ANALYSIS OF STRUCTURAL STRAND ASYMMETRY IN NON-CODING RNAS

JIAYU WEN [*†] and GEORG F. WEILLER

Australian Research Council (ARC) Centre of Excellence for Integrative Legume Research and Bioinformatics Laboratory, Research School of Biological Sciences, Australian National University,
Canberra, ACT 0200, Australia
[†]*E-mail: jiayu.wen@anu.edu.au*
www.anu.edu.au

BRIAN J. PARKER [*]

Life Sciences and Statistical Machine Learning Groups, NICTA, University of Melbourne,
Melbourne, VIC 3010, Australia
E-mail: brian.bj.parker@gmail.com

Many RNA functions are determined by their specific secondary and tertiary structures. These structures are folded by the canonical G::C and A::U base pairings as well as by the non-canonical G::U complementary bases. G::U base pairings in RNA secondary structures may induce structural asymmetries between the transcribed and non-transcribed strands in their corresponding DNA sequences. This is likely so because the corresponding C::A nucleotides of the complementary strand do not pair. As a consequence, the secondary structures that form from a genomic sequence depend on the strand transcribed. We explore this idea to investigate the size and significance of both global and local secondary structure formation differentials in several non-coding RNA families and mRNAs. We show that both thermodynamic stability of global RNA structures in the transcribed strand and RNA structure strand asymmetry are statistically stronger than that in randomized versions preserving the same di-nucleotide base composition and length, and is especially pronounced in microRNA precursors. We further show that a measure of local structural strand asymmetry within a fixed window size, as could be used in detecting and characterizing transcribed regions in a full genome scan, can be used to predict the transcribed strand across ncRNA families.

Keywords: ncRNA; non-coding RNA; structural strand asymmetry; RNA secondary structure

1. Introduction

A variety of functional non-coding RNAs (ncRNAs) have been shown to play key regulatory roles in a number of cellular processes including chromatin modification, transcription initiation, mRNA and protein synthesis, as well as post-translational

[*] Authors contributed equally to this work

RNA modification.[1,2] MicroRNAs (miRNAs) are a class of small ncRNAs that are known to play important roles in gene regulatory networks by influencing the expression of other genes. Systematic identification of ncRNAs is important for understanding complex gene regulatory networks. However, *de novo* ncRNA prediction is a challenge for current bioinformatics due to a lack of statistically reliable characteristics in ncRNA sequences. Glusman et al.[3] has discussed a "third approach" to the problem of noncoding gene detection (in addition to the other two more usual approaches based on sequence similarity and recognition of protein-coding gene features), which involves the detection of transcribed regions by detecting evolutionary signals in the transcribed strand, including base compositional asymmetries. In a similar vein, in this paper we investigate an asymmetry in the structure forming potential between the transcribed and non-transcribed strands of a genomic sequence.

It is widely assumed that the function of a ncRNA depends on its structural features. Previous research has addressed the prospect of stability of RNA secondary structures acting as a statistical signal for RNA identification. Current opinion differs as to whether ncRNAs and/or mRNA can be recognized by their secondary structures. It has been proposed that mRNA secondary structure stability as measured by the predicted minimum free energies (MFE) is more stable than that of randomized sequences with the same base composition.[4] This hypothesis has been questioned by Workman and Krogh[5] who provide evidence that the observed stability signals disappear when sequences are shuffled so as to preserve di-nucleotide frequencies. Also, Rivas and Eddy[6] argue that ncRNA secondary structures are similar to random sequences in their stability, especially while taking local base composition effects into account, and therefore are not useful in ncRNA detection. However, Washietl et al[7] has combined thermodynamic stability and RNA structure conservation to recognize some ncRNAs. In particular, miRNA precursors have been shown to have lower MFEs than is expected by chance.[8]

In our previous work on legume ncRNA transcripts,[9] we introduced the structural strand asymmetry feature for characterizing transcribed regions. In addition to the canonical complementary bases G::C and A::U, RNA secondary structures typically include non-canonical G::U base pairs. The corresponding C::A nucleotides of the complementary strand do not pair. As a consequence, the secondary structures formed depend on the strand transcribed. Sequences that have evolved functional RNA structures should have done so predominantly on the transcribed strand. We thus hypothesize that the differential in potential secondary structures between the two complementary strands may be used as a measure of RNA structure evolution. We showed[9] that local structural strand differential is pronounced in RNAs compared to that of non-transcribed sequences. We further showed that base compositional asymmetries also contribute to distinguishing the transcribed RNA sequences from non-transcribed DNA sequences.

