A Spam Classifier for Biology: Removing Noise from Small RNA Datasets

Diane Wu, CS229 Fall 2010

Background and Motivation

Originally discovered as an antiviral mechanism, RNA interference (RNAi) has since then been shown to play a large role in the natural cellular regulation of endogenous transcripts. The products of the RNAi pathway and the effectors of transcript regulation are small 18-30 nucleotide (nt) small RNAs that are homologous in sequence to their target transcripts. Consequently, discovery of small RNAs that are identical in sequence to an endogenous transcript implies possible role of the RNAi pathway in the regulation of said transcript.

Many groups of small RNAs exist in the cell are encoded by the host genome naturally. Most of the well characterized groups of small RNAs are typically oriented antisense to the target transcripts and are believed to be produced by templating off the endogenous transcript. In the nematode *C. elegans*, this includes a class of 21-23nt small RNAs that have been largely implicated in transcriptional silencing. Two other classes of less-well-characterized small RNAs of sizes of 23-24nt and 26nt, respectively, also appear to have silencing effect. While we will not discuss the mechanistic details regarding the generation of these small RNAs here, we will note that these small RNAs are believed to be produced by templating off the endogenous transcript, hence resulting in an antisense orientation of the small RNAs relative to the transcript (Figure 1).

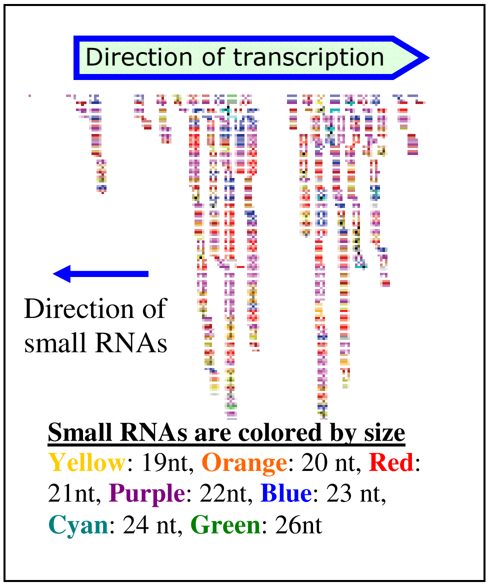

Direction of transcription
Direction of small RNAs
Small RNAs are colored by size
Yellow: 19nt, Orange: 20 nt, Red: 21nt, Purple: 22nt, Blue: 23 nt, Cyan: 24 nt, Green: 26nt

Figure 1: Notice size enrichment of 21-23nt RNAs

But what about small RNAs that align to the sense strand of the transcript? Do these exist, and if so, do have any function? Unfortunately, many confounding factors thwart any ventures to answer this question. While I will not take time here to survey those difficulties, I focus on the problem that small RNA sequencing libraries are often plagued with mRNA degradation, making small RNA datasets hard to evaluate. This problem arises from the fact that small RNAs are extracted using a size-selection method which selects for RNAs of 18-30 nucleotides in length. Hence, any larger transcripts that degrade naturally into this size range by chance (e.g. through random shearing of longer transcripts during the RNA extraction procedure) will also be captured by sequencing. As a result, researchers often ignore looking at small RNAs that align to sense strands of transcripts because they are mostly plagued with contamination. While this is true in many case-by-case analyses, this is a poor assumption to make and may drastically limit our ability to detect novel signals.

The goal of this project is two fold. First, I aim to characterize distinct small RNA signals over loci throughout the genome. Using two unsupervised algorithms, k-means clustering and principal component analysis, I expect to be able to both quantify the representation of known signals throughout the transcriptome as well as potentially identify new small RNA signals. As a second goal, I aim to develop an algorithm to categorize small RNA distributions as "degradation" or as a specific type of signal. In particular. I will build an SVM classifier for this task and experiment with values for the parameters. Training sets will be picked from a combination of exemplary examples from the unsupervised clustering performed previously. In combination, these two goals would allow enhancement of the signal-to-noise ratio in small RNA datasets and provide the potential for discovery of novel small RNA signals from biological datasets.

Methods

Data Set

A collection of 15 small RNA datasets were obtained from the nematode *C.elegans* using whole animals. In brief, small RNAs were sequenced using the Solexa platform, and sequences were aligned to the *C. elegans* transcriptome. RNA counts of each length for each gene were added up. Each dataset was aligned to the genome using the Bowtie short read aligner.

In order to select training points that offer high information content, the total of 49000 transcripts were filtered, for each sample, to include only transcripts that contain at least 20 small RNA read alignments and at least 10 unique sequences.

