A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes

Wanjun Gu[1]☯, **Tong Zhou**[2,3]☯, **Claus O. Wilke**[2,3,4]*

1 Key Laboratory of Child Development and Learning Science of Ministry of Education of China, Southeast University, Nanjing, Jiangsu, China, 2 Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, Texas, United States of America, 3 Section of Integrative Biology, The University of Texas at Austin, Austin, Texas, United States of America, 4 Institute for Cell and Molecular Biology, The University of Texas at Austin, Austin, Texas, United States of America

Abstract

Recent studies have suggested that the thermodynamic stability of mRNA secondary structure near the start codon can regulate translation efficiency in *Escherichia coli*, and that translation is more efficient the less stable the secondary structure. We survey the complete genomes of 340 species for signals of reduced mRNA secondary structure near the start codon. Our analysis includes bacteria, archaea, fungi, plants, insects, fishes, birds, and mammals. We find that nearly all species show evidence for reduced mRNA stability near the start codon. The reduction in stability generally increases with increasing genomic GC content. In prokaryotes, the reduction also increases with decreasing optimal growth temperature. Within genomes, there is variation in the stability among genes, and this variation correlates with gene GC content, codon bias, and gene expression level. For birds and mammals, however, we do not find a genome-wide trend of reduced mRNA stability near the start codon. Yet the most GC rich genes in these organisms do show such a signal. We conclude that reduced stability of the mRNA secondary structure near the start codon is a universal feature of all cellular life. We suggest that the origin of this reduction is selection for efficient recognition of the start codon by initiator-tRNA.

Citation: Gu W, Zhou T, Wilke CO (2010) A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes. PLoS Comput Biol 6(2): e1000664. doi:10.1371/journal.pcbi.1000664

Editor: Berend Snel, Utrecht University, Netherlands

Received October 29, 2009; **Accepted** December 30, 2009; **Published** February 5, 2010

Funding: This work was supported by NIH grant R01 GM088344 to COW and a grant from National Natural Science Foundation of China (No. 30900836) to WG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: cwilke@mail.utexas.edu

☯ These authors contributed equally to this work.

Introduction

Synonymous mutations are frequently used as a neutral baseline to detect selection pressures at the amino-acid level [1]. Yet many mechanisms are now known that cause selection pressure on synonymous sites. Translationally preferred codons are selected for accurate and efficient translation in bacteria, yeast, worm, fly, and even in mammals [2–13]. Selection on synonymous sites acts to increase the thermodynamic stability of DNA and RNA secondary structure [14–18], to improve splicing efficiency [19–21], and to assist protein co-translational folding [22–27].

Synonymous codon choice can also affect translation initiation. Most of the sequence elements that control translation initiation (e.g. the Shine-Dalgarno sequence in prokaryotes and the 5′ cap and Kozak consensus sequence in eukaryotes) are located in 5′ untranslated regions (UTRs) [28–30], where high conservation and AU-richness have been observed [31–34]. Yet Zalucki et al. found a significant bias towards usage of the AAA codon at the second amino acid position in *Escherichia coli* secretory proteins [35]. They proposed that selective pressure for high translation-initiation efficiency accounts for this codon usage bias. Other studies have demonstrated altered expression levels in *E. coli* after changing synonymous codons in the region downstream from the start codon [36–39]. Kudla et al. synthesized a library of 154 genes of green fluorescent protein (GFP) that had random changes at synonymous sites without any change in the amino-acid sequence [40]. They found that the GFP expression level varied 250-fold across the library. In this library, the stability of mRNA secondary structure near the start codon explained more than half of the variation in expression level: mRNAs with more stable local structure in this region had reduced protein expression [40]. These observations suggest that translation initiation is facilitated by a choice of synonymous codons that destabilize local mRNA secondary structure.

Here, we analyzed the local mRNA secondary structure at the 5′ end of the coding region in 340 species, including bacteria, archaea, fungi, plants, fishes, birds, and mammals. We used computational methods to predict the thermodynamic stability of local mRNA secondary structure in sliding windows downstream from the start codon, and used permutation tests to assess deviation from random expectation. We addressed the following questions: (i) Is there a selection pressure on synonymous sites to reduce the stability of local mRNA secondary structure at the translation-initiation region? (ii) Is such a selection pressure a general characteristic for all organisms? (iii) Does 5′ mRNA stability correlate with GC composition, codon usage bias, or gene expression level? (iv) In prokaryotes, does 5′ mRNA stability vary with the optimal growth temperature of the organism?

Results

mRNA stability is reduced near the translation-initiation region

We calculated the local folding energy (ΔG) along the mRNA sequence using a sliding window of 30 nucleotides (nt) in length,

Author Summary

Synonymous mutations are mutations that change the nucleotide sequence of a gene without changing the amino-acid sequence. Because these mutations don't alter the expressed protein, they are frequently also called silent mutations. Yet increasing evidence demonstrates that synonymous mutations are not that silent. In particular, experimental work in *Escherichia coli* has shown that the choice of synonymous codons near the start codon can greatly influence protein production. Codons that allow the mRNA to fold into a stable secondary structure seem to inhibit efficient translation initiation. This observation suggests that selection should prefer reduced mRNA stability near the start codon in many organisms. Here, we show that this prediction generally holds true in most organisms, including bacteria, archaea, fungi, plants, insects, and fishes. In birds and mammals it doesn't hold true genome-wide, but it does hold true in the most GC-rich genes. In all organisms, the extent to which mRNA stability is reduced increases with increasing GC content. In prokaryotes, it also increases with decreasing optimal growing temperature. Thus, it seems that all organisms have to optimize their synonymous sites near the start codon to guarantee efficient protein translation.

moving from the start codon to the 120th downstream nucleotide in steps of 10 nt (for a total of 13 windows). To quantify the deviation from expectation given a gene's amino-acid sequence and codon usage bias, we also calculated ΔG for 1000 permuted mRNA sequences. We obtained permuted sequences by randomly reshuffling synonymous codons within each gene. We then calculated a Z-score, $Z_{\Delta G}$, by comparing the ΔG of the real mRNA segment to the distribution of ΔG values of the permuted sequences (see Materials and Methods). $Z_{\Delta G}$ measures the extent to which local mRNA stability deviates from expectation. A positive $Z_{\Delta G}$ means that local mRNA stability is reduced, and a negative $Z_{\Delta G}$ means that it is increased. For each window, we calculated a genome-wide mean $Z_{\Delta G}$ by averaging the corresponding $Z_{\Delta G}$ values over all genes in a genome.