Here we extend our investigation of the RNA strand structural asymmetry feature by analyzing this feature in sequence sets of known ncRNAs, including miRNAs, and mRNAs. We aim to identify whether this is an independent feature, in addition to structural asymmetries induced by base compositional asymmetries alone. Also, we investigate the effectiveness of local and global measures of structural asymmetry features. Given that the stability of RNA structure depends upon di-nucleotide base stacking energies, we compared them with sequences that were shuffled so as to preserve both mono- and di-nucleotide frequencies.

2. Methods

2.1. *Datasets*

All ncRNA sequences used in this study were obtained from the Rfam database (release 8.0).[10] We retrieved sequences from Rfam.fasta.gz file which filtered Rfam members to $< 90\%$ identity, including several ncRNA families: miRNA precursor, 5S rRNA, 5.8S rRNA, 7SK, Hammerhead ribozyme (type I and type III), Group I and Group II catalytic intron, IRES, RNase MRP, Nuclear RNase P, snoRNA CD-box, snoRNA HACA-box, Eukaryotic type signal recognition particle RNA (SRP), tmRNA, and tRNA. We randomly selected 100 sequences from the miRNA family and up to 50 sequences from each other RNA family; all sequences were retrieved if ncRNA family had less than 50 sequences. To select representative mRNA sequences, we selected proteins in the Swiss-Prot[11] database that were derived from a number of commonly studied organisms including *Arabidopsis thaliana, Bos Taurus, Caenorhabditis elegans, Danio rerio, Drosophila melanogaster, Escherichia coli, Homo sapiens, Mus musculus, Mycoplasma pneumoniae, Oryza sativa, Rattus norvegicus, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Takifugu rubripes, Xenopus laevis,* and *Zea mays.* cDNAs for these proteins were then extracted from EMBL.[12] We randomly sampled 800 sequences and then excluded those with sequence length of more than 1000 bp (to limit computation time) to construct a set of 614 mRNA sequences. Table 1 summarizes the sequence length and GC content of the ncRNA family and mRNA sets used in this study.

2.2. *Sequence randomization*

Each sequence was permutated 100 times to generate 100 shuffled sequences that retained both mono-nucleotide and di-nucleotide base composition and length. Mononucleotide shuffling preserves the same nucleotide frequencies as the original sequences, whereas di-nucleotide shuffling preserves both mono- and di-nucleotide frequencies. We used the EMBOSS program shuffleseq[13] to perform mono-nucleotide shuffling. Di-nucleotide shuffling is particularly important because the stability of RNA secondary structures depends upon the stacking energies.[5] We used the program dishuffleseq.pl which implemented Altschul and Erickson shuffling algorithm[14] to perform di-nucleotide shuffling. See Altschul and Erickson for details of the careful considerations needed for di-nucleotide randomization.

2.3. *RNA secondary structures*

To compute a measure of global structure strand asymmetry, we used RNAfold in the Vienna RNA package (version 1.6.4)[15] to compute RNA global secondary structures separately for the transcribed and complementary non-transcribed strands for each RNA sequence. MFE was measured for both the transcribed and non-transcribed global structures. The values were normalized by the length of the sequences yielding the MFE densities (MFED). The MFED difference between the two strands was taken as $\Delta\text{MFED}_{tr-ntr}$. This measure was calculated on the original RNA sequences and on both mono-nucleotide and di-nucleotide shuffled sequences.

In gene-finding applications, these asymmetry features would typically be applied to local structures computed in a sliding window along the genome. For this reason, it is of interest to examine a measure of local structure strand asymmetry. To compute this, we used RNALfold (in Vienna RNA package) which can compute the MFE of locally stable secondary structures in an RNA sequence, limited to a maximum sliding window size. We used a sliding window size L = 150 (as may be typical for a genome-wide study) to predict a list of RNA local secondary structures for each sequence. The mean MFED over the predicted local structures for each strand was computed, and the difference, $\Delta\text{MFED}_{tr-ntr}$, was taken as the measure of local structure asymmetry.