Feature Selection

The high dimensionality of biological datasets, the limited evidence for making assumptions of independence, and the stochastic process of generating these datasets makes feature selection both a challenging and critical aspect for the success of applying any machine learning algorithm on such datasets. The initial set of features selected for this analysis was based on personal experience with small RNA datasets as well as characterizations of small RNA biogenesis as revealed in the literature. For each gene, I calculated the following possible features:

1. Total number of RNAs aligned
2. Median redundancy of RNAs aligned (i.e. number of aligned RNAs per unique sequence)
3. Proportion of number of RNAs of each length (18bp to 29bp) aligned
4. Proportion of number of RNAs over each 3nt range (eg. 19-21nt, 20-22nt)

Features were evaluated the ability of these features to separate the data using k-means (see below). In particular, I found that simply including (3) seemed to provide the best results; clusters of genes could be separated and identified as possible degradation products based on visual inspection of the small RNAs (using UCSC Genome Browser, alignment visualization tool). Including (1) and (2) expectedly resulted in features that distinguished the data based on expression level and genomic structure, respectively. Surprisingly, including (4) actually reduced the separation of the features without improving robustness of classification.

Unsupervised Learning:

Principal Component Analysis and K-means Clustering

For each sample, highly expressed genes were subject to principal component analysis using the set of features selected from alignment to the sense orientation. Results from each sample was consistent: principal component analysis of sense-oriented small RNA size distributions reveal distinct separation of data points. The first two principal components explained an average of 45% and 27% of the variance, respectively. Analysis of the feature loadings for first two components reveal that these correspond to distributions with high proportions of 22-23nt RNAs and larger (27-29nt) RNAs, respectively. Further, we performed k-means clustering on these features and compared the results with those from PCA (Figure 2), revealing clusters with centroid positions reminiscent of siRNA size distributions (purple, blue, and orange clusters) as well as degradation (green and red clusters.) The choice of K was selected between values of 3 to 7 based on robustness of the clustering for producing similar cluster centroids.

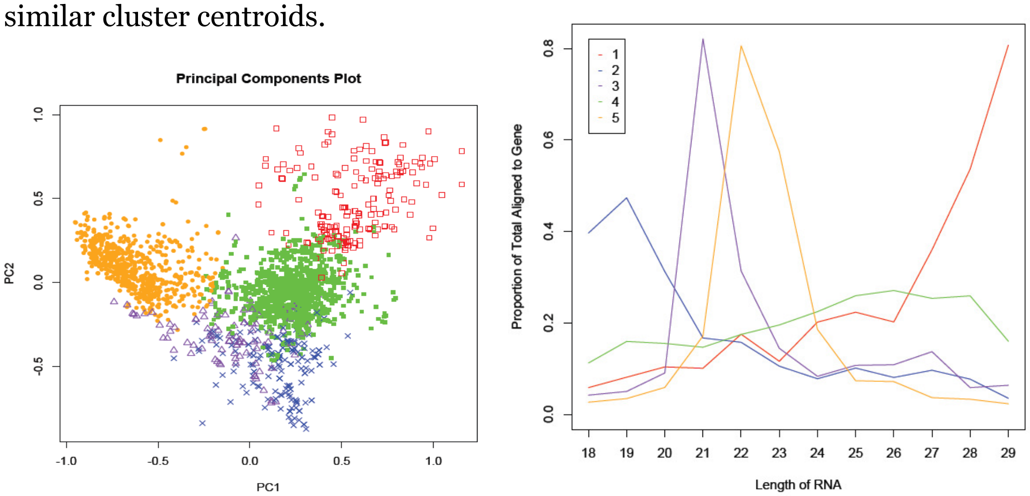


Figure 2: ***(Left)*** PCA of sense-oriented small RNA size distributions. Each data point is a transcript, colored by cluster assignment using k-means clustering with k=5. ***(Right)*** Centroid distributions of clusters.

To test that our method for detecting mRNA degradation is unlikely to misclassify real small RNA signals (based on inherent noise in the small RNA population), we performed the same procedures on a sample of small RNA sequences from a different chemical preparation that is known to select for one class of 21-23nt siRNAs. This reveals that aggregate small RNA profiles for large sets of genes indeed exhibit a strong characteristic and contain very little noise (Supplemental Figure 1). These distributions are consistent regardless of the k used in k-means analysis (data not shown).

Since the populations of small RNAs that we have termed “mRNA degradation” have not been previously characterized as a class of functional small RNAs (nor has specifically studying mRNA degradation signals been of any interest to the small RNA scientific community), we have no convincing evidence from this analysis that these are indeed non-functional or result from degraded mRNAs. However, we do find an indication for this suggestion: genes which exhibit small RNA distributions with the “degradation characteristic” also exhibit high mRNA levels (Supplemental Figure 2), suggesting that they are likely degradation products from these very abundant sources.