We performed the sliding window analysis in 340 species, which included 276 bacteria, 35 archaea, 11 fungi, 2 plants, 2 insects, 4 fishes, 2 birds, and 8 mammals. Figure 1 shows an example of the mean $Z_{\Delta G}$ for 13 windows in *E. coli*. We observed a significant positive deviation of $Z_{\Delta G}$ from zero in the first two windows (t-test: $P \ll 10^{-20}$ in both cases). The positive values of $Z_{\Delta G}$ suggest selection for reduced mRNA stability at the $5'$ end of the coding region. The $Z_{\Delta G}$ values further downstream decrease quickly and we observe negative $Z_{\Delta G}$ values in most downstream windows.

Most species we studied showed a similar pattern to the one we observed in *E. coli* (Figure S1 and Table S1), except for plants and warm-blooded animals (birds and mammals). There was a clear increase in mean $Z_{\Delta G}$ for windows close to the start codon. Because the $Z_{\Delta G}$ at the very start of the coding sequence generally showed the strongest signal of reduced mRNA stability, we will focus on this value for the remainder of this study. In the following, we refer to the $Z_{\Delta G}$ at the very start of the coding sequence also as the $5'$ $Z_{\Delta G}$. In prokaryotes, 262 out of 276 bacteria and 28 out of 35 archaea showed a positive $5'$ $Z_{\Delta G}$ (Figure 2). In eukaryotes, 10 out of 11 fungi, 1 out of 2 plant, both insect species, and all four fish species we analyzed showed this pattern as well. All warm-blooded animals showed a negative $5'$ $Z_{\Delta G}$ throughout the coding sequence. We list the mean and standard error of $Z_{\Delta G}$ for all species and all windows in Table S1.

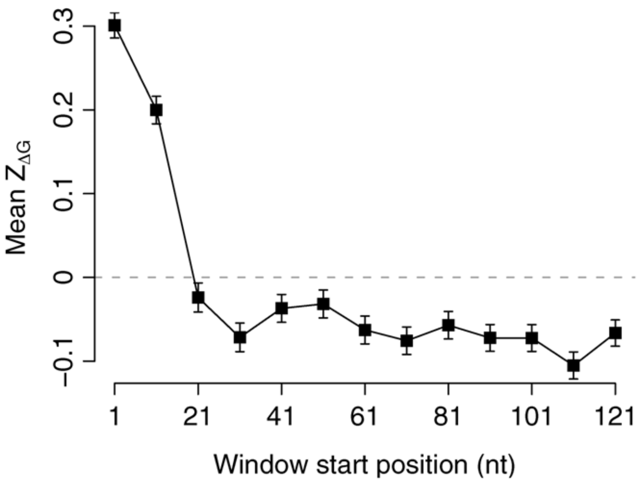

Figure 1. The mean and standard error of $Z_{\Delta G}$ of each sliding window in *E. coli*.
doi:10.1371/journal.pcbi.1000664.g001

To investigate whether window size affected our results, we redid our analysis for four species (two bacteria, one archaeon, one fungus) using sliding windows of 20 nt and 40 nt, respectively. Results for these two window sizes were comparable to those obtained with a window size of 30 nt (Figures S2 and S3). For the same four species, we also recalculated $Z_{\Delta G}$ controlling for dinucleotide content, by using the DicodonShuffle algorithm [41]. The results were virtually unchanged compared to our standard shuffling method. (Figure S4).

Genomic GC composition explains the major variation in $5'$ $Z_{\Delta G}$

We found substantial variation in the mean $5'$ $Z_{\Delta G}$ among different species (Figure 2). Therefore, we next aimed to identify the determinants of $5'$ $Z_{\Delta G}$ in different genomes. We first considered genomic nucleotide content.

We compared the mean $5'$ $Z_{\Delta G}$ in each genome to the genome's GC content in coding sequences. We observed a strong positive correlation between the mean $5'$ $Z_{\Delta G}$ and the genomic GC content (Spearman's $\rho = 0.831$, $P \ll 10^{-20}$) when plants and warm-blooded animals were excluded (Figure 3). Genomes with higher GC content had comparatively less stable mRNA secondary structure at the

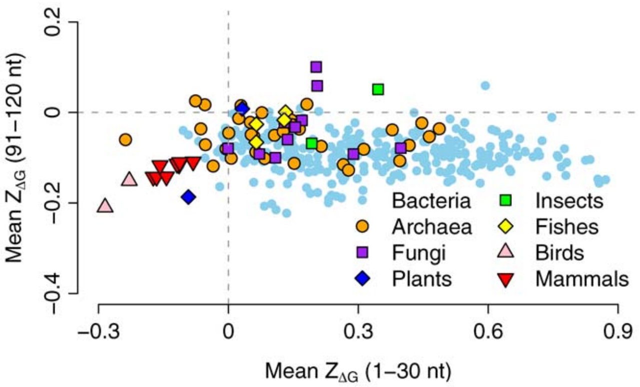

Figure 2. The mean $Z_{\Delta G}$ of the tenth window vs. the mean $Z_{\Delta G}$ of the first window ($5'$ $Z_{\Delta G}$). Each data point represents the entire genome of one organism.
doi:10.1371/journal.pcbi.1000664.g002

translation-initiation region. Since the thermodynamic stability of RNA secondary structure tends to be correlated to the RNA's GC content, we also looked into local deviations in a gene's GC content. We calculated Z_{GC}, which measures the deviation in GC content in a 30 nt window relative to the average in the gene (see Materials and Methods). We found a negative correlation between genomic GC content and the mean Z_{GC} of the first window (Spearman's $\rho = -0.858$, $P \ll 10^{-20}$). Thus, in GC-rich genomes, the sequence regions immediately downstream of the start codon were particularly GC poor (Figure S5).