2.4. *Statistical significance*

To provide evidence for whether any structural strand asymmetry ($\Delta\text{MFED}_{tr-ntr}$) is primarily due to RNA secondary structure conservation, or is primarily a side-effect of base compositional asymmetries, we performed a permutation test using the shuffled versions of the sequences from ncRNA families and mRNAs. This permutation test measures the strength of the structural asymmetry in the original sequence as compared with the structural asymmetry of shuffled versions of the sequence (which maintains base compositional asymmetries). Any overall structure is effectively removed in these shuffled sequences and any remaining structural asymmetry in these shuffled sequences would be mainly due to mono- and di-nuceotide frequency asymmetries between strands. To calculate a p-value for the structural asymmetry, the fraction of the shuffled sequences that achieved a $\Delta\text{MFED}_{tr-ntr}$ as great as the original version was estimated. Z-scores[6] were also calculated as a measure of the structure signal strength above random noise levels: Z-score $= (x-\mu)/\sigma$, where x is the MFED_{tr} values from the original sequences, μ is the mean value of the randomized sequences, and σ is the standard deviation.

An advantage of the structural asymmetry feature in gene-finding applications is that it also inherently provides information on the transcribed strand which is required for annotation of detected transcribed regions. Indeed, global $\Delta\text{MFED}_{tr-ntr}$ has recently been applied successfully for strand orientation detection in ncRNA multiple sequence alignments.[16] Our focus is on characterizing the regions of transcription across the genome and so the accuracy of both global and local strand asymmetry features for transcribed strand detection are computed as an additional measure of the strength of the strand asymmetry features. The accuracy in predicting the transcribed strand was computed using the sign of the differential, where accuracy is defined as the proportion of correct predictions/number of samples.

3. Results and Discussion

We used strand differences in MFEDs ($\Delta\text{MFED}_{tr-ntr}$) to compare the most stable global structures that could be formed by sequences of ncRNA families and mRNAs (See Methods). This measure should not be substantially affected by sequence regularities that equally affect both strands such as overall shifts in the absolute MFE across RNA families. Table 1 gives the MFED on the transcribed strand (MFED_{tr}) and the global structural strand differentials ($\Delta\text{MFED}_{tr-ntr}$) in each ncRNA family and mRNAs. On average, $\Delta\text{MFED}_{tr-ntr}$ shows large negative mean values of -0.0907 kcal/mol·bp for the miRNA set and -0.0379 kcal/mol·bp for the total ncRNA set, compared with a small (and in fact positive) mean value of 0.00365 kcal/mol·bp for the mRNA set. For miRNAs this corresponds to a very substantial 20.8% decrease in MFED from the transcribed strands; for ncRNAs this corresponds a 11.2% decrease; and for mRNAs a 1.2% increase. The preferential use of the transcribed strand for structure formation is substantially higher in miRNAs and ncRNAs than in mRNAs, which suggests a stronger signal for structure evolution in some ncRNA families.

Table 1. RNA secondary structure asymmetry in each ncRNA family and mRNA dataset

RNA type	No.of seqs	length (bp) (mean±sd)	(G+C)%	$\Delta\text{MFED}_{tr-ntr}$ (kcal/mol·bp) mean±sd	MFED_{tr} (kcal/mol·bp) mean±sd	$\frac{\Delta\text{MFED}_{tr-ntr}}{\text{MFED}_{tr}}$
mRNAs	614	499±239	47.9	0.00365 ± 0.042	-0.295 ± 0.062	-1.2%
all ncRNAs	753	191±126	47.7	-0.0379 ± 0.062	-0.337 ± 0.11	11.2%
miRNA	100	87±17	46.0	-0.0907 ± 0.073	-0.435 ± 0.067	20.8%
5.8S rRNA	50	152±11	50.4	-0.0165 ± 0.041	-0.288 ± 0.067	5.7%
5S rRNA	50	116±6	52.8	-0.0237 ± 0.054	-0.336 ± 0.061	7.1%
7SK	50	316±13	51.3	0.0182 ± 0.023	-0.308 ± 0.023	-5.9%
Hammerhead_1	26	48±9	51.8	-0.00459 ± 0.064	-0.244 ± 0.08	1.9%
Hammerhead_3	17	55±2	49.4	-0.00869 ± 0.045	-0.347 ± 0.048	2.5%
Intron_gpI	50	418±89	35.1	-0.0119 ± 0.026	-0.233 ± 0.045	5.1%
Intron_gpII	50	79±14	45.8	-0.116 ± 0.067	-0.341 ± 0.091	34%
IRES	50	286±117	54.1	-0.0250 ± 0.037	-0.359 ± 0.073	6.9%
RNase MRP	21	276±35	43.8	-0.0289 ± 0.036	-0.321 ± 0.061	9%
Nuclear RNase P	39	293±55	55.1	-0.0423 ± 0.043	-0.387 ± 0.052	11%
snoRNA CD-box	50	92±34	41.8	-0.0205 ± 0.049	-0.219 ± 0.08	9.3%
snoRNA HACA-box	50	135±27	45.8	-0.0422 ± 0.047	-0.287 ± 0.055	14.7%
SRP_euk_arch	50	293±13	47.9	-0.0651 ± 0.067	-0.511 ± 0.14	12.7%
tmRNA	50	359±42	47.8	-0.0256 ± 0.026	-0.329 ± 0.08	7.8%
tRNA	50	73±5	47.7	-0.0134 ± 0.053	-0.309 ± 0.097	4.3%

