BIOINFORMATICS DISCOVERY NOTE Vol. 26 no. 21 2010, pages 2651–2655
doi:10.1093/bioinformatics/btq509

Sequence analysis

Advance Access publication September 6, 2010

The mRNA landscape at yeast translation initiation sites

A. Robbins-Pianka[1,2], M. D. Rice[2] and M. P. Weir[1,*]

[1]Department of Biology and [2]Department of Mathematics and Computer Science, Wesleyan University, Middletown, CT 06459, USA

Associate Editor: Alex Bateman

ABSTRACT

Summary: Although translation initiation has been well studied, many questions remain in elucidating its mechanisms. An ongoing challenge is to understand how ribosomes choose a translation initiation site (TIS). To gain new insights, we analyzed large sets of TISs with the aim of identifying common characteristics that are potentially of functional importance. Nucleotide sequence context has previously been demonstrated to play an important role in the ribosome's selection of a TIS, and mRNA secondary structure is also emerging as a contributing factor.

Here, we analyze mRNA secondary structure using the folding predictions of the RNAfold algorithm. We present a method for analyzing these results using a rank-ordering approach to assess the overall degree of predicted secondary structure in a given region of mRNA. In addition, we used a modified version of the algorithm that makes use of only a subset of the standard version's output to incorporate base-pairing polarity constraints suggested by the ribosome scanning process. These methods were employed to study the TISs of 1735 genes in *Saccharomyces cerevisiae*.

Trends in base composition and base-pairing probabilities suggest that efficient translation initiation and high protein expression are aided by reduced secondary structure upstream and downstream of the TIS. However, the downstream reduction is not observed for sets of TISs with nucleotide sequence contexts unfavorable for translation initiation, consistent with previous suggestions that secondary structure downstream of the ribosome can facilitate TIS recognition.

Contact: mweir@wesleyan.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 31, 2010; revised on July 28, 2010; accepted on September 1, 2010

1 INTRODUCTION

In the scanning model of translation initiation (Kozak, 1989), a small ribosomal subunit binds at the m7G cap of an mRNA transcript and scans in the 3′ direction until it reaches the first AUG codon in an appropriate sequence context. The small subunit then assembles with the large ribosomal subunit and initiates translation of the open reading frame, terminating upon reaching the first in-frame stop codon. However, some instances of non-standard translational events such as leaky scanning (leaky initiation), alternative translation initiation and initiation at non-canonical (i.e. non-AUG) start codons fall outside the original model (Kozak, 2002; Touriol *et al.*, 2003; Wang and Rothnagel, 2004).

The recognition/initiation event depends partially on the nucleotide sequence context flanking the initiation codon. It has been noted previously that positions −3 and +4 relative to the initiation codon provide particularly strong cues (Kozak, 1986b). Information-based measures of context based on sequence conservation surrounding aligned TISs suggest that other nucleotide positions also contribute to recognition. The translation relative individual information (TRII) score (Supplementary Material 1.1; M.Weir and M.Rice, submitted for publication) is a convenient and sensitive measure of sequence conservation in sets of aligned sequences. The TRII score of each sequence is the same as its individual information (Schneider, 1997) but incorporates background nucleotide frequencies. By taking into account the relative frequencies of all four nucleotides at each position in an alignment of high-confidence functional start sites, TRII scores quantify an individual sequence's conformance to characteristic nucleotide preferences in the vicinity of the TIS.

It has been shown that mRNA secondary structure can influence scanning and translating ribosomes. In various viruses, *Escherichia coli* and *Saccharomyces cerevisiae*, pseudoknot secondary structures have been associated with frameshifting of translating ribosomes (Jacobs *et al.*, 2007; Tzeng *et al.*, 1992) and stop codon readthrough (Wills *et al.*, 1994). Secondary structure, including pseudoknots, can also influence the selection of a translation initiation site (TIS). Examples include functioning of viral internal ribosome entry sites (IRES) (Balvay *et al.*, 2009) and modulation through riboswitches characterized in prokaryotes; it is speculated that future studies may uncover analogous riboswitch regulation in eukaryotes (Wachter, 2010).

