3S: Shotgun secondary structure determination of long non-coding RNAs

Irina V. Novikova, Ashutosh Dharap, Scott P. Hennelly, Karissa Y. Sanbonmatsu *

Theoretical Biology and Biophysics, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, United States

A R T I C L E   I N F O
Article history:
Available online 6 August 2013

Keywords:
Long non-coding RNA structure
IncRNA structure
IncRNA
Long non-coding RNA
Secondary structure
Steroid receptor RNA activator

A B S T R A C T
Long non-coding RNAs (IncRNAs) have emerged as an important class of RNAs playing key roles in development, disease and epigenetics. Knowledge of IncRNA structure may be critical in understanding function for many IncRNA systems. Due to the enormous number of possible folds for these sequences, secondary structure determination presents a significant challenge, both experimentally and computationally. Here, we present a new strategy capable of determining the RNA secondary structure in the wet lab without significant reliance on computational predictions. First, we chemically probe the entire IncRNA. Next, using a shotgun approach, we divide the RNA into overlapping fragments and probe these fragments. We then compare probing profiles of fragments with the profiles of the full RNA and identify similarities. Sequence regions with profiles that are similar in the fragment and full-length transcript possess only base pairing partners within the fragment. Thus, by experimentally folding smaller and smaller fragments of the full RNA and probing these chemically, we are able to isolate modular sub-domains, dramatically reducing the number of possible folds. The method also eliminates the possibility of pseudo-knots within a modular sub-domain. The 3S technique is ideally suited for IncRNAs because it is designed for long RNA sequences. The 3S-determined secondary structure of a specific IncRNA in one species (e.g., human) enables searches for instances of the same IncRNA in other species.

© 2013 Published by Elsevier Inc.

1. Introduction

The nascent field of long non-coding RNAs is expanding at an exponential pace [1]. These RNAs have been shown to play key roles in development [2], plant biology [3,4], embryonic stem cell pluripotency [5], brain function [6–8], subcellular compartmentalization [9,10], chromatin remodeling [11–13], and cancer [14,15]. Long noncoding RNAs (IncRNAs) are typically defined by two major characteristics: (i) length of the transcript (>200 nts) and (ii) having little potential for translation [16]. Transcript lengths of IncRNAs typically range from 1 to 10 kbp, while systems as large as ~100 kbp have been studied [17–19]. While many IncRNAs originate in intergenic regions, a substantial number are transcribed from the antisense strand of protein-coding genes, overlapping one or more exons. IncRNAs may (1) recruit epigenetic factors to chromatin, (2) act as scaffolding for chromatin, (3) act as decoys for proteins and microRNAs, or (4) propagate allosteric signals. The number of IncRNAs in humans is estimated to be ~15,000 [20]. In light of their large size and relatively recent discovery, few structural studies of intact IncRNAs have been performed [21].

Secondary structure often plays a critical role in RNA mechanism, underpinning overall tertiary architecture by defining sub-domains, helices, bulges, stem-loops, internal loops, and junctions. Secondary structure alone can provide the basis of function when (i) the secondary structure defines a platform for unique protein recognition or (ii) a single sequence produces two competing secondary structures, as in the case of the riboswitch, which represents a quintessential secondary structure based RNA mechanism [22–27]. An example of case (i) is the growth-arrest specific IncRNA (gas5) [28]. This IncRNA contains a hairpin element responsible for regulation of the glucocorticoid receptor protein via a decoy mechanism. An example of case (ii) is a riboswitch RNA [23]. In this case, a single sequence of RNA has two different secondary structures that compete with each other to determine the outcome of gene expression. Often, ligand concentration shifts the equilibrium between secondary structures, resulting in the formation or destruction of a transcriptional terminator helix. In riboswitch studies, the action mechanism (i.e., that the RNA is a ligand-based molecular switch) was determined by studying the secondary structure, without the need for tertiary structure studies. Breaker and co-workers, who have discovered most of the known riboswitches, used chemical probing experiments to determine the secondary structure of the riboswitch in the presence or absence of the ligand. Their technique has been validated in many crystallographic studies [29–32]. Determination of the secondary structure of IncRNAs is a hard problem: the number of possible secondary structure solutions grows exponentially with the transcript length. The number of possible secondary structures $S$ for a transcript of length $L$ with $k$ base pairs is given by:

$S \approx e^{k^2/4L}$

* Corresponding author. Fax: +1 505 655 3493. 
E-mail address: kys@lanl.gov (K.Y. Sanbonmatsu).

