Methods 91 (2015) 48–56

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

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

Walter N. Moss, Joan A. Steitz *

Department of Molecular Biophysics and Biochemistry, Howard Hughes Medical Institute, Yale University School of Medicine, New Haven, CT 06536, USA

ARTICLE INFO

Article history:
Received 6 April 2015
Received in revised form 17 June 2015
Accepted 22 June 2015
Available online 23 June 2015

Keywords:
ncRNA prediction
Viruses
RNA
RNA structure
ncRNA
Sequence analysis

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.

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] 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 Ref. [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

* Corresponding author.
E-mail address: joan.steitz@yale.edu (J.A. Steitz).

http://dx.doi.org/10.1016/j.ymeth.2015.06.015
1046-2023/

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 to 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