Experimentally-guided computational RNA structure modeling

Clarence Cheng
BIOC218
December 9, 2012

Introduction

Determination of the three-dimensional structures of ribonucleic acids (RNA) is critical for understanding many molecular mechanisms of life and disease. In particular, all forms of RNA, from non-coding regulatory RNA (ncRNA) to messenger RNA (mRNA) have been identified as possessing structure that is important for function. A multitude of techniques, both experimental and computational, have been used to gain structural information about RNA. Structures derived from high-resolution experimental techniques, such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, are generally regarded as the 'gold standard' of structure determination, with many structure-based biochemical studies supporting the functional models that such structures suggest. However, these high-resolution techniques are limited in applicability based on the complexity of the macromolecules being studied – for example, large and conformationally heterogeneous RNAs tend to be difficult to crystallize or to analyze by NMR. In addition, NMR and X-ray crystallography require very specialized equipment and large amounts of material to analyze. Thus, experimental techniques that can contribute native structural information about RNAs quickly and less laboriously are of great value to RNA structure determination.

From a computational perspective, *ab initio* prediction of RNA structure is an ideal case, and strides have been made towards understanding the energies and rules that dictate native structures. However, the vast conformational space that unfolded RNAs can explore, along with the presence of non-Watson-Crick base interactions, such as wobble pairs and Hoogsteen

interactions, and the adoption of computationally challenging secondary and tertiary structures, such as pseudoknots, pose significant complications for fully computation-based predictions.

Informing computational techniques for RNA structure modeling using experimental data has proved to be a very powerful approach to this difficult problem. Basic thermodynamic parameters for the energies associated with nearest-neighbor interactions and base mismatches in RNA have been determined by Douglas Turner and colleagues[1,2], which have been critical for RNA structure predictions using energy-based calculations. In addition, information on secondary structure, local dynamics, and tertiary contacts can be obtained from a variety of biochemical methods (Figure 1)[3], which can be incorporated into constraints or 'pseudo-energies' that can bias sampling of modeling algorithms to direct them toward favored native structures[4]. This review will discuss recently developed or refined methods which utilize limited structural information from biochemistry to inform computational approaches to RNA structure modeling. It will focus first on methods for secondary structure determination in RNA and computational tools for analysis of data generated by such techniques, and it will also discuss data-driven techniques for tertiary modeling of RNA structure.

Figure 1.
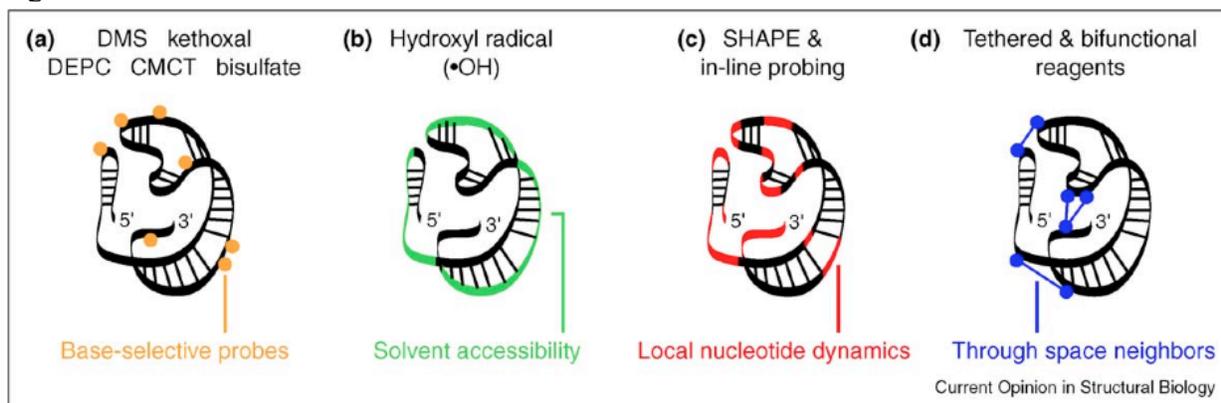
Classes of RNA structure information obtained by chemical probing include (A) base-selective data, (B) solvent-accessibility information, (C) measurements of nucleotide dynamics, and (D) constraints on long-range interactions. From Weeks (2010).

Secondary structure prediction

The principle of nearest-neighbor interactions, in which the behavior of individual nucleotides is dictated by the free energies for motif formation associated with its adjacent bases and base-pairs, forms the basis for many algorithms that are widely used for fast RNA structure prediction, such as mfold[5], Vienna RNA[6], and RNAstructure[7]. These algorithms themselves rely on the basic thermodynamic values for nearest-neighbor interactions which have been previously measured[4,8,9]. However, they cannot account for many types of both secondary and tertiary interactions, such as triple-helices and pseudoknots, because such structures defy basic conditions for the algorithms to operate correctly[10].