1046-2023/$ - see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.ymeth.2013.07.030
The recent development of the selective 2'-hydroxyl acylation by primer extension (SHAPE) technique has produced a tremendous amount of data related to the secondary structures of RNA sequences, providing extremely useful constraints on the secondary structure (Fig. 1) [33,43]. Here, more reactive regions of the RNA are likely to be single stranded and less reactive regions are likely to be base paired. In addition, recent high-throughput techniques may produce chemical probing data on a large scale [44–47]. These chemical and enzymatic probing tools are extremely useful and provide means for initial discrimination between single stranded and base paired regions of a given molecule. These methods alone, however, do not identify base pairing partners of a given residue. Even with experimental probing data in hand, there are still numerous RNA folds consistent with the probing data. In particular, probing data alone cannot distinguish close-range and long-range base pairing interactions. Thermodynamic algorithms, when combined with probing, have been successful for known RNA folds such as the 16S rRNA [33,43]. While an excellent advance forward, these methods have not solved the folding problem for large RNA systems, where one SHAPE profile often corresponds to a large number of possible folds that are not reliably distinguished by current computational methods. The large and fast-evolving nature of lncRNA systems makes the problem even more challenging. For example, the evolutionarily ancient ribosome has thousands of species to draw from, allowing secondary structure determination from a small number of chemical probing experiments and very extensive thermodynamic analysis. For example, the evolutionary ribosome has thousands of species to draw from, allowing secondary structure determination from a small number of chemical probing experiments and very extensive thermodynamic analysis. As most lncRNAs evolved recently (e.g., only ~45 sequences are available for SRA) [21], a more rigorous experimental approach is required.

Our shotgun structure determination methodology applies SHAPE probing in a hierarchical fashion, probing the entire RNA, followed by subsequent probing of overlapping fragments of smaller and smaller size (Fig. 2). The method locates sub-domains by isolating segments of the lncRNA that have very similar chemical probing profiles in isolation and in the context of the full lncRNA. In short, we eliminate possible folds, allowing us to zero in on the correct secondary structure. To help validate the structure,

\[
S(L,k) = \frac{1}{k} \binom{L-k}{k+1} \binom{L-k+1}{k-1}
\]

A rough estimate for a sequence of length 76 with 21 base pairs (e.g., tRNA\textsuperscript{Met}) yields \(\sim 10^{38}\) possible secondary structures, in comparison to \(10^{38}\) for a sequence of length 1000 with 200 base pairs (e.g., a relatively small lncRNA such as the steroid receptor RNA activator). A lncRNA of length 2.2 kbp with 500 base pairs has approximately \(10^{688}\) possible secondary structures. While many computational techniques have been developed to predict secondary structure, predicting the secondary structure of long RNA sequences remains a significant challenge [33–37]. The major methods of structure prediction include free energy approaches, phylogenetic approaches, machine learning approaches, and various combinations of these three techniques [34–36]. Minimum free energy approaches have played an important role in secondary structure prediction and are used widely. For long RNA sequences (e.g., \(N > 500\) where \(L\) is the length of the sequence), these methods tend to produce a few very long helices. Structures with fewer junctions and fewer internal loops will often have significantly lower predicted energies in comparison to structures with many junctions. By comparison, long RNAs with structures based on chemical probing and phylogenetic comparison and validated with X-ray crystallography, such as the ribosome and group I and II introns, have many junctions and helices. Since minimum free energy predictions are often based on thermodynamic rules determined from experiments on large numbers of short duplexes [38–40], these methods are challenged by long sequences. More recently, automated phylogenetic approaches have been employed. These leverage covariance of base pairs across multiple species and incorporate this information into stochastic context-free grammar models [36,41,42]. The computational complexity of the most rudimentary of these algorithms scales as order \(L^3N\), where \(N\) is the number of species and \(N \gg 1\) is required for accurate predictions (e.g., \(N > 1000\) for ribosomes and riboswitches). Thus, for long sequences, these methods are traditionally deemed computationally prohibitive for analysis of the entire sequence. Windowed searches, however, may be used to identify localized structure.

