Title
Comparative and integrative analysis of RNA structural profiling data: current practices and emerging questions

Permalink
https://escholarship.org/uc/item/2qv8q11r

Journal
Quantitative Biology, 5(1)

ISSN
2095-4689

Authors
Choudhary, Krishna
Deng, Fei
Aviran, Sharon

Publication Date
2017-03-01

DOI
10.1007/s40484-017-0093-6

Peer reviewed

Noname manuscript No.
(will be inserted by the editor)

Comparative and integrative analysis of RNA structural profiling data: current practices and emerging questions

Krishna Choudhary* · **Fei Deng*** ·
Sharon Aviran†

Received: date / Accepted: date

Abstract Structure profiling experiments provide single nucleotide information on RNA structure. Recent advances in chemistry combined with application of high-throughput sequencing have enabled structure profiling at transcriptome scale and in living cells, creating unprecedented opportunities for RNA biology. Propelled by these experimental advances, massive data with ever-increasing diversity and complexity have been generated, which give rise to new challenges in interpreting and analyzing these data. We review current practices in analysis of structure profiling data with emphasis on comparative and integrative analysis as well as highlight emerging questions. Comparative analysis has revealed structural patterns across transcriptomes and has become an integral component of recent profiling studies. Additionally, profiling data can be integrated into traditional structure prediction algorithms to improve prediction accuracy. To keep pace with experimental developments, methods to facilitate, enhance and refine such analyses are needed. Parallel advances in analysis will complement profiling technologies and help them reach their full potential.

Keywords RNA structure profiling · high-throughput sequencing · comparative analysis · secondary structure prediction

1 Introduction

RNAs are known to play essential roles in diverse cellular functions, extending well-beyond transfer of information from genes to proteins [1, 2]. For example, small non-coding RNAs such as microRNAs and small interfering RNAs

* These authors contributed equally to this work.

† Corresponding author: saviran@ucdavis.edu.

Author address: Department of Biomedical Engineering and Genome Center, University of California at Davis, Davis, California, 95616, USA.

have regulatory roles in gene expression [3]. Long non-coding RNAs are also widely found in various regulatory roles at both transcriptional and post-transcriptional levels [4]. RNA function is closely linked with its ability to fold into and convert between specific complex structures. In fact, determining structure has become a crucial step in understanding RNA function [5]. Accurate and high-resolution structure models have been traditionally obtained using comparative sequence analysis or experimental techniques, such as X-ray crystallography and nuclear magnetic resonance (NMR) [6]. However, these methods require considerable manual labor and suffer technological limitations, which have precluded their use beyond a small scale [7]. Computational structure prediction from sequence information is a broadly applicable alternative that has been widely used [8, 9], but reported structures often suffer from poor accuracy.

Structure profiling (SP), also known as structure probing or chemical probing, refers to a family of experiments that characterize RNA structure [10, 11]. In these experiments, local structural characteristics are gleaned using structure-sensitive reagents that modify RNAs at nucleotide level. Well-studied reagents include dimethyl sulfate (DMS) [12], kethoxal [13], hydroxyl radicals [14], diethyl pyrocarbonate (DEPC) [15], CMCT [16], lead(II) [17, 18], nucleases [19] and SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) [20]. Until very recently, limitations of probing reagents as well as sequencing and informatics challenges restricted SP to select few RNAs studied individually and primarily under *in vitro* conditions. The newest generation of SP experiments utilizes high-throughput sequencing techniques, which provide unprecedented multiplexing capacity in a cost-effective and automated manner. These advances have been used to study RNAs of varying lengths *in vitro* and *in vivo*, and more recently at transcriptome scale [21–42]. Despite shared principles, experiments differ in the information they extract and in the statistical properties of their measurements. Experimental protocols for SP and their biological applications have been reviewed previously, see for example [11, 43–46].

Sequencing readouts from SP experiments are analyzed to extract structural parameters of interest for each nucleotide, in terms of its *reactivity* to the probing reagent. Nucleotide-level estimates are subsequently used to answer biological questions of interest, which may entail further analysis and interpretation. In this manuscript, we focus on approaches to using reactivity data for comparative and integrative analysis – a central theme in recent studies. Comparative analysis of SP data has revealed structural patterns across different levels, ranging from low-resolution transcriptome level to high-resolution nucleotide level. Each level may require specialized analysis methods. Note that even for the same level, the ideal approach could possibly differ depending on the context. We discuss three different contexts where technical, biological and systematic replicates of SP data are available. In addition to comparative analysis, we also review current progress in data-directed structure prediction, which is the most straightforward application of SP data in structural biology. Unlike X-ray crystallography and NMR, in which RNA structure is explicitly modeled, SP does not directly reveal the pairing state of a nucleotide nor its pairing partner. However, it can complement structure prediction algorithms to enhance their performance [47,48].

This review is organized as follows. We begin with a discussion in Section 2 on shared aspects of SP experiments and devote the bulk of the article to data interpretation and analysis. In Section 3, we review current practices and principles in reactivity calculation. Recent approaches and emerging questions in comparative and integrative analysis are discussed in Section 4, while quality control of large-scale SP data is discussed in Section 5. Algorithms for secondary structure prediction and efforts to leverage SP data for improving their performance are reviewed in Section 6. Recent progress in public repositories, analysis tools and visualization platforms is reported in Section 7.

2 Overview of structure profiling experiments

The general goal of an SP experiment is to obtain nucleotide-resolution structural characteristics of all RNAs in a sample [49]. Structural characteristics in the vicinity of a nucleotide are reflected in local stereochemical properties such as nucleotide dynamics, solvent accessibility and electrostatic environment [11,50]. In particular, pairing state of a nucleotide is known to be correlated with these stereochemical properties [51]. SP experiments utilize reagents that are sensitive to local stereochemistry [11]. These reagents react with nucleotides such that the reactivity to any particular nucleotide depends on its local stereochemistry, which in turn is affected by its pairing state. Hence, SP experiments aim to measure the sequence of reactivities corresponding to nucleotides of each transcript. High and low reactivities are indicative of unpaired and paired nucleotides, respectively [52]. Hence, it is understood that the sequence of nucleotide *reactivities*, henceforth called *reactivity profile*, is a representation of RNA's structure [53].