Supervised Learning: Support Vector Machines

Principal component analysis and K-means clustering are useful for discovery of unanticipated relationships in the features without bias of prior knowledge regarding the data. However, this approach requires supervision to ensure the right choice of K in k-means clustering. Furthermore, k-means clustering is a heuristic algorithm and results may vary largely between replications. Both PCA and K-means also relies on the significant representation of each type of data in the dataset—a signal that is not represented at a high level would be merged into an adjacent cluster in K-means and would not contribute to a significant proportion of the variance in PCA. These above assumptions made by the unsupervised algorithms make them poor candidates for general detection of mRNA degradation signals in novel datasets, particularly datasets from various genetic backgrounds that could drastically perturb the landscape of small RNAs. Hence, we use the insight gained from these techniques to build a support vector machine classifier for detection of mRNA degradation signals.

Since no labeling of positive or negative training sets exist, we define our training set using training points with exemplary traits as defined from our previous analysis.

Selection of Negative Training Examples

To generate negative examples, I took the genes falling into the “degradation” clusters (in green and red from Figure 2), and selected for those genes which exhibit exemplary characteristics to those we have identified as “degradation”. In particular, we chose genes whose small RNA distributions on the **sense** strand are (a) close to the centroid (sum of squared differences <0.4, maximum squared difference of any length <0.2), and (b) have high mRNA levels (>100 mRNA counts). Justification for the choice of cutoff for the first criteria is shown by examining a distribution of the squared differences (Supplemental Figure 3), while justification for the second criteria was noted previously (Supplemental Figure 2).

Selection of Positive Training Examples

Positive training examples were generated using distributions of small RNAs mapping to the **antisense** strand, since these are the canonical siRNAs. Even when k is set to a large number, no centroid with a characteristic “degradation” signal is observed (Figure 3, k=8). Hence, all genes close to centroids in the antisense strand were used as positive training examples (evaluated based on squared differences of proportions as per negative training sets).

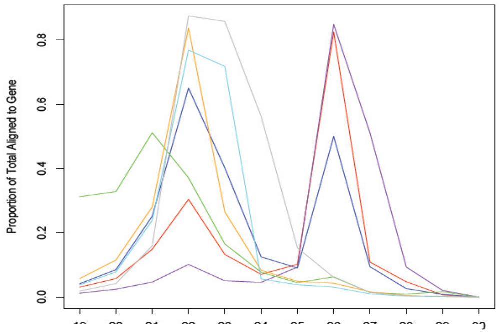


Figure 3. K-means centroids of antisense small RNAs. K=8

Evaluation

The goal of our classifier is to be able to take any dataset and label each loci as degradation or siRNA signal. As such, we are interested in the generalization error of a classifier for labeling the loci in at least one whole small RNA library. Hence, we perform 5-fold cross-validation on the 15 samples to select for optimal parameters for the model (i.e. train on points form 12 samples, test on points from 3 samples). An alternative method for evaluation is k-fold cross-validation of all data points without grouping by sample; however, this provides an underestimation of generalization error. This underestimation is due to the fact that the largest source of variance in small RNA datasets is that between different technical preparations of the RNA libraries; revealing a proportion of points from any library provides indications as to where the remaining points in that library might lie.

SVMs were built for each training set: positive training examples were labeled as "1" and negative training examples (i.e. degradation) were labeled as "0". Different kernels were explored for their performance in the SVM in conjunction with varying parameter values such as those for C, gamma, or degree of polynomial, as pertained to the particular kernel. The polynomial and linear kernels performed worse than the radial kernel for all parameters tested (not shown). The logistic kernel exhibited the worst performance. For the radial kernel, increasing C, the cost value for outliers, resulted in better training and test errors if ɣ is kept low (Table 1). If ɣ and C are increased simultaneously, the model overfits the data, resulting in extremely low training error and high test error. In particular, such values biased towards an unacceptably high false positive rate, labeling all points "1". Optimal values for test error, training error, and a balance of false positive and false negative rates can be achieved with C taking values between 2 to 8 and ɣ taking values between 0.01 and 0.1.