Moreover, a hairpin positioned downstream of a TIS can increase the probability of initiation when the initiation codon is in an otherwise unfavorable nucleotide sequence context. It has been proposed that the effect of this hairpin is to slow the scanning ribosomal subunit, increasing the interaction between the initiation codon and the ribosomal machinery (Kochetov *et al.*, 2007; Kozak, 1990). Conversely, a sufficiently stable hairpin upstream of a TIS can completely abolish initiation at that site, perhaps by interfering with the small ribosomal subunit before it reaches the TIS (Kozak, 1986a).

To evaluate further these roles of RNA secondary structure in TIS recognition, we applied the RNA folding algorithm RNAfold (McCaskill, 1990) to subsets of mRNAs selected based on TRII scores, proteome-scale protein expression studies (Ghaemmaghami *et al.*, 2003) and ribosome-profiling studies (Ingolia *et al.*, 2009). Predicted RNA structures of TISs with high TRII scores ($TRII_{\mathrm{high}}$)

*To whom correspondence should be addressed.

were compared with those of TISs with low TRII scores ($TRII_{\text{low}}$). Similarly, we compared RNA structures for TISs with higher and lower ribosome density ($Ribosome_{\text{high}}$ and $Ribosome_{\text{low}}$, respectively) and genes with higher and lower protein expression. The trends revealed by these comparisons suggest how RNA secondary structure can influence both positively and negatively the functioning of the translation machinery.

2 METHODS

2.1 RNA folding simulations

RNAfold (Hofacker *et al*., 1994) was used to conduct folding simulations on various subsets of a set of 1735 *S.cerevisiae* genes. This *HiConf* set comprised annotated TISs in which we had high confidence (see Supplementary Material 1.1). The RNAfold program is based on an equilibrium partition function (McCaskill, 1990). Rather than output only the minimum free energy structure of an input sequence, RNAfold considers the ensemble of all possible structures and weights each by its energetic favorability (Zuker and Stiegler, 1981). Based on the set of possible structures and their associated weights, RNAfold computes for each nucleotide in the input sequence the probability of base pairing with each other nucleotide in the sequence.

The total base-pairing probability (*TotalBPP*) of a position in a nucleotide sequence is calculated by summing the individual probabilities of every possible pair involving that position. This metric was used to quantify a given mRNA region's overall potential to form secondary structure. The downstream base-pairing probability (*DownBPP*) of a nucleotide position is calculated by summing the individual probabilities of every possible base pair involving that position and any downstream position. The *DownBPP* metric may better reflect the probability that a nucleotide position is involved in a structure when a scanning ribosome encounters it, given that upstream nucleotides protected by the ∼30 nt ribosome footprint (Ingolia *et al*., 2009; Steitz, 1969) may be less likely to be involved in mRNA secondary structure.

Two approaches were used to interpret the RNAfold output. We examined base-pairing probabilities at the resolution of individual positions with the first approach, and we measured regional trends in base-pairing probability in the second approach to provide greater sensitivity when comparing sequence sets. The two approaches are described below.

2.1.1 Position-specific analysis To investigate position-specific trends in predicted structure, sets of genes were aligned at their TISs and mean *TotalBPP* and mean *DownBPP* were calculated at each position. The position-specific approach was used to analyze aligned open reading frame (ORF) sequences. As a control, and to provide a benchmark for comparison, randomly generated sequences were subjected to the same analysis. For analysis of ORF sequences, we used randomly generated sequences with a frequency of 0.25 for each nucleotide ($Random_{0.25}$).