![Fig. 1. Selective 2’-hydroxyl acylation by primer extension (SHAPE) using capillary electrophoresis. After RNA is folded (1), the SHAPE reagent (1M7) is added. The reagent reacts more strongly with nucleotides that are highly mobile (2). Modified nucleotides of RNA cause premature stops in primer extension. Nucleotides within helices are less reactive and display weak or no reactivity with 1M7. A fluorescent primer is used for reverse transcription (3). Capillary electrophoresis is used for analysis. High intensity peaks correspond to highly reactive and highly mobile nucleotides. Peaks with lower intensities correspond to nucleotides with low mobility.](image-url)
we repeat this process with other probing methods (DMS, RNase V1, and in-line probing). We have demonstrated the methodology on the test system, steroid receptor RNA activator (SRA), which is large enough (874 nts) to be considered a typical long non-coding RNA, yet small enough to be tractable for this first study. Our new methodology, however, is scalable to much larger RNA systems. Scalability may be achieved by applying the method in a hierarchical fashion, employing many rounds of probing with successively smaller and smaller RNA fragments that overlap and tile the entire lncRNA.

2. Description of the method

2.1. Step 1

Step 1 is to efficiently design the RNA fragments for shotgun secondary structure determination (Fig. 3). In the example we present, we select fragments randomly. However, a more efficient strategy is depicted in Fig. 3. Here, 5 fragments are chosen to obtain information about modularity of the secondary structure at different levels of resolution. Modification sites produced by treating RNA fragments 1–5 with various chemical reagents will be compared to the modification profile of the full transcript. In addition, the probing signal for fragment 4 will be compared to both the full transcript and fragment 1 to obtain information about region E. Likewise, the probing signal for fragment 4 will also be compared to both the full transcript and fragment 2 to obtain information about region F.

2.2. Step 2

Step 2 involves the preparation of RNA molecules. The full length RNA and its fragments are synthesized by run-off in vitro transcription from dsDNA templates, either chemically synthesized or obtained from available cDNA clones. If cDNA is not available, for longer RNAs (>200 nts), dsDNA templates can be generated from smaller DNA fragments (ultamers of 150–200 nts, IDTDNA) carrying overlapping regions or using other custom gene services.
2.3. Step 3

In step 3, the intact RNA and its fragments are chemically probed. In some cases, it may be beneficial to perform additional probing studies using in-line probing, DMS probing, or RNase V1 digestion.

2.4. Step 4

In step 4, probing sites are analyzed by reverse transcription using multiple site-specific primers. The modification/cleavage sites of the RNA are analyzed by primer extension of fluorophore-labeled primers with the SuperScript III reverse transcriptase from Invitrogen. In particular, 5’-end fluorescently labeled primers are annealed to the RNAs. Reverse transcription generates premature termination stops at positions preceding modified nucleotides (Fig. 1).

2.5. Step 5

In step 5, capillary electrophoresis is used to resolve the cDNA, with a systematic quantification of reactivities for each nucleotide.

2.6. Step 6

In step 6, the probing reactivities of each RNA fragment are compared with the probing reactivities of full-length RNA. If a significant overlap in the probing profile is observed between the fragment and full-length RNAs, then this region of RNA is modular, folding independently of the rest of the sequence via close-range base pairing. This step eliminates the search for other potential secondary structures over the entire full-length transcript. The structure of small-size RNAs (<100 nts) can be reliably predicted when constraints from SHAPE reactivities are incorporated into thermodynamic algorithms [33,43]. If necessary, additional fragmentation can be performed until the size of the fragment is <100 nts. Once all local modular RNA sub-folds are determined via fragmentation, the remainder of the full-length RNA can be searched for evidence of other long-range base pairs. In this manner, modular sub-domains are brought together to define the final secondary architecture.

2.7. Step 7

In step 7, covariance analysis is used to help verify the secondary structure. The covariance analysis is performed by obtaining multiple sequence alignments for the RNA and counting base pair flips (e.g., AU → UA, AU → GC, etc.) in each helix.

