Machine Learning as a Tool for MicroRNA Analysis

Kevin S. Raines (ksraines), Brady Jon Quist (bradyq), and James Gippetti (jgippett)

(Dated: 13 December 2013)

Mature microRNAs (miRNA) are short ($\sim$ 22 nucleotide), single-stranded, noncoding RNA strands that regulate gene expression. MiRNA expression in tissues is increasingly used to classify cell states and show promise as clinical biomarkers for diagnostics. In this report, we summarize our experience using three different Machine Learning (ML) algorithms to classify miRNA expression profiles in two distinct biological models. In each case we use supervised learning and evaluate the performance of a given algorithm through hold-out validation. First we develop classifiers using miRNA expression profiles from a variety of tissues from mouse and human specimens with the goal of species classification. Then we evaluate these classifiers on the harder problem of cancer detection using a database of miRNA expression profiles of cancerous and normal breast tissue obtained by deep sequencing. By comparing and contrasting the performance of different algorithms, we gained insight into the nuances of applying ML to the analysis of miRNA expression profiles.

MicroRNAs are small noncoding bits of the genome that regulate gene expression and thus influence cell state and phenotype. In this study, we report on our findings from using miRNA expression profiles as a biomarker for different biological classification problems.

I. INTRODUCTION

Reliably detecting and classifying cell states is an important problem in cell biology, with applications in developmental biology and human health. Since miRNAs are rapidly emerging as an important player in our understanding of the regulation of gene expression in cells, and since gene expression controls many aspects of phenotype and cell state, there is reason to infer that miRNA expression profiles represent a compact biomarker. However, at this stage, our knowledge of miRNA and gene regulatory networks are incomplete and our measurements of miRNA expression profiles are noisy. Furthermore, since each miRNA can regulate up to hundreds of genes, it is not obvious that miRNA expression profiles can be naively used to classify cell states without additional biological information. Hence the theme of the present study: without large databases or additional biological information, can microRNA expression profiles alone be used to classify cell states?

We approach this problem by studying two simple systems of increasing difficulty. The first is the classification of species from miRNA expression profiles of various tissues within a species. The second system is breast cancer detection from a small labeled database of cancerous and normal breast tissues.

II. SPECIES CLASSIFICATION

We downloaded the miRNA expression levels from a study[1] that investigated RNA-seq experiments on a variety of tissues from different species. The data

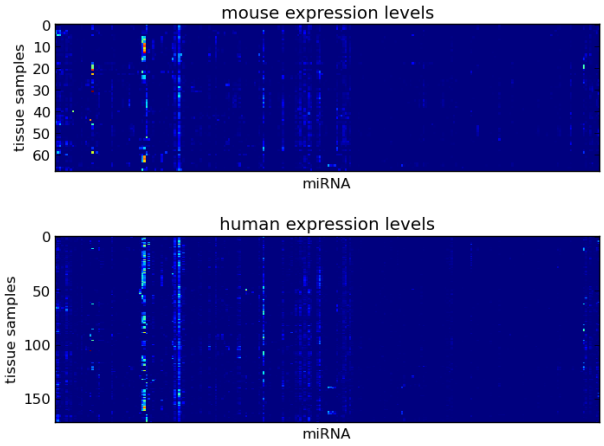

FIG. 1. Data used in this study (log of relative counts).

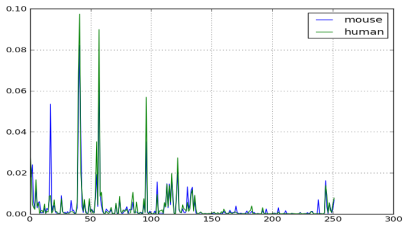

FIG. 2. Mean of data used in this study for each species (log of relative counts).

is available from `http://www.microrna.org/microrna/getDownloads.do` as relative expression levels within given data sets; we restricted our study to the human and mouse data with the goal of correctly classifying species from supervised learning on miRNA expression levels. We constructed attribute vectors from the expression data by intersecting the miRNAs that were measured for both species to construct vectors, indexed by miRNA type, of real numbers that give the relative expression level. We tested three machine learning algorithms and found that reliable species classification is possible.

