Analysis of sequencing data for probing RNA secondary structures and protein-RNA binding in studying post-transcriptional regulations

Authors

*Xihao Hu**

Xihao Hu is a PhD student at the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong

*Yang Wu**

Yang Wu is a PhD student at School of Life Sciences, Tsinghua University, Beijing, China

Zhi John Lu

Zhi John Lu is an Assistant Professor at School of Life Sciences, Tsinghua University, Beijing, China

Kevin Y. Yip

Kevin Y. Yip is an Assistant Professor at the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong

Corresponding author. Kevin Yip, Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. E-mail: kevinyip@cse.cuhk.edu.hk

*These authors contributed equally to this paper

Abstract

High-throughput sequencing has been used to study post-transcriptional regulations, where the identification of protein-RNA binding is a major and fast-developing sub-area, which is in turn benefited by the sequencing methods for whole-transcriptome probing of RNA secondary structures. In the study of RNA secondary structures using high-throughput sequencing, bases are modified or cleaved according to their structural features, which alter the resulting composition of sequencing reads. In the study of protein-RNA binding, methods have been proposed to immuno-precipitate (IP) protein-bound RNA transcripts *in vitro* or *in vivo*. By sequencing these transcripts, the protein-RNA interactions and the binding locations can be identified. For both types of data, read counts are affected by a combination of confounding factors, including expression levels of transcripts, sequence biases, mapping errors, and the probing or IP efficiency of the experimental protocols. Careful processing of the sequencing data and proper extraction of important features are fundamentally important to a successful analysis. Here we review and compare different experimental methods for probing RNA secondary structures and binding sites of RNA binding proteins (RBPs), and the computational methods proposed for analyzing the corresponding sequencing data. We suggest how these two types of data should be integrated to study the structural properties of RBP binding sites as a systematic way to better understand post-transcriptional regulations.

Key Words

- RNA secondary structure
- Protein-RNA interactions
- RNA binding motifs
- High-throughput sequencing
- Data analysis

Introduction

The structures of RNAs are crucial to their functions. It is well established that the functions of some types of RNA, such as snRNAs, snoRNAs, rRNAs and tRNAs, depend highly on their structures. The CRISPR-Cas9 system for genome editing, currently receiving an explosion of extensive studies of its biology and applications, relies on the structures of the crRNA and tarcRNA to recognize the target sequence and interact with the Cas9 protein [1,2]. Even mRNAs, which were traditionally considered as mere messengers of the sequence information, have been suggested to possess structures that affect translational efficiency [3], transcript stability [4] and alternative splicing [5].

Adding to the versatility of RNAs is their ability to interact with various RNA-binding proteins (RBPs). There are more than 1,500 RBPs in human cells that interact with RNAs to perform complex functions [6]. These proteins play important roles in the regulation, localization and functioning of the RNAs [7]. The interactions between RBPs and RNA transcripts depend on both the sequence and structure of the transcripts, albeit with varying degrees of their relative importance in different interactions [8].

It has become possible to probe both RNA structures and RBP-RNA interactions in a high-throughput manner using cutting-edge sequencing methods [9,10]. These methods have enabled large-scale and systematic discoveries of novel structures and interactions. Being high-throughput methods, the data produced are subject to different types of noise and bias, which should be carefully handled during analysis.

In this review, we first briefly describe the current high-throughput sequencing methods for studying RNA structures and RNA-RBP interactions, followed by a discussion of the computational methods for processing and analyzing the resulting data (Figure 1). There are other reviews on high-throughput probing of RNA structures [11,12] and RNA-RBP interactions [8,13]. Here we provide an integrated review of both problems. For each of these two problems, we comprehensively discuss and compare the special characteristics of each type of experimental data due to experimental limitations, biases and noise. Such discussions lay out the background for our subsequent discussions on the computational methods for processing and utilizing these data, which are crucial for a successful analysis and is the focus of this review. In addition, a key message that we would like to convey is the close relationship between the identification of RNA structures and RBP-RNA interactions, that it would be advantageous to utilize both types of data to study RBP binding. We will review current approaches to integrating these two types of data, and point out other possible approaches that could potentially lead to new insights.

[Place of Figure 1]

High-throughput sequencing for probing RNA structures

RNA footprinting is a classical method for studying RNA structures [14]. In this method, chemical reagents or enzymes are used to modify or cleave bases with specific structural features (e.g., unpaired bases), the positions of which can then be determined by gel electrophoresis or capillary sequencing.

RNA footprinting has been combined with high-throughput sequencing for studying RNA structures at the whole-transcriptome scale [15]. We summarize and compare the recent high-throughput RNA structure probing methods in Table 1. These methods mainly differ from each other by the probe they use, which has implications in the bases that can be probed, data resolution, whether *in vivo* probing is possible, sequence and positional biases, and the efficiency and specificity of marking/cleaving bases according to their structural properties.

[Place of Table 1]

Methods based on enzymatic cleavage of bases with specific structural properties

The first method proposed for transcriptome-wide measurement of RNA secondary structures is Parallel Analysis of RNA Structure (PARS) [3]. It uses two kinds of enzyme, namely RNase V1 and nuclease S1, to preferentially cleave paired and unpaired bases in RNAs, respectively. RNA fragments cleaved by V1 and S1 enzymes are separately amplified and sequenced by RNA sequencing (RNA-seq) [16]. For each base, the ratio of corresponding reads in the V1 and S1 data sets indicates the chance that the base is paired in the RNA structure. The PARS experiment was later shown to be applicable for studying RNA folding energies [17] and variations of RNA structures such as riboSNitches [12,18].

Similar to PARS, another method called Fragmentation Sequencing (FragSeq) [19] was proposed to identify structures of non-coding RNAs. This method uses a single enzyme, nuclease P1, to preferentially cleave RNAs at unpaired bases. Without using a second enzyme for paired bases, FragSeq instead uses a control experiment without nuclease treatment to estimate and correct for background biases. Besides PARS and FragSeq, there are also other experimental methods proposed (such as dsRNA-seq) that use RNase ONE to probe unpaired bases [20,21]. One limitation of these enzyme-based methods is that they were developed for *in vitro* probing of RNA structures, which could be different from the *in vivo* structures.

Methods based on chemical modification of bases with specific structural properties

In contrast to treatments using structure-specific enzymes, another class of methods use chemical probes for detecting RNA secondary structures. Selective 2'-Hydroxyl Acylation Analyzed by Primer Extension (SHAPE) is one of these methods that has been integrated with high-throughput sequencing [22]. In this method, a small adduct is preferentially added to flexible unpaired bases in loops, which causes sequencing reads to end at those locations. Due to the use of a small chemical probe, SHAPE-seq has the advantage of inferring structures with a higher precision. Moreover, SHAPE-seq has been extended for *in vivo* usage, leading to a global *in vivo* structural picture of the mouse embryonic stem cell transcriptome [23]. By using individual barcodes to assign identity to RNAs, even highly similar RNAs can be distinguished using SHAPE-seq, but the use of barcodes also makes it difficult to simultaneously study a large number of transcripts genome-wide [12]. Benefiting from a newly designed SHAPE reagent, icSHAPE is currently one of the best methods of probing dynamic RNA secondary structures [23].

In another chemical-based method called DMS-seq, dimethyl sulfate (DMS), an even smaller chemical, was used for high-resolution transcriptome-wide probing of RNA structure *in vivo* [24–26]. A limitation of DMS-seq is that due to the chemistry, mainly only the structural information of adenines and cytosines can be probed. There were also studies that combine DMS and another chemical called CMCT with high-throughput sequencing to probe RNA structures [27].

