Methods 79–80 (2015) 60–66

Contents lists available at ScienceDirect

Methods

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

# Computational analysis of RNA structures with chemical probing data

Ping Ge, Shaojie Zhang *

*Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816-2362, USA*

ARTICLE INFO

*Article history:*
Received 1 November 2014
Received in revised form 16 January 2015
Accepted 9 February 2015
Available online 14 February 2015

*Keywords:*
RNA secondary structure
RNA tertiary structure
chemical probing
parallel sequencing

ABSTRACT

RNAs play various roles, not only as the genetic codes to synthesize proteins, but also as the direct participants of biological functions determined by their underlying high-order structures. Although many computational methods have been proposed for analyzing RNA structures, their accuracy and efficiency are limited, especially when applied to the large RNAs and the genome-wide data sets. Recently, advances in parallel sequencing and high-throughput chemical probing technologies have prompted the development of numerous new algorithms, which can incorporate the auxiliary structural information obtained from those experiments. Their potential has been revealed by the secondary structure prediction of ribosomal RNAs and the genome-wide ncRNA function annotation. In this review, the existing probing-directed computational methods for RNA secondary and tertiary structure analysis are discussed.



## 1. Background

RNA molecules, including both coding RNAs and non-coding RNAs (ncRNAs), play much more vital roles in the biological systems than what was suggested in the central dogma [1–3]. Their functions are not only encoded in the primary sequences [4], but also originate from the secondary and the tertiary structures [5–7]. Some well-known instances are the cloverleaf-like structure of tRNAs and the kink-turn structural motifs which server as important sites for protein recognition. Given the fact that most of transcripts (~90%) in typical eukaryotic genomes are ncRNAs, fully understanding RNAs and their functions is impossible without studying the high-order structures. However, the determination of RNA structures is not a trivial task. The traditional high-resolution techniques, such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, are very time consuming and hard to implement. On the other hand, the RNA structure folding algorithms [8–11] and the RNA functional annotation algorithms [12–14] are not accurate and efficient enough for the large RNAs and the genome-wide data sets.

The chemical probing technique, also named "structure probing" or "footprinting", provides a new way of studying RNA structures. RNAs of interest are treated with the chemical reagents which may modify the specific nucleotides with certain structural features. These modifications can act as stops for the primer extension, and their positions in the sequence can be detected by reverse transcription. Over the last 30 years chemical probing has been adopted for the study of RNA structures [15–17]. Recently more and more new protocols have been proposed to tackle the problems related to RNA structures. One of the most widely used probing experiments is to detect the paired and the unpaired bases. In these experiments, chemical reagents can form stable adducts with the flexible nucleotides in the loop regions, but not the protected bases in the stack regions. Some typical reagent choices are dimethyl sulfate (DMS) [18], kethoxal (KT) [19], diethyl pyrocarbonate (DEPC) [20], and CMCT [21]. None of them can react with all four RNA bases, e.g., DMS can only be applied to N1-adenine and N3-cytidine; KT can only be applied to N1 and N2 of guanine. A new protocol, selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) [22,23], can involve reactions with all bases. Moreover, SHAPE is insensible to the solvent accessibility and RNA size, which makes it an excellent choice for characterizing the structure features of large RNAs. RNase enzyme is another important type of reagent for probing RNA secondary structures. Instead of adducting to nucleotides, RNase catalyzes the degradation of the single- or double-stranded regions into smaller segments [24,25]. As a higher-order conformation which interlinks the packed secondary structure modules with through-space interactions, tertiary structure can also be analyzed with chemical probing experiments. For example, hydroxyl radicals generated by Fe(II)-EDTA catalyst can cleave the specific sites at RNA backbone proximal in space to the location of the bound Fe(II)-EDTA. Hence the long range interactions of the Fe(II) adducted nucleotides can be determined [26,27]. Cross-linking technique adopts a different strategy to detect juxtaposed nucleotides in three-dimensional space. It bridges the nearby nucleotides in an RNA by using bifunctional reagents [28] or UV-irritation [29]. The products of

* Corresponding author.
*E-mail address:* shzhang@eecs.ucf.edu (S. Zhang).

http://dx.doi.org/10.1016/j.ymeth.2015.02.003
1046-2023/

the reaction can be characterized by mass mapping or sequencing experiments.

The introduction of next generation sequencing (NGS) leads to the development of genome-wide RNA structure probing protocols. Many high-throughput protocols, such as SHAPE-seq [30,31], PARS [32–34], FragSeq [35], Map-seq [36], dsRNA-seq [37], CIRS-seq [38], and DMS-based high-throughput sequencing [39,40], have been applied to the transcriptomes of various species. These experiments provide comprehensive insights into the structural features of the coding regions. In addition, the genome-wide sequencing also reveals the structural characterization of substantial ncRNAs, especially the lncRNAs [34,39]. Recent studies show that the mutations and the dysregulations of lncRNAs are directly linked to many human diseases, ranging from neurodegeneration to cancer [41–45]. On the other hand, single-molecule probing has been combined with massive parallel sequencing to target the RNAs with complex structures. For RNA viruses, the functionally active structures are vital during their life cycle [46]. The global and local chemical probing of various viruses, such as the human immunodeficiency virus (HIV) [47], hepatitis C virus [48], influenza A virus [49] and the dengue virus [50], detected several potential regulatory motifs. Considering the limitations of traditional methods for RNA structure analysis, the rapid explosion of probing data coming from the high-throughput sequencing experiments will certainly enhance our understanding of human diseases.

