Published in final edited form as:

Methods. 2015 December ; 91: 48–56. doi:10.1016/j.ymeth.2015.06.015.

***In silico* discovery and modeling of non-coding RNA structure in viruses**

Walter N. Moss^a and Joan A. Steitz^{a,*}

^aDepartment of Molecular Biophysics and Biochemistry, Howard Hughes Medical Institute, Yale University School of Medicine, New Haven, Connecticut 06536, USA

Abstract

This review covers several computational methods for discovering structured non-coding RNAs in viruses and modeling their putative secondary structures. Here we will use examples from two target viruses to highlight these approaches: influenza A virus—a relatively small, segmented RNA virus; and Epstein–Barr virus—a relatively large DNA virus with a complex transcriptome. Each system has unique challenges to overcome and unique characteristics to exploit. From these particular cases, generically useful approaches can be derived for the study of additional viral targets.

Keywords

ncRNA prediction; viruses; RNA; RNA structure; ncRNA; sequence analysis

1. Introduction

This is an explosive period in the progress of Biology, in general, and RNA Biology in particular. Advances in sequencing technology and in bioinformatic analyses of next-generation sequencing data have revealed that the transcriptomes of all living things are immensely complex. Indeed, the rate of sequence accumulation far outpaces that of the discovery of RNA function; and we are left with billions of nucleotides of RNA sequence—much of which is an utter mystery [1]. Though lagging behind, the roles discovered for RNA, beyond simply coding for proteins, are steadily growing. In addition to classical non-coding (nc)RNAs, such as rRNA, tRNA, snRNAs and snoRNAs, a wide array of other ncRNAs have been discovered [2–4]. These ncRNAs are arbitrarily classified based on their length. Small ncRNA are less than 200 nucleotides (nt). They include micro (mi)RNAs, which are ~22 nt RNAs excised from pre-miRNA hairpins, base pair to target mRNAs and affect gene expression [5], as well as a variety of other functionally important ncRNAs [6]. Long non-coding (lnc)RNAs include everything longer than 200 nt. This broad class of molecules includes dozens of functionally important RNAs (e.g. Xist [7] and HOTAIR [8]

*Corresponding author: joan.steitz@yale.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

and hundreds of other lncRNAs of unknown function [9-11]. Functional ncRNAs can also be embedded in coding transcripts (e.g. in introns and untranslated regions), and can even overlap coding sequences [12]. In some cases, the coding capacity of an mRNA is of secondary importance to its non-coding function [12]. Indeed, some groups have proposed the term functional (f)RNA to include both non-coding and coding RNAs [13-15].

Some ncRNAs, e.g. rRNAs and RNase P, are of such importance that they are highly conserved throughout all domains of life [16, 17]; others are only found in a particular clade or species. Regardless, ncRNAs play important roles in every organism. As obligate parasites of cellular life, viruses have evolved their own repertoire of ncRNAs or hijacked cellular ncRNAs for their own use. With their highly size-restricted genomes and the need to mediate many processes to evade host immunity, maintain infection and replicate, viruses make a particularly versatile repertoire of ncRNAs [18]. In addition, many viral genomes are themselves comprised of RNA or require an RNA intermediate during replication. This, coupled to their clear medical importance, makes the study of viral ncRNA structure especially interesting.

Common to all cellular and viral ncRNAs is the importance of RNA folding. All known ncRNA functions are mediated (or affected) by the intramolecular base pairing that comprises RNA secondary structure [6] or by intermolecular base pairing between RNA strands [19]. Therefore, (working backwards) the identification of structured regions of RNA strands can be used to discover functional regions of RNA: either new examples of known structural/functional motifs or completely novel structures or functions. Methods for ncRNA discovery were pioneered in prokaryotes (see the excellent review in reference [20]). This review will focus on methods for detecting functional RNAs in viruses and approaches for modeling their secondary structures. Viral ncRNA discovery presents different challenges and opportunities compared to prokaryotes. Viral genomes are, in general, much smaller than even the simplest prokaryote's; however, unlike those required for cellular life, there are no classes of absolutely conserved "housekeeping" ncRNAs in viruses. For example, all cellular life shares a common ancestor, and deep phylogenetic relationships connect several key cellular ncRNAs: e.g. rRNAs, and tRNAs are shared in all, snoRNAs are shared between eukaryotes and Archaea, and tmRNAs are conserved in prokaryotes. Class-specific tools exist for the detection of bacterial (and other cellular) ncRNAs [21]. Beyond these well-defined and conserved ncRNAs, however, prokaryotes generate numerous small (s)RNAs (ranging from 50–250 nt) that are diverse and not universally conserved. Prediction of bacterial sRNAs is analogous to viral ncRNA prediction; yet, even here, viruses present challenges.

The origin and evolution of viruses is complex and steeped in mystery. Multiple hypotheses are proposed for their origin, but it is likely that different families of viruses arose independently [22]. Additionally, viral polymerases have varying degrees of fidelity and the rates of evolution of viral sequences are clade-specific [23]. Indeed, the evolutionary dynamics of HIV, for example, are so rapid that analysis of sequence evolution within a single host versus a population of hosts can reveal differing evolutionary pressures [24], perhaps even impacting structurally-linked sequence covariation.

In this review, methods will be highlighted that were applied to two very different viral targets: influenza A virus (IAV; a rapidly-evolving RNA virus) and Epstein–Barr virus (EBV; a slowly-evolving DNA virus). These viruses represent two “extreme” cases and the results obtained should be transferrable to almost any target virus. IAV is a relatively-small (~13.5 kb) segmented RNA virus. The IAV genome consists of eight negative-sense viral (v)RNAs that are specifically packaged into virions. The vRNA serves as a template to make positive-sense complementary (c)RNA, itself a template for genome replication, and mRNAs for at least 11 proteins (via alternative RNA splicing) [25]. The virus rapidly proliferates through the host, is shed and spread through populations. IAV evolves rapidly because of the error rate of the viral polymerase [26], the ability of genome segments of different co-infected strains to reassort (antigenic shift), and IAV’s ability to infect and cross between different host species (e.g. humans, pigs and birds), which is responsible for deadly pandemic outbreaks [25]. EBV, on the other hand, is a large (~170 kb) double-stranded DNA virus. In the EBV virion, the viral genome is linear; however, upon infection it circularizes, forming a viral episome. This episome persists in host B cell nuclei during lifelong latent infection. Latent infection proceeds via several distinct programs that re-pattern the cell to make it more hospitable for infection and allow the virus to evade the host immune response [27]. Latency is interrupted by periodic lytic reactivation and shedding of active virions. In both lytic and latent infections, the viral transcriptome is complex; most of the EBV genome is transcriptionally active at some point and transcripts undergo extensive splicing and modification [21, 28, 29]. Each virus presents unique challenges and opportunities for RNA structure discovery and analysis.

