BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Reduced stability of mRNA secondary structure near the translation-initiation site in dsDNA viruses

Tong Zhou[1] and Claus O Wilke[2*]

Abstract

Background: Recent studies have demonstrated a selection pressure for reduced mRNA secondary-structure stability near the start codon of coding sequences. This selection pressure can be observed in bacteria, archaea, and eukaryotes, and is likely caused by the requirement of efficient translation initiation in cellular organism.

Results: Here, we surveyed the complete genomes of 650 dsDNA virus strains for signals of reduced stability of mRNA secondary structure near the start codon. Our analysis included viruses infecting eukaryotic, prokaryotic, and archaeic hosts. We found that many viruses showed evidence for reduced mRNA secondary-structure stability near the start codon. The effect was most pronounced in viruses infecting prokaryotes, but was also observed in viruses infecting eukaryotes and archaea. The reduction in stability generally increased with increasing genomic GC content. For bacteriophage, the reduction was correlated with a corresponding reduction of stability in the phage hosts.

Conclusions: We conclude that reduced stability of the mRNA secondary structure near the start codon is a common feature for dsDNA viruses, likely driven by the same selective pressures that cause it in cellular organisms.

Background

Translation initiation is facilitated by specific nucleotide patterns near the start codon. Upstream of the start codon, sequence features such as the Shine-Dalgarno sequence (in prokaryotes) and the Kozak sequence (in eukaryotes) prime the ribosome to initiate translation [1-7]. Downstream of the start codon, various sequence features promote translation initiation. For example, in *Escherichia coli*, the codon AAA seems to enhance translation initiation [8]. More generally, translation initiation is enhanced if the mRNA downstream of the start codon is AT-rich and does not form a stable secondary structure [9-13].

Experimental and computational work in *E. coli* showed that gene expression levels are correlated to the thermodynamic stability of mRNA secondary structure near the start codon—lower stability implied higher protein abundance [13]. Recent computational studies have shown that the secondary-structure stability of mRNA segments near the start codon is on average lower than expected [14,15]. Tuller et al. found that, in both *E. coli* and *Saccharomyces cerevisiae*, mRNA secondary-structure stability is reduced at the beginning of ORFs [14]. A more comprehensive study by Gu et al. demonstrated that this reduction in stability occurs in most cellular organisms, including bacteria, archaea, fungi, plants, insects, and fishes [15]. The reduction in stability generally increased with increasing genomic GC content. In birds and mammals, the pattern was not found genome-wide but did occur in the most GC-rich genes.

Here, we extended the analysis of Gu et al. [15] to dsDNA viruses. We analyzed the local mRNA secondary structure at the 5' end of the coding region in 650 dsDNA virus strains. We used computational methods to predict the thermodynamic stability of local mRNA secondary structure in sliding windows downstream from the start codon, as described [15]. We addressed the following questions: (i) Is there a selection pressure

* Correspondence: wilke@austin.utexas.edu
[2]Center for Computational Biology and Bioinformatics, Institute for Cell and Molecular Biology, and Section of Integrative Biology, The University of Texas at Austin, Austin, TX 78712, USA
Full list of author information is available at the end of the article

on synonymous sites to reduce the stability of local mRNA secondary structure at the translation-initiation region in dsDNA viruses? (ii) Are overlapping open reading frames confounding the results? (iii) Does 5' mRNA stability correlate with GC composition? (iv) Does the selection pressure depend on the kingdom of the host organism? (v) Does the selection pressure correlate with other host properties, such as the host's GC content?

Results

Reduced mRNA stability at the translation-initiation region in viral genomes

We performed a sliding-window analysis of mRNA secondary-structure stability in 650 fully sequenced dsDNA viruses. For each ORF in each virus, we calculated Z scores $Z_{\Delta G}$. $Z_{\Delta G}$ measures to what extent mRNA secondary-structure stability deviates from random expectation given the amino-acid sequence and codon composition of the ORF [15]. A $Z_{\Delta G} > 0$ indicates that the structure is less stable than expected, and a $Z_{\Delta G} < 0$ indicates the opposite. We calculated $Z_{\Delta G} > 0$ for windows of length 30 nucleotides (nt), and we covered the first 150 nt of each ORF in steps of 10 nt, as described [15]. For each window, we then averaged $Z_{\Delta G}$ over all ORFs in each genome. We refer to this genome-wide average as $\bar{Z}_{\Delta G}$. Below, we test whether this genome-wide average is significantly different from zero. Note that the genome-wide average can be significantly non-zero even if the Z scores for individual genes are relatively small and would individually not be considered significant.