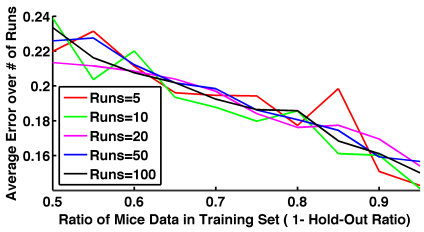

FIG. 3. Decrease in average error as the ratio of data used in training set is increased. It is important to note that the number of human species data points used in the training set was limited by the number of mouse data in order to mitigate the bias from the prior.

A. Naive Bayes

A Naive Bayes classifier was implemented as an initial attempt to classify the miRNA profiles. The validity of the Naive Bayes assumption is certainly in question and at best a rough approximation. However, given that the genetic species identifiers across many tissues are likely to be uncorrelated, and that only a few hundred microRNA regulate thousands of genes, we don't expect too much conditional dependence between the microRNAs.

The first version of the Naive Bayes that we implemented was straightforward, using Laplace smoothing and logarithmic mapping of the posterior function to prevent 'underflow'. The first set of results with hold-out ratio (the fraction of data saved for the test set) of 0.3 yielded an error of 28.77%. However, since there was $\sim 3x$ human tissue examples in the training set than mouse tissue examples, the prior P(Human) = 0.72 had a large enough effect on the prediction to classify all of the tissues in the test set as human. The validity of using Naive Bayes is not just based on the on the conditional independence of the random variables, but also depends upon having a reasonable prior. In this context, the prior doesn't make sense since the amount of training examples of a particular species doesn't predict the amount in the testing set - or at least we don't have reason to assume this. We adjusted the algorithm to effeectively remove the prior, essentially turning it into a maximum likelihood method, which slighly improved the results.

We also extended Naive Bayes to use cross validation and kept the ratio of human to mouse tissues in the randomly generated training matrix constant at 1:1. The algorithm randomly selected a fraction of the data and was run 5 times for $x = [5, 10, 20, 50, 100]$ (figure 3). The results show significant improvement in error as the hold-out ratio is decreased. Building the training matrix with 95% of the available mouse tissues examples and the same amount of human examples, the Naive Bayes implementation correctly classified 85% of the test set on average.

Using the known labels, we evaluated the strongest informal species indicators by comparing the conditional probabilities of observing different miRNA. These indicators are omitted for brevity.

B. SVM with a Gaussian Kernel

In order to study the quality of classification with a standard support vector machine (SVM) algorithm, we used the libsvm library[2]. This library solves the following optimization problem:

$$\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \tag{1}$$

$$\text{subject to} \ \ y^{(i)}(\mathbf{w}^T\phi(x^{(i)}) + b) > 1 - \xi_i \tag{2}$$

$$\xi_i > 0 \tag{3}$$

We found that the Gaussian Kernel or 'Radial Basis function' (RBF) performed the best in initial trials and is recommended by the authors of the library in their online guide[2] as the starting point for the application of SVM to general datasets. The RBF is paramaterized by a single parameter γ and is defined as follows:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma||\mathbf{x}_i - \mathbf{x}_j||_2^2} \ , \ \gamma > 0 \tag{4}$$

Since we were able to achieve excellent results with this kernel, it is the focus of our study. In their guide, the authors of the library recommend linearly scaling the data so that each component of the attribute vector $\mathbf{x}_i$ is in the interval $[0, 1]$. Although the authors advocate this procedure and demonstrated its effectiveness on several datasets, we observed the opposite effect in our data: it reduced the accuracy by about 10%. This is likely in part due to the fact that the data was scaled in a biologically meaningful way and thus altering the scaling diminishes the information content of the data.

