ANALYZING MODULAR RNA STRUCTURE REVEALS LOW GLOBAL STRUCTURAL ENTROPY IN MICRORNA SEQUENCE

Timothy I. Shaw* and Amir Manzour

Institute of Bioinformatics, University of Georgia
Athens,Ga 30605, USA
Email: gatech@uga.edu, amanzour@uga.edu

Russell L. Malmberg

Department of Plant Biology, University of Georgia
Athens,Ga 30605, USA
Email: russell@plantbio.uga.edu

Yingfeng Wang and Liming Cai*

Department of Computer Science, University of Georgia
Athens,Ga 30605, USA
Email: ywang802@uga.edu, cai@cs.uga.edu

Secondary structure remains the most exploitable feature for non-coding RNA (ncRNA) gene finding in genomes. However, methods based on secondary structure prediction may generate superfluous amount of candidates for validation and have yet to deliver the desired performance that can complement experimental efforts in ncRNA gene finding. This paper investigates a novel method, unpaired structural entropy (USE) as a measurement for the structure fold stability of ncRNAs. USE proves to be effective in identifying from the genome background a class of ncRNAs, such as precursor microRNAs (pre-miRNAs) that contains a long stem hairpin loop. USE correlates well and performs better than other measures on pre-miRNAs, including the previously formulated structural entropy[14]. As an SVM classifier, USE outperforms existing pre-miRNA classifiers. A long stem hairpin loop is common for a number of other functional RNAs including introns splicing hairpins loops[25] and intrinsic termination hairpin loops[12]. We believe USE can be further applied in developing ab initio prediction programs for a larger class of ncRNAs.

1. INTRODUCTION

Non-coding RNAs (ncRNAs) carry out many critical functions in living cells[10]. As more functional roles by ncRNAs are being discovered, there is a rapid growth of interest in developing bioinformatics methods that may effectively identify ncRNAs genes from genomic backgrounds. Unlike their protein coding counterparts, ncRNAs do not possess strong statistical signals (e.g., ORFs), making ncRNA identification a computational challenge. For example, programs based on relevant sequential features, such as base composition, are often limited to certain classes of organisms or specific families of ncRNAs[24]. On the other hand, most transcribed single-strand ncRNAs can potentially fold; their secondary structure remains to be the most exploitable feature for a truly successful ncRNA prediction methods. Indeed, such a potential has energized the use of some best secondary structure prediction methods for ncRNA finding[20,22,27]. Nonetheless, the prediction results have generated a

* Corresponding author.

rather unclear picture; in particular, RNAz[26] and EvoFold1[20] generated 30,000 and 48,000 predicted candidates, respectively, on the human and other vertebrate genomes; the overlap of these two sets of candidates is somewhat disappointingly small (7.2%)[27] indicating questionable low sensitivity (for at least one of the programs). In addition, the validation rates for the predicted structural RNAs are low, possibly attributed to high false-positive rates because of the rare low expression levels in known ncRNAs.

The underperformance of structure prediction based methods could be due to the fact that functional RNA may admit alternative structures and random sequences may be fold and be detected (e. g. 11 million hairpin loops were found within the human genome)[3]. This suggests that, in addition to the minimum free energy[5], other rules governing secondary structure folding may be needed for effective ncRNA finding. Fold stability may be one such characteristics as it would help understand the differentials between alternative folds of real ncRNAs and between folds of real ncRNAs and of random sequences. Based on the partition functions[19] that defines a thermodynamic energy ensemble of RNA secondary structures, the fold stability of a given RNA sequence can be measured with the Shannon's entropy[7,19, 23] over various random variables that define base pairings of the sequence. It turns out that some ncRNAs, typically precursor microRNAs (pre-miRNA), have entropy significantly lower than that of their randomly shuffled counterparts, while others do not[8,14]. Independent studies on others measures, such as average free energy[4], self-containment[16], compactness[18], and thermodynamic entropy[31], appears to confirm that precursor miRNAs possess much higher fold stability than other kinds of ncRNAs and such structure characteristics may be exploited to discriminate miRNAs from the genome background.