Figure 1 shows $\bar{Z}_{\Delta G}$ as a function of window position for bacteriophage T7. In this virus, $\bar{Z}_{\Delta G}$ is significantly larger than zero in the first window (t-test, $\bar{Z}_{\Delta G} = 0.35$, $P = 0.005$), and it is not significantly different from zero in windows further downstream. We carried out a similar analysis on all virus strains in our data set. As in cellular organisms, there was substantial variation in the $\bar{Z}_{\Delta G}$ of the first window among different virus strains and somewhat less variation in windows further downstream (Figure 2). Allowing for a false-discovery rate of 5%, we found 181 dsDNA viruses (28%) whose $\bar{Z}_{\Delta G}$ in the first window was significantly non-zero. With the exception of 3 viruses infecting eukaryotes (epiphyas postvittana nucleopolyhedrovirus, cowpox virus, and canarypox virus), $\bar{Z}_{\Delta G}$ was positive in all these cases. For windows further downstream, the number of virus strains with significantly non-zero $\bar{Z}_{\Delta G}$ declined rapidly, and $\bar{Z}_{\Delta G}$ tended to be negative rather than positive (Figures 2 and 3). These results mirror the results of Gu et al. [15], who found that $\bar{Z}_{\Delta G}$ was generally positive near the start codon and negative further downstream.

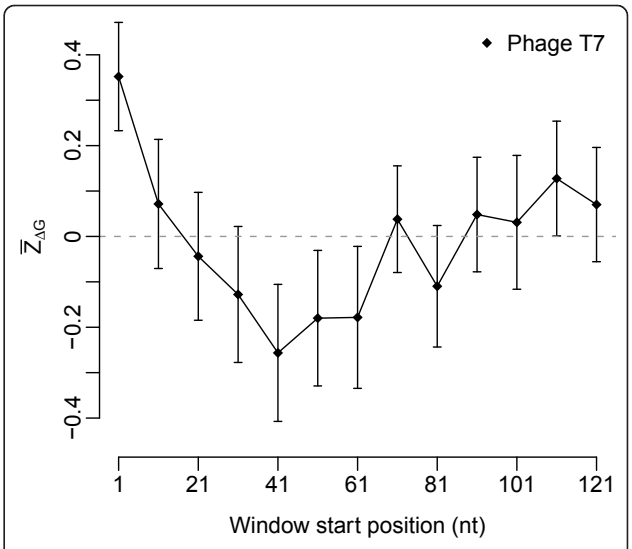

Figure 1 Genome-wide average $\bar{Z}_{\Delta G}$ for Enterobacteria phage T7, as a function of the location of the analysis window. A value of $\bar{Z}_{\Delta G} > 0$ indicates less stable mRNA secondary structure than expected. Error bars indicate the standard error. The value of $\bar{Z}_{\Delta G}$ in the first window is significantly different from zero ($P = 0.005$), the other ones are not.

The main difference in the virus data set is that virus genomes tend to be small, and the error estimates on $\bar{Z}_{\Delta G}$ are consequently large. (E.g., compare Figure 1 of the present study to Figure 1 of [15].)

In aggregate, $\bar{Z}_{\Delta G}$ values for eukaryotic and archaeic viruses were lower than those for prokaryotic viruses.

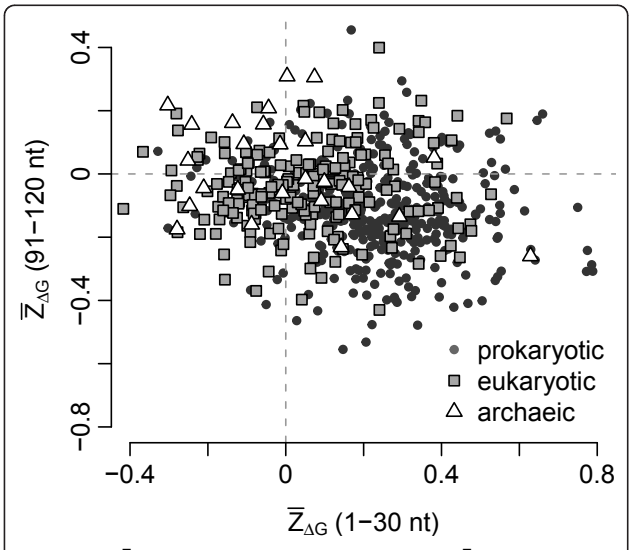

Figure 2 $\bar{Z}_{\Delta G}$ in the 10th window versus $\bar{Z}_{\Delta G}$ in the 1st window. There is a tendency for points to fall into the bottom-right quadrant, indicating decreased mRNA secondary-structure stability near the start codon and increased stability further downstream. However, the effect is much less pronounced than in cellular organisms.