Most sequencing-based SP techniques share a common workflow (Fig. 1) [43,44]. To start with, a sample of RNAs is allowed to react with structure-sensitive reagents, resulting in chemical modifications of nucleotides. Degree of modification can be detected by reverse transcription (RT), which either stops or proceeds but with a mutation at modified nucleotides. The resulting cDNA library is sequenced and reads are mapped to sequences of target RNAs. Then, RT stops or mutations are counted for each nucleotide. To measure background noise in RT stops or mutations, parallel to the experiment, a control assay is performed wherein the RNAs are not treated with reagents. This control assay also yields a count summary for each nucleotide. The counts from experiment and control assays are combined to obtain reactivity profiles for all target RNAs.

Despite the shared principles, measured reactivities are influenced by numerous intertwined factors that all impact the variability of data [54]. In fact, it has been found that single nucleotide variants can lead to significantly different reactivity profiles [55,56] and identical sequences can have different reactivity profiles under different conditions [40,57,58]. Comparison of reactivity profiles reveals that quantitative differences persist even in the absence of structural differences between RNAs from one sample to another [54, 59]. Listed below are factors that influence reactivity profiles.

Technical factors. Numerous technical factors add to variability in observed profiles. First, chemical reactions involved in SP occur in presence of limited quantities of reagents/transcripts. Concentrations of reagents are often controlled deliberately to limiting amounts to achieve desirable reaction kinetics [11]. In addition, many RNAs of interest are present in limited quantities [28]. As such, the reactions feature inherent stochasticity [54,60]. Secondly, these reactions are sensitive to stereochemistry and solvent conditions [11,61]. Nevertheless, they often occur in complex and dynamic solution environments. For example, RNAs often feature a dynamic ensemble of co-existing structures *in vivo* interacting with proteins and other biomolecules [62–64]. However, SP captures only the average profile for all these structures, combining influences from intermolecular interactions [65]. In addition, cDNA library preparation involves numerous steps such as adapter ligation, reverse transcription and PCR, which also introduce stochasticities. Finally, readouts from sequencing machines are also affected by stochasticities [54, 66, 67]. These factors contribute to variance in reactivity profiles. In fact, they contribute to variance in any other parameter of interest that is estimated from data, e.g., Gini index of counts/reactivities [26,40,68]. Variance contribution of technical factors to any parameter of interest can be estimated by performing multiple replicates, called as *technical replicates* of experiment-control study starting from biologically indistinguishable RNA samples. We refer to variance in estimates observed purely due to said technical factors as *technical variation* [69–71].

Biological factors. RNAs with significant structural diversity are subjects of recent studies. For example, ncRNAs are known to be highly structured while mRNAs are thought to have a lesser degree of structure. In fact, within an RNA, structure could significantly vary from one region to another. For example, mRNAs are believed to be less structured in coding regions than in untranslated regions [40]. Additionally, RNA structure is sensitive to factors such as solvent conditions, ligand and salt concentrations, temperature variations and interactions with proteins [61]. Should any of these factors differ between studies, detectable differences in the estimated reactivities may be observed. For example, reactivity profiles for the same transcript have been found to be different between *in vitro* and *in vivo* conditions [40, 68, 72]. We refer to variance of an estimate observed purely due to biological factors as *biological variation*. Additionally, it is to be noted that biological variation might be caused by differences in RNA-protein interactions besides structural differences [73]. Proteins can cover certain stretches of nucleotides on RNA, influencing the reactivities. Two RNA samples known to have come from different biological sources are called *biological replicates* [69–71]. These contain information about biological differences between the samples.

Systematic factors. For biologically identical RNAs, reactivity measurements obtained in one experiment can differ from the profiles obtained through a different experimental protocol [53, 74]. Technical replicates do not capture these variations as they do not differ in protocol steps. Yet, such variations do not originate due to biological factors. Such differences can be attributed to discrepancies in key steps. For example, many current methods differ in choice of probing reagent. In fact, a variety of reagents are available, e.g., DMS, kethoxal, hydroxyl radical, 1M7, NMIA, NAI, NAI-N_3, etc. but each has its pros and cons [11,22,40,75]. These reagents differ in their stereochemical characteristics and reaction mechanisms. Consequently, the reactivity profiles may reflect these differences. In addition, many reagents do not probe all nucleotides and have biases that cause different reactivities depending on nucleotide type even in the absence of structural differences [11]. Besides choice of probing reagent, protocols often differ in priming methods, modification detection method (e.g. stop/mutation), ligation strategies, enrichment, sequencing mode (single/paired-ended), reactivity estimation method among others. These are a few noteworthy steps having equally plausible alternatives. Many of these steps contribute to biases, which interplay with other steps resulting in miscellaneous effects in parameter estimates [54]. Nevertheless, *biologically* identical RNAs can be studied using different protocols to obtain detailed and comprehensive insights [74]. We refer to experiments involving SP of biologically indistinguishable samples using different protocols as *systematic replicates* and variances originating due to differences in protocol as *systematic variation.*

3 Estimation of structure profile

As mentioned earlier, sequenced reads from both experiment and control assays are summarized as count of stops or mutations for each nucleotide. However, per-nucleotide counts are not directly comparable because they can differ in magnitude due to a variety of reasons. Number of reads mapped to a transcript, also known as its *coverage*, varies between transcripts due to the dramatic differences in their relative abundances, which often range over five orders of magnitude [28, 76]. Additionally, priming or ligation biases contribute to sequence-specific variations in counts within the same transcript [22, 54, 59, 77, 78]. Counts may differ due to background noise in RT stops and mutations. In fact, for the same nucleotide between experiment and control, counts may not be comparable due to difference in sequencing depths. For these reasons, counts are processed into normalized reactivities, which are assumed to be comparable across transcripts and replicates.

Reactivity estimation methods differ between studies but share the following conceptual framework (see Fig. 1). 1) Counts are adjusted to account for variations in coverage, yielding two *detection rates* - one for experiment and one for control. 2) Comparison of detection rates yields an estimate for degree of modification, or *raw reactivity.* 3) Raw reactivities are normalized to ensure that values for all transcripts and replicates thereof span the same interval.

Detection rates. Detection rates are calculated to account for variations in coverage. However, variations in coverage exist at all levels. For example, substantial coverage differences have been noted between rRNAs and mRNAs [28]. Significant differences in coverage exist from one transcript to another within the same functional class. Additionally, within a transcript, coverage can be considered on regional basis (e.g., coverage of $5'$ untranslated region or coding region, or $3'$ end, etc.), sequence basis (e.g., more coverage in GC rich regions due to priming bias), or per-nucleotide basis. In general, coverage differences can be noted at all levels of organization. Analysis methods in various studies differ in the level of detail at which they account for coverage variations.