The secondary structure requirement of primary miRNA and precursor miRNAs RNase III drosha and dicer processing consists of a long stem loop might have contributed to the significantly low fold stability entropy[17, 29]. Since such a structural feature is shared by other functional RNAs such as snRNA, introns splicing hairpins loops and intrinsic termination hairpin loops[12,25], it is of acute interest to develop structure stability based methods that can effectively detect such ncRNAs from genomes, especially given the recent discoveries of important roles played by miRNAs[1]. In this paper, we present some preliminary results toward developing an ab initio ncRNA prediction framework based on structure stability of long hairpin stem loops. We propose a novel objective function called Unpaired Structural Entropy (USE), which captures the structural variability for a given sequence. The USE measure was found to be effective in distinguishing miRNAs from its genomic background as well as other ncRNAs. Through the USE objective function, we were able to create a single feature classifier to distinguish miRNAs with a sensitivity of 85% and specificity of 90%, an improvement upon all existing multi-feature miRNA classifiers including the previously investigated structural entropy. Finally, we included the USE along with existing RNA measurements to further improve the performance of an SVM classifier.

2. METHODS

Although generally the minimal free energy is chosen as the predicted structure for a given RNA sequence, many alternative structures also exist. The probability for each of these structures to occur can be calculated through the Boltzmann Partition Function[19] and thus it is possible to calculate the liklihood for base pairings between nucleotides. In this work, we introduce a novel method USE which measures the structure's stability through computing the entropy of the non-pairing probabilities.

2.1. Structural Variation

Here we consider structural variation to be the amount of potential possibilities into which a sequence may fold. Higher structural variation implies a higher number of potential foldings. Figure 1 displays two different NUPACK[7] RNA folds, Figure 1a shows a predicted folding for mir-32 while figure 1b shows the folding for a dinucleotide shuffled mir-32. Although both of them possess a folded structure, the coloring scheme shows more confidence in the pairings for mir-32 than the shuffled sequence.

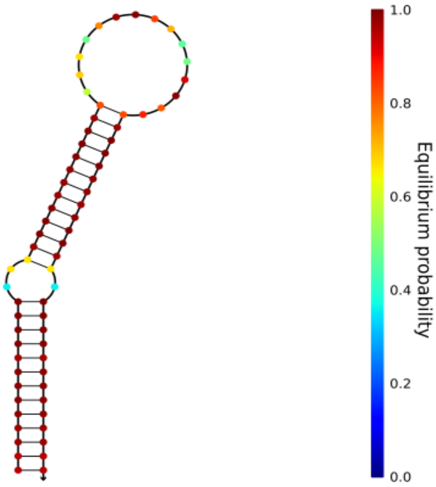


Fig. 1a Structures predicted by NUPACK for pre-mir-32

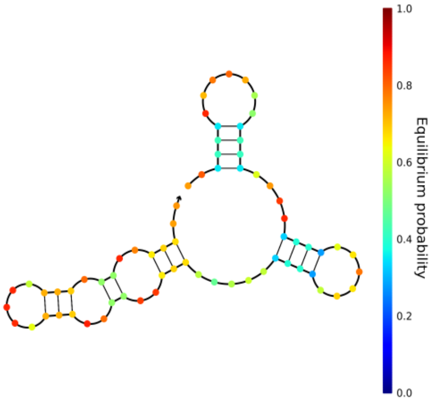


Fig. 1b Structures predicted by NUPACK for dinucleotide shuffling of pre-mir-32

2.2. Base Pairing Probabilities

RNA can exist in an ensemble of structures, and the distribution of these structures can be captured by a Boltzmann distribution. The Boltzmann distribution can allow computation of the partition function (Z) for each substructure. The Partition Function algorithm has been implemented by McCaskill[19] and it calculates the base-pairing probability distribution based on the free energies for each structure within the structural ensemble space Ω. Let s_α be a structure with free energy G_α. Assume that the molar gas constant R=8.31451Jmol^{-1}K^{-1}, and the temperature T, then

$$Z = \sum_{s_\alpha \in \Omega .} e^{-G_\alpha / RT} . \qquad (1)$$

The probability for each structure to occur is the following

$$p(s_\alpha) = \frac{1}{Z} e^{-G_\alpha / RT} \qquad (2)$$

The term $\delta^{\alpha}{}_{ij}$ denotes the occurrence of pairing between nucleotides i and j in s_α. Hence, the probability base pairing probabilityP_{ij} is as follows:

$$p_{ij} = \sum_{s_\alpha \in \Omega} p(s_\alpha)\,\delta^{\alpha}_{ij} \qquad (3)$$

Where p_{i0} corresponds to the non-pairing probability of nucleotide at position i:

$$p_{i0} = 1 - \sum_{j=1}^{N} p_{ij} \qquad (4)$$

N being the length of the sequence.