Because mRNA stability was reduced only near the translation-initiation region, we expected that similarly GC content was reduced only near the start codon. Therefore, the correlation between genomic GC content and mean $Z_{\Delta G}$ should decrease for windows further downstream. We found that indeed the correlation declined continuously and reached approximately zero at the 13th window (Figure S6).

Besides the statistical measure $Z_{\Delta G}$, we also considered ΔG directly (Table S2). We found that the mean ΔG value of the first window ($5'$ ΔG) varied greatly among different species and was largely determined by genomic GC content (Spearman's $\rho = -0.956$, $P \ll 10^{-20}$, Figure S7). As expected, mRNA stability increased with increasing genomic GC content. In fact, we found similar relationships for windows further downstream, but the stability at the $5'$ end of the mRNA was generally lower than the stability further downstream (Figure S8). Moreover, the difference between the mean $5'$ ΔG and the mean ΔG of the downstream windows increased with increasing genomic GC content. As an example, Figure S9 shows the relationship between the mean GC content and the difference in mean ΔG between the first and tenth window (Spearman correlation $\rho = 0.784$, $P \ll 10^{20}$, excluding birds and mammals). In summary, the results for ΔG generally mirrored the ones for $Z_{\Delta G}$.

Optimal growth temperature affects $5'$ $Z_{\Delta G}$ in prokaryotes

For prokaryotes, we analyzed whether the $5'$ $Z_{\Delta G}$ correlated with the optimal growth temperature (Figure 4). We found a significant negative correlation (Spearman's $\rho = -0.365$, $P = 3.0 \times 10^{-4}$). Similarly, the difference in ΔG between the first and the tenth window declined significantly with temperature (Spearman's $\rho = -0.385$, $P = 1.3 \times 10^{-4}$). Thus, prokaryotes living in colder environments tended to have comparatively less stable mRNA secondary structure at the translation-initiation region. We found no correlations between temperature and either genomic GC content (Spearman's $\rho = -0.082$, $P = 0.434$) or the ΔG in the first window (Spearman's $\rho = -0.065$, $P = 0.536$). The lack of a correlation between temperature and genomic GC content agrees with the results of Ref. [42].

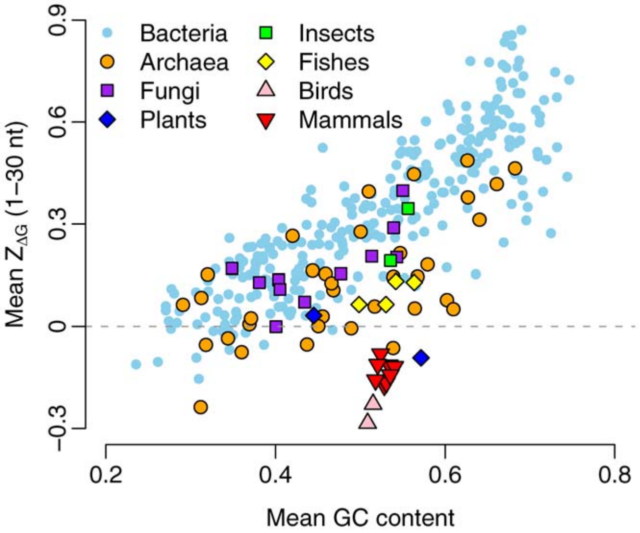

Figure 3. The mean $Z_{\Delta G}$ of the first window as a function of the genomic GC content. Each data point represents one organism.
doi:10.1371/journal.pcbi.1000664.g003

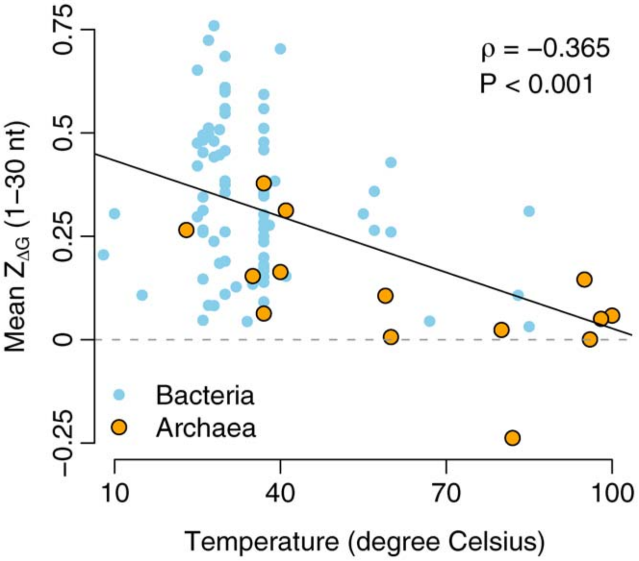

Figure 4. The mean $Z_{\Delta G}$ of the first window as a function of the optimal growth temperature in prokaryotes. Each data point represents one organism.
doi:10.1371/journal.pcbi.1000664.g004

Determinants of $5'$ $Z_{\Delta G}$ within genomes

In the previous subsections, we considered the mean $5'$ $Z_{\Delta G}$ over all genes in a genome. But we expected that there should also be variation in mRNA stability among genes within one genome. Therefore, we next investigated the potential within-genome factors that may affect mRNA stability near the start codon.