Many groups consider coverage variations between transcripts as significant while assuming uniformity of coverage within each transcript. Higher coverages for a transcript may be a result of its over-abundance in sample or priming biases among other factors. In such cases, counts corresponding to nucleotides of the transcript may be assumed to be proportionally higher. Hence, several studies adjust counts by their mean to account for coverage bias [28, 30, 31, 35, 36, 55]. Additionally, Ding *et al.* [28] take the logarithm of counts to make count distribution symmetric. Others note that there could be local biases within the transcripts. For example, Rouskin *et al.* [26] adjusted counts for each nucleotide by maximum counts in a local window. In fact, several studies [22, 25, 40, 60, 79, 80] have accounted for nucleotide-level coverage variations. Through these adjustments, detection rates are estimated for both experiment and control.

Raw reactivities. Detections in control result from noise in RT while detections in experiment result from noise in RT as well as modifications at nucleotides. Hence, it is expected that at any nucleotide, detection rate will be higher in experiment. One assumption is that structure-sensitive modifications contribute additively to a background level of detection rates. Hence, reactivities are calculated by subtracting detection rate in control from that in experiment [22,28,40,60,80]. Alternatively, reactivities have been estimated as odds ratio of experiment to control detection rates [35]. To control the range of reactivities, others take the logarithm of the odds ratio [30,31,36,55,81]. Additionally, sometimes detection rates in experiment are found to be less than control. In such cases, a basal reactivity value of 0 (if subtracting detection rates) or 1 (if taking ratio) is assigned. This is done because the detection rate due to noise is often very low and if detection rates remain comparable or lower in the presence of modifications, it indicates negligible degree of modification.

Normalized reactivities. Profiles from different protocols could span disjoint intervals even for the same RNA. In fact, for different RNAs in the same experiment, profiles could span disjoint intervals because of biological variation. Raw reactivities are not considered comparable in absolute magnitude. Hence, all profiles are normalized such that average reactivity of $\sim 10\%$ of the most reactive nucleotides is 1, excluding few unusually reactive nucleotides that are considered outliers [47]. Outliers can originate in datasets due to a variety of reasons, such as excessive degradation or over-modification at certain nucleotides, or over-representation of certain fragments due to various inherent biases in protocols. In fact, such hyper-reactive sites often appear in datasets [51, 82].

Accordingly, most current approaches to normalization begin with identification of outliers in reactivity estimates [83]. This is done by either box plot analysis whereby reactivities greater than 1.5 times the interquartile range are deemed outliers [47, 82], or by assuming that reactivities beyond a certain percentile are outliers [47]. Outliers are either ignored [47] in the process of calculating normalization constant or winsorized [21, 26, 36]. To estimate normalization constant, one approach is to take the mean of values greater than a certain percentile after removing outliers. For example, 2-8% method assumes that the top 2% of reactivities are outliers and normalizes with mean of the next 8% of highest reactivities [47]. The winsorization approach aims to scale reactivities such that they range from 0-1 for all transcripts. Hence, after winsorization, the highest reactivity is chosen as normalization constant [21,26,36].

In the majority of analysis methods, the above workflow is preceeded by conventional read alignment and counting routines. Recently, these preprocesing steps were integrated with reactivity estimation, such that counting and estimation are resolved simultaneously [79]. This is especially attractive in situations where multi-mapping reads (reads which align to multiple sites in a transcriptome) abound, e.g., in studies of splicing isoforms. While common remedies discard such reads or allocate them uniformly among potential alignments, Li *et al.* [79] expand on prior modeling and statistical inference work in RNA-Seq [84, 85] and SHAPE-Seq analysis [80] to address this issue. Another extension of the said statistical modeling work on SHAPE-Seq has been recently published by Selega *et al.* [81] This method scores significance of modification level from stop counts and nucleotide-level coverages under an assumption that modification states do not randomly switch, i.e., significantly reactive/unreactive nucleotides tend to appear in continuous stretches. The assumption is enforced using a Hidden Markov Model with transition probabilities based on empirically derived expected lengths of reactive and unreactive contiguous stretches.

4 Comparative analysis

Before the advent of high-throughput sequencing, probing was mostly applied to select highly structured ncRNAs under *in vitro* conditions. Recent advances have dramatically expanded the scope of SP and diverse RNAs can now be studied in biologically relevant conditions. In fact, applications of SP to numerous transcripts and transcriptomes have revealed novel insights [2, 44]. Most such applications feature comparative analysis. Several recent examples of such analysis can be noted: 1) Spitale *et al.* [40] compared mRNA profiles and identified conserved patterns around translation start site. 2) Protein-RNA interactions were studied in viral RNA and mammalian ncRNAs and mRNAs by comparing reactivity profiles under different conditions [40, 58, 86, 87], finding that interactions modulate reactivities significantly. 3) Comparison of coding regions of mRNAs revealed a three-nucleotide periodicity pattern in reactivities [28, 30, 40]. 4) Significant structural alterations have been identified for single-nucleotide variants [55,88]. 5) Comparisons of entire transcriptomes at different temperatures identified structure-altering responses [26, 89, 90]. 6) Prevalence of specific noncanonical structural motifs have been found to differ between *in vitro* and *in vivo* conditions [68]. In fact, these studies involve comparisons at different levels such as structure at the level of regions within a transcript, at the transcript level, within functional classes, or at transcriptome level. In this section, we review recent methods and emerging questions in addressing these challenges.

Notably, SP collects data at nucleotide level, but structural dynamics most often involve at least a few nucleotides or even entire functional domains. For example, many of the studies mentioned above seek signals that span protein-binding sites, codons and well-defined local structural motifs. In fact, it is rare for a biological study to home in on isolated single-nucleotide reactivity changes. For this reason, comparative studies must also bridge between the resolution of measurements and that of sought-after effects. This is typically accomplished by integrating nucleotide information for scoping structural effects at various levels of lower resolution and/or by inspecting data-directed secondary structure predictions for detectable changes at that level [40,53,56,91].

4.1 Comparing technical replicates

Agreement between technical replicates indicates high quality of data. Technical replicates can be compared at the level of transcripts or at the level of nucleotides.

Transcript-level comparison. In high-throughput experiments or when studying long transcripts, agreement between replicates of a transcript is commonly evaluated as Pearson correlation coefficient (PCC) for reactivity profiles. Transcripts with low PCC are filtered for biological purposes as their replicates do not agree. For each pair of profiles, PCC quantifies agreement in one number that is invariant to normalization. However, PCC has its limitations as a measure of agreement [92–94].