Confounding factors in high-throughput RNA structure probing data

All these high-throughput structure-probing data contain confounding factors that need to be carefully corrected before downstream analyses. One of the factors unrelated to RNA structures but which affects the read counts most is the abundance levels of the transcripts. Highly abundant transcripts could have orders of magnitude more reads than transcripts with low abundance. These low-abundance transcripts are particularly affected by sampling effects during the sequencing process and errors introduced during sequencing and sequence read alignment. As a result, read counts of the nucleotides of low-abundance transcripts can be poorly correlated with the structural properties of the nucleotides. To deal with this problem, in addition to the obvious solution of increasing the sequencing depth, the DMS-seq and SHAPE-seq procedures can be modified to improve the detection of low-abundance transcripts by gene-specific cDNA amplification [28]. The ability to probe low-abundance transcripts is crucial in the study of RNA structures in general.

Read counts at individual bases are also subject to biases intrinsic to RNA-seq, such as GC content and other intricate sequence features [29]. There have been extensive studies that attempt to model these biases present in general RNA-seq data and correct for their effects [29–32]. It has been proposed that similar corrections should be applied to RNA structure-probing data [12].

There are also other issues specific to the experimental protocols that could lead to non-uniform probing signals not directly caused by RNA secondary structures. One issue is that in some experiments, different types of nucleotide are being modified with different efficiencies. For example, in DMS-seq, chemical probes modify atom N1 in adenines, atom N3 in cytosines, and sometimes atom N7 in guanines, with different efficiencies [15,25,26]. The resulting read count of a nucleotide thus depends partially on its type. If the modification efficiencies of different nucleotides of the same type are expected to be similar, probing signals for each type of nucleotides should be normalized separately.

A more problematic situation is that read counts can also be affected by properties more difficult to measure [33]. For example, when an enzyme is used to cleave the RNA, steric hindrance would bias the enzymes toward cutting at locations that are more accessible [15]. Correcting for such biases requires a good understanding of the accessibility of each base in the experimental condition, which is not always possible.

The two major stages that the confounding factors are handled

In general, both the general (transcript level and sequence bias) and protocol-specific (differential efficiency of base modifications/cleavage and probe size) confounding factors are tackled at two stages. First, at the data production stage, additional data are produced to provide information for correcting the biases. For example, as mentioned above, FragSeq usually involves a control experiment to capture the non-uniform background distribution of read counts [19], whereas in PARS two different enzymes are used to produce data that are assumed to be subject to the same biases, which can be canceled out by comparing read counts from the two resulting sets of data [3].

Second, at the data analysis stage, various computational methods have been proposed to correct for the biases. For example, the original analysis pipeline proposed for SHAPE-seq data involves a sophisticated statistical model that considers termination of reverse transcription due to both the chemical adducts and natural polymerase drop-off [34]. In DMS-seq, read counts were found to increase towards the 3’ end of a transcript [25]. This bias was corrected by normalizing each read count by the largest count within a local window. There are also methods that compute the enrichment of signals by comparing transcript-wise normalized read counts from DMS-seq with those from the control experiment [24].

High-throughput sequencing for identifying RNA-protein interactions

As RNA-binding proteins are known to affect post-transcriptional processes such as splicing and localization, knowing the binding targets of each RBP would help understand these important processes [35]. Besides traditional binding site detection techniques that are largely based on sequence motifs [8], high-throughput methods have also been developed to uncover the RNA functional elements that interact with RBPs at the transcriptome-wide level [36]. Most of these methods are based on immuno-precipitation of protein-bound RNAs, either with cross-linking (CLIP [37]) or without (RIP [38]). They have led to many interesting discoveries, such as classification of RNAs according to their interactions with RBPs [7] and splicing factors of pre-mRNAs [39]. Table 2 gives a list of recently developed high-throughput sequencing techniques for studying RNA-protein interactions on a transcriptomic scale. In addition to including cross-linking as a step of the experimental protocol or not, these techniques also have different cross-linking efficiency, data resolution, specificity, types of RNA that can be studied, and target RBPs.

[Place of Table 2]

The main classes of high-throughput experimental methods

HITS-CLIP [40], which involves CLIP followed by RNA sequencing, was first proposed to determine the RNA binding sites of a splicing factor. It has been used in a wide range of applications. More CLIP-based methods such as iCLIP [41] and PAR-CLIP [42] were later proposed with modified experimental details and additional steps that lead to improved cross-linking efficacy and binding site resolution [36]. Similarly, the RIP procedure that does not involve cross-linking was also combined with high-throughput sequencing to become a method called RIP-seq [43]. RIP-seq has good detection sensitivity, but lacking the cross-linking step makes RIP-seq data more susceptible to false positives due to noise or indirect interactions when compared with CLIP-based alternatives [44].

As immuno-precipitation is protein-specific based on the antibody used, in order to get a global list of binding sites of RNA binding proteins in general, three new protocols select RNA-RBP complexes by targeting the polyadenylic acid (poly-A) tails in RNAs using oligo(dT) [45–47]. Although these methods can produce a global view of RNA binding sites, the scope is restricted to RNAs with poly-A tails, which may miss many functional non-coding RNAs. These methods also do not provide information about the RBPs that bind each identified binding site, and therefore cannot be used for comparing the binding sites of different RBPs directly.

Biases in the high-throughput experimental data

Like other techniques, data from RIP- and CLIP-based experiments contain biases that are usually corrected in multiple steps [44,48,49]. Beginning at the raw sequencing reads, mapping software should be set to allow inexact matches at binding sites if an experimental protocol that introduces insertions, deletions or substitutions is used [50]. Polymerase chain reaction (PCR) duplicates need to be removed, so as to avoid biases in the downstream calculations [51]. Binding site clusters are then identified by peak-calling methods followed by motif discovery algorithms as summarized in two recent reviews [44,49]. During the peak-calling process, it would be useful to normalize the read counts at each peak by the corresponding transcript levels from an RNA-seq data if the read count of each binding site is to be further used for approximating the binding affinity, although it is not a standard step in existing peak calling methods for identifying RBP binding sites. Alternatively, RNA fragments that are likely to contain binding sites can be identified for studying sequence and structural motifs directly, without quantifying binding affinity at signal peaks.

While sequence motifs are commonly identified at RBP binding sites on the RNAs, they are insufficient for determining the binding sites completely. Local RNA structures play a key role in RNA-RBP binding in some cases, which highlight the importance of the ability to systematically probe RNA structures on a large scale. It has also been found recently that N6-methyladenosine could alter local base pairing, resulting in the exposure of buried RNA binding motifs and increased binding of a nuclear RNA-binding protein [52,53]. In general, methylation of RNA bases could introduce structural changes and affect the chance for a binding motif to be an actual binding site. These structural changes can be probed by new experimental methods such as icSHAPE.

Another noticeable example of RNA secondary structure change is caused by single nucleotide variants called riboSNitches that were found to be very common in the human transcriptome [54]. RiboSNitches are known to cause human diseases by having the protein-RNA interactions affected by the structural changes [55], making the identification of these RNA structure changes and their effects to post-transcriptional processes important [56]. As mentioned above, structural changes caused by riboSNitches can be detected by high-throughput structure-probing data. For example, PARS has been used to study riboSNitches by probing the RNA structures of a family trio and calculating changes of the RNA structures among each pair of individuals [12,18]. Based on the PARS data, the performance of 11 RNA folding algorithms in predicting riboSNitches was evaluated [57]. These identified structure changes can be used to help explain differential binding affinity of RBPs at locations with riboSNitches.

Because of these complexities, sophisticated computational methods have been used to model the binding pattern of RBPs by considering both sequence- and structure-related features [8].