2.3. Unpaired Structural Entropy

Shannon Entropy is one of the most fundamental and basic concepts in the field of Information theory; it measures the amount of uncertainty of values taken by a random variable. It also measures the amount of diversity that exists within a set of quantities. Here in this work, we propose Unpaired Structural Entropy (USE) which computes the entropy of the non-pairing probabilities of the nucleotides that are normalized across the sequence:

$$USE(S) = \sum_{i=1}^{N} \frac{p_{i0}}{L_0} \log \frac{L_0}{p_{i0}}, \quad where\ L_0 = \sum_{i=1}^{N} p_{i0} . \qquad (6)$$

Previous attempts have been made to capture the structural variability of a sequence through the entropy of its base pairing probabilities. Huynen et al.[14] have defined the positional entropy (Q) as follows which has been traditionally used in previous research[8]:

$$Q(S) = \frac{1}{N} \sum_{i=1}^{N} E(n_i), \quad where\ E(n_i) = \sum_{j=0}^{N} p_{ij} \log \frac{1}{p_{ij}} . (7)$$

The relationship between Q and USE is as follows:

$$Q(S) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{N} p_{ij}\log\frac{1}{p_{ij}} = \frac{1}{N}\sum_{i=1}^{N} p_{i0}\log\frac{1}{p_{i0}} + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} p_{ij}\log\frac{1}{p_{ij}}$$

$$= \frac{L_0}{N}USE(S) + \log(L_0) + \frac{2L_1}{N}\varphi(S) + \log(L_1)$$

$$where\ \varphi(S) = \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{p_{ij}}{L_1}\log\frac{p_{ij}}{L_1}\log\frac{L_1}{p_{ij}}, L_1 = \sum_{i=1}^{N}\sum_{j>i}^{N} p_{ij}$$

(8)

In Q(S), the entropy of pairing possibilities is calculated for each nucleotide. These individual nucleotide entropies are then averaged across the sequence, making it a local computation of the variability. To calculate USE, only the non-pairing probabilities of nucleotides are taken into consideration. These probabilities are normalized across the sequence before their entropy is calculated; therefore, unlike Q(S), USE(S) is a global computation of the structural entropy. In this study, we will show that USE is more successful than Q as well as other structural features in capturing the stability of long stems structures, in particular pre-miRNAs.

3. RESULTS

A number of tests and analyses were conducted to examine the capabilities of the USE feature in identifying structures with low variability. (1) The USE score was studied and compared across different RNA families. The USE score distribution for miRNAs were significantly lower than other RNA families. (2) We then examined the USE score's ability to distinguish miRNAs from their genomic background. (3) Finally, the performance of the USE score as a classifying feature in distinguishing miRNA from pseudo-miRNA was assessed. The USE score was shown to be highly effective in classifying miRNA from pseudo-miRNA.

3.1. USE Across RNA Families

To investigate the utility of USE, we compared the entropy across different RNA families using the same sequences evaluated by Freyhult et al.[8] USE quartile boxplot of different ncRNA families can be viewed from Figure 2a. The USE score for miRNA was generally much lower than of the other RNA families. The intron sequences possessed the next lowest entropy values, but the graph showed that roughly around 0.80 we have 75% of the miRNA lower than this value and 75% of the intron sequences higher than this value. The RNA family with the third lowest entropy value was the SRP, and the majority of the SRP sequences were above 0.80. In addition from Figure 2b, we included the Q calculation across different RNA families. This feature also showed that miRNA tends to be lower than of the other RNA families; however, if we compared the two graphs, Q as an objective function cannot distinguish miRNA as well from the other families of ncRNA than that of the USE score.

3.2. Detecting pre-microRNAs

We decided to explore the ability of the USE function to identify miRNA precursors by sliding a window across a sequence, then calculating a score within the window. The lengths of known precursor microRNAs (pre-miRNA) in humans usually range from 70 to 100nt. Therefore, we evaluated the behavior of the USE score across different window-sizes and sequences surrounding the precursor miRNA. Figure 3 shows the USE Score of Sliding Window Scan of 500nt upstream and downstream of a human miRNA (mir-30e) the actual window size was the same as the pre-miRNA. The graph indicated that the USE score has a

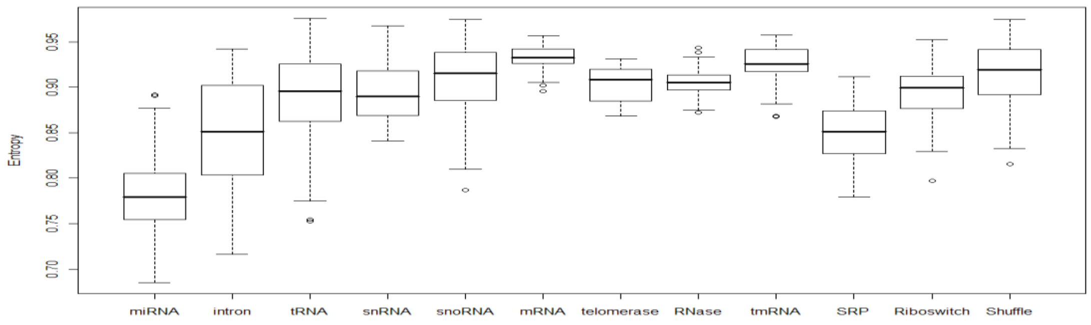


Fig. 2a. A box plot of USE. Box and whisker plots displaying distribution of USE score through quartiles across various ncRNAs. From the graph, the low entropic feature calculation separates miRNA from the other ncRNAs.