First, PCC is sensitive to outliers [92]. PCC is based on the sample means of reactivities in the profiles that it is comparing. Sample means are known to be sensitive to outliers, leading to similar sensitivity of PCC. Indeed, PCC is affected by both magnitude of outliers and the overall proportion of reactivities that is outlier. Hence, PCC is to be used with caution, especially for transcriptome-wide data as outliers have indeed been noted routinely in experiments [47, 59]. From our experience, we have found a common practice in handling of missing information that often systematically leads to outliers in reactivities. While estimating reactivity profiles, poorly covered sites have a bias towards an apparent zero reactivity. This bias significantly adds to the proportion of outliers at the lower extreme – with zero reactivities. However, most studies do not filter outliers while calculating PCC. Hence, PCC may be misleading in evaluating replicate agreement. Second, though PCC can evaluate agreement between profiles for two transcript, it does not quantify agreement at the level of each nucleotide. Finally, PCC only evaluates correlation between two profiles and is unaffected by magnitude differences of nucleotide-level values. Nevertheless, to gauge significance of biological variation found in a study, it is important to quantify technical variation. In fact, since biological variation of interest is often nucleotide-resolution, it is desirable to quantify technical variation at the nucleotide-level.

Nucleotide-level comparison. At the nucleotide-level, replicates have been compared by taking mean and standard deviation of reactivities. In absence of replicates, theoretical formulas and computational methods have been utilized to evaluate technical variation at nucleotide-level [22,59]. However, due to challenges in visualizing technical variation, most such nucleotide-level evaluations were restricted to one or few selected transcripts. In a recent work, Choudhary *et al.* [59] proposed a method to visualize technical variation at nucleotide-resolution for large-scale data based on Signal-to-Noise ratio (SNR). For each nucleotide, its SNR is defined as the ratio of sample mean to standard deviation of reactivities in all replicates. SNR is high when replicates are in quantitative agreement for the nucleotide and low otherwise. SNR values for a transcript could be visualized as box plot to glean replicate agreement for multiple replicates from one plot. Additionally, they proposed mean of SNR as a one-number or point summary for overall transcript's data quality. They found that mean SNR correlates with PCC and transcript's coverage.

Open questions. Nucleotide-resolution comparison of reactivities requires normalization strategies to render values in different replicates comparable. Clearly, normalization methods described in Section 3 require optimizing two criteria – one for identifying outliers and another for selecting reactivities that will be used to estimate normalization constant. However, the proportion of outliers in a dataset could vary depending on the length of transcripts involved, as well as the quality of experiment. Indeed, different labs and even same labs have made different choices for the normalization method for different datasets, though the general principle has been to eliminate outliers and scale reactivities such that they range approximately from 0-2 [39]. These normalization methods have been adopted based either on experience with structure-probing data before high-throughput technologies were developed [47] or validations with structure-prediction [82]. Indeed, the field may benefit from a universal method for normalization, which is assuring enough to dispense with the need for routine optimization of normalization strategy. In fact, before SP became high-throughput, most of the RNAs that were studied with chemical probing were highly structured rRNAs. Heuristic guidelines formulated based on rRNAs may not apply to all transcripts. Also, validation based on structure-prediction itself involves parameter optimization and modeling assumptions as described later. Given the recent advances in SP, methods of normalization warrant a revisit.

4.2 Comparing biological replicates

Comparison of reactivities from different biological replicates could possibly identify significant biological variation. If technical variation is high, statistically significant biological results may not be obtained from data. To estimate significance of biological variation, it has to be examined in comparison with technical variation [69–71]. Indeed, several published studies have reported biological variation at all levels. At transcriptome level, differences in overall structural characteristics have been reported under different conditions and between different strains [26,40]. At transcript-to-transcript level, rRNAs have been described as being more structured than mRNAs. At a finer level, while differences in reactivities can be observed at nucleotide-level, biological variations are assumed to span a stretch of nucleotides [86]. In fact, within transcripts, biological variation has been described between regions. For example, significant differences in structure has been noted between UTRs and coding regions of mRNAs. Here, we review the methods used to measure biological variation.

Transcriptome-level comparison. Current normalization methods as described in Section 3 generally scale the reactivities such that they range from 0 to ∼2 [39]. However, this does not ensure that reactivities of different transcripts are directly comparable. For example, though mRNAs are widely understood to be less structured than the rRNAs [40], current normalization methods scale reactivities for both these classes of RNAs such that they span approximately the same interval. Hence, comparing absolute values of reactivities on a transcriptome scale might be misleading. Differences in lengths of transcripts within the same functional class exacerbate the challenges in comparing profiles due to the need of reliable alignment. To facilitate nucleotide-level comparison of reactivities in case of differences in lengths, particularly for mRNAs, transcripts are often aligned by the start/stop codon and arbitrary lengths (∼ 40-100 nt) are chosen upstream and downstream of start/stop codon in all transcripts to be compared [28,30,36,40,89]. However, functional elements in UTRs differ in sequence and distance from start/stop codon, thus presenting an additional challenge in direct comparisons.

Besides direct nucleotide-level comparisons, another approach invariant to current normalization methods (due to properties as listed below) and applicable for transcripts of different lengths has been utilized. At the transcriptome-level, it has been found that RNAs are, in general, less structured *in vivo* than they are *in vitro* [40]. This conclusion was obtained by examining distribution of Gini indices for reactivity profiles. Gini index is a measure of inequality in a distribution [95]. It has two notable properties - 1) It is a measure of inequality that is high if there is substantial gap in values across the nucleotides. Such high gaps (or inequalities) in distribution of counts and reactivities are expected in case of structured RNAs. Hence, Gini index can serve as a metric to characterize degree of structure in a transcript; 2) It is invariant to scaling, i.e., Gini index does not change as long as relative magnitudes of quantities remain the same. As current normalization methods essentially scale reactivity profiles linearly, scaling invariance is a significant merit of Gini index. It allows application of Gini index without need for optimized normalization strategies.

Transcript-level comparison. Structural similarities are often correlated with sequence and/or functional similarity [96]. Hence, in presence of known sequence and/or functional similarities, it may be reasonable to assume that reactivity profiles should span the same interval. Current normalization schemes do scale reactivity profiles such that they span the same interval from 0 to ∼2 [39]. Hence, for cases with sequence and/or functional similarity, reactivity profiles have been compared by taking difference of normalized reactivities [23, 40, 58, 86]. Additionally, based on models specific to the context, p-values can be calculated to characterize the significance of observed differences. Other approaches to establish statistical significance have also been used. For example, Smola *et al.* [86] used a modified version of Z-factor test [97] instead of p-values to screen for sites with statistically significant differential reactivities. Z-factor is a screening coefficient that identifies nucleotides with biological variation substantially greater than technical variation. Recently, Choudhary *et al.* [59] have described a way using SNR to quantify magnitudes of biological and technical variation. Besides these methods, comparability of profiles under conditions of sequence and/or functional similarity has been assumed when summarizing reactivity profiles for multiple RNAs with the average reactivity profile. For example, mean of reactivities has been used to summarize the general characteristic of mRNA reactivity profiles around the translation start site [26, 28].