2. Methods

2.1 Sequence acquisition

The most powerful support for a functional role of a sequence motif is its evolutionary conservation—this is also true for RNA structural motifs. Indeed, this feature is leveraged in computer algorithms for prediction of functional ncRNAs. Therefore, the starting point for any analysis of RNA structure should be the identification and acquisition of homologous (evolutionarily related) sequences. Due to the variation in evolutionary rates of viruses, each target will differ in what initial sequences to use: e.g. some targets will be limited to one viral species, while others can include additional homologous species. For ncRNA discovery the average pairwise sequence identity (APSI) of sequences should ideally be close to 80% and, barring this, fall within the range of 60–95% APSI. Below this range, sequence alignment becomes unreliable and above this range, the lack of mutations makes identification of structurally-relevant covariation less likely. Additionally, in selecting an initial dataset, it is helpful to avoid biasing this input dataset with multiples of identical or near identical sequences, which will affect consensus structure prediction.

There are three primary nucleotide sequence databases from which sequences can be acquired: GenBank [30], EMBL [31], and the DNA databank of Japan [32]. Within each site are tools for navigating the databases, finding sequences and downloading data. For example, in the study of EBV, a search was made for all herpesvirus reference sequence (RefSeq) genomes in GenBank. (RefSeqs are unique, well-curated entries that can be

thought of as representative biological sequences [33].) For the analysis of EBV, RefSeqs for four strains of EBV and one, closely-related herpesvirus (rhesus lymphocryptovirus [rLCV]) were downloaded. It is also possible to query the database for homologous sequences using the basic local alignment search tool (BLAST; [34]). For large data downloads, the NCBI provides sets of e-tools to query the Entrez Global Query Cross-Database Search System. A detailed guide for using these tools is published on-line (<http://www.ncbi.nlm.nih.gov/books/NBK25501/>).

In addition to the large general databases, there are many targeted databases, some of which are focused on viruses. An up-to-date and comprehensive list of viral databases is compiled in the 2014 NAR Database Summary Paper [35]. For IAV, the NCBI influenza virus resource page [36] provides a number of filters to extract and organize sequence data: by species, strain, year, etc. This makes it possible to analyze particular segments or strains, to track structural mutations over time/species/strain, to identify pandemic strains and much more.

2.2. Sequence alignment

The foundation of almost any study aiming to discover or model RNA structure is a sequence alignment—the rationale being that functional RNA structures will be preserved over evolutionary time and will thus affect sequence evolution (e.g. paired nts will have single point-mutations [consistent mutations] and double point-mutations [compensatory mutations] that can be observed in aligned sequences). The sequences in the alignment must be sufficiently conserved to allow alignment with confidence, but contain enough variation to identify structurally relevant mutations. In practice, ~80% APSI is the ideal number—although the amount will vary depending on the alignment method used [37].

With no *a priori* knowledge of which regions will have conserved RNA structure, alignment based on a known feature, other than nucleotide sequence homology, can allow a more rigorous inference of structural homology. For example, in IAV the majority of each vRNA segment encodes protein. To generate sequence alignments, IAV genome sets were extracted, the short (13 and 12 nt, respectively) 5' and 3' UTRs were removed, and the nucleotide sequences then aligned based on amino acid sequence. Thus, any discovered RNA structural homologies will be founded on the known protein coding function of the sequences analyzed. This also increases the quality of the sequence alignment since the 20 amino acids of protein contain much more information than the 4 nitrogenous bases of RNA. In regions where both RNA structure and amino acid sequence (or other) conservation must be maintained, there are distinct imprints on nucleotide sequence evolution, which can be exploited to discover functional (structured) regions (discussed further below).

A number of programs are available for aligning RNA sequence directly (Table 1), each with benefits and limitations. For a review of alignment methods, see reference [38]. In practice, for sequences with APSI > 80%, the method used is unlikely to have a large impact on alignment quality; however, below this APSI, the approach used becomes more important. For example, for divergent sequences (< ~60%), sequence alignment can be very difficult. Methods that consider putative RNA structure (when homologous structures exist) can be applied to align divergent sequences with structural homology or to discover local

regions of conserved sequence separated by non-conserved “gappy” stretches (Table 1). In EBV, for example, viral genomes were initially aligned using a fast/inexpensive approach (FFT-NS-1; [39]); however, when regions of interest were identified, sequences were re-aligned using slow/expensive methods that depend on the level of sequence divergence of the region analyzed. Even the very best automated aligners can misalign sequences; thus, manual inspection and optimization of alignments are essential.

2.3 Prediction of regions likely to contain ncRNA structure

An important property of many ncRNAs is that, because of their propensity to form RNA structure, they have unusually high thermodynamic stability when compared to random sequences (with the same nucleotide composition) [40]. That is to say, the sequence of a ncRNA is ordered to maintain base pairing and this propensity to form pairing is reflected in the predicted Gibbs free energy of folding (ΔG°). The metric used to estimate this propensity is the thermodynamic z-score. This metric is calculated via the following formula:

$$z - score = \frac{\Delta G^\circ_{native} - \overline{\Delta G^\circ}_{random}}{\sigma}$$

where ΔG°_{native} is the predicted native Gibbs free energy of folding, $\overline{\Delta G^\circ}_{random}$ is the average of the shuffled “mutant” sequences and σ is the standard deviation of the set of fragment sequences. When scrambling the sequence disrupts (likely functional/evolved) base pairs, the energy of the mutant sets will be higher (less favorable) than that of the native sequence and the z-score will be negative. Thus, the z-score measures the increased stability of the evolved native sequence vs. random sequence, where negative numbers indicate increased stability and propensity for forming RNA structure.

A shift toward negative z-scores across a sequence is termed global ordered RNA structure (GORS) and GORS has been discovered in several viruses [41-43]. In IAV, GORS was found to be favored in (+)RNA vs. the viral genomic (−)RNA [44]; GORS was enriched in transcripts from four genome segments that encode key viral genes; and, interestingly, the magnitude of GORS in IAV strongly depends on the host species specificity of the IAV strain [44]. Here, the z-score provided global evidence for RNA structure in IAV that appears to be important for species specificity and, possibly, for the regulation of IAV gene expression. Notably, although only one RNA structure (a pseudoknot in segment 8 (+)RNA [45, 46]) is conserved between IAV and influenza B virus (IBV), GORS is conserved in three homologous gene segments of IBV [47].

Although negative z-scores are hallmarks of ncRNAs, the z-score alone may not be sufficient to identify ncRNA structure. When combined with comparative sequence analysis, however, the z-score is a potent tool for ncRNA discovery—indeed this approach is at the heart of the most successful discovery algorithms (Table 2). For both IAV [48] and EBV [49], the program RNAz [50] was used to deduce structured regions. This program uses primarily the z-score and structure conservation to predict regions likely to generate ncRNA—based on a support vector machine (SVM) trained on known ncRNAs (e.g. known

ncRNAs have low [favorable] z-scores and high structure conservation) [51]. Included with RNAz are Perl scripts for processing alignments into overlapping windows (e.g. 120 nt used for IAV and EBV), optimizing subsets of sequences (e.g. to filter out low quality “gappy” or polymorphic sequences, optimize APSI and more) and to filter and concatenate resulting windows into loci with specified prediction confidence (the SVM RNA-class probability). For detailed protocols for installation and use of RNAz, see references [52, 53].

