorithm and Cross Validation

This project predicted tissue types of samples using the K-Nearest Neighbor algorithm. This algorithm works by computing a similarity measure between a query sample that we wish to classify and a training dataset of samples with known tissue types. In our case, a 1-nearest neighbor approach was used in which the query sample was simply classified as the tissue type of the most similar sample in our training dataset. A 1-nearest neighbor approach was used because when classifying samples based on its RNA-seq expression profiles, the number of different classes of samples may be large. Certain classes may only be represented a few times within the dataset, and in such cases it is harmful to use values of K much larger than 1.

To assess the predictive performance of our classifier, leave one study out cross validation was used. This cross validation technique iterates over every sample in the dataset and considers each sample as a query. The remaining samples in the dataset that do not belong to the same study of the current query sample are then assigned to the training set and used to fit the model. Leave one study out cross validation was used instead of leave one sample out cross validation because samples from the same study tend to be very similar to one another. In a naive approach using leave one sample out cross validation, accuracies were as high as 93.4% since very similar samples from the same study were being returned.

2.3 Similarity Measures Compared

Let x and y represent two gene expression vectors, with x_i and y_i representing the expression of the i^{th} gene or transcript. The following measures were used to compute the similarity between any two samples for our 1-Nearest Neighbor classifier.

1. Euclidean Distance:
   Euclidean distance is one of the most commonly used measures of similarity which computes a straight line distance between the two gene expression vectors.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

2. Pearson Correlation
   Pearson correlation measures the linear relationship between any two gene expression vectors. The Pearson correlation coefficient takes values between -1 to 1, with greater values resulting in higher similarities. The following formula was used to compute the Pearson correlation:

$$Corr(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \tag{2}$$

3. Spearman's Rank Correlation
   Spearmans rank correlation measures the monotonic relationship between any two gene expression vectors. To compute this measure, the TPM values of the gene expressions profiles are first ranked from lowest to highest and then a Pearson correlation shown in equation (2) is computed on the ranks instead of the actual TPM values.

4. Median Absolute Deviation
   Median absolute deviation is a measure of variability based on fold change, a commonly used metric of differential gene expression. For any two samples, the absolute value of a log fold change is computed among each pair of genes. The median of this list of fold changes is then used as the similarity measure. The median, rather than the mean, is robust to extreme outliers and advantageous in the case of gene expression where TPM ranges between several orders of magnitude.

$$\begin{aligned} x_i &= \text{gene i's expression in sample x} \\ y_i &= \text{gene i's expression in sample y} \\ log(\text{Fold Change}) &= log(x_i) - log(y_i) \end{aligned}$$

$$MAD = median(|log(\text{Fold Change})|) \tag{3}$$

5. Poisson Similarity
   The Poisson similarity measure, based on Daniela Witten's paper [5], proposes modeling RNA-seq data using the discrete Poisson distribution rather than a Gaussian distribution, as RNA-seq can be processed to give integer read count data. In contrast to Witten's methods, estimated read counts in

this project were used due to uncertainty in the read mapping process. A number of reads were not mapped uniquely to a single transcript, and in cases of such ambiguity, reads were mapped fractionally to transcripts based on probabilities estimated via a statistical model.

We represented our dataset in matrix notation such that X denotes a n x p matrix with n samples and p features (isoforms or genes). Witten's model then assumes each gene in each sample is Poisson distributed with a unique parameter accounting for variability in both the total number of reads per sample (s_i) and the total number of reads per gene or isoform (g_i). Witten's paper proposes different methods for estimating the size factors in our dataset, and in this work the total count approach was used. In addition, the d_{ij} term allows for the j-th feature to be differentially expressed between samples.

$$X_{ij} \sim Poisson(N_{ij}d_{ij}), \qquad N_{ij} = s_i g_j \tag{4}$$

However, biological replicate data such as RNA-seq tends to be overdispersed, meaning the variance is larger than the mean. As the mean and variance under a Poisson distribution are equal, a power transformation was used to make the data follow the Poisson distribution more closely. The transformation $X_{ij}^{'} \leftarrow X_{ij}^{a}$ is used, where $\alpha \in (0, 1]$ is chosen such that equation (5) holds. This is simply a test of the goodness of fit of the model to the data.