Having introduced the experimental methods for probing RNA structures and RNA-protein interactions and the different types of confounding factors in the corresponding data, we now discuss the computational methods for analyzing these two types of data.

RNA secondary structure prediction guided by high-throughput structure-probing data

High-throughput structure-probing data provide structural information about individual bases, but they are insufficient for determining RNA secondary structures, because they only indicate whether a base is paired, but not the other base that it pairs with. In order to determine the structures, these data have been incorporated into computational RNA secondary structure prediction methods in various ways. We first briefly review two main classes of traditional RNA secondary structure prediction approaches, namely minimum free energy (MFE) and partition-function-based methods. Most partition-function based methods adopt the idea of maximum expected accuracy (MEA). We then discuss how structure-probing data have been incorporated into these algorithms.

Two main classes of RNA secondary structure prediction method

Existing RNA folding algorithms are effective in predicting secondary structures of short RNAs based on sequence information alone [58]. Since RNA secondary structures involve mainly canonical A-U, G-C and wobble G-U pairs, the optimal structure defined as the one with lowest total free energy based on a given energy model can be efficiently predicted using dynamic programming algorithms, when pseudoknots are not considered, by MFE (minimum free energy) algorithms [59]. Improved accuracy has been achieved by using a set of sequence-dependent free energy parameters for each type of sub-structure [60]. More refined models further include chemical modification constraints [33] and enthalpy change [61].

In another type of prediction methods, the algorithms are based on the partition function, which considers the whole ensemble of structures that can be formed from the same RNA sequence. Based on the partition function, an optimal structure can be defined as the one with the maximum expected accuracy (MEA) [62–64]. Alternatively, representative structures can be sampled and clustered from the whole ensemble [65].

These algorithms have been shown highly reliable for predicting the secondary structures of short RNAs. The corresponding software packages also provide a lot of useful functionality. For example, the latest RNAstructure package not only can predict a common structure for multiple RNA sequences but also provides an open programming interface for other researchers to integrate specific algorithms into the prediction process [66]. ViennaRNA is another popular package that has good performance and a user-friendly graphical interface [67]. Some other methods, such as mfold/UNAfold [68,69], are also widely used.

Incorporating high-throughput structure-probing data into RNA secondary structure prediction methods

Given the effectiveness of these algorithms with short RNAs and their well-tested software packages, it has become attractive to incorporate high-throughput structure-probing data to help with the predictions of difficult RNA structures, while reusing the developed algorithms and software modules as much as possible [11,70,71] (Table 3). The incorporation of high-throughput structure-probing data helps shrink the space of possible structures, which could benefit both the accuracy and running time of the prediction algorithms. Different types of high-throughput data have been incorporated into different RNA secondary structure prediction methods, and they have been used in different ways, including the definition of pseudo energy terms, introduction of constraints in the search for RNA structures, or selection of structures from a set of candidates.

[Place of Table 3]

Most of these methods incorporate structure-probing data by modifying the energy model of a traditional MFE algorithm. One of the early approaches is to add pseudo free energy terms that penalize violations of the constraints defined by structure-probing data [72]. Various designs of the pseudo energy term were proposed. The first approach was to add a fixed penalty to the total free energy [33,60,67]. For each base forced to be unpaired according to structure-probing data, a large energy penalty is given if it is paired with another base in a predicted structure. A linear pseudo energy term was later proposed to transform SHAPE constraints into an energy function [73]. The two free parameters in the energy term were optimized using known structures. ShapeKnots [74] extended the energy term for incorporating SHAPE data by adding two more parameters that allow for the modeling of pseudoknots in base-pair stacking regions. Other forms of pseudo-energy terms derived from log-likelihood ratios were also incorporated based on SHAPE [75] and DMS data [76]. This type of pseudo-energy terms allows the quantitative signals from SHAPE [73] or DMS data [76] to be used in pace of the qualitative base-pair constraints.

Pseudo-energy terms can be accommodated with MEA-based algorithms as well. Our RME method [77] adds pseudo-energy terms to base-pairing stacks based on the posterior odds to the partition function, and predicts optimal structures according to the resulting restrained partition function. As the posterior odds can be computed efficiently, the pseudo-energy terms introduced to RME is applicable to diverse types of probing data, including SHAPE-seq, PARS and DMS-seq. To make the probing data applicable to all base types, RNAsc [78] redefines the pseudo energy terms for every structural component with a guarantee that perfect constraints would lead to a correct predicted structure, although the requirement is not always true in SHAPE data [71].

Although the addition of pseudo energy terms is shown to be effective in practice, there is a concern that these added terms do not have well-justified physical-chemical meanings [71]. To overcome this issue, some other approaches incorporate structure-probing data into the prediction process without modifying the energy function. For example, mfold has a long history of allowing flexible folding constraints added directly to the structures [79], which provides a natural way to utilize the structure-probing data. Instead of defining a particular energy term, RME [77] and RNApbfold [80] attempt to minimize the difference between the constraints defined according to structure-probing data and the base-pairing probabilities from the partition function [81].

Instead of modifying the structure prediction algorithms, structure-probing data can also help select suitable predicted structures using a sampling strategy. After candidate structures have been sampled from the structure ensemble and classified into clusters, the ‘sample-and-select’ method finds the structure with the minimum distance to the structure-probing data [71]. SeqFold chooses the structure centroid proximal to the constraints derived from structure-probing data [82]. Different types of structure-probing data are filtered based on their p-values according to a hypergeometric test or some chosen cutoffs, to reduce the negative effects of potentially wrong constraints. Although SeqFold is able to incorporate multiple types of RNA structure-probing data, it has also been suggested that the structural information contained in the data has not been fully utilized [71].

Most existing RNA secondary structure prediction algorithms that incorporate structural probing data assume the data have already been properly cleaned and normalized during the preprocessing steps. Missing or low-quality structural constraints are ignored by setting the corresponding terms in the pseudo free energy [33,60,67,73,77,83,84] or other objective functions [74,78,80,85,68] to zero. Non-uniform signal distribution is not considered in many early tools as the problem was not serious when the input data were from carefully calibrated RNA footprinting at that time. Since high-throughput sequencing protocols introduce more biases, methods that incorporate such data pre-process the raw data by either simple normalizations [3,19,25] or sophisticated statistical models [74,77,86]. If the quality of the raw data is too low, these methods may fail to correct the bias properly. Sampling a set of confident structures rather than getting a single best structure would likely give more robust results [82]. To make it easy to incorporate high-throughput sequencing data, StructureFold provides a user-friendly platform for pre-processing raw RNA structure-probing data and calling external tools to predict RNA structures with the incorporation of the processed data [84].

Limitations of existing structure-probing data-guided RNA secondary structure prediction

High-throughput structure-probing data have offered a great opportunity for RNA secondary structure prediction algorithms to explore previously challenging territories such as the structures of long mRNAs and non-coding RNAs. On the other hand, there is still ample room for better use of such data. Here we propose several aspects of existing structure prediction algorithms that can be further improved.

Some algorithms that incorporate high-throughput structure-probing data rely on the data pre-processing procedure to handle the effects caused by confounding factors in the data. However, some effects cannot be completely eliminated by data pre-processing alone. For example, data obtained from experiments involving enzymatic cleavage suffer from low resolution due to large probe size [15,87], while chemical probing approaches such as DMS can only probe structures of particular types of nucleotide [14]. Therefore, in both cases, even after pre-processing, structure-probing data would still contain uncertainty and missing values. Secondary structure prediction algorithms should explicitly handle these issues by proper modeling the background distribution of read counts at different nucleotides.