A confounding factor in ncRNA discovery can be determining the strandedness of predicted RNA structure. For example, IAV produces (+) and (–) sense RNAs in infected cells; thus, structure can occur in either orientation. In some cases, the metrics of folding can be very similar and, by necessity, conservation will also be high in both orientations. In these cases, a program called RNAstrand [54], another SVM that uses additional parameters to predict which strand is likely to generate structure, is available: for example one strand may have more stabilizing GU pairs than the reverse complement, where GU pairs become destabilizing CA pairs. Indeed, tuning the energetics with GU base pairs may reflect a common adaptive strategy used to evolve strand-specific RNA structures [55]. RNAstrand was used in IAV and also in EBV, where predictions overlapped sites of bidirectional transcription. In cases where the metrics appear favorable for both strands, structure may be present in either strand.

Defining the extent of a structured region is another major difficulty. By dividing the sequence of interest into overlapping windows, predicted structured regions larger than the window used (e.g. 120 nt) will be fragmented in the analysis. Concatenation of overlapping highly-predicted regions should roughly approximate the actual region of structure; however, this may not always be the case: e.g. for very large structures (such as a very large hairpin found in EBV [49, 56]) or where the actual conserved structure is smaller than the window used. To overcome this limitation, the program LOCARNA-P was developed [57]. This approach uses base pair partition function probabilities to estimate the extent of a structured region by defining regions of well-predicted (high partition function probability) pairing. This protocol can sometimes better approximate the extent of RNA structure in a multiple sequence alignment [57]. LOCARNA-P allows the evaluation of pre-existing alignments or the re-alignment and folding, simultaneously, of a set of sequences. Thus this approach can be useful in regions where the nucleotide sequence alignment is less robust (e.g. in divergent regions).

When RNAz and other ncRNA prediction programs fail, they tend to over-predict regions likely to generate structured RNAs [58]. This emphasizes the importance of using complementary data and performing additional sequence comparisons to hone in on the truly functional structured RNAs. In the study of EBV, for example, all but two of the known structured ncRNAs were correctly predicted [49]. For both of the ncRNAs that RNAz failed to identify, the fault was in the nucleotide sequence alignment where more divergent sequences were not aligned correctly. Elsewhere, the majority of predictions were in novel regions. For these putative structured RNAs further comparative sequence analyses were performed to determine which structures were widely conserved and were supported by sequence covariations.

2.4 Prediction of coding regions likely to contain “non-coding” RNA structure

Coding RNAs can also have non-coding functions mediated by RNA structure. This is particularly true for viral RNAs, which arise from highly size-restricted genomes and thus have much pressure to be multifunctional. For example, the HIV-1 genome, while encoding viral proteins, also has extensively modeled RNA structure [59] that is important in packaging [60], regulation of gene expression [61, 62] and more. Work on IAV and EBV also discovered evidence for conserved and unusually stable (likely functional) RNA structure in viral coding regions [48, 49]. In IAV, predicted structures are enriched at or near alternative splice sites [48, 63, 64]. This enrichment was also observed in influenza B and C [65], suggesting that modulation of splicing by RNA structure may be a strategy common throughout influenza species.

If a coding sequence must also maintain evolutionarily conserved structure, the effect on synonymous codon use can be dramatic. For most amino acid codons, the third position (wobble site) in the triplet is degenerate; thus, mutations at this site can maintain the same amino acid in the encoded protein (synonymous mutations). When third codon positions are paired, they are not free to explore the same sequence space as unpaired wobble sites elsewhere in the sequence; thus, in aligned sequences there will be fewer observed mutations between synonymous codons in structured vs. non-structured regions. This effect is termed suppression of synonymous codon use (SSCU) and was applied to identify regions in several viruses [43, 66, 67] likely containing conserved RNA structure—including IAV, where SSCU sites overlapped with RNaz-predicted regions with evidence for conserved and unusually thermodynamically stable RNA folds [48].

The ncRNA prediction tool RNA-Decoder was developed specifically for use in coding regions [66]. This tool uses a stochastic context-free grammar approach that includes parameters such as codon use to make predictions. This method was used, alongside extensive biochemical probing, to facilitate modeling of structure in the HIV-1 genome [59].

2.5 Prediction of regions likely to contain ncRNA structure in highly divergent sequences

All the methods discussed above require high-quality alignments for accurate prediction of ncRNA structure because incorrect alignment of nucleotide sequences makes identification of conserved base pairs impossible. For highly-divergent sequences this becomes problematic, as most aligners perform poorly at low APSIs. As alluded to earlier, this problem is addressed in some alignment programs by considering putative RNA structure in the alignment step [68]. Simultaneous alignment and folding were first used for ncRNA prediction with the program Dynalign [69], which allows pairwise structure prediction/alignment [70, 71]. More recently, an algorithm that can fold/align multiple sequences (Multilign [72]) was applied to ncRNA discovery in the program Multifind (available within the RNAstructure package [73]). In both cases, long sequences are scanned and fragments aligned with respect to RNA folding before ncRNA prediction based on an SVM (similar to RNaz). It is also possible to consider different regions of a long alignment separately. For example, after parsing alignments for ncRNA prediction with RNaz, regions with low APSI can be realigned using a fold/align algorithm such as LOCARNA [57, 74, 75]. Indeed, the

RNAz SVM was also trained on LOCARNA aligned/folded ncRNAs and the program can be run taking this into consideration.

2.6 Modeling RNA structure in predicted regions

The methods discussed above yield overlapping windows likely possessing RNA structure. Predictions are made for base pairing in the selected scanning window based on the provided input sequences; however, structural domains can be (and often are) larger than the scanning window and models can be improved by considering additional (sometimes more divergent) sequences. Overlapping prediction windows can be combined (e.g. RNAz provides tools) to define larger structural domains. To model the structure in RNAz predicted domains, sequences should be re-aligned using more robust algorithms and then manually optimized (when possible). This alignment will form the basis for initial predictions. Modelling is facilitated by using additional homologous sequences. Without an initial structure, the program BlastR can identify likely homologous ncRNA sequences [76]. With a structural hypothesis, the INFERNAL package can be used to identify homologous structures [77, 78]. Detailed protocols for modeling RNA structure can be found in reference [79]; but a brief example is given here for EBV. In EBV, predicted domains were extracted, BLASTed against the NCBI nt database to identify additional sequence homologs, and then realigned using MAFFT [49]. The alignments were manually optimized, and consensus folding was predicted using the program RNAalifold, which is used for structure prediction in RNAz [80, 81]. RNAalifold works by predicting the consensus structure minimum free energy fold—biasing predictions by favoring structure-preserving mutations (with energy bonuses) and penalizing structure-disrupting mutations. There are other algorithms available for folding RNAs individually, in alignments, and for folding and aligning simultaneously. A partial list is given in Table 3.