$$\sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(X_{ij}^{'} - X_{i.}{'}X_{.j}^{'}/X_{..}^{'})^2}{X_{i.}^{'}X_{j.}^{'}/X_{..}^{'}} \approx (n-1)(p-1) \tag{5}$$

where:

$$X_{.j} = \sum_{i=1}^{n} X_{ij} \quad X_{i.} = \sum_{j=1}^{p} X_{ij}$$
$$X_{..} = \sum_{i,j} X_{ij}$$

Once we transformed our data, to compute the similarity measure between any two samples X_{ij} and $X_{i'j}$, we first had to estimate N_{ij} and $N_{i'j}$ under the simpler model using a total size factor estimate. A likelihood ratio test was then conducted that tests for differential expression of the genes between two samples:

$$H_0 : d_{ij} = d_{i'j} = 1 \qquad H_a = d_{ij}, d_{i'j} > 0 \tag{6}$$

We then computed a modified log likelihood ratio under the H_a using the following estimates under Gamma priors.

$$\hat{d}_{ij} = \frac{X_{ij} + \beta}{\hat{N}_{ij} + \beta}, \qquad \hat{d}_{i'j} = \frac{X_{i'j} + \beta}{\hat{N}_{i'j} + \beta} \tag{7}$$

The likelihood ratio used as a measure of dissimilarity is then ultimately computed using the equation:

$$\sum_{j=1}^{p} \hat{N}_{ij} + \hat{N}_{i'j} - \hat{N}_{ij}\hat{d}_{ij} - \hat{N}_{i'j}\hat{d}_{i'j} + X_{ij}\text{log}\hat{d}_{ij} + X_{i'j}\text{log}\hat{d}_{i'j} \tag{8}$$

3 Results

3.1 Commonly Used Similarity Measures

Using our dataset of 1488 samples derived from the Sequence Read Archive (SRA), we built a 1-Nearest Neighbors classifier implementing the different similarity measures at both the isoform and gene level. Our features in the dataset were initially specified at the isoform level, and to analyze our classifier at the gene level, we simply summed the abundances of corresponding transcripts for each gene.

We initially implemented three commonly used similarity metrics: Euclidean distance, Pearson correlation, and Spearman correlation. Using a leave one study out cross validation, we assessed their classification accuracy as shown in Figure 1. Accuracies of these three similarity measures were consistently higher at the gene level representation compared to that of the isoform level. Additionally, using log transformed TPM values outperformed TPM values. In addition, a binary transformation of gene expression, in which TPM values over a threshold were assigned as 1 and 0 otherwise, was implemented under the Euclidean distance measure. After varying the threshold, we found that using TPM > 1 optimized the prediction accuracy under this measure. This binary transformation, simply representing whether a gene is turned on or off, performed comparably to the log TPM values. Overall, Pearson correlation at the gene level using a log transformation performed the best with an accuracy of 84.1%.

Table 1: Performance of 1-Nearest Neighbor classifier implementing the Euclidean, Pearson, and Spearman measures at the gene and isoform level.

Level	Gene Expression Value	Euclidean	Pearson	Spearman
Isoform	TPM	54.2%	57.7%	77.2%
	log TPM	70.6%	78.2%	77.2%
	Binary (TPM > 1)	68.1%		
Gene	TPM	58.5%	60.8%	83.4%
	log TPM	79.6%	84.1%	83.4%
	Binary (TPM > 1)	78.7%		

3.2 Median Absolute Deviation

In an attempt to outperform the commonly used similarity measures, we implemented the median absolute deviation and assessed its performance. This measure utilizes a log fold change, and to address log(0) cases, a constant was added to each gene expression value. From table 2, we noticed that the constant used affected the prediction accuracy of our classifier, and that smaller constant values resulted in higher accuracies.