Biases in standard RNA-seq data that are used for measuring expression levels of genes or transcript isoforms have been extensively studied in the literature [29–32], but how these biases affect high-throughput RNA structure-probing data is still incompletely known. It would be desirable for RNA secondary structure prediction algorithms to consider confidence values of structure-probing data for different bases/transcripts instead of assuming all values equally reliable.

Finally, some RNA secondary structure prediction methods need to be tuned before they can incorporate structure-probing data. For example, prediction algorithms based on Turner's free energy parameters [88], including all algorithms listed in Table 3 that involve an energy model, may need to be adjusted based on the experimental condition [61]. In general, all RNA structure prediction methods that assume an energy model should adjust its parameters based on the conditions of the structure-probing experiments. Moreover, while some existing structure prediction algorithms, such as CONTRAfold [63] and CentroidFold [89], can predict structures represented by a set of sub-optimal structures, whether incorporating structure-probing data would improve their performance is still uncertain. One consideration is that an algorithm should utilize the structure-probing data according to its assumption of whether the probed structure comes from a single "ensemble" structure of multiple sub-optimal structures with certain probabilities. MutualFold [85] is a method that can give a solution for two alternative structures by redefining the optimization problem with some added pseudo energy terms. Results on simulated data showed that the algorithm recovered a structural change often ignored by other methods [85]. Methods that can model even more complex cases are to be seen.

Modeling RBP binding sites using high-throughput sequencing data for probing RNA-RBP interactions and RNA secondary structures

Similar to the high-throughput sequencing data for probing RNA structures, RBP binding sites identified directly from raw high-throughput experiments usually contain false positives and false negatives. False positives, i.e., incorrect binding sites, could come from sequencing errors that mimic nucleotide substitutions caused by the experimental procedure, or background binding due to other RBPs. While sequencing errors could be partially corrected computationally by carefully checking the quality of mapped reads [51,90], correction of background binding requires additional data from control experiments [91], which were unfortunately not performed in some studies. False negatives, on the other hand, are much harder to correct. Binding sites on low-abundance transcripts are likely to be hidden by background noise in experiments with a standard library size. Single nucleotide polymorphisms (SNPs) and incomplete annotation of the transcriptome could also reduce the amount of correctly aligned sequencing reads and lead to false negatives. It has been suggested that CLIP-seq experiments should be paired with a matched RNA-seq experiment [48], which, though not able to recover the binding sites on low-abundance transcripts, could provide useful information for data normalization.

A number of computational methods have been proposed for identifying true RBP binding sites from potentially noisy RIP- and CLIP-based data and analyzing the sequence motifs at these putative binding sites. Table 4 shows a summary of some of the recently proposed methods. These methods differ by the types of high-throughput protein-RNA interaction data incorporated, computational models for the RBP binding sites or the methods for finding them, and the features used for modeling or searching. Many of these methods also consider RNA secondary structure information during the analysis. Such information could conceptually come from either a pure sequence-based secondary structure prediction algorithm, a structure-probing data-assisted prediction algorithm, or structure-probing data directly.

[Place of Table 4]

Methods mainly based on sequence motifs

RNAcontext [92] was originally used to refine RNA binding sites identified from a type of microarray-based experiments called RNAcomplete [93] using a modified motif discovery algorithm. The method is also applicable to data from RIP- or CLIP-based high-throughput experiments. For each RNA transcript in the dataset, RNAcontext obtains an RNA secondary structure profile predicted by Sfold [65] . The predicted structure of each base is represented by a probability that indicates whether the base is predicted to be paired (P) or unpaired and within a certain structural context, namely hairpin loops (L), unstructured regions (U), or miscellaneous regions (M). By encoding both sequence and structure information into the alphabet of pseudo bases, RNAcontext achieved better performance than previous methods that pay more attention to sequence patterns. Since the time cost of Sfold was high for long sequences, the performance of RNAcontext could possibly be improved by a faster structure profile prediction tool such as CapR [94]. Although RNAcontext still adopts the traditional way of motif discovery based on position-specific profiles (of both sequence and structure), it searches for motifs from a pool of training sequences, and further gives a model to estimate RBP binding affinity. Due to the assumption of position-specific profiles, dependencies among different positions are ignored in the model, which limits the power of RNAcontext in capturing base interactions in RNA structures [95].

TEISER [96] adopts a new way of encoding sequence-structure patterns by using a flexible motif model called context-free grammars. Unlike RNAcontext that treats sequence and structure as separate features, special characters are invented to represent the RNA sequence and pairing of nucleotides in the same grammar. The optional grammar for binding motifs is found by an exhaustive search. To reduce the high computational cost due to the great flexibility of the grammars, stringent size constraints are added to seed structural components, such as 4-7 bases for stem regions and 4-9 bases for loop regions. Potential binding motifs are then ranked by the mutual information (MI) between the motifs and candidate binding fragments from the sequencing data. Lastly, motifs with high MI values are refined by proper elongation rules to compensate for the previously added constraints.

The above two methods mainly investigate sequence motifs at the potential binding sites directly detected in the high-throughput data. In contrast, MCarts [97] uses a hidden Markov model (HMM) with a rich feature set to model binding fragments with variable clustered binding motifs obtained from CLIP experiments. HMMs are statistical models that define the probabilities of entering a state initially, transiting from a state to another, and emitting an observation from a state. The states cannot be directly observed (i.e., "hidden"), and the probability for a given observation to come from each state is to be inferred from the observed data. MCarts defines binding regions and several other types of region into six states. The goal is to infer the probability that a genomic region is bound by an RBP based on its features. The feature set contains: a) the distance to the nearest binding site on the same RNA, b) the accessibility of the region in terms of single strandedness, and c) the conservation of the motif as measured by branch length scores (BLS). Applying the trained HMM to the whole genome, half of the genomic regions were defined as binding sites. The use of HMMs allows MCarts to capture complex binding patterns such as clustered RBP-binding motifs, which cannot be modeled by a position weight matrix (PWM) as used by RNAcontext. On the other hand, MCarts can only be applied to study RBPs that recognize binding sites with a number of clustered motifs. MCarts was used to check the prediction performance of different combinations of features. It was found that the full set of all features was the most powerful, indicating that incorporating more types of feature helped better model experimentally derived RBP binding sites.

Methods based on supervised machine learning methods

In early 2014, three new methods were proposed using different learning models and feature sets. Compared with methods for constructing sequence motifs, machine learning methods have greater power to model complex binding motifs such as regions with structural changes. On the other hand, machine learning methods require a sufficiently large set of training data for learning the models, and the learned models can be hard to interpret by having the important sequence and structural features of the binding sites abstracted in the mathematical formulas of the models.

The first method is OliMoSS [98], which uses a support vector machine (SVM) [99] to model the relationships between RBP binding sites and both sequence features and features derived from predicted secondary structures based on PAR-CLIP data. The SVM classifier learns a hyperplane that separates binding sites from other regions in a feature space that can be either the input feature space or a high dimensional space defined by a kernel function. The input feature set for each region includes: a) motif scores obtained from a traditional PWM-based method, b) frequencies of all possible length-4 nucleotide sequences (i.e. tetranucleotides), and c) predicted structural profile such as folding energy, stem density and accessibility. Various combinations of the features were tested, and tetranucleotides alone could give the best performance for many tested RBPs. It should be noted that the sequence motifs identified by this method were rich in adenines and uracils, and it is still unknown to what extent these results were affected by the experimental bias in the binding site sequencing data.