To investigate whether the structural strand differentials are more likely to be primarily due to actual RNA secondary structural signals, or whether they could be explained by base compositional biases alone, we shuffled the original ncRNA and mRNA sequences while maintaining either mono-nucleotide or di-nucleotide base compositions. We then computed $\Delta\text{MFED}_{tr-ntr}$ for the shuffled versions (See Methods). The shuffled version would be expected to show smaller differentials between the two complementary sequence strands if the structural strand asymmetry is primarily caused by RNA secondary structures; this smaller difference in the shuffled sequences being due primarily to base composition biases. Increased G and U content in the transcribed strand can itself lead to increased probability of G::U base pairings and hence structural asymmetries. We therefore compared the original structural strand differentials with the distributions generated by the shuffled versions. Figure 1 shows the distributions of the mean $\Delta\text{MFED}_{tr-ntr}$ values of the shuffled sequences (di-nucleotide shuffling) compared with the mean $\Delta\text{MFED}_{tr-ntr}$ values of the original miRNA, ncRNA and mRNA sets separately. The mean $\Delta\text{MFED}_{tr-ntr}$ values of the original miRNA and ncRNA sets are clearly significantly different from the shuffled distribution, whereas the mean value of the mRNA set is close to the shuffled distribution. The mono-nucleotide shuffled version showed results similar to the di-nucleotide shuffled version (data are not shown).

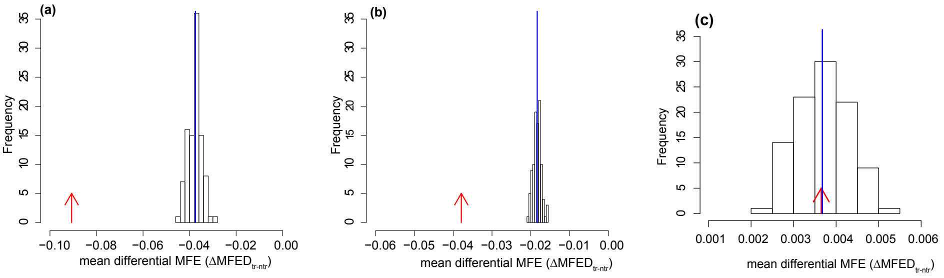

Fig. 1. The distribution of $\Delta\text{MFED}_{tr-ntr}$ values for di-nucleotide shuffled sequences compared to the mean $\Delta\text{MFED}_{tr-ntr}$ values for the original sequences for (a) miRNA set, (b) ncRNA set, and (c) mRNA set. The red arrow shows the mean $\Delta\text{MFED}_{tr-ntr}$ of the original miRNA, ncRNA, and mRNA sequences and blue line shows the mean values of the distribution obtained from the di-nucleotide shuffled sequences.

In particular (shown in table 2), the mean $\Delta\text{MFED}_{tr-ntr}$ values for the original miRNA sequences (-0.0907 kcal/mol·bp; 20.8% decrease from the transcribed strand) is substantially lower than the mean of the di-nucleotide shuffled version (-0.0375 kcal/mol·bp; 15.2% decrease from the transcribed strands). This strand asymmetry difference between the original and the di-shuffled version is statistically significant: over the 100 permutations, no sequence had a value as extreme as this, and so the differences are statistically significant at p-value < 0.01. This strand differential, to a lesser extent, is also shown by the total ncRNA set and its decrease in comparison with its shuffled version: -0.0379 to -0.0165 kcal/mol·bp (11.2% to 6.7%) on the original and di-nucleotide shuffled sequences respectively.