Regional comparison. Reactivity profiles often feature significant variations across the length of the transcript indicating presence of structured and unstructured regions [28, 40]. Several methods have been utilized to scan regions of transcript for structural properties. Overall, the methods differ primarily in the structural characteristic that they scan for. For example, Gini index has been applied to regions within a transcript [26, 40] to identify those with high inequalities in counts/reactivities across nucleotides. While Spitale *et al.* [40] applied it to designated regions (such as UTRs and coding regions of mRNAs), Rouskin *et al.* [26] applied it to rolling windows containing 50 probed nucleotides. Other studies scanned transcripts to identify regions with higher or lower reactivities. Reactivity level in a region can provide an idea about the number of base pairs in that region. To this end, median of reactivities for a region has been used as a robust summary of regional structural characteristics [39,53,98]. Standard statistical tests such as Wilcoxon rank sum test have been used to evaluate statistical significance of differences between centers of reactivity distributions for two regions [36]. Additionally, Siegfried *et al.* [39] utilized Shannon entropy estimates based on pairing likelihood from data-directed ensemble prediction to quantify a region's structural properties. Shannon entropies are low for regions that have well-defined structures or are predominantly single-stranded and high otherwise.

Open questions. Comparative analysis of SP data is in its nascent phase and several issues are yet to be addressed. For several comparisons, the field has resorted to point summary of *structure* (e.g., Gini index of counts). While reactivity profiles are being compared for quantitative differences in structure, standard characteristics of *more* structured sequences have not been discussed adequately. Consequently, multiple metrics for quantifying regional structure have prevailed thus far.

At the transcriptome-level, Gini index has been applied as a point summary of transcripts structure. However, there are several drawbacks of this index. One of the major issues with this index is that it is highly influenced by outliers [99], which again underscores the importance of robust outlier detection methods. Another major issue is that two transcripts could have very different reactivity distributions but the same Gini index making it difficult to interpret this index. For example, consider two transcripts with different distributions - (a) 50% of nucleotides with zero reactivities and rest with equal and high reactivities (or in other words, 50% sites with high reactivity and 50% sites with low reactivity) and (b) 25% of nucleotides with reactivity of 0.11 and rest with reactivity 1 (or in other words, 75% sites with high reactivity and 25% sites with low reactivity). Both these distributions result in Gini index of 0.5, although the underlying structure profiles are significantly different.

4.3 Comparing systematic replicates

Reactivity profiles estimated from systematic replicates may provide more comprehensive insights into structure. For example, collecting and comparing information from multiple probing reagents has traditionally served as means of increasing confidence in structural inference from data. Whereas such approach had been limited in applicability due to cost and labor constraints, as experiments have now become more accessible to the community, it appears to be gaining popularity [74, 100–102]. To date, comparisons of systematic replicates have been mostly performed semi-quantitatively or via PCC [33, 53]. While PCC only informs about agreement of data, it is often desirable to integrate data from systematic replicates. For example, data from systematic replicates could improve the accuracy of data-directed structure prediction if fused appropriately [103], such that correlations and systematic deviations are well-characterized and accounted for. However, systematic replicates often derive from significantly different statistical distributions. So, besides scale, normalizing systematic replicates is done to ensure comparability of statistical properties. For this purpose, Wu *et al.* [104] used quantile normalization to transform reactivities of each RNA in different datasets such that they follow the same distribution. Because the data throughput bottleneck was only recently eliminated, not much has been done to address these emerging needs. Ensuring quantitative comparability and integrability of profiles from systematic replicates remains an open challenge.

5 Screening data for quality

Since its days of inception, SP has moved towards large-scale transcriptome-wide and *in vivo* experiments. Despite significant advances, data quality remains non-uniform across the transcriptome. Data quality is primarily determined by coverage and agreement of technical replicates. Most studies filter out poor-quality data and base the biological insights on high-quality data. Simple criteria based on transcript's coverage per unit length have been utilized to screen for high-quality components of dataset. Several groups have considered transcript's coverage per unit length ≥ 1 as acceptable criterion for quality [26,28,34,36]. Besides transcript's coverage per unit length, others have considered nucleotide-level coverage as an important criterion for good quality [22,39,40]. Several conditions have been used to optimize these criteria. For example, Smola *et al.* recommend nucleotide-level coverage above ~ 2000 for high confidence in reactivity estimates [22]. This value is desired to achieve high accuracy of structure prediction [39]. Another group, Spitale *et al.* optimized their criteria for high coverage such that transcripts satisfying this criteria have high PCC between replicates [40]. On the other hand, Choudhary *et al.* [59] built upon the probabilistic model by Aviran *et al.* [60] to develop Coverage Quality Index (CQI) that quantifies the "goodness" of each nucleotide's coverage. Basically, given a desired level of confidence for an acceptable variation in estimates, CQI is the ratio of desired coverage of a nucleotide to its observed coverage. CQI < 1 is indicative of good quality while CQI > 1 is indicative of poor quality. Including several metrics such as CQI, SNR and others, Choudhary *et al.* put together an SP data quality assessment tool, called SEQualyzer (see Fig. 2 for an example) that includes quality results and visualizations from nucleotide to transcriptome level [105]. Standardized methods for evaluating quality of data and screening data for high-quality components are indeed essential for maturation of the field.

6 Secondary structure prediction

Computational RNA structure prediction has been studied for several decades. Here, we focus on secondary structure prediction; readers are referred to [106] for a recent review on three-dimensional structure modeling. Typically, secondary structure prediction methods fall into three major categories: free energy minimization, ensemble-based prediction and comparative sequence analysis. It is worth noting that most existing methods do not allow pseudo-knots in predicted structures, which will make the problem computationally intractable. Several solutions were developed but with additional constraints on the type of considered pseudo-knots [107–115].

6.1 Free energy minimization

The most widely used method for structure prediction from a single sequence aims to find the structure with minimum free energy (MFE). This method relies on the second law of thermodynamics which states that MFE structure is the most thermodynamically stable and the most prevalent in living cells. Free energy of a structure can be calculated based on a set of nearest-neighbor thermodynamic models (NNTM) parameters, which are obtained using optical melting experiments [116–118].