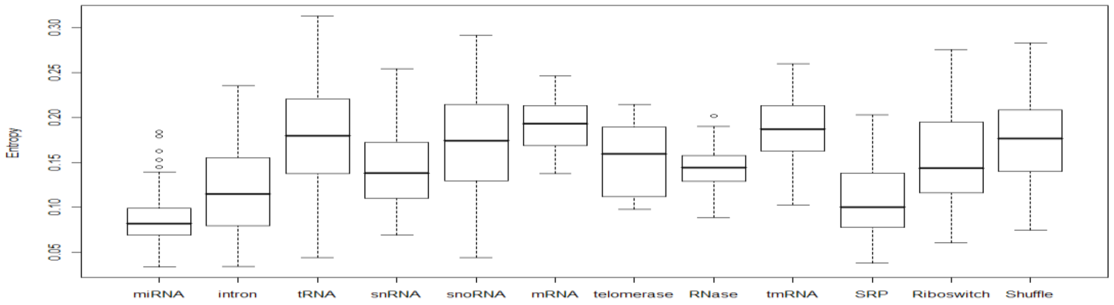


Fig. 2b. A box plot showing the Q distribution across various ncRNA. We also observe that miRNA tends to have a lower entropy; however compared to USE, Q has a harder time distinguishing miRNAs from the other families of ncRNAs.

distinguishing low entropy for the true miRNA in its real genome context.

Since the length of the pre-miRNA sequence was not always known, we performed sliding window scans of different length for each sequence containing the upstream and downstream of 721 pre-miRNAs. To observe the behavior or USE on surrounding sequences of miRNA, we varied the window size by increasing and decreasing in increments of 5 nt and repeated the same process for all pre-miRNA sequences. Figure 4 presented the results of USE scores corresponding to different window lengths and positions, which were averaged across the 721 pre-miRNAs. We showed that for any length of window scan, the lowest average USE values always occurs at the position 0 which correspond to the exact location of the pre-miRNA within the genome.

3.3. USE Correlation with Other RNA measures on Human pre-miRNA

We performed regression on six variables to evaluate the linear relationship between USE across various RNA measurements: Q, miR-CYK, Self Containment (SC), Length, Minimal Free Energy (MFE), and Structural Ensemble as shown in figure 5.

We used the Cocke-Younger Kasami(CYK) algorithm to develop an in house CYK program to perform microRNA gene finding called miR-CYK. CYK in general was used to find the maximum probability alignment of the CFG to the string. Therefore, by defining a Stochastic Context Free Grammar (SCFG) based on the human pre-miRNA structure feature, we used the miR-CYK to score the sequence based on how the predicted likeness of the miRNA structure.

Self Containment (SC) was shown to measure the tendency to retain their structure regardless of the neighboring upstream and downstream sequences. This particular measurement was developed by Kim et al[16]. They took the query sequence and added additional sequences upstream and downstream, and used RNAFold to fold the sequence to examine the structure prior and after the additional sequences. This was done repeatedly to obtain a statistic of the frequency for the structure to retain its shape.

MFE and Ensemble Frequency calculation was based on RNAFold's calculation[11]. Previously Bonnet et al[4]. showed that miRNA compared to other ncRNA tends to have lower MFE, and this might be attributed to the stability of the folding. Ensemble Frequency provided a score of the frequency of the specific structure to occur within the structural space.

Since we were interested in possibly using USE to perform pre-miRNA gene finding, we used the Human pre-miRNA as the sequence to make the comparison. Table 1 presented the correlation coefficient for these six measures. USE was shown to be most closely related to Q, since USE and Q were inherently based on the same type of idea. MiR-CYK and SC both possessed decent correlation to USE while Length, MFE, and Ensemble Frequency possessed the weakest correlation with USE.