Another problem with this approach to scaling the data is that it gives all of the miRNA equal weight. Since miRNA expression levels can be viewed approximately as a two state system (either 'on' or 'off'), then normalizing the data in this way has the potential to turn miRNA in the 'off' state into the 'on' state by amplifying biological and machine noise. We can clearly see in figure 1 that both species share overlapping regions of miRNA which have low expression levels.

The algorithm as described requires two parameters which we searched for via grid search and k-fold cross-validation (figure 4). Using this approach, we were able to consistently predict species labels with $\sim 98\%$ accuracy over a broad range of parameters, indicating robust algorithm performance.

C. Modeling miRNA data as a Poisson Distribution

While the Gaussian Kernel SVM yields excellent results, it is based on a continuous distribution model. Our data, however, are actually from discrete counts acquired from RNA sequencers. We therefore implemented a model[3] that uses a more realistic Poisson distribution.

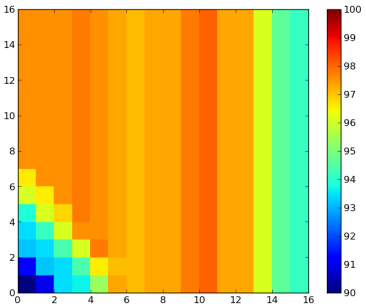

FIG. 4. Grid search results over parameters C (y-axis) and γ (x-axis) using k-fold cross validation (k=10). Colorbar indicates accuracy of prediction; steps are in increments of 2^x, where x is the axis tick. Note that reliable predictions persist over large parameters ranges.

In the implemented method, we have $X_{ij} \sim \text{Poisson}(N_{ij})$, where $N_{ij} = s_i g_j$. Here, X_{ij} is the number of reads mapped from the sample i (of n) to the feature j (of p), s_i is the constant scaling factor from the i$^{\text{th}}$ sample (which can be quite variable due to noise in the RNA sequencers), and g_i represents how often the j$^{\text{th}}$ feature is mapped relative to the other features. A natural extension proposed in this model is given as $X_{ij}|y_i = k \sim \text{Poisson}(N_{ij} d_{kj})$, where $N_{ij} = s_i g_j$, where k represents the classification of a given sample (ie. malignant vs benign tumors, species, etc.), d_{kj} is the parameter that allows for features to be expressed differently based on its class, and y represents the classification.

Due to the highly variable nature of s_i, the method of estimating (or effectively normalizing the i$^{\text{th}}$ row of X) has the potential to drastically affect classification. This is particularly true when a small number of features are much larger than the others and have noticeable variability. Three methods were proposed for estimating s_i to normalize the i$^{\text{th}}$ sample: 1) Total Count Method (dividing the sum of the i$^{\text{th}}$ row of X by the sum of all the elements in X), 2) The Median Ratio (finding the weighted median of all features in the i$^{\text{th}}$ row, divided by the sum of the medians from all other rows), and 3) The Quantile Method (finding the 75th percentile for the i$^{\text{th}}$ row and dividing it by the 75th percentile for all rows). After using training data to compute s_i , $\hat{g}_j$ (which multiplied provide $\hat{N}_{ij}$), and $\hat{d}_{kj}$, the test data is then normalized using each of the three methods above. The log of the conditional distribution is then given as:

$$\log P(y^* = k|\mathbf{x}^*) = \sum_{j=1}^{P} X_j^* \log \hat{d}_{kj} - \hat{s}^* \sum_{j=1}^{p} \hat{g}_j \hat{d}_{kj} \quad (5)$$

$$+ \log P(y = k) + c \quad (6)$$

The model then classifies the test data x^* to be the class for which $\log P(y^* = k|\mathbf{x}^*)$ is largest. We also experimented with soft thresholding as discussed in the paper

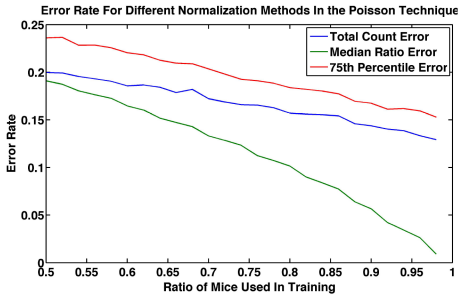

FIG. 5. The error rates for the three normalization techniques used in our Poisson model are shown above. We clearly see that the Median Ratio method has significantly less error than the other methods.

without substantial impact upon the results.