The stronger structural strand asymmetry signals in ncRNAs, especially pro-

Table 2. Analysis results of structural strand asymmetry features[a]

RNA type	original/ di-shuffle	$\Delta\text{MFED}_{tr-ntr}$ (mean±sd) (kcal/mol·bp)	$p-$value[b]	$\frac{\Delta\text{MFED}_{tr-ntr}}{\text{MFED}_{tr}}$	mean Z-score of MFED_{tr}	Accuracy for detecting transcribed strand		
						$\frac{f_G+f_U}{f_A+f_C}$ [c]	$\Delta\text{MFED}_{tr-ntr}$ (global)	$\Delta\text{MFED}_{tr-ntr}$ (local L=150)
mRNA	original	0.00365 ± 0.042	0.49	-1.2%	-0.177	0.33	0.49	0.64
	di-shuffle	0.00368 ± 0.042	—	-1.2%	—			
all ncRNAs	original	-0.0379 ± 0.062	< 0.01	11.2%	-2.3	0.64	0.73	0.70
	di-shuffle	-0.0183 ± 0.047	—	6.7%	—			
miRNA	original	-0.0907 ± 0.073	< 0.01	20.8%	-5.54	0.94	0.87	0.87
	di-shuffle	-0.0375 ± 0.048	—	15.2%	—			
5.8S rRNA	original	-0.0165 ± 0.041	< 0.01	5.7%	-0.637	0.46	0.68	0.52
	di-shuffle	-0.0062 ± 0.041	—	2.3%	—			
5S rRNA	original	-0.0237 ± 0.054	< 0.01	7.1%	-1.12	0.44	0.64	0.64
	di-shuffle	-0.0109 ± 0.043	—	3.6%	—			
7SK	original	0.0182 ± 0.023	1	-5.9%	0.733	0.70	0.20	0.72
	di-shuffle	0.00735 ± 0.023	—	-2.29%	—			
Hammerhead_1	original	-0.00459 ± 0.064	< 0.01	1.9%	-1.07	0.08	0.50	0.50
	di-shuffle	0.0284 ± 0.063	—	-15.2%	—			
Hammerhead_3	original	-0.00869 ± 0.045	0.41	2.5%	-3.19	0.53	0.65	0.65
	di-shuffle	-0.00577 ± 0.048	—	2.8%	—			
Intron_gpI	original	-0.0119 ± 0.026	< 0.01	5.1%	-1.52	0.10	0.68	0.70
	di-shuffle	-0.000384 ± 0.021	—	0.2%	—			
Intron_gpII	original	-0.116 ± 0.067	< 0.01	34%	-3.10	0.82	1.00	1.00
	di-shuffle	-0.0491 ± 0.045	—	21.6%	—			
IRES	original	-0.0250 ± 0.037	< 0.01	6.9%	-1.38	0.66	0.76	0.56
	di-shuffle	-0.0148 ± 0.041	—	4.4%	—			
RNase MRP	original	-0.0289 ± 0.036	< 0.01	9.1%	-3.36	0.86	0.81	0.57
	di-shuffle	-0.0121 ± 0.031	—	4.5%	—			
Nuclear RNase P	original	-0.0423 ± 0.043	< 0.01	11%	-1.50	0.79	0.87	0.85
	di-shuffle	-0.0305 ± 0.033	—	8.5%	—			
snoRNA	original	-0.0205 ± 0.049	0.02	9.3%	-0.575	0.72	0.64	0.64
CD-box	di-shuffle	-0.0111 ± 0.047	—	5.5%	—			
snoRNA	original	-0.0422 ± 0.047	< 0.01	14.7%	-0.99	0.80	0.86	0.80
HACA-box	di-shuffle	-0.0193 ± 0.043	—	7.4%	—			
SRP_euk_arch	original	-0.0651 ± 0.067	< 0.01	12.7%	-6.78	0.80	0.84	0.80
	di-shuffle	-0.0529 ± 0.059	—	13.3%	—			
tmRNA	original	-0.0256 ± 0.026	< 0.01	7.8%	-1.97	0.45	0.84	0.49
	di-shuffle	-0.00874 ± 0.024	—	2.9%	—			
tRNA	original	-0.0134 ±0.053	0.88	4.3%	-1.84	0.66	0.60	0.60
	di-shuffle	-0.0195 ± 0.058	—	8.2%	—			