We first considered gene GC content. We compared the mean $5'$ $Z_{\Delta G}$ between genes with the highest 5% and the lowest 5% GC content in each species. In almost all genomes, including birds and mammals, the mean $5'$ $Z_{\Delta G}$ in GC-rich genes was higher than it was in GC-poor genes (Figure 5 and Table S3). The differences became weaker as we considered windows further downstream (Figure S10). Interestingly, even though the whole-genome mean $5'$ $Z_{\Delta G}$ was negative in birds and mammals, GC-rich genes in these animals showed a positive $5'$ $Z_{\Delta G}$ (Figure 5).

Next we considered codon usage bias. We used the effective number of codons (*ENC*) to measure the codon usage bias of each gene [43]. Lower *ENC* values indicate stronger bias. By comparing the bottom 5% of genes with the lowest *ENC* to the top 5% of genes with the highest *ENC*, we found that, in most species, genes with stronger codon bias had higher $5'$ $Z_{\Delta G}$ (Figure S11 and Table S3).

Finally, we tested whether the reduction in $5'$ mRNA stability increased with gene expression level. We compared the mean $Z_{\Delta G}$ between genes with the highest 5% and the lowest 5% expression level in *E. coli*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, and *Homo sapiens*. In all species except *H. sapiens*, the mean $Z_{\Delta G}$ for the highest-expressed genes tended to be higher than that for the genes with the lowest expression level (Figure 6).

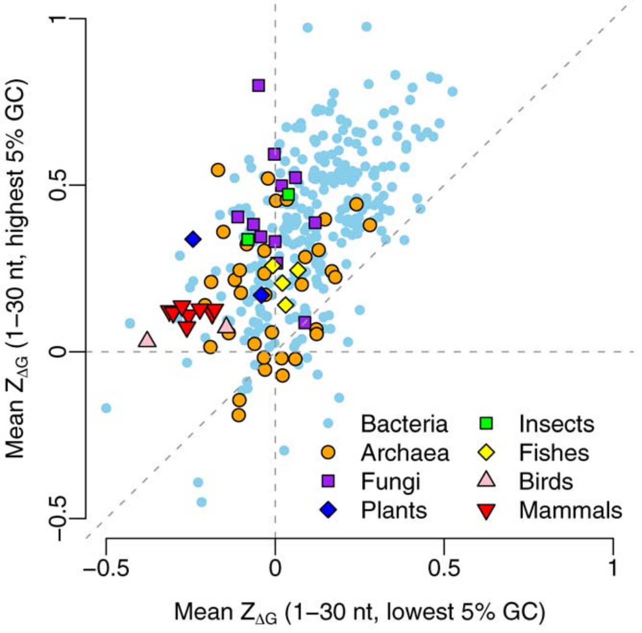

Figure 5. Comparison of the mean 5′ $Z_{\Delta G}$ between genes with the highest 5% and the lowest 5% GC content within each genome. Each data point represents one organism.
doi:10.1371/journal.pcbi.1000664.g005

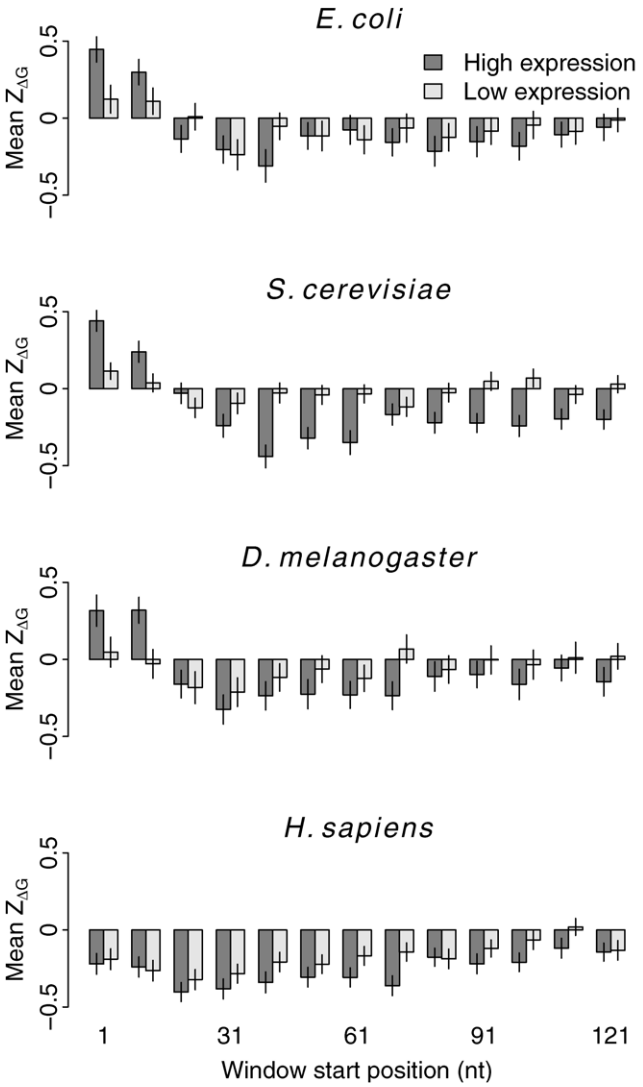

Figure 6. Comparison of the mean 5′ $Z_{\Delta G}$ between genes with the highest 5% and the lowest 5% expression level in *E. coli*, *S. cerevisiae*, *D. melanogaster*, and *H. sapiens*.
doi:10.1371/journal.pcbi.1000664.g006

Since GC content, codon bias, and expression level all correlated with $Z_{\Delta G}$, we tried to determine whether these quantities are independent sources of variation. We carried out a principle component regression [44] and found that GC content, *ENC*, and gene expression level contributed nearly equal to variation in $Z_{\Delta G}$ in *E. coli*, *S. cerevisiae*, and *D. melanogaster* (Figure S12). In human, GC content and *ENC*, but not gene expression level, contributed to the variation.