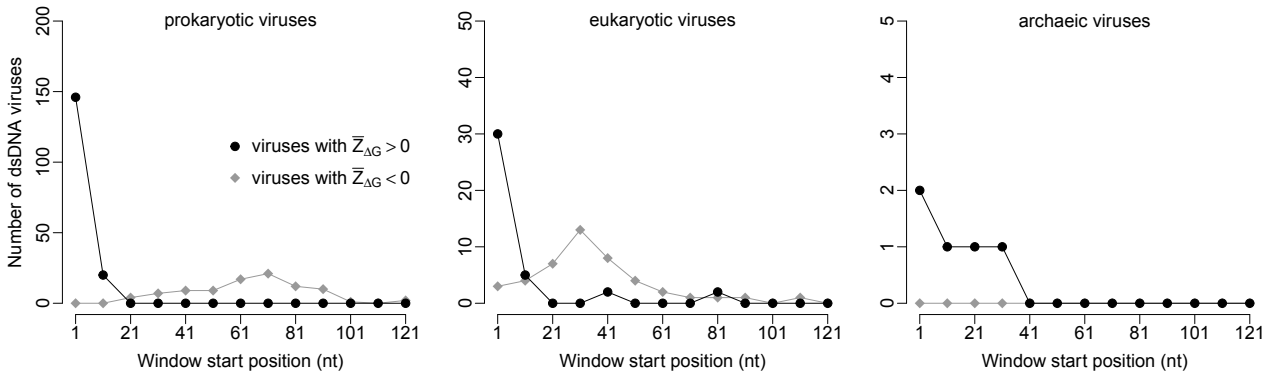

Figure 3 Number of dsDNA virus strains with significantly non-zero $\bar{Z}_{\Delta G}$ after false-discovery-rate correction for multiple testing, as a function of the location of the analysis window. The analysis was carried out on 419 dsDNA viruses with prokaryotic hosts, 204 with eukaryotic hosts, and 27 with archaeic hosts. A value of $\bar{Z}_{\Delta G} > 0$ indicates less stable mRNA secondary structure than expected.

On average, archaeic viruses had a $\bar{Z}_{\Delta G}$ of 0.0049, not significantly different from zero (*t*-test, $P = 0.91$). Eukaryotic viruses had an average of 0.057, which was significantly different from zero (*t*-test, $P = 3.2 \times 10^{-5}$). Prokaryotic viruses had an average of 0.22, also significantly different from zero (*t*-test, $P < 10^{-10}$). The $\bar{Z}_{\Delta G}$ distribution for prokaryotic viruses was significantly different from those for eukaryotic viruses (*t*-test, $P < 10^{-10}$) and archaeic viruses (*t*-test, $P = 2.1 \times 10^{-5}$).

Since we obtained $Z_{\Delta G}$ by shuffling codons within genes, we implicitly assumed that there is no sub-stantial, site-specific selection on synonymous sites outside the focal analysis window. This assumption is violated in regions where reading frames overlap and synonymous sites are primarily determined by the amino-acid sequence of the overlapping reading frame. Therefore, we tested whether our $Z_{\Delta G}$ values were confounded by overlapping reading frames. For all ORFs in all virus genomes, we determined whether they overlapped with any other ORFs and classified them into overlapping and non-overlapping ORFs. We found that on average 50% of the ORFs in a virus genome were overlapping, with a standard deviation of 15.9 percentage points. We then tested for each window in each virus genome whether the mean $Z_{\Delta G}$ for overlapping genes was different from the mean $Z_{\Delta G}$ for non-overlapping genes, using *t* tests. We found that this was generally not the case. Allowing for a false-discovery rate of 5%, not a single virus genome showed a significant difference between overlapping and non-overlapping ORFs in the first four windows. Over all 13 windows, there were only two cases were we could reject the null hypothesis of no difference, invertebrate iridescent virus 3 in window 5 and clanis bilineata nucleopolyhedrosis virus in window 11.

However, when pooling data from all viruses into a single analysis, we found a small shift towards lower $\bar{Z}_{\Delta G}$ values for overlapping ORFs in eukaryotic and archaeic (but not prokaryotic) viruses (Additional File 1 Figure S1). We concluded that there was no evidence that overlap influences $\bar{Z}_{\Delta G}$ in prokaryotic viruses, and weak evidence that it does so in the other two types of viruses. Since gene overlap certainly does not increase $\bar{Z}_{\Delta G}$, we concluded that including both overlapping and non-overlapping ORFs in our analysis was a conservative approach. Therefore, we freely mixed overlapping and non-overlapping ORFs throughout our analysis. The reduced $\bar{Z}_{\Delta G}$ in eukaryotic and archaeic viruses compared to prokaryotic viruses was not due to the inclusion of overlapping ORFs; $\bar{Z}_{\Delta G}$ values for all virus groups were nearly identical regardless of whether overlapping ORFs were included or not (not shown).

For a few select virus species, we also tested whether our results were confounded by dinucleotide frequencies. We calculated alternate $Z_{\Delta G}$ values using reshuffled sequences in which all dinucleotide frequencies had been held constant. We found that our standard shuffling method and the dinucleotide shuffling method resulted in nearly identical *Z* scores (Additional File 1 Figure S2). Note that in the dinucleotide shuffling method, we shuffled synonymous codons such that both amino-acid sequence and dinucleotide frequencies were held constant. The algorithm to perform this shuffling is fairly computationally expensive and runs approximately 24 times slower than regular codon shuffling.