Table 1. Average Test and training errors (5-fold validation) using the Radial Kernel for varying values of C and ɣ

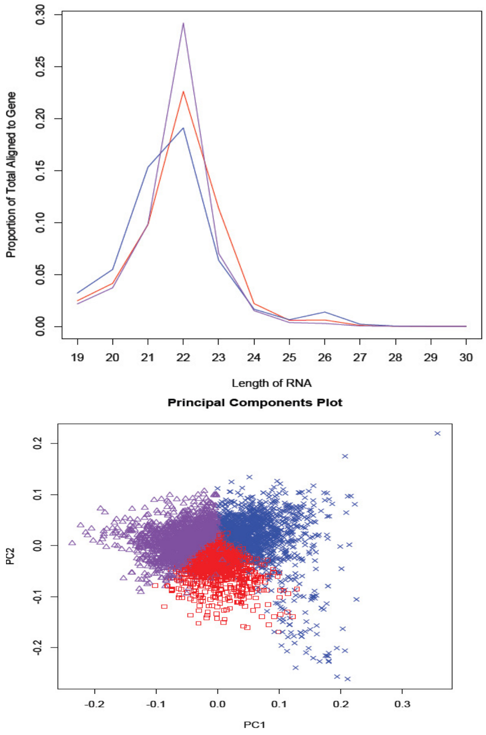
		Training Error			Test Error		
C	**gamma**	**Avg % Error**	**Avg %FP**	**Avg %FN**	**Avg% Error**	**Avg%FP**	**Avg%FN**
0.005	0.0156	1.81	35.42	0.05	2.87	47.51	0.06
0.005	0.0312	1.60	31.58	0.03	2.76	46.43	0.04
0.005	0.0625	1.77	35.41	0.02	3.11	52.34	0.01
0.005	0.125	3.42	68.43	0.00	5.10	87.09	0.00
0.005	0.25	5.04	100.00	0.00	5.68	100.00	0.00
0.005	0.5	5.04	100.00	0.00	5.68	100.00	0.00
0.02	0.0156	0.99	19.44	0.03	1.79	27.14	0.18
0.02	0.0312	0.82	15.93	0.03	1.50	22.59	0.17
0.02	0.0625	0.74	14.41	0.03	1.41	21.28	0.18
0.02	0.125	1.02	20.45	0.02	2.12	35.30	0.07
0.02	0.25	3.16	63.22	0.00	4.92	83.73	0.00
0.02	0.5	5.03	99.76	0.00	5.68	100.00	0.00
0.5	0.0156	0.42	7.23	0.06	0.86	10.84	0.23
0.5	0.0312	0.36	6.02	0.06	0.83	10.04	0.25
0.5	0.0625	0.30	5.08	0.05	0.81	9.75	0.24
0.5	0.125	0.23	3.81	0.03	0.83	10.00	0.24
0.5	0.25	0.16	2.81	0.02	0.88	10.63	0.25
0.5	0.5	0.12	2.23	0.01	2.43	39.34	0.09
2	0.0156	0.34	5.53	0.07	0.76	8.65	0.25
2	0.0312	0.26	4.34	0.05	0.76	8.39	0.26
2	0.0625	0.19	3.02	0.04	0.76	8.56	0.25
2	0.125	0.13	2.19	0.02	0.78	9.15	0.25
2	0.25	0.09	1.51	0.01	0.83	9.87	0.26
2	0.5	0.04	0.79	0.00	1.56	22.07	0.22
8	0.0156	0.28	4.43	0.06	**0.74**	7.92	0.27
8	0.0312	0.19	2.81	0.05	**0.79**	8.50	0.28
8	0.0625	0.13	2.04	0.03	**0.76**	8.42	0.26
8	0.125	0.08	1.35	0.01	0.78	8.87	0.27
8	0.25	0.03	0.64	0.00	0.82	9.60	0.27
8	0.5	0.01	0.28	0.00	1.59	22.14	0.24