2.1.2 Rank-ordered analysis The above position-specific approach has the disadvantage that unless positional base pairing is well aligned with the alignment landmark (TIS), high-probability base-pairing predictions in individual sequences will tend to be diluted in the computation of the mean. To address this problem, we used an alternative approach (Fig. S1) that focused on the number and probability of base-pairing events (but not their specific locations) in a selected region of the sequence alignment.

Sets of genes aligned at their TISs were analyzed for the overall degree of predicted secondary structure in a region chosen relative to the TIS. Positional *TotalBPP* values of each sequence in the alignment were sorted high-to-low (i.e. ranked) and equivalently ranked values were averaged (Fig. S1). This approach was used to analyze predicted secondary structure of 5′UTR sequences. Two sets of randomly generated sequences provided a baseline for comparison: $Random_{0.25}$ (see above) and $Random_{\text{UTR}}$ (which reflected nucleotide frequencies found in observed 5′UTRs (Miura *et al*., 2006; Nagalakshmi *et al*., 2008) excluding the region from −40 to −1; see Section 3).

3 RESULTS AND DISCUSSION

3.1 Upstream of AUG

3.1.1 Neighborhood assessment using TRII scores The eukaryotic consensus sequence surrounding a TIS—C C (A|G) C C AUG G—is typically considered to comprise a relatively narrow window of positions (Kozak, 1987). Similarly, the TIS in budding yeast has a narrow consensus sequence

$$\text{Consensus}_{0.5,-0.5} = \text{N A\U A\(C|U) N\G A\U AUG G C\(A|U))}$$

defined by the frequency weight matrix for *HiConf* TISs (Fig. 1), where A\(G|U) denotes 'A and not G and not U' (M.Weir and M.Rice, submitted for publication). The consensus is dominated by excluded nucleotides which have negative weight matrix values (Fig. 1). Computation of TRII scores over the interval −5 to +5 reveals that the score distribution for 1735 *HiConf* start sites can be distinguished from the distribution for 2000 $Random_{\text{UTR}}$ artificial sites (Fig. 2A). Although the distinction is more subtle, *HiConf* sites can also be distinguished from $Random_{\text{UTR}}$ sites by comparing the distributions of TRII scores computed over the interval −20 to −6 (Fig. 2A). This suggests that there are characteristics upstream of functional start sites that may contribute to recognition and initiation.

3.1.2 Nucleotide frequency profiles To investigate the upstream region and the nature of its apparent contribution to TIS recognition, positional nucleotide frequencies were calculated for an alignment of the 713 *HiConf* genes (Fig. 2B) having 5′UTR ≥ 100. A strong trend was noted for the *HiConf* subset beginning at approximately position −40. The frequency of A increases to about 42% while the frequency of U decreases to about 26%. In contrast, the frequencies of G and C remain fairly uniform over this region, although G is a little depressed compared to C. While the AU and GC content are fairly uniform, there is selection for A over U and for C over G in the coding strand of double-stranded DNA upstream of the TIS.

nt position	-20	-19	-18	-17	-16
A	0.33	0.37	0.31	0.35	0.42
C	0.15	-0.06	-0.05	-0.01	-0.07
G	-0.12	-0.14	-0.10	-0.25	-0.20
U	-0.46	-0.35	-0.30	-0.31	-0.40

nt position	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6
A	0.34	0.41	0.46	0.34	0.44	0.29	0.33	0.37	0.31	0.18
C	-0.14	-0.10	-0.22	-0.07	-0.01	-0.16	0.03	-0.09	-0.19	-0.18
G	-0.16	-0.19	-0.14	-0.29	-0.20	-0.13	-0.27	-0.15	0.06	0.14
U	-0.25	-0.36	-0.40	-0.23	-0.47	-0.18	-0.29	-0.32	-0.31	-0.18

nt position	-5	-4	-3	-2	-1	1	2	3	4	5
A	0.22	**0.53**	**1.02**	0.47	**0.58**	1.63	-10.76	-10.76	-0.11	**-0.60**
C	0.15	0.16	**-1.16**	0.16	-0.12	-10.76	-10.76	-10.76	-0.47	**1.26**
G	-0.18	-0.30	0.20	**-0.52**	-0.15	-10.76	-10.76	2.62	**0.77**	-0.11
U	-0.26	**-0.74**	**-2.10**	-0.49	**-0.71**	-10.76	1.58	-10.76	-0.18	**-0.75**
CONSENSUS	N	**A\U**	**A\(C|U)**	**N\G**	**A\U**	A	U	G	**G**	**C\(A|U)**