Relationship between genomic GC composition and mean 5' $Z_{\Delta G}$

For the remainder of this work, we refer to the $\bar{Z}_{\Delta G}$ at the very start of the coding sequence (in sliding window #1) as the 5' $\bar{Z}_{\Delta G}$. To explain the variation observed in the 5' $\bar{Z}_{\Delta G}$, we correlated it with the mean GC content in coding sequences, since this quantity is a good predictor of the 5' $\bar{Z}_{\Delta G}$ in cellular organisms [15].

Because different virus strains are evolutionarily related, a relationship between 5' $\bar{Z}_{\Delta G}$ and GC content may be confounded by the viral phylogeny [16]. We can avoid this issue by correlating phylogenetically independent contrasts (PIC), which are differences of variables among organisms [16]. We found that the PIC of the 5' $\bar{Z}_{\Delta G}$ were well correlated with the PIC of the GC content in coding sequences (r = 0.53, $P = 10^{-31}$ for prokaryotic viruses, r = 0.54, $P = 10^{-17}$ for eukaryotic viruses, and r = 0.49, P = 0.009 for archaeic viruses, see also Figure 4). Genomes with higher GC content had comparatively less stable mRNA secondary structure near the start codon.

Because mRNA stability was reduced only at the translation-initiation region, we expected that the correlation between PIC of $\bar{Z}_{\Delta G}$ and PIC of GC content should decrease when $\bar{Z}_{\Delta G}$ was caluclated for windows further downstream. Thus, we calculated the corresponding correlation coefficient for all windows. We found that indeed the correlation declined continuously and was consistently near zero (for eukaryotic viruses) or negative (for prokaryotic or archaeic viruses) from the 5th window onwards (Additional File 1 Figure S3).

Since the thermodynamic stability of RNA secondary structure tends to be correlated to the RNA's GC content, we also considered 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 GC content in the gene (see Materials and Methods). We found a negative correlation between PIC of genomic GC content and PIC of $\bar{Z}_{GC}$ in the first window (r = -0.67, $P = 10^{-84}$ for prokaryotic viruses, r = -0.68, $P = 10^{-58}$ for eukaryotic viruses, and r = -0.74, $P = 10^{-35}$ for archaeic viruses, see also Additional File 1 Figure S4). Thus, in GC-rich viruses, the sequence regions immediately downstream of the start codon have undergone stronger GC reduction.

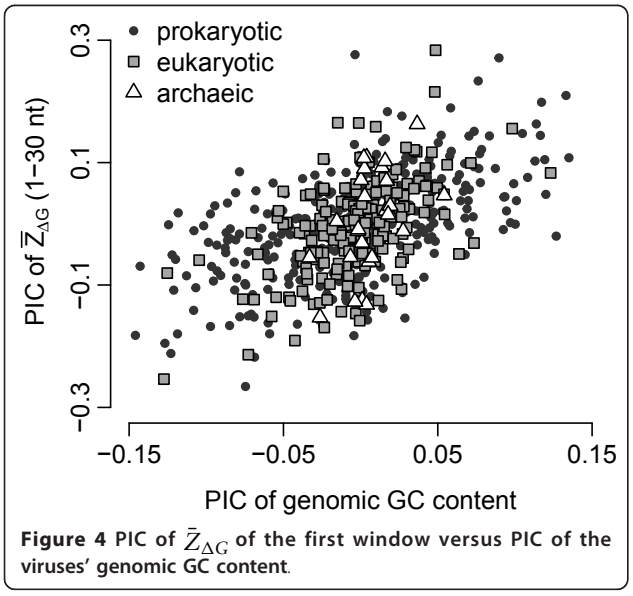

Figure 4 PIC of $\bar{Z}_{\Delta G}$ of the first window versus PIC of the viruses' genomic GC content.

We also analyzed to what extent ΔG (rather than its deviation from expectation, as measured by $Z_{\Delta G}$) varied with GC content. Our results mirrored those we had previously found for cellular organisms [15]. There was a strong negative correlation between the mean ΔG and GC content (Additional File 1 Figure S5). The higher the GC content, the more stable the mRNA secondary structure, even in the first window. We also tested for a correlation between GC content and the difference in stability between the first window and the tenth window, and found that this difference increased in viruses infecting prokaryotic or eukaryotic hosts, but not in those infecting archaeic hosts (r = 0.68, $P < 10^{-15}$ for prokaryotic viruses, r = 0.37, $P = 10^{-7}$ for eukaryotic viruses, and r = 0.37, P = 0.06 for archaeic viruses, see also Additional File 1 Figure S6). These results remained unchanged when correcting for phylogeny (not shown).