We note that we set $P(y = k) = 0.5$ even though that wasn't reflected in our test data. This was done to prevent bias as we wanted to test data that did have an equal distribution of humans and mice.

Figure 5 shows the error for each normalization method versus the ratio of mice data used in the training set (with the number of human samples being the same as the number of mice samples). All other data was then tested. These points were calculated by taking 10^3 iterations of randomly selected samples for each ratio of data. The plots clearly show the superior performance of the Median Ratio when compared to the other methods. This was not expected, as the data presented with the model[3] showed very little variation based on normalization method. We believe that our results differ because the data were pre-normalized since the sum of the miRNA across the ssample was unity. This is inconsistent with the Poisson model and therefore led to divergent results.

III. CANCER DATA

We downloaded the data from a deep sequencing study of breast cancer[4]. This data has several important differences compared to the species data. It is highly imbalanced with respect to the two classes: with only 11 normal tissue samples compared to 151 invasive breast carcinoma samples, the ratio of the two sample sizes differs by a factor of almost 14. Furthermore this data was unscaled and thus a scaling method had to be chosen. Although accuracy is an important quality measure, especially for comparision to the species study, it is not ideal in this case. Due to the imbalance of the sizes of the data sets, it is possible for the model to predict that everything is cancer and obtain a high accuracy rate. We therefore analyzed false positives as well as false negatives. In a hypothetical clinical situation, we imagine that false negative would be the quality measure that carried the most weight. Finally, the imbalance in the size of the classes and the large number of carcinoma

samples made k-fold cross validation impractical. Hence we used hold-out random sampling to test the different algorithms.

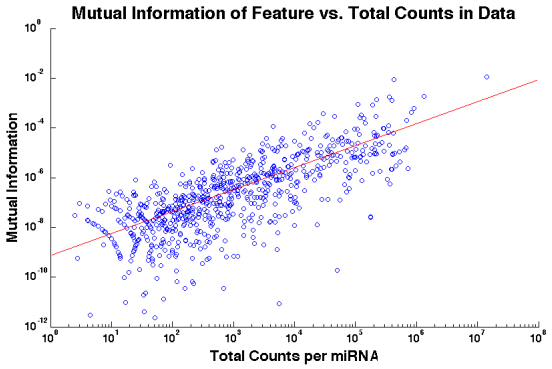

FIG. 6. This plot shows the correlation between the miRNA's Mutual Information value and its total number of counts in the data set. There is a clear trend indicating that the miRNA that are more represented in the data set are of greater value for cancer classification. The slope of the linear fit to the log-scaled data is 0.87, which means that $\mathrm{MI}(\mathrm{miRNA}_j) \sim n_j^{0.87}$.

In order to analyze the informational content of the data, we used the feature selection approach of Mutual Information, assigning relative values to each miRNA feature, which represented the correlation between the miRNA and the cancer state. There was a definite trend between the number of miRNA counts in the data and the informational value of the miRNA features (figure 6). The slope of the linear fit to the log-scaled data was 0.87, which means that, if we denote the number of counts for the j^{th} miRNA as n_j, then the mutual information MI is roughly proportional to $n_j^{0.87}$. The majority of the miRNA had MI values extremely close to zero, corresponding to a negligible correlation or divergence. We were able to apply our findings from this Mutual Information analysis by only including miRNA with significant counts in the data (i.e. the top 250 out of ~ 1100).