3. Experimental results

3.1. Chemical probing of the intact long RNA

We performed SHAPE probing on the intact RNA and analyzed with reverse transcription and capillary electrophoresis. Processed reactivities for the entire RNA are shown in Fig. 4. High reactivities correspond to single-stranded nucleotides; low reactivities correspond to base paired nucleotides. Overall, the traces reveal a high degree of structure in the RNA with many different single-stranded and base paired regions; however, the SHAPE reactivity of the intact RNA alone is not sufficient to determine proper base pairing partners of the nucleotides inside a base pairing region. These nucleotides may pair with other nucleotides within the same base pairing region or with nucleotides in distant base pairing regions.

3.2. Fragment analysis to determine sub-domains

A useful strategy to determine the structural fold is to eliminate alternative folds. If structural elements are formed via

![Fig. 4. A worked example: shotgun secondary structure (35) analysis of the human steroid receptor RNA activator lncRNA (SRA). Similarities in SHAPE reactivities between the full RNA and fragments delineate sub-domains that fold into modular units and do not base pair with nucleotides outside of these sub-domains. (A) SHAPE probing reactivities of full intact RNA transcript. (B) SHAPE probing reactivity of fragment 1 (nucleotide positions 213–409). (C) SHAPE probing reactivity of fragment 2 (nucleotide positions 260–480). (D) SHAPE probing reactivity of fragment 3 (nucleotide positions 460–680). (E) SHAPE probing reactivity of fragment 4 (nucleotide positions 651–874). (F and G) Examples of modular regions with corresponding modular folds. Modularity was demonstrated by the similarity between the SHAPE reactivity profiles of regions in the boxes of (C and D) in comparison with (A). Orange, high reactivity; yellow, medium reactivity; grey, low reactivity; no color, very little reactivity. Shaded boxes (B and E), examples of regions with substantial differences in comparison to (A), representing non-modular regions of the lncRNA. Nucleotides in these regions pair with nucleotides outside of these regions in the context of the full RNA. The fragment profiles suggest that these regions exist in a different fold when studied as isolated fragments.](image-url)
close-range base pairing in the context of the entire RNA, we expect them to also base pair in the context of smaller fragments. By performing SHAPE experiments on smaller fragments of the SRA sequence, we are able to determine modular structural elements and simultaneously eliminate a large number of other possible folds. To test this strategy, we randomly chose four fragment sequences (~200 nts each) that overlap regions of SRA sequence between positions 213–409, 260–480, 460–680 and 651–874 (Fig. 4). The second fragment (260–460) has significant overlap in SHAPE reactivity with the SHAPE reactivity profile of the full RNA for positions 363–446. The overlap in SHAPE data suggests that the region occupying positions 363–446 forms an autonomous secondary structure corresponding to helices H12 and H13 in the SRA secondary structure. The third fragment (positions 460–680) was almost entirely modular, corresponding to helices H15, H16, H17, H19, H20, and H21. The SHAPE reactivities of this stretch of RNA are very similar to the SHAPE reactivities of this sequence in the context of the entire SRA. Performing SHAPE probing on fragments of the intact RNA provided information critical to the solution of the secondary structure of the entire RNA.

3.3. Secondary structure determination

Using a combination of probing and fragment analysis, we present a secondary structure in Fig. 5, annotated with SHAPE reactivity values. Looping regions are extensively modified by the chemical probes, targeting single-stranded nucleotides. Our structure consists of four sub-domains, with various secondary elements ranging from autonomous, small helical stems such as H1 and H2 of domain I to larger structures formed via long-range base pairing. The first modular region identified by fragment analysis comprises well-defined helices H12 and H13 of Domain II. The second modular region, identified with fragment analysis, comprises helices H15, H16, H17, H18, H19, H20 and H21 of domain III, which base pair in a close-range and are limited to this sequence. Domain I contains helices H1, H2, H3, H4, H5, H6 and H7. Domain II contains helices H10, H11, H12, H13 and H14. Domain III contains H8, H9, H15, H16, H17, H18, H19, H20 and H21. Domain IV contains helices H22, H23, H24 and H25. The overall helical composition of the lncRNA accounts for 48% of the total number of SRA nucleotides with a helical density of one helix per thirty-four residues.