Table 1. Correlation Coefficient R^2 between USE and other RNA Measures

Measure	Correlation with **USE**
Q	0.638
SC	0.460
miR-CYK Score	0.514
MFE	0.068
Length	0.164
Frequency in Ensemble	0.235

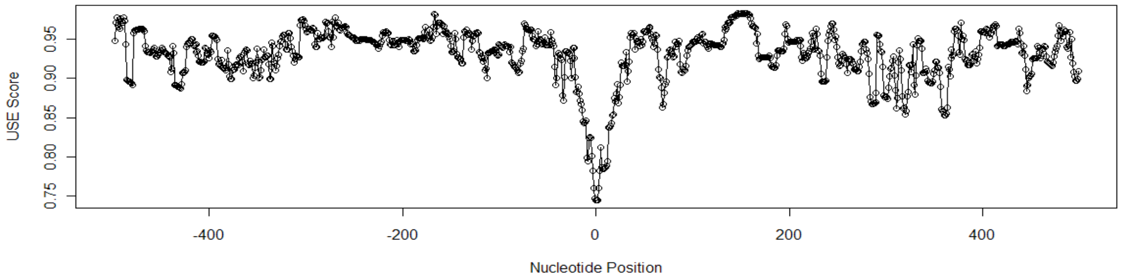
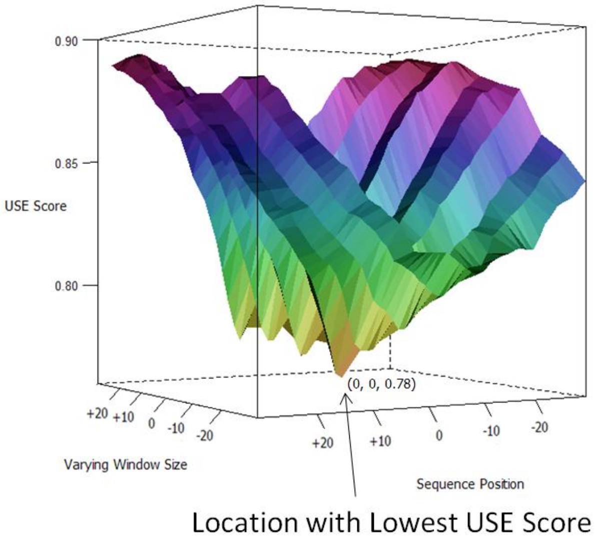


Fig. 3. Sliding window scan of USE score across 500nt upstream and downstream of has-mir-30e with window size 93nt (the length of the mir-30e sequence).



◀ **Fig. 4.** Average USE value of all miRNAs: Any point on the graph corresponded to a specific window length and position and represents the USE score averaged across all 721 Human microRNA sequences. The labels on the sequence-position axis represented the relative upstream/downstream position from the location of the actual microRNA. The window-size axis represented the amount of increments/decrements of the window-length relative to the length of the actual miRNA. The lowest averaged USE values were aligned in position 0 for all window lengths.

3.4. Classifying Human pre-miRNA via USE

We also evaluated the performance of the USE function as a classifier. In order to do so, we used 721 Human miRNAs as the positive datasets and 8494 pseudo miRNA as the negative datasets. The Human miRNA sequences were downloaded from miRbase[9]. The negative set of pseudo-miRNAs, was obtained from Xue et al[28] derived from a set of sequences with hairpin loops located within the coding region.

The ROC for various RNA objective function are plotted in figure 6. We also included a Paired Structural Entropy (PSE). The probability of each nucleotide to pair can be defined as (1 – Pi0), and PSE was the same entropy calculation as USE but calculated only on the paired nucleotide region. Figure 6 showed that the entropy of the non-pairing probability has more classification power than that of the pairing probability. A true positive was defined as a Human microRNA that was below the cutoff, and a false positive was defined as a Human microRNA that is above the cutoff. A true negative was defined as a pseudo microRNA that was above the cutoff, and a false negative was defined as a pseudo microRNA that was below the cutoff. The graph showed that USE, CYK, and SC’s performance was relatively similar.

3.5. Comparisons Across different microRNA Classifiers

Most existing miRNA prediction programs rely on a machine learning algorithm trained with a variety of features in primary and secondary structure for classification. To assess the power of USE Classifer, we compared its performance to four other SVM microRNA classifiers: TripletSVM[28], Virgo[15] miRFinder[13], and microPred[2]. Triplet-SVM, an ab initio algorithm uses the local contiguous base-pairing structures as features for the SVM classification[28]. Virgo, a viral miRNA detector that was trained on human miRNA sequences[15]. miRFinder used the pre-miRNA structural characteristic and structural mutation information for the classification[13]. MicroPred attempted to improve the prediction through effective machine learning techniques[2]. To allow a valid fair comparison, we used the 8494 pseudo-microRNA as the negative dataset. This particular negative dataset was used in all of the programs. Using a larger negative dataset satisfied a requirement of the miRNA gene finder that it not produce many false positives, one of the primary difficulties in miRNA detection. To evaluate each program's performance, we chose to use Sensitivity, Specificity, and the Mathews Correlation Coefficient which were calculated:

Specificity = TN / (TN + FP)
Sensitivity = TP / (TP + FN)

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Table 2 contained the sensitivities, specificities, and MCC for the above classifiers. Two different cutoffs were chosen for the USE model, to demonstrate the diversity of the sensitivity and specificity over a range of cutoff values. The USE model cutoff 1 was a rough estimation based on the comparison over the distribution of USE score between miRNA and different RNA families (See Figure 2a). For cutoff 2, the threshold was more stringent than cutoff 1 in attempt to reduce the number of false positives. The SVM USE Model corresponded to an SVM model that integrated USE score, SC, and CYK, and we trained the SVM using cross validation. We saw that the MCC for such a classifier was significantly higher than other classifiers.

Table 2. Performance comparison across different miRNA gene finding models.

	Sensitivity	Specificity	MCC
SVM with USE Model	**0.777**	**0.974**	**0.724**
USE Model (Cutoff 1)	0.845	0.906	0.560
USE Model (Cutoff 2)	0.760	0.950	0.620
miRFinder	0.809	0.906	0.538
Virgo	0.823	0.712	0.306
triplet-SVM	0.739	0.914	0.510
microPred	0.908	0.733	0.363

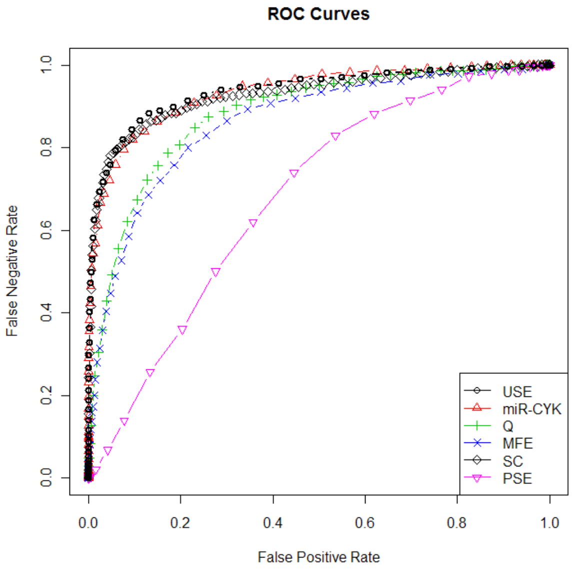


Fig. 6. ROC plot of for prediction of classifying miRNA through various RNA measurements USE, CYK, Q, MFE, and SC.

4. DISCUSSION AND CONCLUSION

In this work, a novel objective function called **USE** is presented that utilizes the variability based on unpaired probabilities of nucleotides. USE can be interpreted as a measurment of uncertainty for a nucleotide in a structure to be unpaired (i. e. bulge or loop). MicroRNAs generally possess a stereotypical long-stem structure making them relatively less flexible and more stable than other ncRNAs. Nucleotide sequence variations within microRNA was shown not to affect the drosha and dicer processing[6]. In fact, the most important secondary structure determinants for miRNA were found to be greater than 16bp stem, lower number and reduced size of bulges and internal loops.[21, 29]. From this we can infer that structures with long stem, fewer bulges and shorter loop tend to be more stable. Figure 2 indicates that the USE score can be an acceptable criterion in distinguishing miRNA structures from other ncRNA families. If we compare our miRNA USE score distribution to Q, USE can distinguish more miRNAs than Q, demonstrating the novelty and statistical power of the USE function.