Discussion

We have completed a broad survey of mRNA stability near the translation-initiation region of protein-coding genes. We have considered the complete genomes of 340 species, including bacteria, archaea, fungi, plants, insects, and vertebrates. We have found a general tendency for reduced mRNA stability in the first 30–40 nt of the coding sequence. In this region, mRNA stability tends to be less than expected given a gene's amino-acid sequence and codon-usage bias. Experimental work had previously suggested that increased local mRNA stability at the translation-initiation region could prevent efficient translation initiation and hence decrease gene expression level [38,40].

We have found that there is variation in the extent to which mRNA stability is reduced both among and within genomes. Among genomes, GC content of coding sequences is a major predictor of the reduction in mRNA stability. The higher the GC content, the larger the reduction in mRNA stability at the 5′ end of the coding sequence (i.e., the larger 5′ $Z_{\Delta G}$). For prokaryotes, the optimal growth temperature also predicts 5′ $Z_{\Delta G}$. The lower the optimal growth temperature, the larger the reduction in mRNA stability. Within genomes, 5′ $Z_{\Delta G}$ also increases with increasing GC content. In addition, it increases with increasing codon usage bias and gene expression level.

The region with reduced mRNA stability is located right downstream from the start codon and has a length of 30 to 40 nt (the first two windows in our analysis). This region is similar to the one identified by Kudla et al. [40]. Kudla et al. studied primarily a library of sequences encoding green fluorescent protein, but they also carried out a computational analysis of mRNA stability across the *E. coli* genome. They found that across the genome, ΔG was significantly more positive (indicating reduced stability) in the region from nt -4 to $+37$ than immediately downstream [40]. Our work shows that Kudla et al.'s observation applies to most organisms with known genomes, including bacteria, archaea, and both single- and multi-celled eukaryotes. Further, by focusing on Z scores relative to the expectation in permuted sequences, our analysis excludes biases such as amino-acid content or preferred-codon usage as the cause of this signal.

Past the first two windows, $Z_{\Delta G}$ decayed quickly towards a negative asymptotic value. Thus, mRNA stability near the start codon is less than expected, but elsewhere in the gene it is generally higher than expected. The latter result is comparable to observations made by Chamary and Hurst [16] in the mouse genome and by Seffens and Digby [15] in individual genes from several species. Interestingly, for many organisms, $Z_{\Delta G}$ in windows

3 to 5 dips below the negative asymptotic value further downstream (Figure S1 and Table S1). This behavior seems to reflect a selection pressure for particularly stable local mRNA structure right after the translation-initiation region. This increased stability may compensate for the reduced mRNA stability in the translation-initiation region.

Previous works have identified AT-biased translation enhancers in prokaryotes [37,45–47] and preferred nucleotide sequences regulating translation in eukaryotes [48] within the first 30 to 40 nt of the coding sequence. The mechanism by which these sequence motifs work is not currently known. We suggest that the primary mechanism may be destabilization of the mRNA structure near the start codon. By contrast, some motifs work by known mechanisms unrelated to RNA secondary structure. For example, alanine is preferred at the second amino-acid position in highly expressed proteins in several organisms [49] and its codon might bind to a complementary sequence in the 18S ribosomal RNA [50].

We found that the higher the GC content of a genome, the more was mRNA stability reduced at the translation-initiation region. This finding makes thermodynamic sense. GC-rich RNAs tend to fold into more stable structures than AU-rich RNAs, simply because a GC pair has three hydrogen bonds whereas an AU pair has only two. Thus, assuming that selection targets the same low 5′ mRNA stability in all organisms, we would expect that the decrease in stability is larger in GC-rich RNAs, simply because they start from a more-stable baseline. Whether selection actually targets the same low 5′ mRNA stability cannot be determined by our analysis. We found that the mean 5′ ΔG increased with increasing GC content. This increase could imply either that organisms with higher GC content can tolerate a higher 5′ mRNA stability or that the selection pressure to reduce 5′ mRNA stability in those organisms is counterbalanced by other selective forces or mutation pressures that increase GC content.

For prokaryotes, we addressed the question whether the optimal growth temperature affects 5′ $Z_{\Delta G}$. Thermodynamics predict that the lower the temperature at which an organism grows, the stronger should mRNA stability interfere with translation initiation. In agreement with this prediction, we found that the optimal growth temperature correlated negatively with 5′ $Z_{\Delta G}$. The organisms growing at the lowest temperatures showed the biggest reduction in mRNA stability at the beginning of the coding sequence. This result was independent of the relationship between 5′ $Z_{\Delta G}$ and genomic GC content. In our data set, the optimal growth temperature was not correlated with GC content. Even though some authors have argued that GC content correlates with temperature [51,52], more recent studies have disputed this finding [42,53]. Our results agree with these more recent studies.

Within individual genomes, the reduction of mRNA stability at the translation-initiation region was greater in GC-rich genes than in GC-poor ones. Besides GC content, we found that codon usage bias and gene expression level correlated with 5′ $Z_{\Delta G}$. Because codon usage bias is correlated with gene expression level, in particular in fast-growing microbes [2,3,8,11], these two correlations likely reflect the same underlying effect. The correlation with expression level mirrors the general observation that evolutionary constraints tend to increase with gene expression level [8,11,13,54–58]. Whether expression level, codon usage bias, and GC content contribute independently to 5′ $Z_{\Delta G}$ is unclear. These three quantities tend to all be correlated with each other, and we cannot easily disentangle which of these quantities is most important for reduced 5′ $Z_{\Delta G}$. For example, in mammals, high GC content in genes can increase mRNA levels through increased efficiency of transcription or mRNA processing [59]. Using principal component regression, we showed that in *E. coli*, yeast, and fly, the three quantities codon usage bias, GC content, and gene expression level all contribute equally to reduced 5′ $Z_{\Delta G}$, whereas in humans only GC content and codon-usage bias seem to contribute.