However, after addressing the log(0) scenario, the best confusion matrix using $k = .0001$ shown in Figure 1 shows that a majority of samples were being predicted as brain samples. This was occurring as a result of an abundance of non-expressed genes in both samples. These specific non-expressed genes had a log fold change of 0, and as a result the median absolute deviation was being skewed towards 0. Therefore, we implemented a threshold in which a gene had to be expressed over a certain threshold in either sample to be included in the median fold change computation. Using the constant $k = .0001$ which resulted in our highest initial accuracy, we introduced a threshold of TPM > 1 that improved the accuracy of the classifier from 57.1% to 74.7% as shown in table 2.

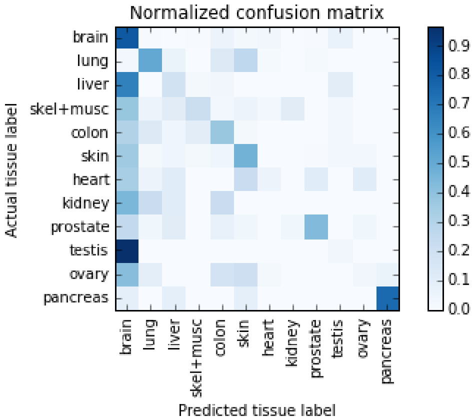

Figure 1: Confusion matrix for median absolute deviation with constant k = .0001 added.

Table 2: Performance of 1-Nearest Neighbor classifier implementing the median absolute deviation similarity measure with varying constraints at the gene level.

Level	Gene Expression Value	Constraints	Median Absolute Deviation
gene	TPM	k = 5	46.0%
		k = 1	56.4%
		k = .0001	57.1%
		k = .0001, (TPM > 1)	74.7%
		k = .0001, (TPM > 3)	74.3%
		k = .0001, (TPM > 5)	74.6%

3.3 Poisson Similarity

Our early attempts to implement the Poisson similarity measure based on the methods proposed by Witten were unsuccessful. When attempting to classify samples in our cross validation, certain samples in the training set were found to dominate the nearest neighbors across a majority of predicted samples. This was shown early on using a subset of 97 samples instead of the entire dataset. When we initially implemented our classifier, we identified a single sample dominating the nearest neighbor for a majority of our other samples. In an attempt to identify potentially bad samples, we removed this particular sample and ran our classifier on the remaining samples. However, this next iteration encountered the sample problem, albeit with a different sample. As a result, we repeated the process of removing this sample and rerunning the classifier. Through this process, 8 such samples were identified that when removed, the classifier stopped being dominated by a single sample.

To figure out why this set of samples was dominating a majority of nearest neighbors in our classifier, we looked at a few features: total read count sum, number of genes with zero read counts, and the distribution of gene read counts. After comparing this set of 8 quirky samples with the others, we found that these samples tended to have a higher number of genes with zero read counts and lower total read count sum as shown in Figure 2.

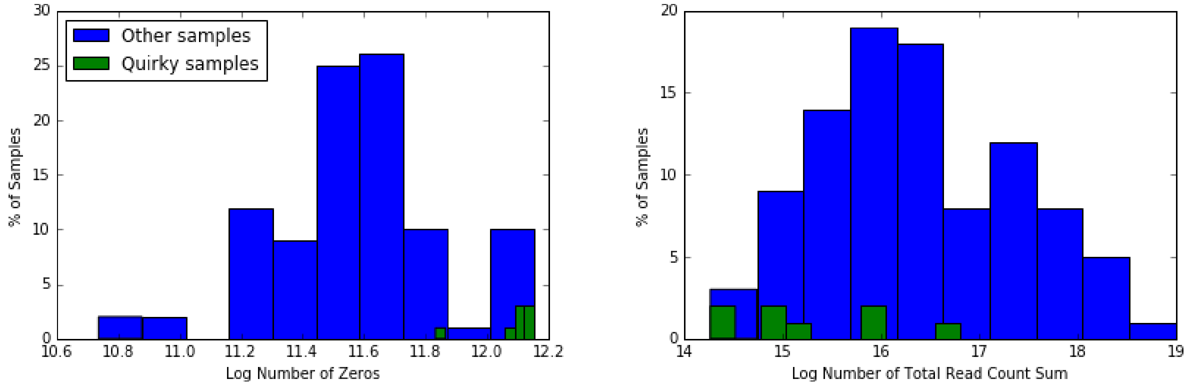

Figure 2: Histograms showing that quirky samples had a higher number of genes with zero read counts and lower total read count sums.