GraphProt [95] also uses SVM but has a more sophisticated feature encoding scheme with a graph kernel. For an RNA fragment, each region identified by CLIP-seq is extended to 150 nucleotides, and the probable secondary structures of the extended region is sampled and represented by a high-level shape abstraction using RNAshapes [100] The RNA sequence and its predicted structure are then encoded by a graph based on GraphClust [101]. The feature set, represented in graphs, is further mapped onto a kernel space using a neighborhood subgraph pairwise distance kernel function [102]. Models learnt from many CLIP-based data sets were successfully confirmed with similar motifs from the literature, and the models were further used to predict binding affinities with a high accuracy, even affinity data were not used in training.

Different from the aforementioned methods that use predicted structures to model protein-RNA interactions, ProbRNA [86] extract useful structural features from PARS data. Specifically, ProbRNA uses a mixture of Poisson linear model that fits raw V1 and S1 read counts from PARS experiments. For each nucleotide, its hidden state in the mixture indicates its local structural preference, which was shown to be more useful than some other computationally predicted structural features in predicting RBP binding sites. ProbRNA then constructs a Random Forest model [103] that uses the extracted structural feature together with sequence information in each local window to predict whether the window overlaps an RBP binding site. A Random Forest model involves a set of decision trees constructed from different samples of the training data. The final prediction is made by combining the predictions of the individual decision trees. The use of ensemble of trees helps avoid over-fitting, a situation in which the constructed model only fits the training data but predicts poorly on unseen data. It was shown that ProbRNA could effectively distinguish RBP binding regions from shuffled regions that maintain the same nucleotide composition, suggesting that the model captured genuine properties of RBP binding sites rather than biases caused by the confounding factors. The results highlight the contribution of RNA structural information extracted from high-throughput probing data in identifying protein binding sites.

Considerations when incorporating structural information into the identification of RBP binding sites

As far as we know, ProbRNA is currently the only method that uses RNA structure-probing data to model binding sites of RBP directly without explicitly predicting a complete RNA secondary structure. Due to the confounding factors in existing high-throughput structure-probing data as discussed above, it has been shown in one of the former studies that predicted secondary structures of a pure sequence-based RNA structure prediction algorithm could be more useful than *in vitro* RNA structure-probing data when identifying *in vivo* RBP binding sites [8]. The results of ProbRNA, on the other hand, suggest that the *in vitro* structure-probing data could also be useful in identifying *in vivo* RBP binding sites, although the useful structural signals need to be extracted using specifically-designed methods.

Since structural changes can affect binding [53], structure-probing data obtained from the same experimental condition as the RBP binding data would be most ideal for studying the RNA structural features of RBP binding sites. *In silico* structures predicted based on sequence information alone and *in vivo* structures in the native cell environment represent two extremes on the horizon of RNA structures, while *in vitro* and *in vivo* structure-probing data could hopefully provide information for determining RNA structures closer to the *in vivo* ones that are recognized by RNA binding proteins.

Some recent studies have looked for structural motifs at RBP binding sites using both experimental and computational methods. The hiCLIP method identifies RNA secondary structures interacting with RBPs transcriptome-wide [104]. Secondary structure contexts significantly enriched in RNAcomplete experiments were also cataloged [105].

Conclusion and future directions

High-throughput sequencing has greatly extended the power of traditional RNA footprinting methods. Experimental methods for probing RNA structures and RBP binding sites are now available for large-scale studies at the whole-transcriptome level.

Based on the discussions above, here we give some concrete recommendations to three types of people who deal with these two types of high-throughput sequencing data.

For experimentalists, it would be good to understand the limitations of the corresponding computational methods, and produce the additional experimental data necessary for effective data processing, such as those from control experiments for subtracting background signals and from RNA-seq experiments for normalizing read counts according to transcript levels.

For data analysts, it would be good to understand the properties and confounding factors of the experimental data, such as the nucleotide modification preference of DMS and the effect of RNA abundance on probing efficiency. Based on these properties, an analyst should choose appropriate analysis tools that can properly handle them. When tools specifically designed for a particular type of data are available, they are usually preferred over standard tools for general sequencing experiments.

For developers of new analysis methods, it would be good to have a general framework that facilitates the analysis of different types of structure-probing (such as DMS, PARS and SHAPE) or RBP-binding (such as HITS-CLIP, iCLIP and PAR-CLIP) data, but at the same time allows algorithms specific to certain data types to be plugged in as modules to particular steps of the pipeline. It would also be good to have different modes of running the software depending on the availability of additional data, and provide clear explanations to the user regarding the significance of these additional data and the limitations of the analysis methods when they are unavailable.

Regarding future directions, one important new direction is the combined study of both RNA structures and RBP binding. In recent protein interaction profile sequencing (PIP-seq) experiments [9,10], RNA secondary structures and RNA binding sites are probed in the same pipeline by applying different treatments on the same RNA sample. Applying footprinting on both bound and unbound RNAs offers a broader view of the function of RNA secondary structures in post-transcriptional regulations.

More generally, it would be desirable to integrate multiple sources of RNA structure-probing data and RBP binding data, existing RNA structure prediction algorithms and motif discovery algorithms all together to maximize the accuracy of computational models. For example, when modeling protein-RNA binding sites, structural features can be derived from the RNA structure-probing data and RNA structure prediction algorithms, while sequence features can be derived from motif discovery algorithms. The noisy RBP binding sites obtained from high-throughput RBP binding data can serve as an initial training set for learning a model of these binding sites using the derived features. From the high-confidence predictions of the model, structural and sequence features commonly shared by these predicted regions could in turn be discovered, which could iteratively refine both the model and the useful features for characterizing RBP binding sites.

New data processing and modeling tools are required to make better use of the large amount of data that could suffer from technical biases and experimental noise. Careful correction of these confounding factors will be the key to successful data processing methods. A lot of improvements have been seen in the latest work for integrating RNA structure-probing data into classical RNA secondary structure prediction algorithms. The same trend of incorporating high-throughput experimental data into existing computational algorithms is expected to become more and more popular for modeling RBP binding sites.

Key Points

- High-throughput sequencing techniques extend traditional RNA foot-printing methods to provide data for whole-transcriptome study of RNA secondary structures.
- These high-throughput RNA structure-probing data can be used alone or integrated into RNA secondary structure prediction algorithms to infer RNA secondary structures.
- Likewise, immuno-precipitation and crosslinking coupled with high-throughput sequencing can help identify protein-RNA interactions and the exact binding sites.
- Both types of sequencing data contain various types of biases, and therefore require computational methods for normalizing the data and extracting useful information for determining RNA structures and RNA-protein interactions.
- Since RNA-protein binding could depend on the structures of the RNAs and in turn structure-probing efficiency/accessibility could also be affected or even biased by protein binding *in vivo*, integrating the two types of sequencing data could lead to better modeling of RNA-protein interactions *in vivo*, providing a broader view of the landscape of post-transcriptional regulation.

Funding

This work was supported by CUHK Direct Grant [2050479], the Hong Kong Research Grant Council Theme-based Research Scheme [T12-401/13-R], National Natural Science Foundation of China [31271402, 31522030], Tsinghua University Initiative Scientific Research Program [2014z21045] and Computing Platform of National Protein Facilities (Tsinghua University).