Host-specific patterns in bacteriophages

Finally, we asked to what extent sequence features in viruses correlated with corresponding features in their hosts. Previous work has shown that synonymous codon usage in bacteriophages exhibits significant bias towards host-preferred codons [17], and that genomic GC content in some phages is close to the genomic GC content of their host organisms [18,19]. Thus, we would expect more generally that both GC content and 5' $\bar{Z}_{\Delta G}$ in viruses correlate with the same quantities in the appropriate host organisms.

For all bacteriophages in our data set, we identified the corresponding host organism based on the information provided by RefSeq. We then compared GC content in phages and hosts. We found that the GC content in phages correlated strongly with the GC content of the host (correlation coefficient r = 0.89, $P \ll 10^{-15}$, Figure 5). More specifically, Figure 5 suggests that the phage's GC content places a lower limit on the GC content of host organisms the phages can infect. Nearly all data points fall below the dashed line indicating identical GC content for phages and hosts. Moreover, some phages with low GC content are associated with hosts with high GC content, but phages with high GC content are never associated with hosts with low GC content.

The correlation between phage and host GC content may, however, be confounded by phylogeny, as explained in the previous subsection. One complication here is that the phylogeny of phages is not necessarily the same as the phylogeny of the hosts. We are not aware of any method that can correctly compare two data sets with distinct covariance structures. Therefore, we opted for two strategies. On the one hand, we

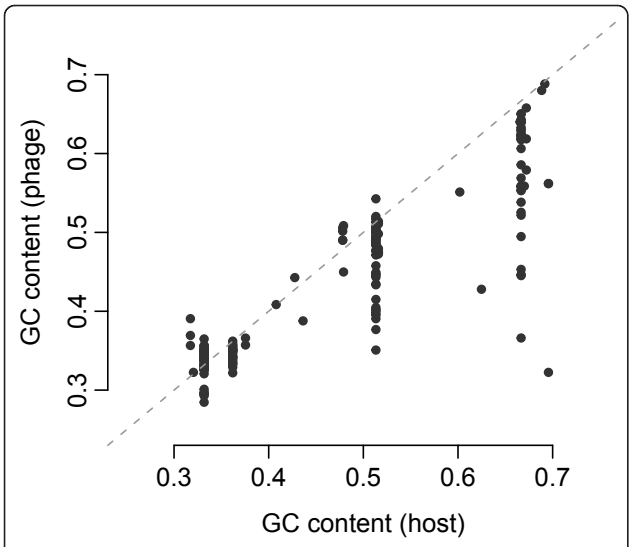

Figure 5 Mean GC content of coding sequences in bacteriophages versus the same quantity in the associated host organisms. The dashed line indicates positions of indentical GC content for phages and hosts.

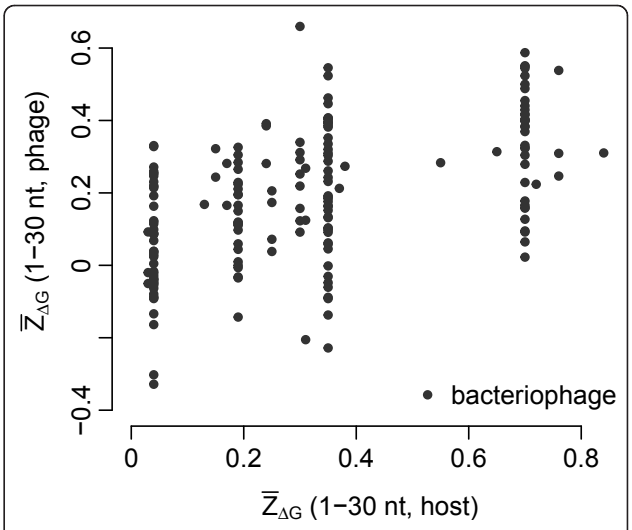

Figure 6 5' $\bar{Z}_{\Delta G}$ in bacteriophages versus the same quantity in the associated host organisms.

calculated correlations without considering phylogeny at all, and thus obtained the value $r = 0.89$ cited above. On the other hand, we considered the GC content of the host a measurement on the virus, and thus used the virus phylogeny to calulate PIC for both virus and host GC content. The correlation we obtained in this way was nearly indistinguishable from the one obtained not controlling for phylogeny at all (correlation coefficient $r = 0.88$, $P = 10^{-60}$, Additional File 1 Figure S7).

We also analyzed whether the 5' $\bar{Z}_{\Delta G}$ in bacteriophages correlated with that in their hosts. The 5' $\bar{Z}_{\Delta G}$ values for the phage hosts were obtained from [15]. We found a significant positive correlation, both when ignoring the phylogeny (correlation coefficient $r = 0.53$, $P \ll 10^{-12}$, Figure 6) and when using the phage phylogeny to calculate PIC for both phage and host 5' $\bar{Z}_{\Delta G}$ (correlation coefficient $r = 0.50$, $P \ll 10^{-12}$, Additional File 1 Figure S8).