With an initial structure model and alignment, it then becomes possible to identify more divergent sequences that may have structural homology. One of the most successful strategies for undertaking this search is implemented in the INFERNAL package [77, 78]. Here an input alignment and structure are used to build a covariance model (to describe variation in the RNA sequences), which allows searching databases of query sequences for sequences/structures that might be homologous. This process is analogous to a BLAST search, but for structure in addition to sequence. In a number of cases, INFERNAL searches enabled the identification of homologous structures beyond EBV strains in more distant herpesviruses [49]. With additional sequences it is possible to add to and refine the structure alignment, build a new covariance model and iteratively search for novel hits; this process was used to model the structures in EBV.

2.7 Modeling RNA pseudoknot structures

One major limitation of all methods discussed so far is their inability to predict RNA pseudoknots, which play many important roles in biology [82]—particularly in viruses (e.g. frameshift pseudoknots, internal ribosomal entry sites, etc. [83]). This difficulty is due to the high level of computational expense in pseudoknot prediction and the necessity for using complex heuristics to reduce the complexity of dynamic programming approaches to RNA folding [82]. Once a structural “hot spot” is identified, however, it is possible to apply these

relatively expensive calculations to predict pseudoknots. A list of prediction algorithms that identify pseudoknots is given in Table 3. In IAV, for example, a region that had strong evidence for SSCU overlapped a 3' splice site in a viral mRNA. The program DotKnot [84] was then used to predict a stable pseudoknot in this region, which was subsequently found to be highly conserved across IAV strains [48, 63, 64].

2.8 Statistical significance of RNA structure

With a robust model for an RNA fold in hand, it is possible to test for significance (if there are sufficient sequences/variation) using statistics. If two sites within a sequence are evolving together (e.g. to maintain structure), this will be reflected in a high mutual information (MI) score [85]. MI can be calculated using the following equation:

$$MI(i, j) = \sum f(n_i n_j) \cdot \log_4 \left[\frac{f(n_i n_j)}{f(n_i) \cdot f(n_j)} \right]$$

where $n_i, n_j \in [A, G, C, U]$, $f(n_i)$ and $f(n_j)$ are the individual nt frequencies at a site and $f(n_i n_j)$ is the frequency of a nt combination at positions i and j . The summation is taken over all 16 pair combinations. The MI is 1 for perfect correlation between sites (e.g. if all observed combination frequencies are equal and maintain base pairs) and 0 for no correlation (random changes at sites).

In the case of IAV, thousands of sequences are available, which allows MI calculations to be used. The analysis of IAV predicted structures discovered multiple positions with very high MI scores indicating structural constraints on evolution [86, 87]. Even in cases where there are not enough sequences to calculate MI scores, analyses of sequence alignments can be used to identify sequence changes that preserve structure. For example, the Vienna RNA package includes visualization scripts to generate annotated consensus structures where consistent and compensatory mutations are annotated. Such changes can also be manually identified by inspecting alignments (e.g. BioEdit [88] allows RNA structure-annotations of alignments where single point [consistent] mutations and double [compensatory] point mutations can be identified). Such mutations provide strong evidence that a modeled base pair reflects the actual native RNA fold.

3. An example from IAV genome segment 5

To illustrate the methods described above, we will analyze the IAV genome segment 5, which encodes viral nucleoprotein (NP). This segment was previously shown to contain two regions with unusually stable and conserved predicted structure that overlap codon constrained sites (with SSCU) [48]. Here we will redo the discovery calculations and extend them to model the structure of one particular region.

3.1. Sequence acquisition and alignment

For the prediction of likely structured regions, the six available RefSeq segment 5 sequences were obtained from the NCBI Influenza Virus Resource [36] (GenBank #s NC_004905.2, NC_007381.1, NC_007369.1, NC_002019.1, NC_026426.1, NC_007360.1). Coding

sequences were aligned based on amino acids using Clustal [89] (as implemented in the BioEdit alignment editor [88]). UTRs were aligned using MAFFT (G-INS-i strategy; [39, 90]), then appended to the ends of the codon-aligned nucleotide sequences. The alignment was manually inspected before use in predictions. For the analysis of conservation of structure and SSCU, a set of all unique segment 5 sequences were downloaded from the Virus Resource site (18,500 sequences). A sequence alignment was performed using the MAFFT FFT-NS-1 strategy [39, 90], which is very fast and able to handle massive datasets. An example of the command-line used is

```
$ mafft --retree 1 --maxiterate 0 SEG5.fa > SEG5.mafft.fa
```

3.2 Prediction of regions likely to contain RNA structure

To predict structured regions, the alignment of RefSeqs was converted to a Clustal format alignment using the NIH file conversion tool (<http://genome.ncbi.nlm.nih.gov/tools/reformat.html>). This alignment was processed using Perl scripts provided with the RNAz program [50]. First, the alignment was split into windows of 120 nt with a 10 nt step size.

```
$ perl rnazWindow.pl --window=120 --slide=10 SEG5.mafft.aln > SEG5.windows
```

This processing also allows for filtering of the alignment windows. In this case, the following default filtering parameters were used: maximum fraction of gaps, 0.25; minimum APSI, 0.5; minimum sequences, 2; optimum APSI, 0.8.

Each window was next evaluated on both the forward and reverse strands using the RNAz SVM prediction program with filtering of output based on calculated RNA class probability (> 0.5).

```
$ RNAz --both-strands --cutoff=0.5 SEG5.windows > SEG5_RNAz.out
```

Next, prediction windows (shown in Fig. 1A) were compiled into overlapping “loci” with RNA class probability > 0.9 using another Perl program provided with RNAz.

```
$ perl rnazCluster.pl --cutoff=0.9 --html SEG5_RNAz.out > SEG5_RNAz.loci
```

This process generates text files with loci coordinates, RNA class probabilities, z-scores, and detailed information about the loci (e.g. thermodynamic stability, structure conservation index, GC content, etc., as well as html files for visualizing the results).

The predicted windows/loci are on both the forward (+) and reverse (−) strands. For example, the overlapping windows at the 3′ end are all on the forward strand (Fig. 1A), but it is worth checking with the program RNAstrand. The aligned nts spanned by overlapping windows were extracted and submitted to RNAstrand for analysis using a webserver [54]. The resulting calculation showed that all the metrics considered by the SVM (e.g. z-score, folding energy, GU vs. CA pairs, etc.) favored structure in the (+) strand (data not shown). It is worth noting that although the SVM class probability highly favors (+)RNA structure, the potential for (−)RNA structure should not be definitively ruled out. The results simply suggest that the initial focus should be on (+)RNA structure.

3.3 Identification of regions with suppressed codon usage

To identify regions with SSCU, subsets of 300 random sequences were pulled from the alignment of all available unique segment 5 sequences. Coding sequences were extracted and aligned via amino acid codons using Clustal. The aligned coding regions were opened in the Simmonics Sequence Editor [91] and analyzed for SSCU using a module within the editor. The resulting data were plotted in Excel (Fig. 1A). Two regions of significant (< 0.25) SSCU were observed that also overlapped RNAz windows with high probability of forming RNA structure—the most extensive SSCU occurred toward the 3' end of the sequence. SSCU calculations were repeated with different subsets of 300 sequences, and the results were unchanged.