At the core of MFE prediction is a dynamic programming algorithm put forth in [119, 120] and was first proposed in the context of maximizing the number of predicted base pairs in [119,121]. It was then extended in [122,123] by incorporating free energies of different structure motifs. This algorithm has been implemented in popular software packages such as UNAFold [124], RNAstructure [125] and ViennaRNA package [126]. For algorithmic details on various MFE prediction algorithms, readers are referred to the comprehensive reviews [9, 127–131].

While MFE predictions have been well studied and widely used, they often suffer from low prediction accuracies when utilizing sequence information alone, especially for long RNAs [132]. One possible reason is that the assumption that RNA folds into the MFE structure may not always hold [47]. On the other hand, RNA can interact with other biomolecules in the cell, stabilizing specific non-MFE conformations. In addition, the existing set of NNTM parameters are not perfect although they have been improved over the years. The free energy of some structure motifs, such as multi-branch loops, are still not well understood and are thus obtained using simplified models [118].

In addition to the MFE structure, many programs have the option to also report a set of suboptimal structures. This is also a computational solution to the imperfect situation mentioned above. Such information is valuable for many downstream analysis applications. For example, one could generate energy dot plots from optimal and suboptimal structures, which could then used to find frequent structure motifs [133].

6.2 Ensemble-based predictions

Prediction of suboptimal structures is complementary to the MFE structure. However, it is worth noting that suboptimal structures could be quite different than the corresponding MFE structure, even when the differences between their free energies are very small. Take the aspartic acid tRNA in yeast as an example (Fig. 3). The energy of the predicted MFE structure and its closest suboptimal structures differ by 0.1 (-28 vs. -27.9), but their sensitivity differ quite a lot (76.2% vs. 33.3%); see Section 6.4 for a formal definition of sensitivity. Furthermore, MFE predictions are very sensitive in the sense that a minor change in NNTM parameters or experimental conditions may lead to a switch between the MFE and suboptimal structures; see [134] for a discussion on ribosomal 30S subunit structure revealed in [135].

A natural extension of suboptimal structures is to consider all possible structures. This can be accomplished by computing a partition function, which models the contribution of all structures weighted by their Boltzmann probabilities [62, 136, 137]. For a given sequence, the partition function, Q, can be calculated as

$$Q = \sum_k e^{-\Delta G_k/RT},$$

where ΔG_k is the free energy of the k-th possible secondary structure, R is the gas constant and T is temperature. Furthermore, the probability of a base pair formed by nucleotide i and j can be calculated as

$$p_{ij} = \frac{\sum_{k_{ij}} e^{-\Delta G_{k_{ij}}/RT}}{Q},$$

where the sum occurs over structures that include base pair i-j.

Several algorithms that utilize the statistical nature of partition function calculations have been proposed for structure predictions. The Sfold program first samples a user-specified number of structures from the Boltzmann ensemble. It then computes a centroid structure based on base-pair distances between structures [138]. Another type of approach predicts secondary structure by maximizing the expected base-pair accuracy (MEA). Briefly, MEA looks for the structure that maximizes the sum of base-paired and single-stranded probabilities. This objective function matches well with the observation that base pairs with high pairing probabilities are more likely to be present in the known reference structure [136]. MEA was first proposed in CONTRAfold, which learns probabilistic parameters from a set of known structures based on conditional log-linear models [139]. Later, Lu *et al.* implemented another MEA approach that directly depends on base pairing probabilities derived from partition function of the given sequence [140]. A relevant work that considers pseudo-expected accuracy is reported in [141].

It is most common for prediction algorithms to report a single optimal structure. However, some RNAs are known to have multiple functional structures in living cells. The function of these RNAs not only depends on these conformations but also on their ability to inter-convert [142]. For example, riboswitches can adopt different structures upon binding by a small molecule in order to control gene expression. [5, 143]. Similar to riboswitch, a single nucleotide variant (SNV) can alter the structure of riboSNitchs [88], which is critical to understand the effect of polymorphic loci, notably in humans. For such studies, analysis of structure ensembles are the natural choice compared to MFE prediction.

6.3 Comparative sequence analysis

The structures of many RNAs, such as tRNAs and rRNAs, are usually highly conserved, despite possible discrepancies in their primary sequences [144]. Comparative sequence analysis aims to find a consensus structure from a set of homologous sequences [7, 9, 145]. This approach is highly accurate and has been widely used to study the structures of several RNAs, e.g., rRNAs [146]. Overall, three approaches currently exist to implement comparative analysis.

Align then fold aligns sequences first and then predicts the consensus structure [110, 147, 148]. Two of the widely used programs in this category are RNAalifold [149] and Pfold [150]. RNAalifold aims to find the minimum energy structure that are formed by a set of aligned sequences. It also supports the computation of partition function and the centroid structure, which is the structure with minimum base pair distance to other structures in the ensemble. Here, distance is defined based on base-pairing probabilities. Pfold uses the stochastic context-free grammar (SCFG) [151,152] to combine an evolutionary model of sequences with a probabilistic model for secondary structures.

Fold and align simultaneously aligns and folds input sequences [153–156]. This idea was first proposed by Sankoff [153], which basically is a dynamic programming algorithm. The Sankoff algorithm has a time complexity of $O(n^{3m})$ for m sequences with maximum length n, and thus is computationally expensive to apply to large inputs. By posing extra restrictions on the problem, several variations of the Sankoff algorithm with feasible complexity have been developed [155, 157–159].

Fold then align predicts a structure from each input sequence, followed by alignment of structures. This method is particularly useful in scenarios where input sequences are not sufficiently conserved for direct alignment. Representatives of this method are reported in [160, 161].

Although comparative sequence analysis is highly accurate, it has been successfully applied only to a limited number of RNAs with rich phylogenetic information available. This is because, analogous to many phylogenetic studies, high accuracy can only be achieved when input sequences are sufficiently divergent to contain enough co-variation information. At the same time, sequences need to be sufficiently similar in order to be aligned properly; otherwise it becomes unfeasible to find a good consensus [47].

6.4 Performance measures

The accuracy of a predicted structure can be measured by comparing it to the known reference structure, which is typically obtained through crystallography experiments or comparative sequence analysis [145]. Sensitivity and PPV are the two most commonly used metrics for this purpose. Sensitivity is the fraction of base pairs in the reference structure that are correctly predicted, while PPV is the fraction of correctly predicted base pairs in the predicted structure. Matthews correlation coefficient (MCC) is another widely used metric that combines sensitivity and PPV. Some studies approximate it using the geometric mean of sensitivity and PPV [145]. For partition function-based predictions, one can measure the reliability of a prediction by calculating ensemble diversity and positional entropy, as proposed by [48].