Discussion

We have studied the mRNA stability at the translation-initiation region of protein-coding genes in 650 genomes of dsDNA viruses. We have found for many of these viruses that there is a 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. We have also found that 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' $\bar{Z}_{\Delta G}$). For bacteriophage, the 5' $\bar{Z}_{\Delta G}$ also correlates positively with the 5' $\bar{Z}_{\Delta G}$ in the host organisms.

Experimental and computational work had previously shown that increased local mRNA stability in the translation-initiation region impaired translation initiation in *E. coli* [11,13]. Two computational studies suggested that this effect exists more broadly in both prokaryotes and eukaryotes [14,15]. Here, we have shown that similar selection pressures exist in the viral kingdom.

As in cellular organisms, 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). Past the first two windows, $\bar{Z}_{\Delta G}$ tends to be zero or slightly negative. In cellular organisms, $\bar{Z}_{\Delta G}$ is consistently negative downstream from the start codon [15]. The lack of a negative $\bar{Z}_{\Delta G}$ in most virus genomes likely reflects lack of statistical power, a consequence of the small genomes of viruses. The strong positive correlation between genomic GC composition and the reduction of mRNA stability at the translation-initiation region is in agreement with the finding by Gu et al. [15].

In contrast to cellular organisms, viruses frequently have overlapping ORFs. In fact, nearly all viruses in our analysis had at least one overlapping ORF. Our codon-shuffling approach conserves the amino-acid sequence of the focal ORF, but does not conserve the amino-acid sequence of any second ORF that overlaps with the focal one. Thus, overlapping sequences will experience additional selective constraint that our approach does not accurately take into account. In principle, this issue could cause spurious results. However, we found that

there is little difference in $\bar{Z}_{\Delta G}$ values in overlapping and non-overlapping ORFs. At worst, $\bar{Z}_{\Delta G}$ values in overlapping ORFs are reduced compared to those in non-overlapping ORFs (Additional File 1 Figure S1). Therefore, treating overlapping ORFs as non-overlapping ORFs, as we have effectively done throughout much of this work, is a conservative approach when looking for elevated $\bar{Z}_{\Delta G}$ values.

To understand why $\bar{Z}_{\Delta G}$ increased with increasing GC content, we also considered the raw ΔG values. One can envision two extreme cases of how ΔG might depend on GC content. On the one hand, the ΔG in the first window might be required to be at a fixed low value, independent of GC content, to enable efficient translation. The ΔGs further downstream would be expected to decrease with increasing GC content, due to the higher thermodynamic stabiltiy of GC bonds. On the other hand, the ΔG in the first window might always differ by a fixed amount from ΔGs further downstream, independent of GC content. We found the reality to be somewhere in between these two extreme cases. Even though the ΔG in the first window showed a strong negative correlation with GC content, the difference in ΔG was not constant for prokaryotic or eukaryotic viruses, for which it increased strongly and moderately, respectively. For archaeic viruses, however, it did not significantly increase. Since the correlation between $\bar{Z}_{\Delta G}$ and GC was of comparable magnitude for all three groups, we infer that two separate mechanisms are at play. First, for prokaryotic and eukaryotic viruses, the requirement for decreased stability in the first window increases with increasing GC content. Second, in general, the $\bar{Z}_{\Delta G}$ measure seems to become more powerful for sequences with increased GC content, because the higher the GC, the less likely it is that a reshuffled sequence shows reduced stability.

For bacteriophages, we addressed the question whether the requirement of low mRNA secondary-structure stability in host genomes affects the 5' $\bar{Z}_{\Delta G}$ in phages. Because phages share the cellular environment and translation machinery with their hosts, we would expect that phages are optimized for the expression machinery of their hosts. We found a significant positive correlation between the 5' $\bar{Z}_{\Delta G}$ in phage genomes and that in their hosts. We also observed an even stronger correlation between the genomic GC content in phages and that in their hosts. Moreover, we found that a phage's GC content seems to impose a lower limit on the GC content of the hosts it can infect (Figure 5). These host-specific results are consistent with previous reports that synonymous codon usage in bacteriophage mimicks that of their hosts [17] and that viral and host GC content are similar in certain cases (*Mycobacterium tuberculosis*, 63.6% phage vs. 65.6% host, [18]; *Staphylococcus aureus*, 33.7% phage vs. 32.9% host, [19]).