3.4 Modeling and evaluation of extent of structure

RefSeq aligned sequences—from the (+)RNA strand—corresponding to the locus spanning predicted windows at the 3' end of segment 5 were extracted. These sequences were re-aligned and folded using LOCARNA-P [57], which also predicts a base pair partition function. The resulting consensus sequence/structure is shown in Figure 1B with annotations indicating conservation of base pairing, compensating mutations, and the base pair partition function probability. Within the predicted consensus structure, only nts 1428–1530 have high probability base pairs; thus, only this region was used for further analyses. Within nts 1428–1530 are two short and one longer hairpins. All structures are well conserved in the six RefSeqs analyzed. Mutations occur only in the longer hairpin, where they maintain base pairing in the predicted model—indeed, there are one consistent and two compensatory mutations in the longer hairpin.

The INFERNAL package was used to search for homologous structure in the alignment of all IAV segment 5 sequences, as well as in collections of influenza B (IBV) and C (ICV) viruses. The alignment and consensus structure from LOCARNA-P was converted into Stockholm format (manually) and used to build a covariance model (cm) with INFERNAL (for details see INFERNAL manual).

```
$ cmbuild SEG5_1428-1529.cm SEG5_1428-1529.sto
```

Next, the covariance model was calibrated.

```
$ cmcalibrate SEG5_1428-1529.cm
```

Finally, the calibrated model was used to search IAV sequences (as well as IBV and ICV) for homologous structures.

```
$ cmsearch SEG5_1428-1529.cm SEG5_IAV.fa > SEG5_1428-1529_INFERNAL.out
```

Homologous structures can be found in all IAV sequences but not in IBV or ICV sequences, suggesting that structure in this region is unique to IAV. Subsets of returned IAV fragments were aligned and folded using LOCARNA-P, as well as two other folding algorithms (implemented in the RNAstructure package): TurboFold [92] and Multalign [72]. The only part of the model that was consistently predicted was the longest hairpin. Conservation of base pairing was determined by counting the observed pairs between the *i*-*j* nts from the

LOCARNA-P model structure in the alignment. The overall structure is very well conserved (99.1% preservation of base pairing) and, when mutations occur, they support the model. Not only the two compensatory and one consistent mutation found in the analysis of the RefSeq data, but also additional supportive mutations were identified (Table 4). The frequency of these mutations, however, could be low in comparison to the predominantly-observed pair. In all but one case, the frequency of consistent mutations is larger than compensatory mutations; this is due to the entire hairpin falling within a coding region where only two paired sites (the pair between nts 1479-1527 and 1476-1530) are formed by nts at wobble (3rd codon) positions. Thus, compensatory changes are harder to form elsewhere in the hairpin because of the need to maintain coding potential and (presumably) RNA structure.

To evaluate the statistical significance of mutations in the proposed model, the MI was calculated between all pairs in the alignment region spanning 1428–1530. Only one pair, between nts 1479:1527 (both wobble sites), had a significant MI value: 0.35 (Fig. 1C). The MI score is maximized (value of 1.00) when the frequencies of canonical pairs are equal: $f(\text{AU}) = f(\text{GC}) = f(\text{UA}) = f(\text{CG}) = 0.25$. The majority pair at 1479-1527 is CG (77.6%) followed by the compensatory change to UA (18.9%) and a smaller fraction of the consistent mutation to UG (3.2%). Compensatory mutations in other sites do not have high enough frequencies to affect the MI score. For example the pair between nts 1476-1530 has a negligible MI score (0.02) because the observed compensatory changes are too low in abundance vs. the predominant pair (Table 4 and Fig. 1C)—a UG wobble pair that is not accounted for in the MI calculation.

The distribution of observed pairs can be challenging to interpret since the IAV data, in addition to being in coding regions, are highly skewed—some strains (e.g. virulent ones) are massively overrepresented in the sequence database. Because of this skew in IAV sequences, the MI score was normalized by the entropy (H) at each aligned residue in the pair. The ratios $R(i,j) = \text{MI}(i,j)/H(i)$ and $\text{MI}(i,j)/H(j)$ allow one to estimate correlation at two sites from biased datasets [87]. Using this approach, the only pair with significant correlation still appears between nt 1479:1527 (R values of 0.69 and 0.75).

4. Concluding remarks

This review discusses some of the approaches that have been successfully used to deduce structured RNAs in viruses. A particular example from IAV was used to illustrate how a region likely to contain conserved, stable and functional RNA structure can be identified. Here, we recapitulated previous work, then expanded it to show how a site can then be “pruned” to deduce the extent of the likely functional motif and how to model its structure. Near the 3' end of the (+)RNA for IAV segment 5 a hairpin fold was modeled that is almost perfectly conserved in all available IAV sequences. Synonymous codon use is highly suppressed in this region, indicating structural constraints on codon evolution. Compensatory and consistent changes support the model and a statistically significant correlation was discovered between nts undergoing compensatory changes; thus, structure constrains both codon use and nucleotide sequence, strongly suggesting that this hairpin performs some function in IAV. With this evidence and the model, interesting questions can

be asked (e.g. how are compensatory and consistent mutations correlated with virulence, species-specificity, etc.). Furthermore, the sequence/structure provides a basis for additional experimental analyses to identify protein binding partners, introduce structurally-relevant mutations and track viral fitness, among others. Although we have discussed EBV and IAV (in particular), the methods presented here are general and can be applied to any virus.

Acknowledgments

We would like to thank Dr. Kazimierz Tycowski for his helpful reading of this manuscript. Our work was supported by National Institutes of Health grant CA16038. Walter N. Moss is funded through an American Cancer Society postdoctoral fellowship (grant 126117-PF-62814-139-01-MPC) and Joan A. Steitz is an investigator of the Howard Hughes Medical Institute.