We found reduced mRNA stability near the start codon in a wide range of organisms, including both prokaryotes and eukaryotes. Yet warm-blooded animals (birds and mammals) showed no such trend on the whole-genome level, even though their genomic GC content is well within the range in which we found reduced mRNA stability in bacteria, archaea, fungi, insects, and fishes. We believe that our finding for birds and mammals was caused by the isochore structure of their genomes [60]. Gene GC content in these organisms ranges from 20% to 95% and is much more varied than in organisms without isochores. The whole-genome average of 5′ $Z_{\Delta G}$ may not be meaningful in organisms with isochores. When we considered only to top 5% most GC-rich genes, we did find a moderate signal of reduced mRNA stability in these organisms as well.

What is the biological mechanism that links mRNA stability near the start codon to efficient protein translation? There are two possibilities. First, strong local mRNA secondary structure could interfere with ribosome binding. Second, it could interfere with start-codon recognition. We believe that the currently available evidence favors the latter explanation. In prokaryotes, ribosome binding occurs at the Shine-Dalgarno sequence, located a few nucleotides upstream from the start codon [28]. Kudla *et al.* [40] showed that synonymous mutations near the start codon can regulate protein expression. They concluded from computational modeling that the primary determinant of protein expression was the stability of local mRNA secondary structure near the start codon, not occlusion of the Shine-Dalgarno sequence by RNA secondary structure [40]. In eukaryotes, translation initiation follows a scanning mechanism. The 40S ribosomal subunit enters at the 5′ end of the mRNA and migrates linearly until it encounters the first AUG codon [30]. If synonymous mutations near the start codon could affect ribosome entry at the 5′ cap, there should be a correlation between 5′ UTR length and mRNA stability near the start codon. The further away the start codon is from the 5′ cap, the less should local mRNA stability near the start codon affect ribosome entry. However, we did not find such a relationship, neither within genomes nor among genomes (data not shown). Therefore, we suggest that both in prokaryotes and in eukaryotes, reduced mRNA stability at the translation-initiation region primarily facilitates efficient start-codon recognition.

Materials and Methods

Genomic data

We collected the genomes for 276 bacteria, 35 archaea, 11 fungi, 2 plants, 2 insects, 4 fishes, 2 birds, and 8 mammals. The genomic sequences of the bacteria, archaea, fungi, plants, and insects were downloaded from the NCBI FTP server (ftp://ftp.ncbi.nih.gov/), while the sequences of the vertebrates were obtained from Ensembl (http://www.ensembl.org/). We only considered coding sequences longer than 50 codons.

Expression data

We collected previously published expression data for four species: for *E. coli*, we obtained gene expression levels measured in mRNAs per cell from Ref. [61]; for *S. cerevisiae*, we used expression data from Ref. [62]; for *D. melanogaster*, we used as expression level the geometric mean of expression data from different tissues obtained in Ref. [63]; and for *H. sapiens*, we also measured expression level as the geometric mean of expression among different tissues [64].

Optimal growth temperature

We obtained optimal growth temperature data for 80 bacteria and 14 archaea from Ref. [42], which is a collection from multiple sources, including original publications, American Type Culture Collection, German Collection of Microorganisms and Cell Cultures, and Prokaryotic Growth Temperature Database.

RNA secondary structure folding

We calculated RNA folding energies using the RNAfold program in the Vienna package [65,66]. We used default settings: folding occurred at 37°C; GU pairs were allowed; unpaired bases could participate in at most one dangling end; energy parameters were as reported in Ref. [67]. We evaluated only the minimum-free-energy structure. ΔG is the change in free energy from the unfolded state to this structure.

mRNA randomization

If synonymous selection acts on mRNA folding near the start codon, then on average the secondary structure in this region should be less stable for the naturally occurring sequence than for permuted sequences. For each gene, we randomly reshuffled synonymous codons among sites with identical amino acids, to control for amino-acid sequence, codon usage bias, and GC content. We repreated this process 1000 times to obtain 1000 permuted sequences for each gene. For the wild-type sequence and each permuted sequence, we then calculated local mRNA folding energies in a sliding window of 30 nt (20 nt and 40 nt were also used in some species). To determine the deviation of the wild-type sequence from the permuted ones, we calculated the Z-score of the local mRNA stability ($Z_{\Delta G}$) for each sliding window by:

$$Z_{\Delta G} = \frac{\Delta G_{\mathrm{N}} - \overline{\Delta G_{\mathrm{P}}}}{\sqrt{\sum_{i=1}^{n} \frac{(\Delta G_{\mathrm{P}i} - \overline{\Delta G_{\mathrm{P}}})^2}{n-1}}}. \tag{1}$$

Here, ΔG_{N} is the folding free energy for the naturally occurring sequence in the window under consideration, ΔG_{P_i} is the folding energy of the corresponding window of the i^{th} permuted sequence, and $\overline{\Delta G_{\mathrm{P}}}$ is the mean of $\Delta G_{\mathrm{P}i}$ over all permuted sequences. The variable n represents the total number of permuted sequences. Here, $n = 1000$.