We used independent contrasts to assess whether $\bar{Z}_{\Delta G}$ correlated with GC content. The independent contrasts method requires an accurate phylogeny of the organisms under study. Such a phylogeny is difficult to obtain for viruses, because viruses have either arisen multiple times independently or their common ancestor is extremely ancient [20-23]. In our analysis, we separately considered viruses infecting eukaryotes, prokaryotes, and archaea, and used phylogenetic trees derived from the taxonomic classification of these viruses. The branch lengths in these trees reflect simply the number of taxonomic levels that two viruses are separated by. Therefore, the branch lengths are almost certainly incorrect. Nevertheless, these trees should at a minimum remove any major biases that might arise if some groups of viruses were more heavily sampled than others. We found generally that the results based on PIC were nearly identical to results calculated on the raw data (not shown). Therefore, we believe that our results are not strongly confounded by phylogeny and that the correction for phylogeny we employed was sufficient.

In our comparison of viruses with their hosts, we encountered the added complication that virus and host trees will in general not be identical. We are not aware of any method that can calculate correct correlations in this scenario. We addressed this issue by considering both the raw data and PIC based on the virus trees (because properties of the virus host can be considered as a measurement on the virus). Again, both methods produced nearly identical results. Thus, it is unlikely that the results are strongly confounded by phylogeny.

Conclusions

Many dsDNA viruses show evidence for reduced mRNA secondary-structure stability near the start codon. The effect is the strongest in viruses infecting prokaryotes, but exists also in viruses infecting eukaryotes and archaea. For bacteriophage, the reduction tends to co-occur with a corresponding reduction of stability in the phage hosts. Thus, the same selective pressures that cause reduced stability of mRNA secondary structure in cellular organisms likely also act on the viruses infecting these organisms.

Methods

We collected virus genomes from the NCBI RefSeq project ftp://ftp.ncbi.nih.gov/refseq/release/viral/. We only considered coding sequences longer than 50 codons. We also excluded virus genomes that had 10 or fewer genes (overlapping reading frames were considered as different genes). We ended up with 650 genomes for dsDNA

viruses (419 with prokaryotic hosts, 204 with eukaryotic hosts, and 27 with archaeic hosts). To test whether overlapping ORFs confounded our analysis, we classified all ORFs into overlapping and non-overlapping ones. We defined the ORFs that had no genome region shared with any other ORFs as non-overlapping ORFs. All the other ORFs were considered as overlapping ones.

We analyzed the stability of local mRNA secondary structure exactly as described [15]. In brief, we calculated the local folding energy (ΔG) along the mRNA sequence using a sliding window of 30 nucleotides (nt), moving from the start codon to the 120th downstream nucleotide in steps of 10 nt (for a total of 13 windows). We calculated ΔG using the RNAfold program in the Vienna package [24,25] under default settings: folding occurred at 37°C; GU pairs were allowed; unpaired bases could participate in at most one dangling end; energy parameters were obtained from [26]. We evaluated only the minimum-free-energy structure.

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, as described [15]. $Z_{\Delta G}$ 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. We also evaluated the difference in local GC composition between the actual and randomized sequences via a Z-score Z_{GC}, as described [15]. For a few select virus species, we also tested whether our results were confounded by dinucleotide frequencies. We randomized virus mRNA sequences using a dinucleotide shuffling algorithm [27]. This algorithm preserves the dinucleotide composition as well as the codon use frequency in the reshuffled sequence.

We corrected for phylogenetic relationship among viruses by calculating phylogenetically independent contrasts (PIC), using the R library ape, version 2.5-1. We used three separate phylogenetic trees, one for viruses infecting eukaryotes, one for viruses infecting prokaryotes, and one for viruses infecting archaea. Since widely diverged viruses cannot be aligned, the phylogenetic trees we used were constructed purely based on taxonomic classification, as provided by NCBI's taxonomy tool http://www.ncbi.nlm.nih.gov/taxonomy.

We matched viruses to hosts using the "host" attribute in the "source" feature of the genbank file provided by RefSeq. We obtained quantities for hosts (such as GC content, $Z_{\Delta G}$) from our previous study [15].

We carried out all statistical analyses using R, version 2.10.1. Our R scripts plus accompanying raw data files are provided as supplementary data [Additional Files 2 and 3].

Additional material

Additional file 1: Supplementary Figures. A single pdf file containing Supplementary Figures S1-S8.

Additional file 2: Supplementary Data Part 1. A zip file containing raw data plus R scripts to reproduce all analyses.

Additional file 3: Supplementary Data Part 2. A zip file containing additional raw data used by the R scripts in Additional File 2.

Acknowledgements
This work was supported by NIH grant R01 GM088344.

Author details
[1]Section of Pulmonary, Critical Care, Sleep and Allergy, Department of Medicine, and Institute for Personalized Respiratory Medicine, University of Illinois at Chicago, Chicago, IL 60612, USA. [2]Center for Computational Biology and Bioinformatics, Institute for Cell and Molecular Biology, and Section of Integrative Biology, The University of Texas at Austin, Austin, TX 78712, USA.

Authors' contributions
TZ and COW designed the study, carried out analyses, prepared figures, and wrote the manuscript. Both authors read and approved the final manuscript.