References

1. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell 2014; 157:1262–1278

2. Doudna JA, Charpentier E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science 2014; 346:1258096

3. Kertesz M, Wan Y, Mazor E, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature 2010; 467:103–107

4. Goodarzi H, Zhang S, Buss CG, et al. Metastasis-suppressor transcript destabilization through TARBP2 binding of mRNA hairpins. Nature 2014; 513:256–260

5. Barash Y, Calarco JA, Gao W, et al. Deciphering the splicing code. Nature 2010; 465:53–59

6. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat. Rev. Genet. 2014; 15:829–845

7. Tuck AC, Tollervey D. A transcriptome-wide atlas of RNP composition reveals diverse classes of mRNAs and lncRNAs. Cell 2013; 154:996–1009

8. Li X, Kazan H, Lipshitz HD, et al. Finding the target sites of RNA-binding proteins. Wiley Interdiscip. Rev. RNA 2014; 5:111–130

9. Silverman IM, Li F, Alexander A, et al. RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. Genome Biol. 2014; 15:R3

10. Gosai SJ, Foley SW, Wang D, et al. Global analysis of the RNA-protein interaction and RNA secondary structure landscapes of the Arabidopsis nucleus. Mol. Cell 2015; 57:376–88

11. Ge P, Zhang S. Computational analysis of RNA structures with chemical probing data. Methods 2015;

12. Wan Y, Kertesz M, Spitale RC, et al. Understanding the transcriptome through RNA structure. Nat. Rev. Genet. 2011; 12:641–655

13. McHugh CA, Russell P, Guttman M. Methods for comprehensive experimental identification of RNA-protein interactions. Genome Biol. 2014; 15:203

14. Weeks KM. Advances in RNA structure analysis by chemical probing. Curr. Opin. Struct. Biol. 2010; 20:295–304

15. Fourmy D, Yoshizawa S. Protein-RNA footprinting: An evolving tool. Wiley Interdiscip. Rev. RNA 2012; 3:557–566

16. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 2009; 10:57–63

17. Wan Y, Qu K, Ouyang Z, et al. Genome-wide Measurement of RNA Folding Energies. Mol. Cell 2012; 48:169–181

18. Wan Y, Qu K, Zhang QC, et al. Landscape and variation of RNA secondary structure across the human transcriptome. Nature 2014; 505:706–9

19. Underwood JG, Uzilov A V, Katzman S, et al. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat. Methods 2010; 7:995–1001

20. Zheng Q, Ryvkin P, Li F, et al. Genome-Wide Double-Stranded RNA Sequencing Reveals the Functional Significance of Base-Paired RNAs in Arabidopsis. PLoS Genet. 2010; 6:e1001141

21. Li F, Zheng Q, Ryvkin P, et al. Global Analysis of RNA Secondary Structure in Two Metazoans. Cell Rep. 2012; 1:69–82

22. Lucks JB, Mortimer SA, Trapnell C, et al. Multiplexed RNA structure characterization with selective 2’-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc. Natl. Acad. Sci. U. S. A. 2011; 108:11063–8

23. Spitale RC, Flynn RA, Zhang QC, et al. Structural imprints in vivo decode {RNA} regulatory mechanisms. Nature 2015; 519:486–490

24. Ding Y, Tang Y, Kwok CK, et al. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 2014; 505:696–700

25. Rouskin S, Zubradt M, Washietl S, et al. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature 2014; 505:701–5

26. Talkish J, May G, Lin Y, et al. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. RNA 2014; 20:713–720

27. Incarnato D, Neri F, Anselmi F, et al. Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. Genome Biol. 2014; 15:491

28. Kwok CK, Ding Y, Tang Y, et al. Determination of in vivo RNA structure in low-abundance transcripts. Nat. Commun. 2013; 4:2971

29. Li J, Jiang H, Wong WH. Modeling non-uniformity in short-read rates in RNA-Seq data. Genome Biol. 2010; 11:R50

30. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010; 38:e131

31. Roberts A, Trapnell C, Donaghey J, et al. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011; 12:R22

32. Wu Z, Wang X, Zhang X. Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. Bioinformatics 2011; 27:502–508

33. Mathews DH, Disney MD, Childs JL, et al. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci. U. S. A. 2004; 101:7287–7292

34. Aviran S, Trapnell C, Lucks JB, et al. Modeling and automation of sequencing-based characterization of RNA structure. Proc. Natl. Acad. Sci. U. S. A. 2011; 108:11069–11074

35. Chaturvedi P, Neelamraju Y, Arif W, et al. Uncovering RNA binding proteins associated with age and gender during liver maturation. Sci. Rep. 2015; 5:9512

36. Milek M, Wyler E, Landthaler M. Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing. Semin. Cell Dev. Biol. 2012; 23:206–212

37. Ule J, Jensen K, Mele A, et al. CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods 2005; 37:376–86

38. Zhao J, Sun BK, Erwin JA, et al. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 2008; 322:750–6

39. Baejen C, Torkler P, Gressel S, et al. Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition. Mol Cell 2014; 55:745–757

40. Licatalosi DD, Mele A, Fak JJ, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 2008; 456:464–9

41. Huppertz I, Attig J, D'Ambrogio A, et al. iCLIP: Protein-RNA interactions at nucleotide resolution. Methods 2014; 65:274–287

42. Hafner M, Landthaler M, Burger L, et al. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. Cell 2010; 141:129–141

43. Zhao J, Ohsumi TK, Kung JT, et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell 2010; 40:939–53

44. Reyes-Herrera PH, Ficarra E. Computational Methods for CLIP-seq Data Processing. Bioinform. Biol. Insights 2014; 8:199–207

45. Baltz AG, Munschauer M, Schwanhäusser B, et al. The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. Mol. Cell 2012; 46:674–690

46. Castello A, Fischer B, Eichelbaum K, et al. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell 2012; 149:1393–1406

47. Freeberg MA, Han T, Moresco JJ, et al. Pervasive and dynamic protein binding sites of the mRNA transcriptome in Saccharomyces cerevisiae. Genome Biol. 2013; 14:R13

48. Kishore S, Jaskiewicz L, Burger L, et al. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat. Methods 2011; 8:559–564

49. Wang T, Xiao G, Chu Y, et al. Design and bioinformatics analysis of genome-wide CLIP experiments. Nucleic Acids Res. 2015; gkv439

50. Corcoran DL, Georgiev S, Mukherjee N, et al. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biol. 2011; 12:R79

51. Chen B, Yun J, Kim MS, et al. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. Genome Biol. 2014; 15:R18

52. Schwartz S, Agarwala SD, Mumbach MR, et al. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. Cell 2013; 155:1409–21

53. Liu N, Dai Q, Zheng G, et al. N6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. Nature 2015; 518:560–564

54. Wan Y, Qu K, Zhang QC, et al. Landscape and variation of RNA secondary structure across the human transcriptome. Nature 2014; 505:706–9

55. Solem AC, Halvorsen M, Ramos SB V, et al. The potential of the riboSNitch in personalized medicine. Wiley Interdiscip. Rev. RNA 2015; 6:517–32

56. Halvorsen M, Martin JS, Broadaway S, et al. Disease-associated mutations that alter the RNA structural ensemble. PLoS Genet. 2010; 6:e1001074

57. Corley M, Solem A, Qu K, et al. Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. Nucleic Acids Res. 2015; 43:1859–68

58. Eddy SR. How do RNA folding algorithms work? Nat. Biotechnol. 2004; 22:1457–1458

59. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981; 9:133–48

60. Mathews DH, Sabina J, Zuker M, et al. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 1999; 288:911–40

61. Lu ZJ, Turner DH, Mathews DH. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. Nucleic Acids Res. 2006; 34:4912–24

62. Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res. 2003; 31:3423–3428

63. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics 2006; 22:e90–8

64. Lu ZJ, Gloor JW, Mathews DH. Improved RNA secondary structure prediction by maximizing expected pair accuracy. RNA 2009; 15:1805–13

65. Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. 2003; 31:7280–7301

66. Mathews DH. RNA Secondary Structure Analysis Using RNAstructure. Curr. Protoc. Bioinformatics 2014; 46:12.6.1–12.6.25

67. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, et al. ViennaRNA Package 2.0. Algorithms Mol. Biol. 2011; 6:26

68. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003; 31:3406–15

69. Markham NR, Zuker M. UNAFold: Software for nucleic acid folding and hybridization. Methods Mol. Biol. 2008; 453:3–31

70. Kwok CK, Tang Y, Assmann SM, et al. The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. Trends Biochem. Sci. 2015; 40:221–232

71. Eddy SR. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. Annu. Rev. Biophys. 2014; 43:433–56

72. Hofacker IL, Fontana W, Stadler PF, et al. Fast folding and comparison of RNA secondary structures. Monatshefte fur Chemie Chem. Mon. 1989; 125:167–188

73. Deigan KE, Li TW, Mathews DH, et al. Accurate SHAPE-directed RNA structure determination. Proc. Natl. Acad. Sci. U. S. A. 2009; 106:97–102

74. Hajdin CE, Bellaousov S, Huggins W, et al. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. Proc. Natl. Acad. Sci. U. S. A. 2013; 110:5498–503

75. Rice GM, Leonard CW, Weeks KM. RNA secondary structure modeling at consistent high accuracy using differential SHAPE. RNA 2014; 20:846–854

76. Cordero P, Kladwang W, VanLang CC, et al. Quantitative Dimethyl Sulfate Mapping for Automated RNA Secondary Structure Inference. Biochemistry 2012; 51:7037–7039

77. Wu Y, Shi B, Ding X, et al. Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. Nucleic Acids Res. 2015; 43:7247–59

78. Zarringhalam K, Meyer MM, Dotu I, et al. Integrating Chemical Footprinting Data into RNA Secondary Structure Prediction. PLoS One 2012; 7:e45160

79. Zuker M, Mathews DH, Turner DH. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. 1999; 70:11–43

80. Washietl S, Hofacker IL, Stadler PF, et al. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. Nucleic Acids Res. 2012; 40:4261–72

81. Tacker M, Stadler PF, Bornberg-Bauer EG, et al. Algorithm independent properties of RNA secondary structure predictions. Eur. Biophys. J. 1996; 25:115–130

82. Ouyang Z, Snyder MP, Chang HY. SeqFold: Genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. Genome Res. 2013; 23:377–387

83. Cordero P, Lucks JB, Das R. An RNA mapping data base for curating RNA structure mapping experiments. Bioinformatics 2012; 28:3006–3008

84. Tang Y, Bouvier E, Kwok CK, et al. StructureFold: genome-wide RNA secondary structure mapping and reconstruction in vivo. Bioinformatics 2015; 31:2668–75

85. Zhong C, Zhang S. Simultaneous folding of alternative RNA structures with mutual constraints: an application to next-generation sequencing-based RNA structure probing. J. Comput. Biol. 2014; 21:609–21

86. Hu X, Wong TKF, Lu ZJ, et al. Computational identification of protein binding sites on RNAs using high-throughput RNA structure-probing data. Bioinformatics 2014; 30:1049–1055

87. Mauger DM, Weeks KM. Toward global RNA structure analysis. Nat. Biotechnol. 2010; 28:1178–1179

88. Turner DH, Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res. 2010; 38:D280–2

89. Hamada M, Kiryu H, Sato K, et al. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics 2009; 25:465–73

90. Sievers C, Schlumpf T, Sawarkar R, et al. Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. Nucleic Acids Res. 2012; 40:e160

91. Friedersdorf MB, Keene JD. Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. Genome Biol. 2014; 15:R2

92. Kazan H, Ray D, Chan ET, et al. RNAcontext: A new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS Comput. Biol. 2010; 6:28

93. Ray D, Kazan H, Chan ET, et al. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat. Biotechnol. 2009; 27:667–70

94. Fukunaga T, Ozaki H, Terai G, et al. CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. Genome Biol. 2014; 15:R16

95. Maticzka D, Lange SJ, Costa F, et al. GraphProt: modeling binding preferences of RNA-binding proteins. Genome Biol. 2014; 15:R17

96. Goodarzi H, Najafabadi HS, Oikonomou P, et al. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. Nature 2012; 485:264–268

97. Zhang C, Lee KY, Swanson MS, et al. Prediction of clustered RNA-binding protein motif sites in the mammalian genome. Nucleic Acids Res. 2013; 41:6793–6807

98. Livi CM, Blanzieri E. Protein-specific prediction of mRNA binding using RNA sequences, binding motifs and predicted secondary structures. BMC Bioinformatics 2014; 15:123

99. Vapnik VN. The Nature of Statistical Learning Theory. SpringerVerlag New York 1995; 8:

100. Steffen P, Voss B, Rehmsmeier M, et al. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics 2006; 22:500–3

101. Heyne S, Costa F, Rose D, et al. GraphClust: alignment-free structural clustering of local RNA secondary structures. Bioinformatics 2012; 28:i224–i232

102. Grave FC and KD. Fast Neighborhood Subgraph Pairwise Distance Kernel. Proc. 27th Int. Conf. Mach. Learn. 2010; 255–262

103. Breiman L (University of C. Random forest. Mach. Learn. 1999; 45:

104. Sugimoto Y, Vigilante A, Darbo E, et al. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. Nature 2015; 519:491–494

105. Ray D, Kazan H, Cook KB, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature 2013; 499:172–7

106. Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from genome-wide studies. Nat. Rev. Genet. 2014; 15:469–79

Figures and Tables

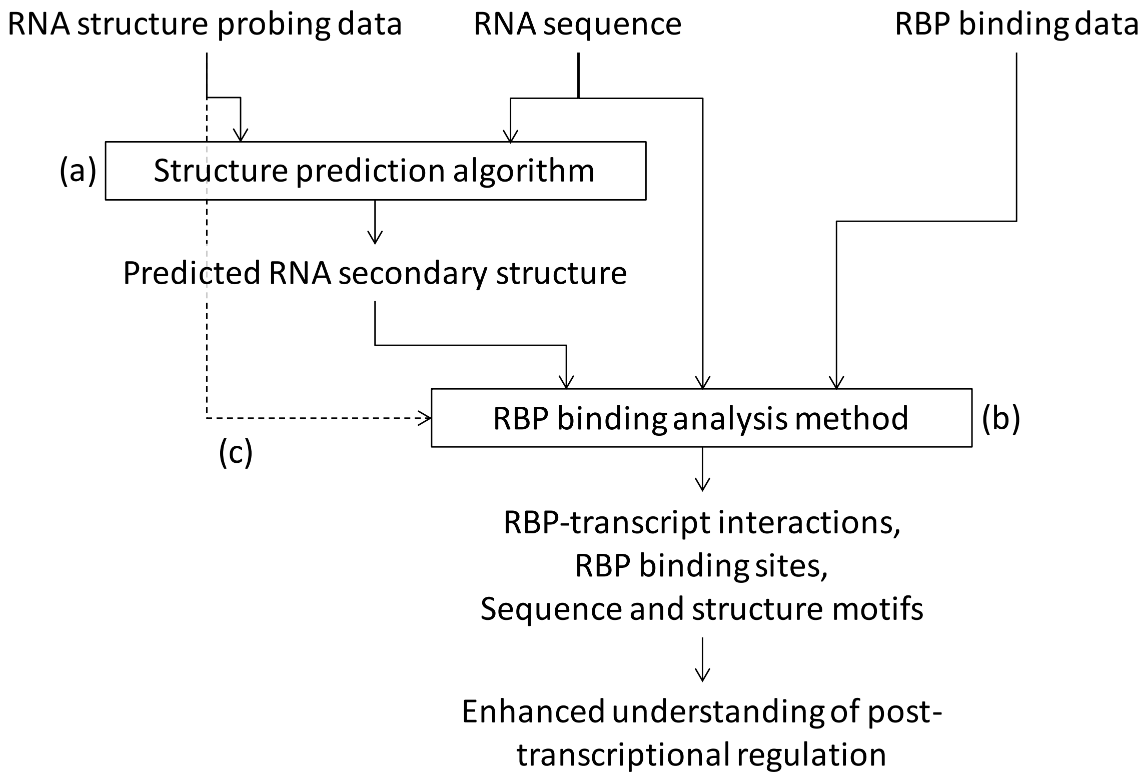
RNA structure probing data
RNA sequence
RBP binding data
(a)
Structure prediction algorithm
Predicted RNA secondary structure
RBP binding analysis method
(b)
(c)
RBP-transcript interactions,
RBP binding sites,
Sequence and structure motifs
Enhanced understanding of post-transcriptional regulation