Fig. 1. TIS frequency weight matrix. The figure shows a weight matrix for the *HiConf* set of 1735 genes. Each entry in the weight matrix is $\log_2(\text{freq}_{\text{observed}}/\text{freq}_{\text{background}})$ where the background frequencies are those of *HiConf* 5′UTRs excluding −40 to −1 (A: 0.322; C: 0.182; G: 0.162; U: 0.334). Matrix values ≥0.5 and ≤−0.5 (bold black and bold gray, respectively) define a $\text{consensus}_{0.5,-0.5}$. Matrix values between −0.5 and −0.1 (shaded) are dominated by nucleotides G and U. Depression of G and U reduces RNA secondary structure.

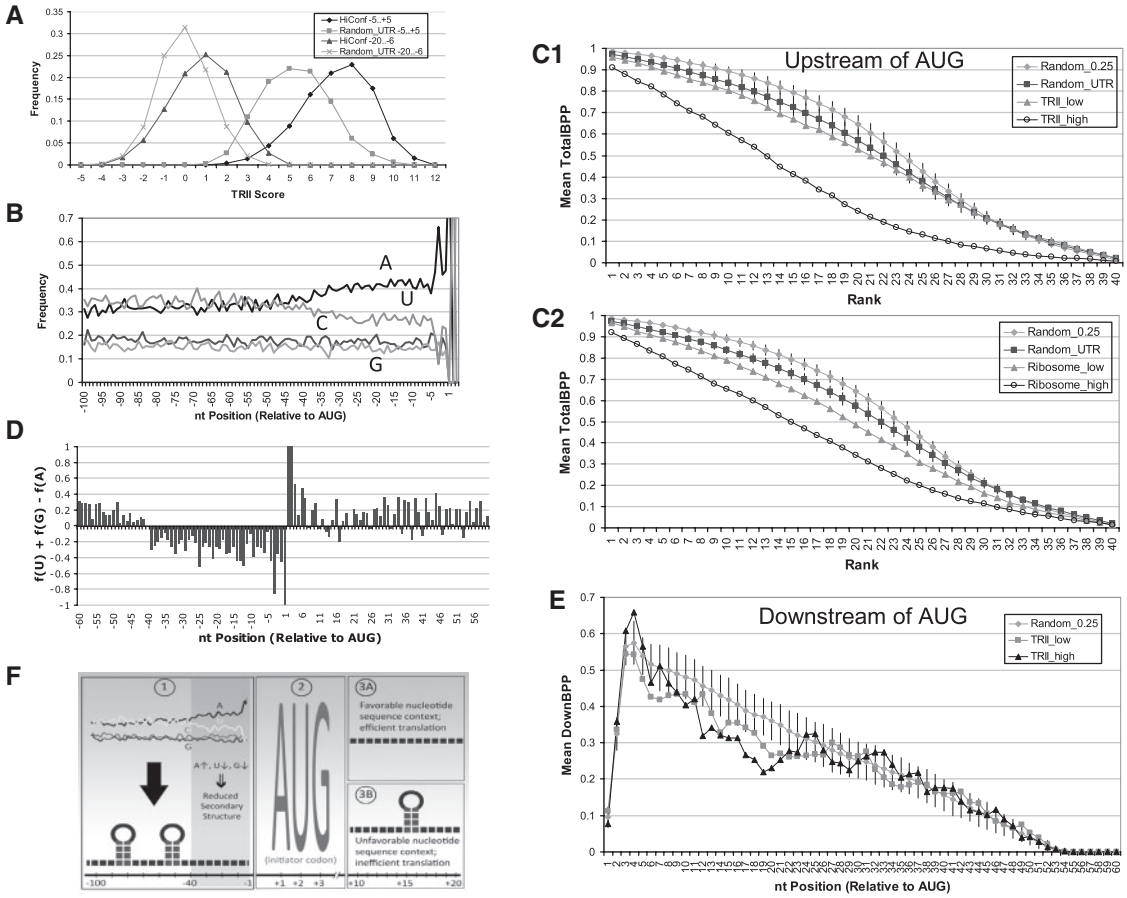

Fig. 2. (**A**) TRII score distributions for aligned high-confidence TISs can be distinguished from AUG triplets flanked by random sequences using either the canonical region −5 to +5 (c.f. $HiConf_{-5..5}$ with $Random_{UTR-5..5}$) or the interval −20 to −6 upstream of the TIS (c.f. $HiConf_{-20..-6}$ with $Random_{UTR-20..-6}$) (χ^2 goodness-of-fit tests $P < 0.01$). (**B**) The set of 713 aligned *HiConf* start sites with observed 5′UTR length≥100 exhibits a striking divergence of A and U frequencies over the interval −40 to −1. (**C1** and **C2**) The *HiConf* set was partitioned by TRII score into subsets $TRII_{high}$ and $TRII_{low}$ (C1; see Supplementary Material 1.1), and by ribosome density into subsets $Ribosome_{high}$ and $Ribosome_{low}$ (C2). Each set was analyzed for secondary structure upstream of the TIS using the rank-ordering approach (see Fig. S1). The $TRII_{high}$ and $Ribosome_{high}$ subsets (unfilled circles) exhibited a significant reduction of predicted secondary structure while the $TRII_{low}$ and $Ribosome_{low}$ subsets (gray triangles) exhibited a slight reduction compared with sets of random