We ran the same modified implementation of Naive Bayes developed with the species data on the cancer data set. The same algorithm yielded much worse performance on the cancer data; the maximum average accuracy ranged from 58% with a hold-out ratio of 90% to 61% with a hold-out ratio of 10%. This decrease in performance may be due to a lack of normal tissue data, as well as the increased correlation between the miRNA behavior in these carcinogenic and normal human tissues. In contrast to the species classification problem, where we had reason to assume conditional independence between the miRNA levels that differentiated species, here it seems likely that the miRNA involved in tumorigenesis are conditionally correlated. This invalidates the Naive Bayes assumption and explains the poor performance of that algorithm on these data.

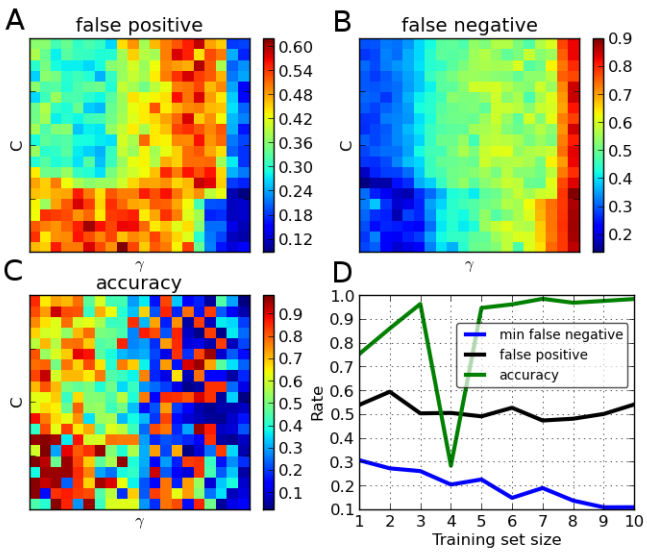

FIG. 7. A.-C. Grid search over SVM parameters using three different quality measures. Note that there is not overlap between minima. The quality landscape is much more rough compared to the species data; D. The false negative rate shows improvement with increasing training set size, as well as concomitant improvement in the other measures. The dip in the accuracy illustrates the difficulty associated with the non-overlapping landscapes: the values plotted are for the parameters that optimized the false negative rate only.

A. SVM with a Gaussian Kernel

Significant differences to the species classification problem were observed with the Gaussian Kernel SVM. The most obvious difference is the decline in the peformance across all measures. In figure 7, we observe that the exponentially-varying parameter space is not only more sensitive to parameter values but also noisier - that is, small changes in parameter values cause abrupt changes in quality. In contrast to the species classification problem, which was insensitive with respect to the regularization parameter (C), the cancer data showed much more variation. This suggests that the cancer data is not as well separated as the species data.

Perhaps the most significant feature of the grid parameter search is the non-overlapping minima of the different quality measures. Minimization of, say, the false negatives does not yield an optimal value for the false positives. The second observation is that all of the measures were sensitive to the data partitioning between training and testing sets. We used 100 random samples for each pixel in the grid search to reduce sampling noise, but the variation between samples was substantial, which suggests a need for larger datasets. We experimented with significantly unequal numbers of class representatives in the training data to compensate for the small number of normal tissue representatives, but this decreased performance. Performance also decreased, and computa-

tion time significantly increased, when we used all of the miRNA features; following the MI analysis, all results shown used only the 250 most prominant miRNA.

Finally, we note the declining rate of false negatives with training set size. Given that the largest training set used only had ten representatives from each tissue type, the false negative rate of ~ 0.1 is quite promising.