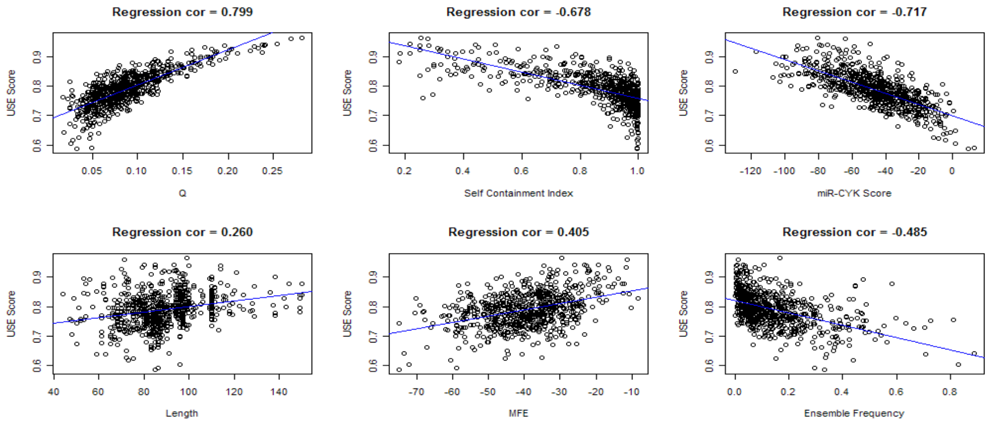


Fig. 5. Correlation plot of the USE score compared to Q, Self Containment Index, miR-CYK, Length, MFE, and Ensemble Frequency.

Although RNA families are characterized by their sequence and structure, this does not imply that size is preserved across the RNA family. This is especially observable within pre-miRNAs of different length while having similar hairpin loop structures. Therefore, it is important for a RNA-family detector to be robust against assumptions on the sequence length. Here, the window scan of USE calculation is done for different window lengths in order to evaluate the USE window scan's performance. Figure 3 shows that the USE score is less dependent of the window size (i. e. the length of the microRNA to be found), since the USE score is always the lowest at the miRNA position, regardless of the length of the sliding window.

Furthermore, Lee and Kim[16] have demonstrated that miRNAs have a tendency to retain their structure regardless of the neighboring upstream and downstream sequences. Their finding indicates that the sequence containing the miRNA and additional upstream and downstream sequence has as relatively low structural variability as the original miRNA, and they have termed this phenomenon as self-containment (SC). The USE scores of microRNAs are also observed to have a similar behavior, since they tend to stay relatively low even when upstream and/or downsream sequences are added to the actual microRNA. This suggests a high correlation between the USE Score and the SC index which can also be observed from figure 5.

Capturing structural features as well as other RNA measurements has always played a significant role in classifying different RNA sequences. The challenge is to select the features that are specific to a category of ncRNA. The power of such features can be assessed through various machine learning techniques. As we have discussed earlier, the low structural variability of miRNAs distinguishes them from other families of ncRNAs as well as from their background. Table 2 is a comparison of the different miRNA classifiers, and our single feature USE classifier's sensitivity, specificity, and MCC outperforms all existing SVM methods. The two cutoffs of the USE classifier demonstrate that a higher specificity can be achieved without sacrificing the sensitivity. Finally, the inclusion of USE with existing features results in an even better performance with a higher MCC value suggesting that a lot of information is contained in the USE structural feature. For our SVM method we purposely chose a high cutoff to have a stringent specificity, since this is the major difficulty in computational detection of miRNA.

In conclusion, microRNA molecules possess low structural variability compared to other families; USE successfully captures this low global variability, offering a substantial improvement to the current state of miRNA gene finding. Since USE is able to better quantify the structural variability of a sequence of long stems and small bulges, we believe USE can be further applied to develop ab initio prediction programs for a larger class of ncRNAs, or be applied to study stem-loop structures within viral sequences. A limitation of our study is its dependency on the NUPACK's secondary structural model. Looking to future applications, there is potential for the USE feature to be applied in the prediction and validation of various tertiary models by quantifying the structural variability of a sequence or be applied to the identification of miRNA gene targets. The USE method seems to work well on the long stem loop of pre-miRNAs but not on long stem loops of random sequences nor that of some other ncRNAs like snoRNAs. This could indicate some more intrinsic nature of ncRNAs that have yet to be discovered. Such a phenomenon offers an opportunity for future investigation on techniques for detecting other ncRNAs.

5. AVAILABILITY

A webserver inferface for computing USE is available at http://www.uga.edu/RNA-Informatics/?f=software&p=StructuralEntropy

6. ACKNOWLEDGMENTS

We acknowledge, and are grateful for, funding by NIH BISTI R01GM072080-01A1grant and NIH ARRA Administrative Supplement to this grant. We thank Bram Sebastian for suggesting the negative training dataset, Ying-Wai Li's valuable discussion on the Entropy calculation, Joseph Robertson's revision of the paper, and Eric Talevich's comments on the paper.