When comparing different prediction algorithms, studies often use a benchmark dataset with multiple RNAs and compare their average performances. It is pointed out in [134] that this simple metric is not informative enough, as it is heavily biased by performances of short RNAs. To resolve this issue, this study proposed to use the "sequence-length-weighted average" (SLW-average) to replace the simple average. Intuitively, the SLW-average takes sequence length into consideration when averaging the performances of multiple RNAs.

7 Data-directed secondary structure prediction

In this section, we review data-directed prediction methods. While most methods seek a single optimal structure, they differ in their interpretation of SP data and/or in how they integrate it with computation.

7.1 Pseudoenergy-based approaches

The idea of converting SHAPE data into a pseudoenergy was first proposed by Deigan *et al.* [82]. Serving as *ad hoc* energy modifications, pseudoenergies are incorporated into MFE predictions to find the structure that minimizes the sum of NNTM free energy and pseudoenergy. For a given reactivity α, its pseudoenergy is calculated using a linear-log formula, $m(1 + \alpha) + b$, where m and b are two parameters determined on a training set of RNAs with known reference structures using grid search. Note that the optimal values of m and b could differ quite a lot between different data sets [33, 162], as they depend on the statistical properties of the data as well as on its dynamic range. This method was first implemented in the RNAstructure package [125], and was recently included in the ViennaRNA package [48]. It is also part of data analysis pipeline for some genome-wide SP experiments [163].

Deigan *et al.*'s approach has been widely used by the community and proved to significantly improve predictions for several RNAs [28, 48, 164, 165]. For example, it has been included in RNAalifold program in the new version of the ViennaRNA package [48, 166], which predicts the MFE structure and centroid structure given a set of aligned sequences. As another example, this approach is also at the core of the experimental 3S technique for secondary structure determination of long non-coding RNAs [167]. 3S, also called shotgun SHAPE, is motivated by the observation that traditional thermodynamic-based prediction algorithms often have limited accuracy. It probes an entire RNA along with its shorter overlapping segments. By comparing reactivity profiles of short segments with that of the entire RNA, modular sub-domains are identified, whose structures are then predicted using Deigan *et al.*'s approach. However, it is worth noting that this linear-log model was not built based on biological assumptions but rather using heuristics [131,168]. Initially developed and optimized for SHAPE chemistry data, it is unknown how well this model fits other types of SP data. In fact, Deng *et al.* showed, using mock-probe simulations, that Deigan *et al.*'s approach can give relatively poor performance when input data deviate from its assumed model [134]. To alleviate this problem, several other methods have been developed. Most methods follow the "training and prediction" paradigm, where a model is first trained on SP data with known reference structures. The trained model is then used to direct structure prediction on new data. In an earlier work, pseudoenergies are derived from the log likelihood ratio of a nucleotide being paired versus unpaired given its reactivity [74]. Benchmarked on DMS data, this work uses two gamma distributions to model paired and unpaired likelihood separately.

Motivated by the log likelihood ratio in [74], the RME program converts reactivities into posterior probabilities before deriving pseudoenergies from them [104]. Pseudoenergies are then used to direct partition function calculation and to further obtain an MEA structure, in contrast to the MFE structure in [74,82]. Note that in RME, SP data are not only involved in the initial calculation of partition function, but also in the post-calibration of base pairing probabilities, both in the form of posterior probabilities.

Eddy pointed out that Deigan *et al.*'s model actually signifies a base-pairing likelihood ratio [52]. Furthermore, he proposed a principled and broadly applicable framework that directly derives from statistical modeling of SP data. Under the assumption that reactivities are only depedend on structural contexts (e.g., paired, unpaired, stacked, helix-end), the pseudoenergy of a reactivity for a given structural context can be derived from its likelihood. This framework has been implemented and extended in the RNAprob package for MFE prediction [134]. RNAprob investigates two different resolutions of structure context, with a low resolution distinguishing between paired and unpaired nucleotides while the higher resolution further dividing paired nucleotides into stacked and helix-end, resulting into three contexts. In RNAprob, pseudoenergies are applied once to each nucleotide, regardless of its structural context. In contrast, they are applied to every nearest-neighbor stack in [74,82,104]. Consequently, pseudoenergies are applied 0, 1 and 2 times for each unpaired, helix-end and stacked nucleotide respectively. Note that RNAprob is implemented within the programming infrastructure of RNAstructure package [125], while providing enhanced applicability.

Similar to RNAprob, RNAsc includes pseudoenergies for all nucleotides, featuring two structural contexts (paired and unpaired) [169]. Unlike the aforementioned likelihood- and posterior-based pseudoenergy derivation, RNAsc first converts each reactivity i into p_i, the probability of being unpaired. A pseudoenergy is then computed for each of the two structural contexts as $\beta|x_i - p_i|$, where β is a user-specified scaling factor, $x_i = 0$ and 1 for unpaired and paired nucleotides, respectively.

RNApbfold extends the idea of pseudoenergy into perturbations in the context of partition function, without explicitly converting SP data into *ad hoc* pseudoenergies [170]. Specifically, it aims to find a perturbation vector that minimizes the discrepancy between predictions and SP data. This perturbation vector applies only when SP data disagree with predictions based on the thermodynamic model.

7.2 Non-pseudoenergy-based approaches

While pseudoenergy-based approaches have attracted much attention in recent years, alternative data-directed prediction approaches have gained much progress. SeqFold adopts the "sample and select" strategy [171]. It first samples a set of structures from the whole structure ensemble of a given sequence, which are then clustered using Sfold [63]. One of the clusters is then selected based on the distance of each sampled structure to the input SP profile, from which a consensus structure is further computed. The accuracy of this approach is largely determined by its ability to sample the "correct" structure. However, as the number of possible structures is huge, there is no guarantee that the correct structure will be sampled. Ideas of sample and select can also be found in [65].

PPfold 3.0 extends the *Pfold* package [150] by combining phylogeny with SP data [172]. It uses 1) a stochastic context-free grammars (SCFGs) to model structures; 2) a phylogeny model to compute the likelihood of input alignments and 3) a probabilistic model to include SP data. In a more recent work, ProbFold proposes to combine SCFGs with probabilistic graphical models [173]. While SCFGs give prior knowledge over structures as in PPfold 3.0, the probabilistic graphical models account for sequence and SP data.