3.4. Validation with other methods of probing

Since SHAPE probing, in-line probing, DMS probing and RNase V1 digestion each rely on different chemical mechanisms, each method yields complementary information about the RNA structure. While SHAPE probing is highly sensitive to nucleotides with moderate mobility, in-line probing often produces higher contrast between base paired and non-base paired nucleotides. Both SHAPE and in-line probing detect mobility and solvent exposure of the 2'-OH group. DMS probing detects exposure of the base. RNase V1 digests helical regions, yielding complementary information to SHAPE, in-line and DMS probing. Our results show the four methods to be consistent and complementary with each other, producing a consistent secondary structure.

3.5. Validation with covariance analysis

To perform covariance analysis, we obtained a multiple sequence alignment of SRA across 45 species from the ENCODE database [48]. Using an algorithm similar to that employed by Hofacker and co-workers, we found covariant base pairs in the vast majority of helices in the structure, verifying our fold [49].

4. Detailed protocol

4.1. Step 1: Design of RNA fragments for shotgun analysis

An efficient strategy for fragment analysis is shown in Fig. 3

1. The full length transcript (fragment 0) should be bisected once, producing fragments 1 and 2.
2. The resulting two fragments should then be bisected again. Fragment 3 runs from the 5'-end of the full transcript to the midpoint of fragment 1. Fragment 5 runs from the midpoint of fragment 2 to the 3'-end of the full transcript.
3. Obtain the final fragment. Fragment 4 runs from the midpoint of fragment 1 to the midpoint of fragment 2. We note that the fragmentation approach has been successful for RNAs of lengths ~1 kb. For significantly longer RNAs, additional rounds of fragmentation can be performed.

4.2. Step 2: RNA synthesis

1. Use dsDNA templates in run-off transcription. We utilize the T7-scribe Standard RNA IVT kit from CELLSKRIPT.
2. Extract the RNA products with phenol–chloroform and precipitate with the addition of one volume of 5 M ammonium acetate and 2.5 volumes of ethanol.
3. Check the integrity of the RNA on agarose and polyacrylamide gels.

4.3. Step 3: Chemical probing

1. Denature RNA in water at 94 °C for 2 min and snap-cool on ice.
2. Fold RNA (250 nm) in 1X HMK buffer (50 mM HEPES-NaOH pH 8.0, 100 mM KCl, and desired concentration of MgCl₂) for 30 min at 37 °C. Note that, for in-line probing, a slightly different protocol is used (described below).
3. For SHAPE probing, synthesize 1M7 from 4-nitroisatoic anhydride using protocol developed Mortimer and Weeks. Adjust folded RNA with 1M7 (dissolved in DMSO) to a final concentration of 3 mM and incubate at 25 °C for 20 min. Ideally, the final concentration of 1M7 should be optimized for each new RNA through titration experiments. Treat parallel RNA samples with the same amount of pure DMSO to obtain the blank. Collect modified RNAs using the standard sodium acetate/ethanol precipitation technique.
4. For DMS probing, add 1/20th volume of 10% DMS in ethanol (or pure ethanol for blank trace) to the folded RNA and incubate for 1 h on ice. Quench with the addition of one volume of stop solution (1 M Tris–HCL pH = 8.0; 1 M B-mercaptoethanol, 1 M sodium acetate). To precipitate the alkylated RNAs, add 2.5 volumes of ethanol to the mixture followed by incubation at −80 °C and centrifugation.
5. For in-line probing, fold RNA in a 1X in-line probing buffer (50 mM Tris–HCL pH = 8.3; 20 mM MgCl₂, 100 mM KCl) and incubate for 46 h at 25 °C. Precipitate the products of the in-line cleavage using sodium acetate/ethanol.
6. For RNase V1 digestion, perform serial dilutions of RNase V1 (Ambion) to optimize the conditions of the cleavage. After optimization, perform digestion reactions for 20 min at 25 °C, followed by the addition of the precipitation/inactivation buffer supplied with the RNase V1 enzyme.
4.4. Step 4: Primer design and reverse transcription

1. Design the primers with the desired Tm. Primers should target regions of the IncRNA separated by approximately 200 nts or greater.