Secondary structural information can also be gleaned from chemical mapping techniques, which provide orthogonal structural information to thermodynamic algorithms that rely on nearest-neighbor free energy calculations. Chemical modification reagents can be used that react with the Watson-Crick faces of certain nucleotides, such as dimethyl sulfate (DMS), which reacts with unpaired adenine bases[11]. Alternative reagents react with the 2' hydroxyl of flexible regions in the RNA backbone[12]. Primer extension from one end of the modified RNA produces DNAs of various lengths, which can be separated by sequencing techniques based on size to provide information on regions of the RNA that can be modified by the reagents, and are therefore flexible or unprotected by base-pairing (Figure 2)[13]. Individual chemical mapping reagents can provide information about the specific form of flexibility or base-pairing present at its chemical targets within an RNA, but predictions based on these have been shown to allow for significant errors in secondary structure prediction[14]. Thus, application of a suite of chemicals to any single RNA of interest may be the most powerful method to determine its secondary structure.

Various computational tools have been developed to permit analysis of electrophoresis data for chemical mapping. In particular, these require pipelines of analysis steps that assign the RNA sequence to the sequence of bands, quantify the intensity of the signal at each nucleotide position, and normalize these signals to reaction probabilities, which can be mapped onto the sequence of the RNA[13,15] and used to predict a secondary structure by biasing the probabilities of helix formation toward unreactive regions of the RNA.

A standing problem that remains unaddressed by chemical mapping techniques is distinguishing between the ensembles of native structures that many RNAs adopt. Most chemical mapping information is 1-dimensional, meaning it provides reactivity information at each position in a single sequence of RNA. In this context, all ensemble RNA structures are convolved into a single set of signals. However, the recently-developed 'mutate-and-map' approach[16] has the potential to deconvolve these structural ensembles. In mutate-and-map, all nucleotide positions in an RNA are mutated to bases that would disrupt a native Watson-Crick interaction at that position. Then chemical mapping is performed on all of these RNAs in a multiplexed manner. When the reactivity at each nucleotide position is plotted in relation to the location of the mutated residue, long-range interactions can be observed as changes in reactivity at nucleotide positions distal

Figure 2.
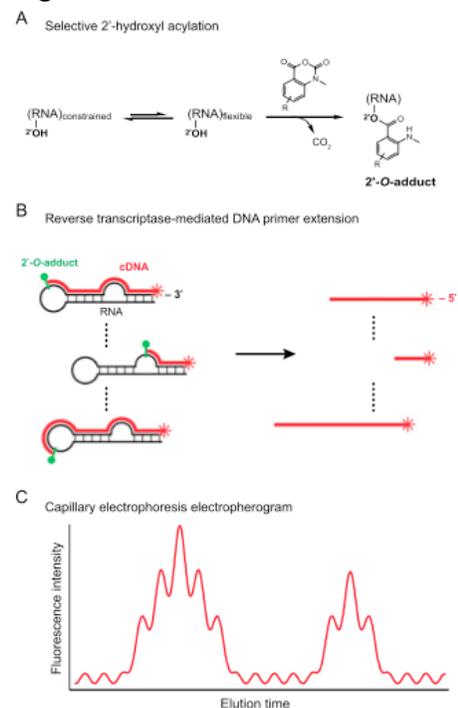
SHAPE chemical probing of RNA structure. (A) Mechanism of SHAPE chemistry. (B) Extension of fluorescently labeled primers by reverse transcriptase from the 39 end of an RNA to the site of the first adduct generates a population of fluorescently labeled cDNA molecules. (C) Capillary electrophoresis yields an electropherogram trace that quantitatively reflects cDNA molecules of various lengths, thus indicating the positions of flexible nucleotides in the RNA molecule. Figure from Karabiber *et al.* (2012).[1]

from the location of the mutation. In some cases, the reactivity profile of the RNA completely changes based on a single mutation, indicating a switch to a distinct conformation. Current work in the Das lab seeks to computationally calculate the thermodynamic favorability of members of an ensemble using factor analysis, which may be able to apply mutate-and-map data to determine the relative stabilities of RNAs with several native structures. In particular, this will be useful in analyzing the structures of RNAs that switch structures based on ligand binding or other interactions, which include regulatory riboswitches in cells.