Similarly, we evaluated the difference between the local mRNA GC composition of the wild-type sequence and the permuted sequences. The Z-score of local mRNA GC content (Z_{GC}) for each window can be expressed as:

$$Z_{\mathrm{GC}} = \frac{GC_{\mathrm{N}} - \overline{GC_{\mathrm{P}}}}{\sqrt{\sum_{i=1}^{n} \frac{(GC_{\mathrm{P}i} - \overline{GC_{\mathrm{P}}})^2}{n-1}}}. \tag{2}$$

The definitions for GC_{N}, $GC_{\mathrm{P}i}$, and $\overline{GC_{\mathrm{P}}}$ are analogous to ΔG_{N}, $\Delta G_{\mathrm{P}i}$, and $\overline{\Delta G_{\mathrm{P}}}$ but refer to GC content rather than to free energy of folding.

Supporting Information

Table S1 Mean and standard error of $Z_{\Delta G}$ for each window.
Found at: doi:10.1371/journal.pcbi.1000664.s001 (0.08 MB PDF)

Table S2 Mean and standard error of ΔG for each window.
Found at: doi:10.1371/journal.pcbi.1000664.s002 (0.08 MB PDF)

Table S3 Intra-genome comparison on mean $Z_{\Delta G}$ of the first window.
Found at: doi:10.1371/journal.pcbi.1000664.s003 (0.06 MB PDF)

Figure S1 Distribution of the mean $Z_{\Delta G}$ as a function of window start position in different groups of species.
Found at: doi:10.1371/journal.pcbi.1000664.s004 (0.05 MB EPS)

Figure S2 The mean and standard error of $Z_{\Delta G}$ of each sliding window (20 nt) in four species.
Found at: doi:10.1371/journal.pcbi.1000664.s005 (0.03 MB EPS)

Figure S3 The mean and standard error of $Z_{\Delta G}$ of each sliding window (40 nt) in four species.
Found at: doi:10.1371/journal.pcbi.1000664.s006 (0.03 MB EPS)

Figure S4 Comparison of the mean $Z_{\Delta G}$ obtained by different mRNA randomization procedures. Codon Shuffle algorithm permutes sequences by randomly reshuffling synonymous codons within each gene, which may change dinucleotide composition of the sequence; while DicodonShuffle algorithm shuffle codons preserving the dinucleotide composition.
Found at: doi:10.1371/journal.pcbi.1000664.s007 (0.05 MB EPS)

Figure S5 The mean Z_{GC} of the first window as a function of the genomic GC content. Each data point represents one organism.
Found at: doi:10.1371/journal.pcbi.1000664.s008 (0.14 MB EPS)

Figure S6 Spearman correlation coefficient ρ between the mean $Z_{\Delta G}$ and the genomic GC content for each window in bacteria, archaea, and fungi.
Found at: doi:10.1371/journal.pcbi.1000664.s009 (0.02 MB EPS)

Figure S7 The mean ΔG of the first window as a function of the genomic GC content. Each data point represents one organism.
Found at: doi:10.1371/journal.pcbi.1000664.s010 (0.14 MB EPS)

Figure S8 The mean ΔG of the tenth window vs. the mean ΔG of the first window. Each data point represents the entire genome of one organism.
Found at: doi:10.1371/journal.pcbi.1000664.s011 (0.14 MB EPS)

Figure S9 The difference in mean ΔG between the first and the tenth window as a function of the genomic GC content. Each data point represents one organism.
Found at: doi:10.1371/journal.pcbi.1000664.s012 (0.14 MB EPS)

Figure S10 Distribution of the difference in mean $Z_{\Delta G}$ between genes with the highest 5% and the lowest 5% GC content for the entire data set of all 340 species.
Found at: doi:10.1371/journal.pcbi.1000664.s013 (0.11 MB EPS)

Figure S11 Comparison of the mean 5′ $Z_{\Delta G}$ between genes with the highest 5% and the lowest 5% *ENC* within each genome. Each data point represents one organism.
Found at: doi:10.1371/journal.pcbi.1000664.s014 (0.14 MB EPS)

Figure S12 Principal component regression of 5′ $Z_{\Delta G}$ against GC content, expression level, and *ENC* value. PC1, PC2, and PC3 denote the first, second, and third principal component.
Found at: doi:10.1371/journal.pcbi.1000664.s015 (0.02 MB EPS)

Author Contributions

Conceived and designed the experiments: WG TZ COW. Performed the experiments: WG TZ. Analyzed the data: WG TZ COW. Contributed reagents/materials/analysis tools: WG TZ. Wrote the paper: WG TZ COW.