2. Primers can be ordered directly with attached fluorophore at the 5'-end or labeled in-house. For example, in-house labeling with Alexa 488 can be achieved by using DNA oligos synthesized with an amino moiety on their 5'-end (IDTDNA) and an Alexa Fluor 488 amine reactive ester (Invitrogen).

3. Purify the fluorophore-labeled primers on reverse phase HPLC.

4. We prefer to set up reverse transcription reactions in small volumes of 8–15 µl. For the example of an 8 µl reaction, mix 6 pmol of RNA (2 µl) with 2 pmol of site-specific primer (1 µl), 1 µl of water and 1 µl of dNTP mix (2.5 mM).

5. Heat for 5 min at 65 °C and place on ice.

6. Add 2 µl of 4X Reverse Transcription buffer and 1 µl of Superscript III (200 U/ml). Prepare 4X Reverse Transcription buffer by combining 4 parts of 5X First Strand buffer and 1 part of 0.1 M DTT supplied with the enzyme.

Fig. 5. Secondary structure of the steroid receptor RNA activator (SRA) IncRNA determined with the shotgun secondary structure determination method (35). In addition to probing the full RNA, SHAPE analysis of fragments of the RNA eliminates alternative folds. Nucleotides are colored by SHAPE reactivity: Orange, high reactivity; yellow, medium reactivity; grey, low reactivity; black, very little reactivity; green, 5'- and 3'-end regions. Insets display SHAPE profiles for boxed regions.
7. Incubate for 40 min at 55 °C followed by additional 15 min at 70 °C for enzyme inactivation. The optimal incubation temperature depends on the Tm of the primers. The optimal range is 50–55 °C.
8. Dilute with water to a final volume of 50 µl and desalt with micro Bio-Spin columns filled with Bio-gel P6 (Bio-Rad Life Science). Desalting is critical for obtaining high resolution traces by capillary electrophoresis.
9. Mix the desired amount of desalted mix with formamide, denature at 94 °C for 2 min and load onto capillary. For example, one possibility is 0.045 pmol of cDNA in 20 µl of formamide.
10. Perform two dideoxy sequencing reactions in parallel (A-sequencing and C-sequencing), where 1 µl of water is substituted with 1 µl of 1 mM ddNTP (ddTTP for A-sequencing and ddGTP for C-sequencing).

4.5. Step 5: Capillary electrophoresis and trace processing

1. Run sequencing and probing primer extension reactions on ABI PRISM 3100-Avant genetic analyzer using either 50 cm or 80 cm capillaries loaded with POP-6 polymer. The traces can be manually aligned.
2. Integrate peaks by fitting Gaussians to each peak.
3. Correct probing reactivity traces for exponential decay using statistical model implemented in ShapeFinder software.
4. Subtract reverse transcription stops observed in the blank traces from the probing traces.
5. Normalize the traces by the average reactivities for highly reactive nucleotides such that the range extends from 0 to 1.5, as recommended by Vasa and co-workers.
6. As an alternative to the above processing steps, traces can be aligned and integrated using the ShapeFinder or QuSHAPE packages [50,51].

4.6. Step 6: Obtain the secondary structure by first identifying regions of the RNA with modular structure

1. To identify large sub-domains, compare (i) the probing signals of fragment 0 with fragment 1, (ii) the probing signals of fragment 0 with fragment 2 and (iii) the probing signals of fragment 0 with fragment 3. Regions of similarity are modular sub-domains. Specifically, similarity between two signal profiles occurs when the two profiles are consistent with nearly identical secondary structures. A more detailed analysis of the pattern similarity could be obtained with signal processing algorithms such as dynamic time warping. In general, while the 3S method does not identify tertiary contacts, our method does eliminate the possibility of tertiary contacts (including pseudoknots) within a modular domain.
2. To identify smaller modular regions, compare (i) the probing signals of fragment 0 with fragment 0, (ii) the probing signals of fragment 0 with fragment 5, (iii) the probing signals of fragment 0 with fragment 2 and 4, and (iv) the probing signals of fragment 0 with fragment 1 and 4. Regions of similarity represent smaller modular regions.
3. For the smaller modular regions, search for Watson–Crick pairs in low reactivity regions. Software such as RNAstructure can be used [33,43]. Because each of these regions is modular, nucleotides in each modular region only pair to other nucleotides in the same region. Therefore, these nucleotides do not need to be considered for pairing in the next sub-step (4).
4. Search for larger sub-domains. Only nucleotides outside of the regions identified in sub-step 3 need to be searched for base pairing partners.
5. Search for Watson–Crick pairs between regions connecting modular regions identified in sub-step 4.
6. If multiple folds persist, repeat probing with even smaller fragments.