sequences $Random_{0.25}$ (gray diamonds) and $Random_{UTR}$ (gray squares). (**D**) A nucleotide secondary structure (NSS) index of a position in an alignment was calculated by subtracting the frequency of A at that position from the sum of the frequencies of U and G. A higher NSS index indicates a greater propensity for secondary structure formation; shown is the NSS index profile of $TRII_{high}$ genes. A strong depression is evident over −40..+1, and a more subtle depression is evident in the critical region +10 to +20. (**E**) To examine secondary structure downstream of the TIS, mean *DownBPP* values were computed for each position in alignments of $TRII_{high}$ (black) and $TRII_{low}$ (gray), which in this case comprised the top 10% and bottom 5–15% of TRII scorers, respectively. Both subsets exhibit a depression in the critical region +10 to +20 compared to $Random_{0.25}$ (light gray), and $TRII_{high}$ exhibits a more pronounced depression than $TRII_{low}$. Similar effects were observed monitoring *TotalBPP* instead of *DownBPP* (Fig. S9B,C) although *TotalBPP* does not accommodate protection of upstream positions by the ribosome footprint. Error bars show 1.96 SD (95% thresholds) for 100 sets of 87 (C1, C2) or 174 (E) random sequences. (**F**) Hypothetical model of the RNA landscape surrounding a TIS.

The base-pairing potential of U is higher than A. While A can base pair only with U, the U nucleotide can base pair with either A or G, the latter through non-canonical G:U pairing. Hence, we speculated that the observed elevated A and depressed U and G indicated that there is selection for reduced RNA structure in the region upstream of the TIS. To investigate this possibility, we examined predicted RNA structure upstream of the TIS.