Received: 20 July 2010 Accepted: 7 March 2011
Published: 7 March 2011

References
1. Shine J, Dalgarno L: **Determinant of cistron specificity in bacterial ribosomes.** *Nature* 1975, **254**:34-38.
2. Kozak M: **An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs.** *Nucleic Acids Res* 1987, **15**:8125-8148.
3. Kozak M: **Regulation of translation via mRNA structure in prokaryotes and eukaryotes.** *Gene* 2005, **361**:13-39.
4. Yamagishi K, Oshima T, Masuda Y, Ara T, Kanaya S, Mori H: **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* 2002, **9**:19-24.
5. Shabalina SA, Ogurtsov AY, Rogozin IB, Koonin EV, Lipman DJ: **Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals.** *Nucleic Acids Res* 2004, **32**:1774-1782.
6. Komarova AV, Tchufistova LS, Dreyfus M, Boni IV: **AU-rich sequences within 5' untranslated leaders enhance translation and stabilize mRNA in *Escherichia coli*.** *J Bacteriol* 2005, **187**:1344-1349.
7. Vimberg V, Tats A, Remm M, Tenson T: **Translation initiation region sequence preferences in *Escherichia coli*.** *BMC Genomics* 2007, **8**:100.
8. Zalucki YM, Power PM, Jennings MP: **Selection for efficient translation initiation biases codon usage at second amino acid position in secretory proteins.** *Nucleic Acids Res* 2007, **35**:5748-5754.
9. Chen H, Pomeroy-Cloney L, Bjerknes M, Tam J, Jay E: **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* 1994, **240**:20-27.
10. Qing G, Xia B, Inouye M: **Enhancement of translation initiation by A/T-rich sequences down-stream of the initiation codon in *Escherichia coli*.** *J Mol Microbiol Biotechnol* 2003, **6**:133-144.
11. Griswold KE, Mahmood NA, Iverson BL, Georgiou G: **Effects of codon usage versus putative 5'-mRNA structure on the expression of *Fusarium***

solani cutinase in the *Escherichia coli* cytoplasm. *Protein Expres Purif* 2003, **27**:134-142.
12. Gonzalez de Valdivia EI, Isaksson LA: **A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli*.** *Nucl Acids Res* 2004, **32**:5198-5205.
13. Kudla G, Murray AW, Tollervey D, Plotkin JB: **Coding-sequence determinants of gene expression in *Escherichia coli*.** *Science* 2009, **324**:255-258.
14. Tuller T, Waldman YY, Kupiec M, Ruppin E: **Translation efficiency is determined by both codon bias and folding energy.** *Proc Natl Acad Sci USA* 2010, **107**:3645-3650.
15. Gu W, Zhou T, Wilke CO: **A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes.** *PLoS Comput Biol* 2010, **6**:e1000664.
16. Felsenstein J: **Phylogenies and the comparative method.** *Am Nat* 1985, **125**:1-15.
17. Lucks JB, Nelson DR, GR GRK, Plotkin JB: **Genome landscapes and bacteriophage codon usage.** *PLoS Comput Biol* 2008, **4**:e1000001.
18. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR Jr, Hendrix RW, Hatfull GF: **Origins of highly mosaic mycobacteriophage genomes.** *Cell* 2003, **113**:171-182.
19. Kwan T, Liu J, DuBow M, Gros P, Pelletier J: **The complete genomes and proteomes of 27 Staphylococcus aureus bacteriophages.** *Proc Natl Acad Sci USA* 2005, **102**:5174-5179.
20. Iyer LM, Aravind L, Koonin EV: **Common Origin of Four Diverse Families of Large Eukaryotic DNA Viruses.** *J Virol* 2001, **75**:11720-11734.
21. Bamford DH: **Do viruses form lineages across different domains of life?** *Res Microbiol* 2003, **154**:231-236.
22. Forterre P: **The origin of viruses and their possible roles in major evolutionary transitions.** *Virus Res* 2006, **117**:5-16.
23. Koonin EV, Senkevich TG, Dolja VV: **The ancient Virus World and evolution of cells.** *Biology Direct* 2006, **1**:29.
24. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P: **Fast folding and comparison of RNA secondary structures.** *Monatshefte f Chemie* 1994, **125**:167-188.
25. Hofacker IL, Stadler PF: **Memory efficient folding algorithms for circular RNA secondary structures.** *Bioinformatics* 2006, **22**:1172-1176.
26. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.** *J Mol Biol* 1999, **288**:911-940.
27. Katz L, Burge CB: **Widespread selection for local RNA secondary structure in coding regions of bacterial genes.** *Genome Res* 2003, **13**:2042-2051.

doi:10.1186/1471-2148-11-59
Cite this article as: Zhou and Wilke: Reduced stability of mRNA secondary structure near the translation-initiation site in dsDNA viruses. *BMC Evolutionary Biology* 2011 **11**:59.