Note: [a] global strand asymmetries unless specified otherwise; [b] p-value using permutation test; [c] nucleotide frequency nounced in miRNA precursors, may indicate that secondary structures have predominately evolved on the transcribed strands and that the stable structure is important for the function of these ncRNAs. The MFE calculated on the transcribed strand ($MFED_{tr}$) of miRNA sequences showed a particularly pronounced negative shift from that of its shuffled version (Figure 2). We calculated Z-scores to indicate the thermodynamic stability of an RNA structure compared to the distribution of the shuffled versions. For each original RNA sequence, a Z-score was calculated for the transcribed sequence strand $MFED_{tr}$ (See Methods) and the average Z-score values are shown in Table 2. Specifically, the Z-score of $MFED_{tr}$ in the miRNA set is very substantial (mean values of -5.54) , which is consistent with the results of Bonnet et al.[8] The Z-score over all ncRNA is -2.3, which is consistent with the results of Rivas et al.[6] Figure 3 gives the Z-score distributions of $MFED_{tr}$ in the miRNA, ncRNA, and mRNA datasets.

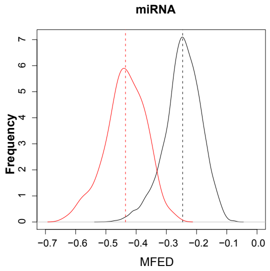

Fig. 2. The distribution comparison for the MFEDs calculated on miRNA transcribed strand (red) and that on miRNA di-nucleotide shuffled version (black). The dashed lines are the mean values of MFEDs.

As discussed above, the permutation test results indicate that the potential structural strand asymmetry as measured by $\Delta MFED_{tr-ntr}$ is unlikely to be a result of global base compositional biases. We conjecture that this structural differential between strands is caused by RNA structural constraints.

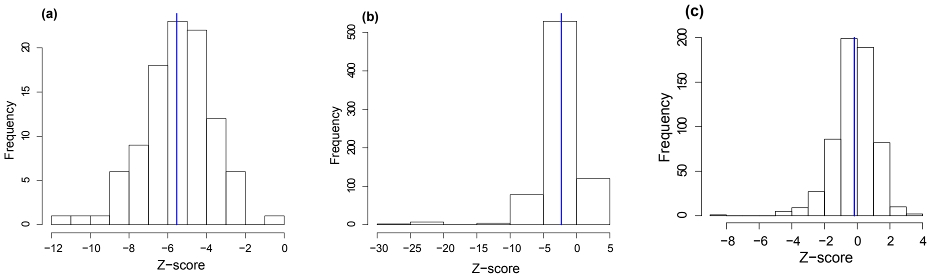

Fig. 3. The histogram of Z-score values for the di-nucleotide shuffled sequences in (a) miRNA set, (b) ncRNA set, and (c) mRNA set. The blue line shows the mean z-score.

To measure how consistently $\Delta MFED_{tr-ntr}$ favours the transcribed strand, as we conjecture it will, we measured the classification accuracy of this feature for predicting the transcribed strand (see Methods). Results are shown in Table 2. Using the global strand asymmetry feature gave a 73% overall accuracy for ncRNA. The local strand asymmetry feature for L=150 shows accuracy for ncRNAs within 3% of the global measure despite being limited to a fixed window size. Dinucleotide-shuffling the ncRNA set reduced the accuracy from 70% to 61%. For mRNAs, using global $\Delta\text{MFED}_{tr-ntr}$ gave a random level of 49%, and for local (L=150) $\Delta\text{MFED}_{tr-ntr}$, accuracy improved above random levels to 64% accuracy, which is statistically significant ($p < 0.01$). Using a di-nucleotide shuffled version of mRNA gave random accuracy results of 51%, indicating that the 64% accuracy result is unlikely to be due to global base compositional asymmetries alone.

Base composition biases can also be useful features for classification of RNA. For comparison with the structural asymmetry feature, the transcribed strand prediction results for a base composition feature are presented in Table 2. We use $\frac{f_G+f_U}{f_A+f_C}$, which as described in[3,17] shows a strand asymmetry in transcribed regions caused by differences in base mutation rates. We note that the increased G and U content in the transcribed strand is also consistent with increased structure–forming potential in this strand. This feature shows some discriminability, although inferior accuracy relative to the local structural strand asymmetry feature.