3.1.3 Predicted RNA secondary structure Given that the region −40 to −1 has been implicated in TIS recognition, we used the RNAfold algorithm to predict RNA secondary structures. Because the nucleotide frequency profiles (Fig. 2B) did not suggest that the precise positions of base pairing were likely to be important, we decided to use a rank-ordering filter to assess the predictions. This approach (Fig. S1) indicates the number of high-probability base-pairings over a region. After running RNAfold on mRNA sequences from −60 to +60, we applied the rank-ordering filter to the set of output probabilities (*TotalBPP*; see Section 2) for positions −40 to −1.

For analysis, we chose to compare $TRII_{high}$ with $TRII_{low}$ (see Supplementary Material 1.1; Fig. S2). As controls, RNAfold was run

on $Random_{UTR}$ and $Random_{0.25}$ sequences. Compared to both sets of controls, the $TRII_{high}$ set showed highly depressed predicted base pairing, whereas the $TRII_{low}$ subset has less pronounced depression (Figs. 2C1, S3A). An analogous assessment of positions −8 to −1 (Kochetov *et al*., 2003) similarly showed a negative correlation between TIS quality and *TotalBPP*. This result is consistent with the idea that the TRII score, which was applied to a window of −40 to +40 relative to the AUG, incorporates sequence preferences that tend to reduce RNA secondary structure. Indeed, the weight matrix used to compute TRII scores generally has small positive entries for A and small negative entries for U upstream of −6 (similar to Fig. 1, shaded weight matrix entries), consistent with this interpretation. These trends are also illustrated by the profile of a nucleotide secondary structure index, NSS = (freq U + freq G − freq A), which has a pronounced depression between −40 and 1 ($TRII_{high}$, Fig. 2D; compare with $TRII_{low}$, Fig. S6C), providing a striking delineation of the TIS.

Since TRII scores indicate how well sequences conform to TIS consensus sequences, we wished to test whether genes with high protein expression similarly show depressed predicted secondary structure upstream of the TIS. Recent ribosome profiling analysis using deep sequencing (Ingolia *et al*., 2009) has suggested that ribosome tag densities are an excellent indicator of protein expression. Therefore, we selected from the *HiConf* set a subset of 165 genes with high ribosome tag densities ($Ribosome_{high}$, see Supplementary Material 1.1). Compared to both sets of controls ($Random_{UTR}$ and $Random_{0.25}$), the $Ribosome_{high}$ set showed highly depressed predicted base pairing (Figs. 2C2, S3B) and NSS index (Fig. S6A), suggesting that depression of RNA structure in the region preceding the TIS facilitates efficient initiation and translation. Equivalent depression of secondary structure was observed in transcripts that exhibited a high-density ribosome profile in only one of the two growth conditions tested, rich or amino acid-starved media (Fig. S5), suggesting that low RNA structure facilitates the potential for regulated high protein expression. We also tested a subset of the *HiConf* set with 165 genes with detectable but low ribosome tag densities ($Ribosome_{low}$). $Ribosome_{low}$ exhibited much less pronounced depression in predicted secondary structure (Fig. 2C2) and NSS index (Fig. S6B), suggesting that there is less selection against secondary structure for lower expression genes.

Equivalent results were observed for subsets of *HiConf* partitioned based on protein expression (Fig. S4) rather than ribosome densities. Based on proteome-scale western analysis of TAP and GFP tagged proteins (Ghaemmaghami *et al*., 2003), we generated subsets of *HiConf* with high ($Protein_{high}$) and low but detectable ($Protein_{low}$) protein expression. Like the $Ribosome_{high}$ set, $Protein_{high}$ showed depressed secondary structure upstream of the TIS whereas $Protein_{low}$ had much less pronounced depression.

Transcript sequences for genes with high-confidence TISs were also examined in *Drosophila melanogaster* and *Schizosaccharomyces pombe*. Depression of NSS index (Fig. S10) and predicted secondary structure (Fig. S11A) were observed in the region upstream of the TIS and this was more pronounced for high-TRII score and high protein expression TISs; suppression of secondary structure was also observed downstream of the AUG start codon (Fig. S11B; see Section 3.2). These effects were more pronounced in *D.melanogaster* than *S.pombe*, suggesting that the contributions of mRNA secondary structure to translation initiation may vary between species.