Conclusions and Future Directions

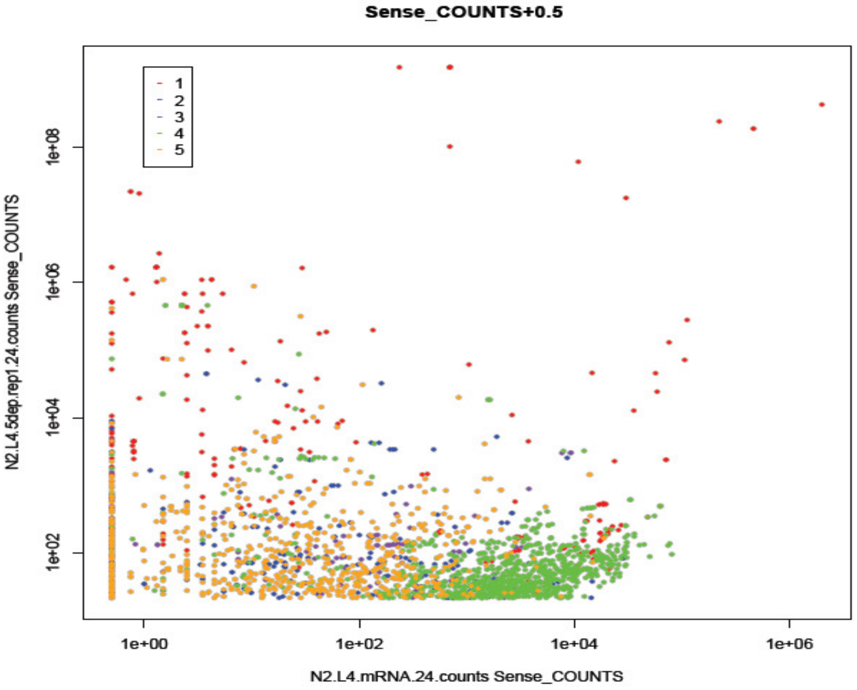
By building a model to filter mRNA degradation signals from sense-oriented small RNA profiles, we provide an opportunity for further exploration novel pathways and functions of small RNA regulation. Comparisons of small RNA profiles of particular genes between samples in various genetic backgrounds would be the logical next step in deciphering the function and regulation of these sense-oriented small RNAs. Characteristic differences may exist between sense-oriented small RNAs and antisense-oriented small RNAs that we have not explored in this work. Nevertheless, the results from this study illuminate the large class of sense-oriented small RNAs that have previously been ignored.

Supplemental Figures

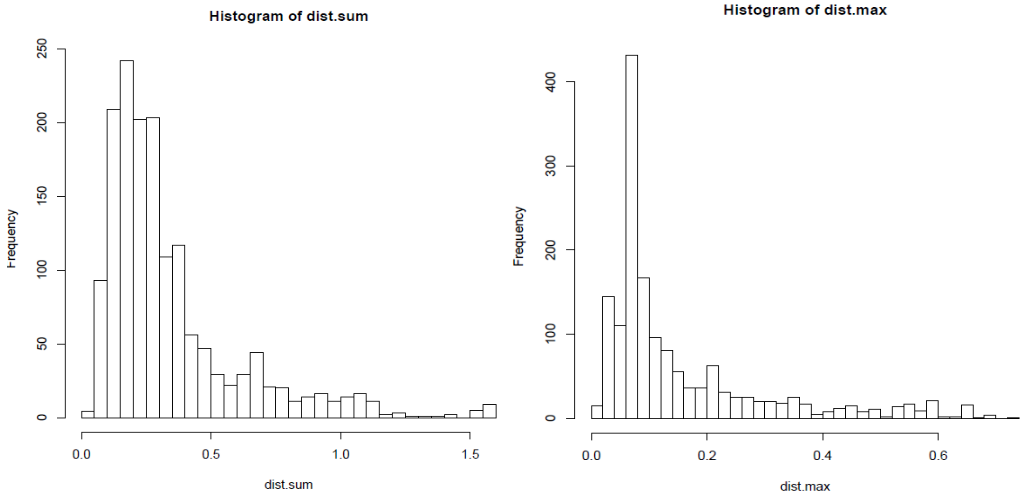
Supplemental Figure 1. K-means and PCA of small RNA distributions from different RNA cloning procedure that selects for canonical siRNAs.



Supplemental Figure 2. mRNA Seq data from same sample was used to compare to small RNA data. Higher counts indicate higher expression level. Genes from degradation cluster (colored as in Figure 2) are highly expressed in mRNA, confirming implications of degradation in siRNA signal.



Supplemental Figure 3. ***(Left)*** Distribution of sum-of-squared differences between genes in the orange "degradation" cluster (Figure 3) and their centroid. Alternatively, we can calculate the maximum-squared differences of any RNA length between each gene and its centroid ***(right).***



References

Karatzoglou A, Meyer D, and Hornik K. Support Vector Machines in R. *Journal of Statistical Software*. April 2006, Volume 15, Issue 9.

Chang CC, Lin CJ (2001). "libsvm: A Library for Support Vector Machines." URL http://www.csie.ntu.edu.tw/~cjlin/libsvm

Pak J and Fire A. Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. *Science*.2007 Jan 12;315(5809)

Han T, Manoharan AP, Harkins TT, Bouffard P, Fitzpatrick C, Chu DS, Thierry-Mieg D, Thierry-Mieg J, Kim JK. 26G endo-siRNAs regulate spermatogenic and zygotic gene expression in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A*. 2009 Nov 3;106(44)