4. Conclusion

This study has not focused on the RNA optimal minimum-energy folded structures, but rather on the strand asymmetry of RNA secondary structures, extending our previous study[9] to known ncRNA datasets. We have shown that there exists a substantial asymmetry in RNA structure potential between the complementary sequence strands in ncRNAs (including miRNAs), and that this bias is in addition to that due to base compositional strand asymmetries. We conjecture that this is due to structural constraints on the transcribed strand of functional RNA sequences.

This structural strand asymmetry should be useful as an independent feature in helping to distinguish transcribed regions, including transcription orientation, for gene-finding (particularly ncRNA) purposes. This approach can be applied across an entire genome as the local structural asymmetry feature can be easily computed in this case. The non-coding gene prediction framework of Glusman et al[3] is an example which could be extended with this feature. It will be required to combine with additional statistical features to achieve higher discriminability for ncRNA prediction. The possible candidate features include base compositional biases and conserved ncRNA elements. Both our previous work[9] and other studies[18] have suggested that base compositional biases may serve as indicators of ncRNAs. Also, RNA structural conservation using comparative sequence analysis has also shown promise for ncRNA prediction.[7,19] In future work, we will investigate combining these features using machine learning approaches and apply to whole genomes, including UTR, intergenic and intronic regions.

References

1. Storz, G., An expanding universe of noncoding RNAs., *Science*, 296(5571):1260–1263, 2002.
2. Mattick, J. S., and Makunin, I. S., Small regulatory RNAs in mammals., *Hum. Mol. Genet.*, 14(Spec No 1):R121-R132, 2005.
3. Glusman, G., Qin, S., El-Gewely, M. R., Siegel, A.F., Roach, J.C., Hood, L, and Smit, A. A third approach to gene prediction suggests thousands of additional human transcribed regions., *PLOS Computational Biology*, 2(3):160–173, 2006.
4. Seffens, W., and Digby, D., mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences., *Nucleic Acids Res.*, 27(7):1578–1584, 1999.
5. Workman, C., and Krogh, A., No evidence that mRNAs have lower folding free energies than random sequences with the same di-nucleotide distribution., *Nucleic Acids Res.*, 27(24):4816–4822, 1999.
6. Rivas, E., and Eddy, S. R., Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs., *Bioinformatics*, 16(2):583–605, 2000.
7. Washietl, S., Hofacker, I. L., and Stadler, P. F., Fast and reliable prediction of noncoding RNAs., *Proc. Natl. Acad. Sci. U. S. A.*, 102(7):2454–2459, 2005.
8. Bonnet, E., Wuyts, J., Rouz, P., and Van de Peer, Y., Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences., *Bioinformatics*, 16(2):2911–2917, 2004.
9. Wen, J., Parker, B. J., and Weiller, G. F., *In silico* identification and characterization of mRNA-like noncoding transcripts in *Medicago truncatula*., *In Silico Biology*, 7, 0034. </isb/2007/07/0034/>
10. Griffiths-Jones,S., Moxon S., Marshall,M., Khanna,A., Eddy,S.R., Bateman, A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, 33(Database issue):D121-D124, 2005.
11. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, 31:365–370, 2003.
12. Kulikova, T., Akhtar, R., Aldebert, et al. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res.*, 35:(Database issue):D16-20, 2006.
13. Rice, P., Longden, I., and Bleasby, A., EMBOSS:The European Molecular Biology Open Software Suite. *Trends Genet.*, 16(6):276-277, 2000.
14. Altschul, S. F., and Erickson, B. W., Significance of nucleotide sequence alignments: A method for random sequence permutation that preserves di-nucleotide and codon usage. *Mol. Biol. Evol.*, 2:526–538, 1985.
15. Hofacker, I. L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker,M., Schuster, P. Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie*, 125:167–188, 1994.
16. Reiche, K., Stadler P. F., RNAstrand: reading direction of structured RNAs in multiple sequence alignments., *Algorithms Mol. Bio.*, 2:6, 2007.
17. Green, P., Ewing, B., Miller, W., Thomas, P. J., NISC Comparative Sequencing Program, Green, E. D., Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.*, 33:514–517, 2003.
18. Schattner P., Searching for RNA genes using base-composition statistics. *Nucleic Acids Res.*, 30:2076–2082, 2002.
19. Rivas, E., Eddy, S. R., Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8, 2001.