References

1. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, et al. The reality of pervasive transcription. *PLoS biology*. 2011; 9(7):e1000625. discussion e1001102. [PubMed: 21765801]
2. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A. Rfam 11.0: 10 years of RNA families. *Nucleic acids research*. 2013; 41(Database issue):D226–232. [PubMed: 23125362]
3. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. *Nucleic acids research*. 2003; 31(1):439–441. [PubMed: 12520045]
4. Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*. 2014; 157(1):77–94. [PubMed: 24679528]
5. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*. 2004; 116(2):281–297. [PubMed: 14744438]
6. Mattick JS, Makunin IV. Non-coding RNA. *Human molecular genetics*. 2006; 15:R17–29. Spec No 1. [PubMed: 16651366]
7. Brown EA, Zajac AJ, Lemon SM. In vitro characterization of an internal ribosomal entry site (IRES) present within the 5' nontranslated region of hepatitis A virus RNA: comparison with the IRES of encephalomyocarditis virus. *Journal of virology*. 1994; 68(2):1066–1074. [PubMed: 8289336]
8. Woo CJ, Kingston RE. HOTAIR lifts noncoding RNAs to new levels. *Cell*. 2007; 129(7):1257–1259. [PubMed: 17604716]
9. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. The transcriptional landscape of the mammalian genome. *Science*. 2005; 309(5740):1559–1563. [PubMed: 16141072]
10. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNAdb: a reference database for long noncoding RNAs. *Nucleic acids research*. 2011; 39(Database issue):D146–151. [PubMed: 21112873]
11. Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic acids research*. 2015; 43(Database issue):D168–173. [PubMed: 25332394]
12. Mattick JS, Rinn JL. Discovery and annotation of long noncoding RNAs. *Nature structural & molecular biology*. 2015; 22(1):5–7.
13. Carter RJ, Dubchak I, Holbrook SR. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic acids research*. 2001; 29(19):3928–3938. [PubMed: 11574674]
14. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS computational biology*. 2006; 2(4):e33. [PubMed: 16628248]

15. Babak T, Blencowe BJ, Hughes TR. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC bioinformatics*. 2007; 8:33. [PubMed: 17263882]
16. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*. 1990; 87(12):4576–4579. [PubMed: 2112744]
17. Pace NR, Smith DK, Olsen GJ, James BD. Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA--a review. *Gene*. 1989; 82(1):65–75. [PubMed: 2479592]
18. Conrad NK, Fok V, Cazalla D, Borah S, Steitz JA. The challenge of viral snRNPs. *Cold Spring Harbor symposia on quantitative biology*. 2006; 71:377–384.
19. Morris KV, Mattick JS. The rise of regulatory RNA. *Nature reviews Genetics*. 2014; 15(6):423–437.
20. Backofen R, Amman F, Costa F, Findeiss S, Richter AS, Stadler PF. Bioinformatics of prokaryotic RNAs. *RNA biology*. 2014; 11(5):470–483. [PubMed: 24755880]
21. Stadler PF. Class-specific prediction of ncRNAs. *Methods Mol Biol*. 2014; 1097:199–213. [PubMed: 24639161]
22. Koonin EV, Senkevich TG, Dolja VV. The ancient Virus World and evolution of cells. *Biology direct*. 2006; 1:29. [PubMed: 16984643]
23. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. *Nature reviews Genetics*. 2008; 9(4):267–276.
24. Lemey P, Rambaut A, Pybus OG. HIV evolutionary dynamics within and among hosts. *AIDS reviews*. 2006; 8(3):125–140. [PubMed: 17078483]
25. Bouvier NM, Palese P. The biology of influenza viruses. *Vaccine*. 2008; 26(Suppl 4):D49–53. [PubMed: 19230160]
26. Elena SF, Sanjuan R. Adaptive value of high mutation rates of RNA viruses: separating causes from consequences. *Journal of virology*. 2005; 79(18):11555–11558. [PubMed: 16140732]
27. Young LS, Rickinson AB. Epstein-Barr virus: 40 years on. *Nat Rev Cancer*. 2004; 4(10):757–768. [PubMed: 15510157]
28. Concha M, Wang X, Cao S, Baddoo M, Fewell C, Lin Z, Hulme W, Hedges D, McBride J, Flemington EK. Identification of new viral genes and transcript isoforms during Epstein-Barr virus reactivation using RNA-Seq. *Journal of virology*. 2012; 86(3):1458–1467. [PubMed: 22090128]
29. O'Grady T, Cao S, Strong MJ, Concha M, Wang X, Bondurant SS, Adams M, Baddoo M, Srivastav SK, Lin Z, et al. Global Bidirectional Transcription of the Epstein-Barr Virus Genome During Reactivation. *Journal of virology*. 2013
30. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic acids research*. 2015; 43(Database issue):D30–35. [PubMed: 25414350]
31. Baker W, van den Broek A, Camon E, Hingamp P, Sterk P, Stoesser G, Tuli MA. The EMBL nucleotide sequence database. *Nucleic acids research*. 2000; 28(1):19–23. [PubMed: 10592171]
32. Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic acids research*. 2002; 30(1):27–30. [PubMed: 11752245]
33. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*. 2007; 35(Database issue):D61–65. [PubMed: 17130148]
34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of molecular biology*. 1990; 215(3):403–410. [PubMed: 2231712]
35. Fernandez-Suarez XM, Rigden DJ, Galperin MY. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic acids research*. 2014; 42(Database issue):D1–6. [PubMed: 24316579]
36. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D. The influenza virus resource at the National Center for Biotechnology Information. *Journal of virology*. 2008; 82(2):596–601. [PubMed: 17942553]
37. Wilm A, Mainz I, Steger G. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms for molecular biology : AMB*. 2006; 1:19. [PubMed: 17062125]

38. Thompson JD, Linard B, Lecompte O, Poch O. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS one*. 2011; 6(3):e18093. [PubMed: 21483869]
39. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*. 2002; 30(14):3059–3066. [PubMed: 12136088]
40. Clote P, Ferre F, Kranakis E, Krizanc D. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*. 2005; 11(5):578–591. [PubMed: 15840812]
41. Simmonds P, Tuplin A, Evans DJ. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. *RNA*. 2004; 10(9):1337–1351. [PubMed: 15273323]
42. Davis M, Sagan SM, Pezacki JP, Evans DJ, Simmonds P. Bioinformatic and physical characterizations of genome-scale ordered RNA structure in mammalian RNA viruses. *Journal of virology*. 2008; 82(23):11824–11836. [PubMed: 18799591]
43. Simmonds P, Smith DB. Structural constraints on RNA virus evolution. *Journal of virology*. 1999; 73(7):5787–5794. [PubMed: 10364330]
44. Priore SF, Moss WN, Turner DH. Influenza A virus coding regions exhibit host-specific global ordered RNA structure. *PloS one*. 2012; 7(4):e35989. [PubMed: 22558296]
45. Gultyaev AP, Heus HA, Olsthoorn RC. An RNA conformational shift in recent H5N1 influenza A viruses. *Bioinformatics*. 2007; 23(3):272–276. [PubMed: 17090581]
46. Gultyaev AP, Olsthoorn RC. A family of non-classical pseudoknots in influenza A and B viruses. *RNA biology*. 2010; 7(2):125–129. [PubMed: 20200490]
47. Priore SF, Moss WN, Turner DH. Influenza B virus has global ordered RNA structure in (+) and (-) strands but relatively less stable predicted RNA folding free energy than allowed by the encoded protein sequence. *BMC research notes*. 2013; 6:330. [PubMed: 23958134]
48. Moss WN, Priore SF, Turner DH. Identification of potential conserved RNA secondary structure throughout influenza A coding regions. *RNA*. 2011; 17(6):991–1011. [PubMed: 21536710]
49. Moss WN, Steitz JA. Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA. *BMC genomics*. 2013; 14:543. [PubMed: 23937650]
50. Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF. Rnaz 2.0: Improved Noncoding Rna Detection. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*. 2010; 15:69–79. [PubMed: 19908359]
51. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(7):2454–2459. [PubMed: 15665081]
52. Washietl S. Prediction of structural noncoding RNAs with RNAz. *Methods Mol Biol*. 2007; 395:503–526. [PubMed: 17993695]
53. Washietl S, Hofacker IL. Identifying structural noncoding RNAs using RNAz. *Current protocols in bioinformatics / editorial board, Andreas D Baxeavanis [et al]*. 2007 Chapter 12:Unit 12 17.
54. Reiche K, Stadler PF. RNAstrand: reading direction of structured RNAs in multiple sequence alignments. *Algorithms for molecular biology : AMB*. 2007; 2:6. [PubMed: 17540014]
55. Ivica NA, Obermayer B, Campbell GW, Rajamani S, Gerland U, Chen IA. The paradox of dual roles in the RNA world: resolving the conflict between stable folding and templating ability. *Journal of molecular evolution*. 2013; 77(3):55–63. [PubMed: 24078151]
56. Moss WN, Lee N, Pimienta G, Steitz JA. RNA families in Epstein–Barr virus. *RNA biology*. 2014; 11(1):1–8.
57. Will S, Joshi T, Hofacker IL, Stadler PF, Backofen R. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA*. 2012; 18(5):900–914. [PubMed: 22450757]
58. Washietl S, Will S, Hendrix DA, Goff LA, Rinn JL, Berger B, Kellis M. Computational analysis of noncoding RNAs. *Wiley interdisciplinary reviews RNA*. 2012; 3(6):759–778. [PubMed: 22991327]

59. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW Jr, Swanstrom R, Burch CL, Weeks KM. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*. 2009; 460(7256):711–716. [PubMed: 19661910]
60. Stephenson JD, Li H, Kenyon JC, Symmons M, Klennerman D, Lever AM. Three-dimensional RNA structure of the major HIV-1 packaging signal region. *Structure*. 2013; 21(6):951–962. [PubMed: 23685210]
61. Charbonneau J, Gendron K, Ferbeyre G, Brakier-Gingras L. The 5' UTR of HIV-1 full-length mRNA and the Tat viral protein modulate the programmed -1 ribosomal frameshift that generates HIV-1 enzymes. *RNA*. 2012; 18(3):519–529. [PubMed: 22286970]
62. Vallejos M, Carvajal F, Pino K, Navarrete C, Ferres M, Huidobro-Toro JP, Sargueil B, Lopez-Lastra M. Functional and structural analysis of the internal ribosome entry site present in the mRNA of natural variants of the HIV-1. *PloS one*. 2012; 7(4):e35031. [PubMed: 22496887]
63. Moss WN, Dela-Moss LI, Kierzek E, Kierzek R, Priore SF, Turner DH. The 3' splice site of influenza A segment 7 mRNA can exist in two conformations: a pseudoknot and a hairpin. *PloS one*. 2012; 7(6):e38323. [PubMed: 22685560]
64. Moss WN, Dela-Moss LI, Priore SF, Turner DH. The influenza A segment 7 mRNA 3' splice site pseudoknot/hairpin family. *RNA biology*. 2012; 9(11):1305–1310. [PubMed: 23064116]
65. Dela-Moss LI, Moss WN, Turner DH. Identification of conserved RNA secondary structures at influenza B and C splice sites reveals similarities and differences between influenza A, B, and C. *BMC research notes*. 2014; 7(1):22. [PubMed: 24405943]
66. Pedersen JS, Meyer IM, Forsberg R, Simmonds P, Hein J. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic acids research*. 2004; 32(16):4925–4936. [PubMed: 15448187]
67. Tuplin A, Evans DJ, Simmonds P. Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequences of hepatitis C virus by RNase cleavage and novel bioinformatic prediction methods. *The Journal of general virology*. 2004; 85:3037–3047. Pt 10. [PubMed: 15448367]
68. Mathews DH, Moss WN, Turner DH. Folding and finding RNA secondary structure. *Cold Spring Harbor perspectives in biology*. 2010; 2(12):a003665. [PubMed: 20685845]
69. Uzilov AV, Keegan JM, Mathews DH. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC bioinformatics*. 2006; 7:173. [PubMed: 16566836]
70. Fu Y, Sharma G, Mathews DH. Dynalign II: common secondary structure prediction for RNA homologs with domain insertions. *Nucleic acids research*. 2014; 42(22):13939–13948. [PubMed: 25416799]
71. Mathews D. Predicting the secondary structure common to two RNA sequences with Dynalign. *Current protocols in bioinformatics / editorial board, Andreas D Baxevanis [et al]*. 2004 Chapter 12:Unit 12 14.
72. Xu Z, Mathews DH. Multilign: an algorithm to predict secondary structures conserved in multiple RNA sequences. *Bioinformatics*. 2011; 27(5):626–632. [PubMed: 21193521]
73. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*. 2010; 11:129. [PubMed: 20230624]
74. Smith C, Heyne S, Richter AS, Will S, Backofen R. Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA. *Nucleic acids research*. 2010; 38(Web Server issue):W373–377. [PubMed: 20444875]
75. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS computational biology*. 2007; 3(4):e65. [PubMed: 17432929]
76. Bussotti G, Raineri E, Erb I, Zytynicki M, Wilm A, Beaudoin E, Bucher P, Notredame C. BlastR--fast and accurate database searches for non-coding RNAs. *Nucleic acids research*. 2011; 39(16):6886–6895. [PubMed: 21624887]
77. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. 2013

78. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. *Bioinformatics*. 2009; 25(10):1335–1337. [PubMed: 19307242]
79. Moss WN. Computational prediction of RNA secondary structure. *Methods in enzymology*. 2013; 530:3–65. [PubMed: 24034313]
80. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC bioinformatics*. 2008; 9:474. [PubMed: 19014431]
81. Hofacker IL. RNA consensus structure prediction with RNAalifold. *Methods Mol Biol*. 2007; 395:527–544. [PubMed: 17993696]
82. Liu B, Mathews DH, Turner DH. RNA pseudoknots: folding and finding. *F1000 biology reports*. 2010; 2:8. [PubMed: 20495679]
83. Brierley I, Pennell S, Gilbert RJ. Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat Rev Microbiol*. 2007; 5(8):598–610. [PubMed: 17632571]
84. Sperschneider J, Datta A. DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model. *Nucleic acids research*. 2010; 38(7):e103. [PubMed: 20123730]
85. Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic acids research*. 1992; 20(21):5785–5795. [PubMed: 1454539]
86. Gulyaev AP, Tsyganov-Bodounov A, Spronken MI, van der Kooij S, Fouchier RA, Olsthoorn RC. RNA structural constraints in the evolution of the influenza A virus genome NP segment. *RNA biology*. 2014; 11(7):942–952. [PubMed: 25180940]
87. Gulyaev AP, Olsthoorn RC, Spronken MI, Fouchier RA. Detection of structural constraints and conformational transitions in the influenza virus RNA genome using structure predictions and mutual information calculations. *Proceedings IWBBIO 2014*. 2014
88. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*. 1999; 41:95–98.
89. Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*. 1988; 73(1):237–244. [PubMed: 3243435]
90. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research*. 2005; 33(2):511–518. [PubMed: 15661851]
91. Simmonds P. SSE: a nucleotide and amino acid sequence analysis platform. *BMC research notes*. 2012; 5:50. [PubMed: 22264264]
92. Harmanci AO, Sharma G, Mathews DH. TurboFold: iterative probabilistic estimation of secondary structures for multiple RNA sequences. *BMC bioinformatics*. 2011; 12:108. [PubMed: 21507242]

Highlights

A review of methods for the *in silico* discovery and modeling of non-coding RNA structure in viruses is presented.