4.7. Step 7: Covariance analysis

1. Obtain multiple sequence alignment for full transcript.
2. For each base pair of each helix, search for base pairs that (i) are conserved across species and (ii) vary in sequence (e.g., AU → UA, AU → GC).

5. Concluding remarks

Long non-coding RNAs (lncRNAs) have been classified into guide/targeting RNAs, decoy RNAs, control device RNAs and scaffolding RNAs [1,52,53]. Structure of the lncRNA may play a role in action mechanism for each of these classes. Because scaffolding lncRNAs are thought to stabilize chromatin structure, the structure of these lncRNAs may be directly related to their function. In this article, we have presented a new experimental strategy (3S, or Shotgun secondary structure) for determining RNA secondary structures of long sequences with low sequence conservation. The method compares chemical probing profiles of the intact full-length transcript with profiles of fragments of the transcript, allowing identification of modular sub-domains, eliminating a large number of possible secondary structures from consideration. In past RNA studies, secondary structure has played an important role in elucidating mechanism. In the case of riboswitch RNAs, for example, the switch mechanism was determined entirely by secondary structures without need for tertiary structure determination. In the case of the ribosome, secondary structure studies provided a framework for tRNA binding studies, ribosomal protein binding studies, and preliminary tertiary structure studies. An important long-term goal is to obtain the kinetics of IncRNA–protein interactions, which has been achieved for other RNA–protein systems [54]. Secondary structure studies of IncRNAs create an essential framework for such mechanistic studies as well as for more detailed three-dimensional structural studies. In addition to IncRNAs, our method will be useful for determining secondary structures of mRNAs [55], including the 5′-UTR, 3′-UTR and coding regions.

In addition to mechanistic understanding, these experimentally determined secondary structures will be useful benchmarks for improving RNA secondary structure prediction algorithms. To date, most prediction algorithms have been limited to the secondary structure of the 16S rRNA to test the accuracy on long RNA sequences. However, the 16S rRNA is not necessarily representative of large RNA systems. For example, messenger RNAs are thought to be significantly less structured in comparison to ribosomal RNAs [45]. The 3S method now allows rigorous experimental determination of large RNA secondary structures that can be used to validate existing and new prediction algorithms.

In the future, an obvious extension to the 3S method is the probing of smaller and smaller fragments, eliminating all need of computational prediction. Validation in vivo represents a second avenue of future research. A challenge that remains to be addressed is disentangling populations of transcripts containing many different folds simultaneously. We note that an important method of validation of the 3S structure is covariance analysis. The 3S-determined structure of a specific IncRNA represents an excellent starting point enabling the search for instances of the same IncRNA in other species. Due to relatively low sequence identity of many functional RNAs (e.g., U2 and U4 spliceosomal RNAs, 5S rRNA, group I intron and riboswitches), it is often difficult to ascertain the degree of conservation from sequence alone.
For example, the TPP riboswitch has a nucleotide sequence identity of only 57%. Its secondary structure, however, is conserved across 11,197 sequences [56]. BLAST searches for similar RNAs return either (i) no results or (ii) an enormous number of false positives, depending on the threshold used. Searches incorporating secondary structure, however, can produce meaningful results, allowing one to more accurately assess the level of conservation secondary structure, however, can produce meaningful results, understanding their mechanism and establishing a framework for therapeutic studies.

Acknowledgement

The work was performed under the auspices of the US Department of Energy via LANL LDRD.

References