7. REFERENCES

1. V. Ambros, "The functions of animal microRNAs," Nature, vol. 431, pp. 350-355, Sep 2004.
2. R. Batuwita and V. Palade, "microPred: effective classification of pre-miRNAs for human miRNA prediction," Bioinformatics, vol. 25, pp. 989-995, 2009.
3. I. Bentwich, et al., "Identification of hundreds of conserved and nonconserved human microRNAs," Nature Genetics, vol. 37, pp. 766-770, Jul 2005.
4. E. Bonnet, et al., "Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences," Bioinformatics, vol. 20, pp. 2911-2917, 2004
5. P. Clote, et al., "Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency," Rna-a Publication of the Rna Society, vol. 11, pp. 578-591, 2005.
6. S. Diederichs and D. A. Haber, "Sequence variations of microRNAs in human cancer: Alterations in predicted secondary structure do not affect processing," Cancer Research, vol. 66, pp. 6097-6104, 2006.
7. R. M. Dirks, et al., "Thermodynamic analysis of interacting nucleic acid strands," Siam Review, vol. 49, pp. 65-88, 2007.
8. E. Freyhult, et al., "A comparison of RNA folding measures," Bmc Bioinformatics, vol. 6, 2005.
9. S. Griffiths-Jones, "MiRBase: The MicroRNA sequence database," Methods in Molecular Biology, pp. 129-138, 2006.
10. S. Griffiths-Jones, "Annotating noncoding RNA genes," Annual Review of Genomics and Human Genetics, vol. 8, pp. 279-298, 2007.
11. A. R. Gruber, et al., "The Vienna RNA Websuite," Nucleic Acids Research, vol. 36, pp. W70-W74, 2008.
12. I. Gusarov and E. Nudler, "The mechanism of intrinsic transcription termination," Molecular Cell, vol. 3, pp. 495-504, 1999.
13. T. H. Huang, et al., "MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans," Bmc Bioinformatics, vol. 8, 2007.
14. M. Huynen, et al., "Assessing the reliability of RNA folding using statistical mechanics," Journal of Molecular Biology, vol. 267, pp. 1104-1112, 1997.
15. S. Kumar, et al., "Prediction of viral microRNA precursors based on human microRNA precursor sequence and structural features," Virology Journal, vol. 6, 2009.
16. M. T. Lee and J. Kim, "Self Containment, a Property of Modular RNA Structures, Distinguishes microRNAs," PLoS Computational Biology, vol. 4, 2008.
17. Y. Lee, et al., "The nuclear RNase III Drosha initiates microRNA processing," Nature, vol. 425, pp. 415-419, 2003.
18. S. N. K. Loong and S. K. Mishra, "Unique folding of precursor microRNAs: Quantitative evidence and implications for de novo identification," Rna-a Publication of the Rna Society, vol. 13, pp. 170-187, 2007.
19. J. S. McCaskill, "The Equilibrium Partition-Function and Base Pair Binding Probabilities for RNA Secondary Structure," Biopolymers, vol. 29, pp. 1105-1119, 1990.
20. J. S. Pedersen, et al., "Identification and classification of conserved RNA secondary structures in the human genome," PLoS Computational Biology, vol. 2, pp. 251-262, 2006.
21. W. Ritchie, et al., "RNA stem-loops: To be or not to be cleaved by RNAse III," Rna-a Publication of the Rna Society, vol. 13, pp. 457-462, 2007.
22. E. Rivas and S. R. Eddy, "Noncoding RNA gene detection using comparative sequence analysis," Bmc Bioinformatics, vol. 2, pp. 1-19, 2001.
23. C. E. Shannon, "The mathematical theory of communication," M D Computing, vol. 14, pp. 306-317, 1997.
24. Schattner, P. "Computational gene-finding for noncoding RNAs", (2003) in Noncoding RNAs: Molecular Biology and Molecular Medicine, ed. Barciszewski and Erdmann, Kluwer Academic/Plenum Publishers.

25. N. A. Smith, et al., "Gene expression - Total silencing by intron-spliced hairpin RNAs," Nature, vol. 407, pp. 319-320, Sep 2000.
26. S. Washietl, et al., "Fast and reliable prediction of noncoding RNAs," Proceedings of the National Academy of Sciences of the United States of America, vol. 102, pp. 2454-2459, Feb 2005.
27. S. Washietl, et al., "Structured RNAs in the ENCODE selected regions of the human genome," Genome Research, vol. 17, pp. 852-864, 2007.
28. C. H. Xue, et al., "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine," Bmc Bioinformatics, vol. 6, 2005.
29. Y. Zeng and B. R. Cullen, "Structural requirements for pre-microRNA binding and nuclear export by Exportin 5," Nucleic Acids Research, vol. 32, pp. 4776-4785, 2004.
30. Y. Zeng and B. R. Cullen, "Recognition and cleavage of primary MicroRNA transcripts," Methods in Molecular Biology, pp. 49-56, 2006.
31. M. Zuker. The entropy of the Boltzman distribution of RNA folding, Benasque Workshop on Computational Methods for RNA Analysis, Benasque, Spain. 2009.