The embedded structural information in the probing data can be quantified, and then incorporated into the computational method. The first breakthrough was in the field of RNA secondary structure folding. By integrating reactivities as extra pseudo energy terms into the nearest neighbor energy model, the secondary structure prediction accuracy of *Escherichia coli* 16S rRNA can be increased greatly [51]. This successful application suggests the great potential in using reactivities to assist the computational analysis of RNA structure. This review will introduce the existing chemical probing-directed computational methods and their applications (Fig. 1). In the discussion section we will also propose the possible directions of future research.

## 2. The computation of reactivities

In the chemical probing experiments, the modifications on the flexible nucleotides can be located by the 5′-end labeled primer extension. The lengths of the cDNA fragments imply the positions of the modified sites, and the number of the mapped fragments at each site indicates its reactive degree [47]. Traditionally, gel electrophoresis (GE) had been utilized to visualize the results of probing experiments. Analyzing the gel images is a tedious work, so computational methods are required to automate and accelerate the procedure. SAFA [52] (https://simtk.org/home/safa) is a semi-automated analyzing tool for gel quantification. The users only need to edit the intermediate results guided by a graphic user interface. In recent years, most of the single-molecule probing protocols began to make use of capillary electrophoresis (CE) for sequencing. However, the traditional algorithms designed for DNA CE sequencing may not be suitable for quantifying structural probing reactivities [53]. The major issues are signal decay correction, *x*-axis and *y*-axis scaling, signal alignment, sequence alignment and peak fitting. CAFA [54] (https://simtk.org/home/cafa) offers a CE analyzing method for the chemical probing experiment which focuses on peak detection and fitting. ShapeFinder [53] (http://giddingslab.org/software) adds a peak and sequence alignment step to refine the fitting. It still requires users to select parameters and adjust the alignment manually. FAST [55] (http://glennlab.stanford.edu/software.html) improves the efficiency of CE analysis by automating the *x*-axis and *y*-axis scaling. QuShape [56] (http://www.chem.unc.edu/rna/qushape) is presented as an updated version of the ShapeFinder by introducing new alignment and scaling algorithms. To align hundreds of capillaries together, two tools are provided by the Das lab: HiTRACE [57] (https://simtk.org/home/hitrace) and HiTRACE-Web [58] (http://hitrace.org). The output intensity data of HiTRACE can be further processed with a likelihood-based framework [59,60]. Recently, HiTRACE is also extended to allow CE processing standardization [61].

Compared to the reactivity computation of CE traces, the processing of reads generated by high-throughput sequencing-based probing protocols is more straightforward. First, the mapping of reads to the reference genome infers the sites of modification (normally 1nt upstream of the mapped reads). Second, the number of reads mapped to each site indicates its reactivity. Based on the two features, there are two groups of methods that quantify the read counts to reactivity values. The first group of methods normalizes the read counts directly. For examples, the raw read count of each site can be normalized by that of the most reactive base in a given window [39]; FragSeq computes pseudo counts based on the raw counts in a transcript, and then the pseudo counts are normalized such that they sum up to 1 [35]; PARS normalizes all read counts by sequencing depth, and then the log ratios of normalized counts for V1 RNase (cleaves the double-stranded RNA) and for S1 RNase (cleaves the single-stranded RNA) are computed. Notice that normalization is a general idea and can be applied to almost all the scenarios. On the other hand, the second type relies on the sophis-

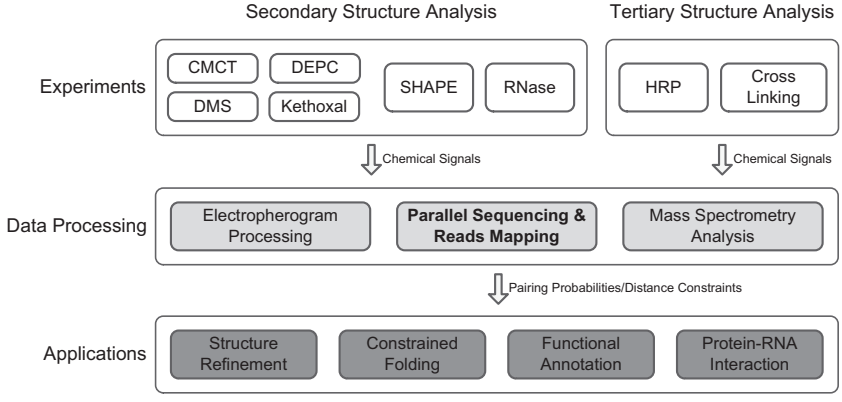


**Fig. 1.** The hierarchical overview of the RNA high-order structure analysis with probing-based computational methods. The white blocks represent chemical experiments and the shaded blocks represent the computational processing modules. Secondary structure and tertiary structure analysis adopt different protocols with different reagents. The output signals of the top-layer experiments are converted into reactivities, indicating pairing probabilities in secondary structure analysis or distance constraints in tertiary structure analysis, at the mid-layer. Finally the reactivities are incorporated into traditional algorithms at the bottom-layer.