3.2 Downstream of AUG

Our observations above indicate that genes with high TRII scores, high ribosome densities and high protein expression benefit from depressed RNA secondary structure upstream of the TIS. We were interested whether depression of RNA structure downstream of the TIS is also beneficial. This was of particular interest since previous studies (Kochetov *et al*., 2007; Kozak, 1990) have suggested that elevated secondary structure in the region 13–17 nucleotides downstream of the TIS can facilitate TIS recognition by the ribosome, particularly for poorer context TISs.

We applied the RNAfold algorithm to the subsets $TRII_{high}$ and $TRII_{low}$ but in this case, the algorithm was applied to +1 to +60 relative to the TIS. We used a modified version of RNAfold in which we only collected the probabilities of base pairing downstream of each nucleotide position (*DownBPP*; see Section 2). In considering scanning of the ribosome from 5′ to 3′, we were interested in the probability of downstream base pairing given that upstream positions would likely be protected by the ribosome footprint. Also, unlike our earlier analysis, we did not apply the rank-order filter to the RNAfold output because we wished to preserve nucleotide position information given that previous studies have implicated as important the region near nucleotides 13–17 downstream of the TIS. Instead of using the rank-order filter, we computed for each nucleotide position the mean *DownBPP* in each test set $TRII_{high}$, $TRII_{low}$ and $Random_{0.25}$.

When compared to $Random_{0.25}$, we observed a depression in predicted RNA secondary structure in the critical region +10 to +20. This depression was most pronounced for highly expressed genes—$TRII_{high}$ (Figs. 2E, S9) and $Ribosome_{high}$ (Fig. S7)— and least pronounced for low expression genes—$TRII_{low}$ (Figs. 2E, S9) and $Ribosome_{low}$ (Fig. S7). These results suggest that efficient translation is facilitated by reduced RNA structure in the region immediately downstream of the ribosome (approximately nucleotide positions +10 to +20 downstream of the TIS). Our results agree with previous observations (Kochetov *et al*., 2007; Kozak, 1990) that poorer sequence context TISs have elevated predicted RNA structure downstream of the TIS. However, it may be more appropriate to consider the predicted RNA structure in $TRII_{low}$ sequences to be less depressed rather than elevated—$TRII_{low}$ sequences show depressed predicted RNA structure when compared with $Random_{0.25}$ sequences, but less depressed RNA structure when compared to $TRII_{high}$. Hence, for suboptimal start sites, there may be a balance between two opposing selection forces: (i) selection for depressed secondary structure to increase translation rates and (ii) selection for elevated secondary structure in the downstream region to create a partial barrier to scanning and thereby increase the likelihood of translation initiation.

Other studies have suggested that in addition to RNA structure properties, efficiency of translation is influenced by selection for codons with high tRNA adaptation indices (tAI) (Tuller *et al.,* 2010a, b) and clustering of codons on mRNAs that use the same tRNA (Cannarrozzi *et al.*, 2010). Examination of mean tAI indices for $TRII_{high}$ and $TRII_{low}$, as well as $Ribosome_{high}$ and $Ribosome_{low}$, and $Protein_{high}$ and $Protein_{low}$ (Fig. S12) suggests that, as expected, the higher expression genes show selection for higher tAI index

codons, which are better represented in the cell's tRNA pool. The higher expression genes also show selection for clustering of codons that utilize the same tRNA (Fig. S13).

3.3 Conclusions

For highly expressed genes, there is selection against RNA secondary structure upstream of the TIS (positions −40 to −1) and in a critical region downstream of the TIS (positions +10 to +20). Reduction in secondary structure may facilitate efficient initiation by the ribosome. Elevation of A and depression of U or G content appear to contribute to reduced RNA structure (Figs. 2F, S8).