Figure 1: Analysis workflow of RNA structures at RBP-binding sites. (a) Some contemporary RNA secondary structure prediction algorithms can incorporate both the information in RNA sequences and high-throughput RNA structure probing data in their prediction process. (b) Many methods for analyzing high-throughput RBP binding data take RNA sequences and/or predicted RNA structures to correct for the biases in the binding data, and identify both RBP-transcription interacting pairs as well as the binding locations of the RBPs on their bound

transcripts. (c) A possible alternative approach is to input RNA structure probing data directly into RBP binding analysis methods without predicting complete RNA secondary structures.

Table 1: High-throughput sequencing methods for probing RNA structures. The methods are ordered by the size of the probe they use.

Probe (size in *Dalton*)	**Method**	**Protocol attributes**		**Data attributes**	
		Advantages	**Limitations**	**Technical biases**	**Experimental noise**
DMS (126Da)	DMS-seq [25], structure-seq [24], Mod-seq [26]	High resolution and able to work both *in vivo* and *in vitro*	Mainly probing A and C bases only [25]	Signals are most enriched at solvent accessible positions [25]	Different chemical modification efficiencies for different types of nucleotide [15]
IM7 (222Da)	SHAPE-seq [22]	High resolution and able to work both *in vivo* and *in vitro*	Difficult for genome-wide screening [12]	Weak signals near 5' end due to natural polymerase drop-off [34]	Non-uniform distribution of reactive bases [34]
RNase V1 (15,900Da)	PARS [3,17,18], ssRNA-seq [21]	Able to probe paired bases and estimate folding energies	For *in vitro* studies only [3,106]	No signals at 3' end due to the blind tail [3,18]	Low resolution due to steric hindrance [15,87]
RNase ONE (27,000Da)	dsRNA-seq [20,21]	Able to probe unpaired bases			
Nuclease S1 (32,000Da)	PARS [3,18]				
Nuclease P1 (36,000Da)	FragSeq [19]				

Table 2: High-throughput sequencing protocols for identifying RBP binding sites

Method	Brief technical description	Advantages	Limitations
RIP-seq [43]	RNA immuno-precipitation (RIP)	High sensitivity	Containing indirect interactions, high noise level [49]
HITS-CLIP [40]	UV cross-linking (CL) and immuno-precipitation (IP)	Wide applications	May contain nucleotide biases [8,48]
PAR-CLIP [42]	Photoactivatable-ribonucleoside (PAR) with CLIP	Longer UV wavelength and better cross-linking efficiency	
iCLIP [41]	CLIP at individual-nucleotide resolution	High resolution	More steps in the experimental protocol [49]
(Baltz et al., 2012) [45]	Photocrosslinking and RNA pull-down by oligo(dT) beads	Identifying binding sites of all RBPs	Only for RNAs with poly-A tails and cannot tell the RBP that binds each binding site
(Castello et al., 2012) [46]	PAR and complementary cross-linking, using oligo(dT) to select RNAs		
gPAR-CLIP [47]	Using oligo(dT) to select RNAs, and biotin to select proteins		

Table 3: Structure probing data integrated into RNA secondary structure prediction methods

Sequencing data for RNA structures	Structure prediction approach	Software /module	Backend package	Way of using structure probing data	Special features
DMS	Minimum free energy (MFE)	Fold --constraint [33]	RNAstructure	Free energy penalty	Using DMS signals to define binary (paired or unpaired) constraints
DMS	MFE	Fold --DMS [76]	RNAstructure	Pseudo energy term	Can utilize quantitative signals of DMS data
DMS or SHAPE	MFE	StructureFold [84]	RNAstructure and ViennaRNA	Pseudo energy term	Providing functions to pre-processing structure-probing data
SHAPE	MFE	Fold --SHAPE [73]	RNAstructure	Pseudo energy term	Can utilize quantitative SHAPE signals
SHAPE	Maximize expected accuracy (MEA)	RNApbfold [80]	ViennaRNA	Discrepancy minimization	Treating structure-probing data as probabilistic inputs
SHAPE	MEA	RNAsc [78]	RNAstructure	Pseudo energy term for all nucleotides	Enforcing constraints to all bases
SHAPE	MFE	ShapeKnots [74]	RNAstructure	Pseudo energy term for nucleotides and pseudoknots	Can predict structures with pseudoknots
SHAPE	MFE	MutualFold [85]	-	Mutual constraints	Can predict alternative RNA structures
SHAPE or Enzymatic cleavage	Sample and select	SeqFold [82]	Sfold	Centroid of the closest structure cluster	Can handle very noisy structure-probing data
Enzymatic cleavage	MFE	Fold --constraint [60]	RNAstructure	Free energy penalty	Using structure-probing data to define binary constraints
Enzymatic cleavage	MFE	Parameter AUX [68]	Mfold	Force constraints	Allowing constraints to be

					specified in flexible ways
Enzymatic cleavage	MFE	RNAfold --constraint [67]	ViennaRNA	Pseudo energy term	Using structure-probing data to define binary constraints
DMS, SHAPE or Enzymatic cleavage	MEA	RME [77]	RNAstructure	Pseudo energy term and discrepancy minimization	Treating structure-probing data as probabilistic inputs

Table 4: Large-scale RBP binding data used in RBP binding motif modeling methods

Sequencing data for RBP binding sites	Software	Computational models/ methods	Features for modeling	Advantages	Limitations
RIP-seq or CLIP-seq	RNAcontext [92]	Cost function minimization	Sequences, and predicted structural preferences (Sfold)	Easy to understand the motifs	Using only one-dimensional structural profiles, leaving pairing relationships between bases unused [95]
HITS-CLIP	TEISER [96]	Mutual information	Context-free grammars for both sequence and structural information	Use of a flexible context-free grammar framework, exhaustive search of motifs	Finding only short local motifs [8]
HITS-CLIP	MCarts [97]	Hidden Markov model	Distance to neighbor sites, accessibility of tetramers, and conservation (BLS)	Flexible model, can capture clustered motifs	Limited to RBPs that recognize binding sites with clustered motifs
gPAR-CLIP or PAR-CLIP	ProbRNA [86]	Random forest	Sequences, predicted secondary structure (RNAfold), and PARS scores (MPL)	Can utilize RNA-structure probing data to derive structural features	Not trivial to understand the sequence and structural features of the binding sites abstracted in the machine learning models
PAR-CLIP	OliMoSS [98]	Support vector machine	Sequences, motifs, and predicted secondary structure (RNAfold)	Able to utilize known motifs and discover new motifs	
HITS-CLIP, PAR-CLIP, or iCLIP	GraphProt [95]	Support vector machine with a graph kernel	Sequence, Predicted secondary structure	Involving a rich set of features and capture relational	

			(RNAshape)	information using a graph kernel	