References

1. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32–43.
2. Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2: 13–34.
3. Sharp PM, Tuohy T, Mosurski K (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14: 5125–5143.
4. Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. Genetics 136: 927–935.
5. Stenico M, Lloyd AT, Sharp PM (1994) Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucl Acids Res 22: 2437–2446.
6. Akashi H, Eyre-Walker (1998) Translational selection and molecular evolution. Curr Opin Genet Dev 8: 688–693.
7. Duret L (2002) Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 12: 640–649.
8. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23: 327–337.
9. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7: 98–108.
10. Stoletzki N, Eyre-Walker A (2007) Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol Biol Evol 24: 374–381.
11. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352.
12. Higgs PG, Ran W (2008) Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol 25: 2279–2291.
13. Zhou T, Weems M, Wilke CO (2009) Translationally optimal codons associate with structurally sensitive sites in proteins. Mol Biol Evol 26: 1571–1580.
14. Vinogradov AE (2003) DNA helix: the importance of being GC-rich. Nucleic Acids Res 31: 1838–1844.
15. Seffens W, Digby D (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res 27: 1578–1584.
16. Chamary JV, Hurst LD (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol 6: R75.
17. Hoede C, Denamur E, Tenaillon O (2006) Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. PLoS Genetics 2: e176.
18. Stoletzki N (2008) Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. BMC Evol Biol 8: 224.
19. Parmley JL, Chamary JV, Hurst L (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol 23: 301–309.
20. Parmley JL, Hurst LD (2007) Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. Mol Biol Evol 24: 1600–1603.
21. Warnecke T, Hurst LD (2007) Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. Mol Biol Evol 24: 2755–2762.
22. Thanaraj TA, Argos P (1996) Ribosome-mediated translational pause and protein domain organization. Protein Sci 5: 1594–1612.
23. Komar AA, Lesnik T, Reiss C (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Lett 462: 387–391.
24. Cortazzo P, Cervenansky C, Marin M, Reiss C, Ehrlich R, et al. (2002) Silent mutations affect in vivo protein folding in *Escherichia coli*. Biochem Biophys Res Commun 293: 537–541.
25. Goymer P (2007) Synonymous mutations break their silence. Nat Rev Genet 8: 92.
26. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, et al. (2007) A "silent" polymorphism in the mdr1 gene changes substrate specificity. Science 315: 525–528.
27. Zhang G, Hubalewska M, Ignatova Z (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nature Struct Mol Biol 16: 274–280.
28. Shine J, Dalgarno L (1975) Determinant of cistron specificity in bacterial ribosomes. Nature 254: 34–38.
29. Kozak M (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res 15: 8125–8148.
30. Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 361: 13–39.
31. Yamagishi K, Oshima T, Masuda Y, Ara T, Kanaya S, et al. (2002) Conservation of translation initiation sites based on dinucleotide frequency and codon usage in *Escherichia coli* K-12 (W3110): non-random distribution of A/T-rich sequences immediately upstream of the translation initiation codon. DNA Res 9: 19–24.
32. Shabalina SA, Ogurtsov AY, Rogozin IB, Koonin EV, Lipman DJ (2004) Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. Nucleic Acids Res 32: 1774–1782.
33. Komarova AV, Tchufistova LS, Dreyfus M, Boni IV (2005) AU-rich sequences within 5′ untranslated leaders enhance translation and stabilize mRNA in *Escherichia coli*. J Bacteriol 187: 1344–1349.
34. Vimberg V, Tats A, Remm M, Tenson T (2007) Translation initiation region sequence preferences in *Escherichia coli*. BMC Genomics 8: 100.
35. Zalucki YM, Power PM, Jennings MP (2007) Selection for efficient translation initiation biases codon usage at second amino acid position in secretory proteins. Nucleic Acids Res 35: 5748–5754.
36. Chen H, Pomeroy-Cloney L, Bjerknes M, Tam J, Jay E (1994) The influence of adenine-rich motifs in the 3′ portion of the ribosome binding site on human IFN-gamma gene expression in *Escherichia coli*. J Mol Biol 240: 20–27.
37. Qing G, Xia B, Inouye M (2003) Enhancement of translation initiation by A/T-rich sequences downstream of the initiation codon in *Escherichia coli*. J Mol Microbiol Biotechnol 6: 133–144.
38. Griswold KE, Mahmood NA, Iverson BL, Georgiou G (2003) Effects of codon usage versus putative 5′-mRNA structure on the expression of *Fusarium solani* cutinase in the *Escherichia coli* cytoplasm. Protein Expres Purif 27: 134–142.
39. Gonzalez de Valdivia EI, Isaksson LA (2004) A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli*. Nucl Acids Res 32: 5198–5205.
40. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. Science 324: 255–258.
41. Katz L, Burge CB (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res 13: 2042–2051.
42. Zeldovich KB, Berezovsky IN, Shakhnovich EI (2007) Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol 3: e5.
43. Wright F (1990) The 'effective number of codons' used in a gene. Gene 87: 23–29.
44. Mandel J (1982) Use of the singular value decomposition in regression analysis. American Stat 36: 15–24.
45. Etchegaray JP, Inouye M (1999) Translational enhancement by an element downstream of the initiation codon in *Escherichia coli*. J Biol Chem 274: 10079–10085.
46. Stenstrom CM, Isaksson LA (2002) Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3′ side. Gene 288: 1–8.
47. Brock JE, Paz RL, Cottle P, Janssen GR (2007) Naturally occurring adenines within mRNA coding sequences affect ribosome binding and expression in *Escherichia coli*. J Bacteriol 189: 501–510.
48. Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K (2008) Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. Nucleic Acids Res 36: 861–871.
49. Tats A, Remm M, Tenson T (2006) Highly expressed proteins have an increased frequency of alanine in the second amino acid position. BMC Genomics 7: 28.
50. Sanchez J (2008) Alanine is the main second amino acid in vertebrate proteins and its coding entails increased use of the rare codon GCG. Biochem Biophys Res Commun 373: 589–592.
51. Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 44: 632–636.
52. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, et al. (2004) Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett 573: 73–77.
53. Wang HC, Susko E, Roger AJ (2006) On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors. Biochem Biophys Res Commun 342: 681–684.
54. Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol 17: 68–74.
55. Pal C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158: 927–931.
56. Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL (2005) Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol 22: 1345–1354.
57. Wolf YI, Carmel L, Koonin EV (2006) Unifying measures of gene function and evolution. Proc R Soc B 273: 1507–1515.
58. Eames M, Kortemme T (2007) Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mrna level and protein abundance. Structure 15: 1442–1451.
59. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol 4: e180.
60. Eyre-Walker A, Hurst LD (2001) The evolution of isochores. Nat Rev Genet 2: 549–555.
61. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429: 92–96.
62. Holstege FCP, Jennings E, Wyrick JJ, Lee TI, Hengartner CJ, et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95: 717–728.
63. Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, et al. (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science 306: 655–660.
64. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101: 6062–6067.

65. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, et al. (1994) Fast folding and comparison of RNA secondary structures. Monatshefte f Chemie 125: 167–188.

66. Hofacker IL, Stadler PF (2006) Memory efficient folding algorithms for circular RNA secondary structures. Bioinformatics 22: 1172–1176.

67. Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288: 911–940.