Methods based on free energy minimization are discussed.

Methods based on sequence analysis are discussed.

Examples are given from previous work on influenza A and Epstein–Barr virus.

A brief protocol is presented using influenza A segment 5 as an example.

Figure 1.

Analysis of IAV segment 5. (A) At the top is a plot of the calculated SSCU across 300 IAV segment 5 sequences. Low SSCU values indicate lower than expected variation at the third codon position, which suggest constraints on codon evolution (due to RNA structure or other constraints). Below the SSCU plot are results for RNAz predictions. Each 120 nt prediction window with class probability > 0.5 is shown. Darker blue color shows higher class probability where the darkest window is 1.00 (maximal value). The arrow point indicates strand (rightward is forward (+)RNA strand and rearward is the reverse (–)RNA strand). (B) Consensus secondary structure for the RNAz prediction window spanning nts 1420–1530. Prediction based on the six RefSeqs used in RNAz via LOCARNA-P. Base pairs are colored based on their variation: red is invariant, yellow has two different base pairs, and green has three different base pairs. The opacity of the color indicates conservation: fully opaque color is 100% conserved, while lower opacity indicates that base pairing is not conserved (base pairs formed within nts 1428–1530 are 100% conserved). Consistent and compensatory mutations are indicated by circles around the mutated base. (C) Structure model of a hairpin formed using nts 1476–1530. The sequence shown is from the H5N1 avian flu strain (GenBank # NC_007360.1). All nts in the hairpin fall within the coding region for IAV NP protein. Red bases are at third codon positions. Two base pairs are boxed and MI scores are given as are the percentage of observed base pairs in the alignment of all available IAV sequences (18,500 sequences).

Table 1

Partial list of sequence alignment programs

Name	Description	Access
ClustalW	Progressive alignment *	www.clustal.org
MUSCLE	Progressive/iterative alignment	www.drive5.com/muscle/
T-Coffee	Progressive alignment	www.tcoffee.org/Projects/tcoffee/
Multi-LAGAN	Dynamic programming alignment; good for long (e.g. genomic) sequences	http://lagan.stanford.edu/lagan_web/citing.shtml
MAFFT (FFT-NS-1)	Progressive alignment - very fast; good for many sequences	http://mafft.cbrc.jp/alignment/software/
MAFFT (E-INS-i)	Iterative alignment - very slow; good for <200 sequences with conserved domains	"
MAFFT (L-INS-i)	Iterative alignment - very slow; good for <200 sequences with one conserved domain	"
MAFFT (G-INS-i)	Iterative alignment - very slow; good for <200 sequences with global homology	"
MAFFT (Q-INS-i)	Iterative alignment - extremely slow; good for <200 sequences with RNA structure	"

* Can also be run through BioEdit alignment Editor, where sequences can be toggled between DNA/RNA/Protein to allow for amino acid based alignments of coding nucleic acids.

Table 2

Partial list of ncRNA discovery programs

Name	Description	Access
RNAz	SVM - scans aligned nt sequences	www.tbi.univie.ac.at/~wash/RNAz/
RNAstructure (Multifind)	SVM - scans and re-aligns for structure	http://rna.urmc.rochester.edu/RNAstructure.html
EvoFold	SCFG - scans aligned nt sequences	https://users.soe.ucsc.edu/~jsp/EvoFold/
QRNA	SCFG - scans aligned nt sequences	http://selab.janelia.org/software.html
RNA-DECODER	SCFG - scans aligned coding nt sequences	www.e-rna.org/rnadecoder/help.cgi
CMfinder	SCFG - predicts for local alignments	http://bio.cs.washington.edu/yzizhen/CMfinder/

Table 3

Partial list of RNA folding programs

Name	Description	Access
Vienna RNA package (RNAfold)	Dynamic programming - single sequence folding	http://www.tbi.univie.ac.at/RNA/
RNAstructure (Fold)	Dynamic programming - single sequence folding	http://rna.urmc.rochester.edu/software.html
Mfold	Dynamic programming - single sequence folding	http://mfold.rna.albany.edu/?q=mfold/download-mfold
DotKnot	Dynamic programming - single sequence folding of pseudoknots	http://dotknot.csse.uwa.edu.au
RNAstructure (ProbKnot)	Dynamic programming - single sequence folding of pseudoknots	http://rna.urmc.rochester.edu/software.html
Vienna RNA package (RNAalifold)	Dynamic programming - consensus folding fixed alignment	http://www.tbi.univie.ac.at/RNA/
RNAstructure (Multalign)	Dynamic programming - consensus folding and alignment	http://rna.urmc.rochester.edu/software.html
LOCARNA	Sankoff's algorithm - consensus folding and alignment	http://www.bioinf.uni-freiburg.de/Software/

Table 4

Analysis of IAV structured region

i	j	GC	CG	AU	UA	GU	UG	BP cons.	MI(i,j)
1476	1530	0	4927	0	1542	0	11940	99.5%	0.02
1477	1529	0	0	0	18492	0	2	100.0%	0.00
1478	1528	0	18477	0	0	0	11	99.9%	0.00
1479	1527	1	14239	0	3470	58	586	99.2%	0.39
1480	1526	0	0	0	18488	0	3	100.0%	0.00
1481	1525	0	1	0	24	0	18468	100.0%	0.00
1482	1524	0	0	0	18007	0	57	97.7%	0.00
1483	1523	18473	0	0	0	2	0	99.9%	0.00
1485	1522	1	15697	0	2	0	2261	97.1%	0.00
1486	1521	0	0	18437	0	18	0	99.8%	0.00
1487	1520	0	0	0	18407	0	54	99.8%	0.00
1490	1518	16132	0	0	0	12	0	87.3%	0.00
1491	1517	0	0	0	17924	0	4	96.9%	0.00
1494	1513	0	3204	0	1	0	14910	97.9%	0.00
1495	1512	18375	0	0	0	120	0	100.0%	0.00
1496	1511	0	0	18489	0	5	0	100.0%	0.00
1497	1510	0	0	16127	0	2348	0	99.9%	0.00
1498	1509	18463	0	0	0	20	0	99.9%	0.00
1499	1508	9	0	2	0	18486	0	100.0%	0.00
1500	1507	0	0	14060	0	4207	0	98.7%	0.00