4 Discussion

In this work, we attempted to improve upon commonly used similarity measures including Euclidean distance, Pearson correlation, and Spearman correlation in the context of transcriptome-based phenotyping. These measures had been previously implemented in work completed by Bernstein using the same initial dataset as this work. His work identified the Pearson Correlation measure at the gene level to be the most effective with an accuracy rate of 81% [4]. However, in this work, our dataset differed since 43 samples were identified to have total read counts under 1,000,000 and therefore excluded in the analysis. Additionally, when log transformed TPM values were computed in this work, each TPM value was incremented by k=0.5 to address log(0) cases while the work done by Bernstein used k=1. Though our datasets and method for log transformation differed slightly, we also identified the Pearson Correlation to optimize prediction accuracy of tissue type with an accuracy of 84.1%.

The first alternative measure we implemented was median absolute deviation. After introducing a constant to avoid log(0) cases and a TPM threshold to avoid non-expressed genes, this measure performed comparably to the other measures. In particular, adding k = .0001 and using the threshold of TPM > 1, the median absolute deviation achieved its highest accuracy of 74.7%. We found that smaller added constant values improved prediction accuracies, while different TPM thresholds did not have as much of an impact. It is possible that different combinations of constants and thresholds may ultimately improve the prediction accuracy of this measure. However, for all constants and thresholds tested, the simpler measures using either a log or binary transformation outperformed this measure as shown in Table 1.

The other alternative measure we implemented was the Poisson-based similarity measure. However, using this measure with the 1-nearest neighbor approach was not successful due to the non-robust nature of the measure to what were identified as quirky samples. When attempting to classify samples in our cross validation experiment, certain samples in the training set dominated the nearest neighbors across a majority of query samples. Using a subset of 97 samples, we repeatedly implemented our classifier and identified a single sample dominating the nearest neighbor on each run. Through this process 8 such samples were identified, and when these were removed, the classifier stopped being dominated by a single sample. Upon closer inspection of these 8 samples, we found that they tended to have larger numbers of genes with zero read counts as well as lower total read count sums. However, we can see from Figure 2 that there were other samples on both ends of these extremes. Therefore, it is hard to conclude that these two features are what the Poisson similarity measure is not robust to. Future work is needed to determine exactly what features are causing certain samples to heavily affect the Poisson similarity measure.

This work utilized a 1-Nearest Neighbor algorithm approach, in which the most similar sample is returned. It would be interesting to see if different values of K used for the nearest neighbors algorithm would help to address the non-robustness of the Poisson similarity measure since the predicted label would depend on more samples. However, as stated earlier, additional steps would need to be taken to ensure that each phenotype class is represented by more samples than the value of K used in the algorithm. Otherwise, samples from underrepresented classes would be unlikely to be predicted correctly because the algorithm returns the label of the majority of a sample's K neighbors.

5 Acknowledgements

I would like to thank Dr. Colin Dewey and Matthew Bernstein for their guidance and support throughout this research project. In addition, I'd like to thank Whitney Sweeney for supporting and accommodating our CBB subgroup.

This work was supported by the IBS-SRP and the Department of Biostatistics and Medical Informatics with funding from the CPCP/BD2K Grant U54 AI117924.

References

[1] Wang, Z., Gerstein, M., & Snyder, M. *RNA-Seq: a revolutionary tool for transcriptomics.* Nature Reviews Genetics 10, 57-63, 2009.

[2] Li, B., Ruotti, V., Stewart, R., Thomson, J., Dewey, C. *RNA-Seq gene expression estimation with read mapping uncertainty.* Bioinformatics 26, 493-500, 2009.

[3] Jill Adams *Transcriptome: connecting the genome to gene function.* Nature 1, 2008.

[4] Mattthew Bernstein. *Deriving network-based features for predicting phenotype from RNA-seq.* Unpublished manuscript, 2015.

[5] Daniela Witten. *Classification and clustering of sequencing data using a poisson model.* The Annals of Applied Statistics 5, 2493-2518, 2011.