Tertiary structure

Determining the tertiary structures of RNAs presents an additional level of complexity to the problem of secondary structure modeling, because it requires positioning of both primary and secondary structural elements with respect to each other in three dimensions. Techniques for tertiary structure mapping have taken advantage of solvent accessibility of small strand-scission catalyzing reagents, such as hydroxyl radicals, for surface probing by analysis of fragmented RNA[17]. In addition, distance constraints can be derived from tethered reagents, as is used in multiplexed hydroxyl radical cleavage analysis (MOHCA)[18]. MOHCA uses an iron atom tethered to the RNA backbone by a covalently linked chelating agent to generate hydroxyl radicals within a defined three-dimensional space, cause correlated strand scission events which can be used to map regions of the RNA that are within the radius of diffusion of the highly-reactive radical species. In the original application of MOHCA, helical-resolution information was obtainable by modeling distance constraints determined by two-dimensional polyacrylamide gel electrophoresis (PAGE) in the fragment assembly of RNA (FARNA) methodology, which relies on experimental information to sample backbone and side-chain conformations to build *de*

novo tertiary structures for RNA[19]. Future improvements on this technique aim to use deep-sequencing technology and further quantification of tethered radical source reactivities to obtain higher-resolution information from this experimental paradigm.

In addition to direct measurements of chemical strand scission of RNAs, recent work has leveraged less direct information for modeling RNA tertiary structures. In particular, secondary structure information, from thermodynamic or direct experimental measurements, as well as sequence covariation and crosslinking data, can be used for all-atom tertiary structure modeling[20]. NMR ^{1}H chemical shifts, which are not traditionally used for RNA tertiary structure determination, have also been used to model non-canonical RNA motifs.

Conclusion

In the last decade, a plethora of techniques have arisen to assess RNA secondary and tertiary structure by computation supplemented by experiments. Though thermodynamic techniques for nearest-neighbor modeling of RNA secondary structure opened the door to many computational approaches, chemical mapping methods, which provide direct readouts of native RNA structures and ensembles, as well as computational techniques that exploit commonly available but underutilized data are pioneering the next stages of RNA structure determination. Future developments are expected to expand on these approaches, in particular with respect to both expanding the catalog of experimental techniques that provide secondary and tertiary information-rich data and developing creative computational approaches to dissect these data.

References

1.   Freier, S. M. *et al.* Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 9373–9377 (1986).

2.   Wu, M., McDowell, J. A. & Turner, D. H. A Periodic Table of Tandem Mismatches in RNA - Biochemistry (ACS Publications). *Biochemistry* (1995).

3.   Weeks, K. M. Advances in RNA structure analysis by chemical probing. *Curr. Opin. Struct. Biol.* **20**, 295–304 (2010).

4.   Mathews, D. H. *et al.* Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7287–7292 (2004).

5.   Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).

6.   Hofacker, I. L. RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc Bioinformatics* **Chapter 12**, Unit12.2 (2009).

7.   Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).

8.   Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940 (1999).

9.   Xia, T. *et al.* Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719–14735 (1998).

10.   Aigner, K., Dreßen, F. & Steger, G.in *RNA 3D Structure Analysis and Prediction* **27**, 19–41 (Springer Berlin Heidelberg, 2012).

11.   Peattie, D. A. & Gilbert, W. Chemical probes for higher-order structure in RNA. *... of the National Academy of Sciences* (1980).

12.   Wilkinson, K. A., Merino, E. J. & Weeks, K. M. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat Protoc* **1**, 1610–1616 (2006).

13.   Karabiber, F., McGinnis, J. L., Favorov, O. V. & Weeks, K. M. QuShape: Rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA* (2012).doi:10.1261/rna.036327.112

14.   Kladwang, W., VanLang, C. C., Cordero, P. & Das, R. Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry* **50**, 8049–8056 (2011).

15.   Yoon, S. *et al.* HiTRACE: high-throughput robust analysis for capillary electrophoresis. *Bioinformatics* **27**, 1798–1805 (2011).

16.   Kladwang, W., Cordero, P. & Das, R. A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. *RNA* **17**, 522–534 (2011).

17.   Ding, F., Lavender, C. A., Weeks, K. M. & Dokholyan, N. V. Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat Methods* **9**, 603–608 (2012).

18.   Das, R. *et al.* Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proceedings of the National Academy of Sciences* **105**, 4144–4149 (2008).

19.   Das, R. & Baker, D. Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14664–14669 (2007).

20.   Seetin, M. G. & Mathews, D. H. Automated RNA tertiary structure prediction from secondary structure and low-resolution restraints. *J Comput Chem* (2011).doi:10.1002/jcc.21806