For lower expression genes, there is less selection against RNA secondary structure, both upstream of the TIS and in the region predicted to be immediately downstream of the footprint of an initiating ribosome (positions +10 to +20). The presence of secondary structure downstream of the AUG is thought to facilitate ribosome pausing over the start site region, increasing the likelihood of translation initiation.

These results suggest that future algorithms for prediction of protein translation initiation will benefit from inclusion of classifiers that incorporate predicted RNA structure upstream and downstream of the TIS.

ACKNOWLEDGEMENTS

We thank Danny Krizanc, Robert Lane and the reviewers for careful reading of the manuscript.

Funding: This work was supported in part by funds from the Howard Hughes Medical Institute to support undergraduate initiatives in the life sciences.

Conflict of Interest: none declared.

REFERENCES

Balvay,L. *et al.* (2009) Structural and functional diversity of viral IRESes. *Biochim. Biophys. Acta*, **1789**, 542–557.

Cannarozzi,G. *et al.* (2010). A role for codon order in translation dynamics. *Cell*, **141**, 355–367.

Ghaemmaghami,S. *et al.* (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.

Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, **125**, 167–188.

Ingolia,N.T. *et al.* (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.

Jacobs,J.L. *et al.* (2007) Identification of functional, endogenous programmed -1 ribosomal frameshift signals in the genome of Saccharomyces cerevisiae, *Nucleic Acids Res.*, **35**, 165–174.

Kochetov,A.V. *et al.* (2003). Interrelations between the efficiency of translation start sites and other sequence features of yeast mRNAs. *Mol. Genet. Genomics*, **270**, 442–447.

Kochetov,A.V. *et al.* (2007) AUG_hairpin: prediction of a downstream secondary structure influencing the recognition of a translation start site. *BMC Bioinformatics*, **8**, 318.

Kozak,M. (1978) How do eukaryotic ribosomes select initiation regions in messenger RNA? *Cell*, **15**, 1109–1123.

Kozak,M. (1986a) Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. *Proc. Natl Acad. Sci. USA*, **83**, 2850–2854.

Kozak,M. (1986b) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, **44**, 283–292.

Kozak,M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, **15**, 8125–8148.

Kozak,M. (1989) The scanning model for translation: an update. *J. Cell Biol.*, **108**, 229–241.

Kozak,M. (1990) Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proc. Natl Acad. Sci. USA*, **87**, 8301–8305.

Kozak,M. (2002) Pushing the limits of the scanning mechanism for initiation of translation. *Gene*, **299**, 1–34.

McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Miura,F. *et al.* (2006) A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc. Natl Acad. Sci. USA*, **103**, 17846–17851.

Nagalakshmi,U. *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Schneider,T.D. (1997). Information content of individual genetic sequences. *J. Theor. Biol.*, **189**, 427–441.

Steitz,J.A. (1969) Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. *Nature*, **224**, 957–964.

Touriol,C. *et al.* (2003) Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biol. Cell*, **95**, 169–178.

Tuller,T. *et al.* (2010a). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell, **141**, 344–354.

Tuller,T. *et al.* (2010b). Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl Acad. Sci. USA*, **107**, 3645–3650.

Tzeng,T.H. *et al.* (1992) Ribosomal frameshifting requires a pseudoknot in the Saccharomyces cerevisiae double-stranded RNA virus. *J. Virol.*, **66**, 999–1006.

Wachter,A. (2010) Riboswitch-mediated control of gene expression in eukaryotes. *RNA Biol.*, **7**, 1–10.

Wang,X.Q. and Rothnagel, J.A. (2004) 5'-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. *Nucleic Acids Res.*, **32**, 1382–1391.

Wills,N.M. *et al.* (1994) Pseudoknot-dependent read-through of retroviral gag termination codons: importance of sequences in the spacer and loop 2. *EMBO J.*, **13**, 4137–4144.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.