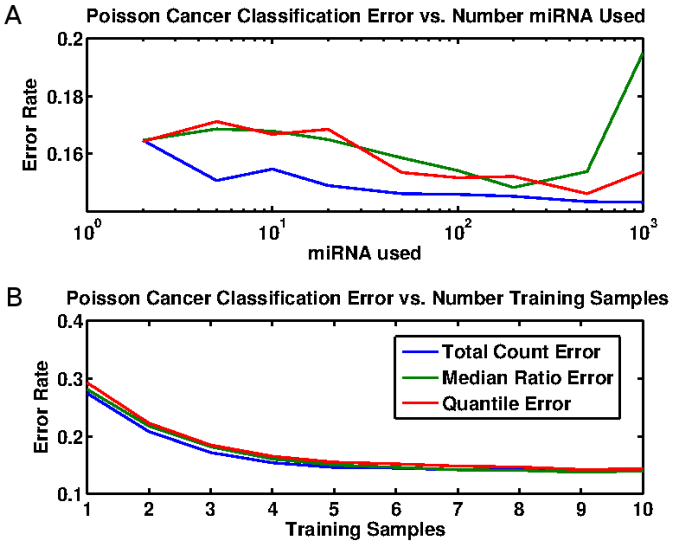

FIG. 8. A. Here we visualize the Poisson cancer classification error verses the number of miRNA used for classification. We see that the number of RNA samples used can have a significant impact on the classification, with some normalization methods even getting worse when all miRNA are used. B. The Poisson cancer classification error verses the number of training samples used for each cancer and normal tissue. Here we see that the normalization method had little effect on the classification accuracy.

B. Poisson Model

We performed cancer classification by supervised learning while modeling the data as a Poisson distribution as with the species data. Figure 8.a shows the cancer classification error versus the number of most prominent miRNA used in classification. It is interesting to note that the performance deteriorated significantly when we included all of the miRNA for the Median Ratio normalization method. This is consistent with the MI analysis.

Figure 8.b shows the Poisson cancer classification error as a function of the training samples used when considering only the 250 most prominent miRNA. Here we see little difference between normalization techniques, especially as the number of training samples increases. This is consistent with the findings in (3), but is in contrast to the species classification where the Median Ratio normalization was significantly better than the other techniques. We believe this is because the cancer data were not normalized in advance, which allowed them to be more closely modelled by a Poisson distribution.

For both analyses, the same number of cancer and normal tissues were randomly chosen for both training and testing, thus removing the possibility that misclassifying normal tissue would have little effect on the displayed accuracy. Furthermore, 10^4 iterations were performed for each parameter value to reduce noise. The classification analysis performed here shows that the Poisson classification method performs quite well for the amount of training data and would presumably continue to improve with more data.

CONCLUSION & OUTLOOK

We began this study with the goal of understanding how to apply ML to the classification of cell states by miRNA expression profiles. By beginning with a simple system, the classification of distinct species from various tissue profiles, we encountered the main issues that persisted throughout our investigation. The species analysis demonstrated that the ratio of species in the training data had a significant impact on the performance of all three classification algorithms. We found that by constructing a training set with a 1:1 ratio of species, each algorithm performed better. This observation proved useful in the cancer detection problem where the same approach improved our results.

Our analysis showed that while cancer detection using miRNA samples is promising, it is significantly harder than species differentiation. This is consistent with the intuition that the underlying biological difference between two species should be greater than a given tissue in a normal and disease state. Additionally, the relatively small size of the normal breast tissue data limited our ability to train the classifiers.

While our results were promising, it is likely that they can be improved with more data and by incorporating additional biological knowledge, such as correlating miRNA abundance with other biological markers to construct more detailed cell state fingerprints. In conclusion, we found that miRNA expression profiles are surprisingly flexible biomarkers and, in combination with ML tools, can characterize cell states across dramatically different biological models without significant prior biological knowledge or extensive datasets.

[1] P. Landgraf, "A mammalian microrna expression atlas based on small rna library sequencing." Cell **129**, 1401–1414 (2007).
[2] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011).
[3] D. M. Witten, "Classification and clustering of sequencing data using a poisson model," The Annals of Applied Statistics **5**, 2493–2518 (2011).
[4] H. Farazi, Thalia A, "Microrna sequence and expression analysis in breast tumors by deep sequencing." Cancer research **71**, 4443–4453 (2011).