The above data-directed structure prediction methods all utilize SP data from a single experiment. The mutate-and-map (M2) strategy developed by the Das lab provides 2D SP data [174]. For a sequence of length N, M2 performs $N+1$ SP experiments: one for the wild type and others for each of the N point-mutated sequences. Basically, M2 is based on the assumption that mutation of a nucleotide may result in local or global structural changes, which in turn result in reactivity change. M2 data can be converted into Z-scores and then plug in to RNAstructure package as extra energy bonus for MFE structure prediction. Recently, M2 data have been used to predict multiple functional structures as well as their relative proportion in the REEFFIT program [142]

7.3 Information content of SP data

The addition of SP data to better predict RNA structure proved to be successful on a variety of RNAs. A natural question that arises is "Do all reactivities contribute equally to drive structure prediction?". This question was recently addressed in the context of SHAPE data [134]. Instead of evaluating the relative contribution (information content) of each single reactivity in a SHAPE profile, they are divided into five equally populated subsets (a.k.a quintiles). The information content of each quintile is then quantified using a combination of leave-one-in and leave-one-out analysis. In the leave-one-in analysis, only a selected quintile is used to direct structure prediction, whereas in the leave-one-out analysis, all quintiles except for the selected one are used. Benchmarked on a set of 23 RNAs, this study showed that the top 20% reactivities are the major driving force for structure prediction, followed by the lowest 20%. In contrast, middle-range reactivities are less informative and have marginal contribution to improving prediction. Furthermore, the study showed, by a thought experiment, that middle-range reactivities are key to further improving predictions (Fig. 4). Briefly, this experiment is done by inputting perfect information (0 and 1.6 for paired and unpaired nucleotides, respectively in [134]) to a given quintile, while leaving reactivities of other quintiles unchanged. Note that while it remains unknown if the conclusions reported above will hold for other types of SP data, these analytical methods will work for any SP datasets.

Understanding information content of SP data gives us some practical guideline on data-directed predictions. For example, one may choose to use selective reactivities that are informative and ignore reactivities that are ambiguous. In addition, this also facilitates new models with more discriminative power, which can possibly reduce the number of less informative reactivities and in turn improve structure prediction.

7.4 Open questions

Structure prediction has been greatly facilitated by the rapid development of SP technologies. Studies have shown that data-directed predictions often lead to better predictions. However, it is worth noting that the extent of improvement in prediction accuracy varies substantially among RNAs and appears to be sequence dependent. It sometimes can have minor or even negative effect on resulting predictions [134, 175]. On the other hand, regardless of the existence of various strategies to incorporate SP data, currently no method is universally better than the others [134]. As such, further improvement is desired and can be possibly approached from the following aspects: 1) The pseudoenergy-based methods give good performance in practice. We anticipate that better performances can be achieved with pseudoenergy derivation models that are more biological and statistical meaningful. 2) As in [171–173], pseudoenergies are not the only way to use SP data and it will be interesting to explore alternative strategies for modeling SP data. 3) Recent development of novel transcriptome wide methods to probe RNA structures experimentally presents us with massive data of unprecedented complexity and diversity. This data have the potential to lead to better structure prediction, while presenting challenge on how to integrate information from multiple SP data in current algorithms. An attempt in this direction is reported in [103]. Availability of probabilistic methods, such as RNAprob and ProbFold, will certainly help efforts in this direction.

8 Software infrastructure

The rapid development of SP has brought massive amounts of diverse data. As for many other sequencing-based studies, tools for data sharing and analysis are two major needs. Here, we review recent progress towards these aims.

8.1 Databases and visualization tools

Structure Surfer [176], RNAex [177] and FoldAtlas [178] are three recent tools for data sharing, covering DMS-seq [26], structure-seq [28], icSHAPE [40], PARS [55], ds/ssRNA-seq [179], etc. In addition, they all provide a set of useful inspection and visualization tools. Specifically, Structure Surfer allows to visually compare different data sets, while RNAex and FoldAtlas support visualization of predicted secondary structures. RNAex also supports annotated RNA editing, RNA modification and SNP sites in predicted structures. A recent tool, SEQualyzer, for SP data quality screening is reported in [105].

8.2 Data preprocessing

Data analysis usually entails five major steps: 1) *Data cleaning* removes adapters, PCR duplicates or other undesired sequences; 2) *Read mapping* maps reads to a reference set of transcripts; 3) *Count summarization* at nucleotide level; 4) *Reactivity calculation* ; and 5) *Data-directed secondary structure prediction*. Steps 2, 3 and 4 are routinely featured in all platforms, while steps 1 and 5 are supported by a subset of tools.

Most recent SP protocols are adjoined by specialized analysis pipelines. Spats processes reads from SHAPE-Seq experiments [33]. Reactivities are calculated using a maximum likelihood estimate model [60,80,180]. ShapeMapper and SuperFold are two separate analysis pipelines for SHAPE-MaP experiments [39]. ShapeMapper converts raw sequencing reads into mutational profiles, which are then used as input to SuperFold for secondary structure prediction. They also allow *de novo* identification of well-defined and stable structure regions. Other specialized pipelines include Mod-Seeker [35], MAPseeker [38] and icSHAPE [40].

Tools designed with broader applicability in mind include StructureFold [163], RSF [181] and PROBer [79]. Deployed as part of the Galaxy platform [182], StructureFold supports covertion of reads into reactivities and supports structure prediction, each of which is provided as a separate module. It implements the reactivity calculation method proposed in [28]. Another modular pipeline, RNA Structure Framework (RSF), supports similar funtionality as well as data cleaning. Additionally, it offers flexibility in choosing certain reactivity calculation methods [26,28] and normalization strategies (2-8%, 90% winsorizing and box plot). In contrast to the former two, PROBer is a closed-box solution that implements the statistical model-based approach of

Li *et al.* (see Section 3). However, it applies to a broader class of experiments, encompassing a number of recent techniques beyond SP, which share a common workflow, as follows. 1) Chemical modification of nucleotides encodes a signal of interest. 2) Signal is detected via RT termination. 3) cDNA products are sequenced and mapped to estimate modification intensities per nucleotide. Examples of biological signals that can be studied within this framework include protein-RNA interactions [183, 184], posttranscriptional RNA modifications [185–191] and sites of unique structural motifs such as RNA G-quadruplexes [42]. This unified view not only lends iteself to shared analysis tools but also allludes to plausible commonalities in comparative and integrative analysis. Methods that approach these emerging challenges from a broader perspective may then reach a wider research community and potentially exert greater impact.

9 Conclusion

We reviewed current practices and emerging questions in comparative and integrative analysis of SP data. However, there are other emerging applications that we have not touched upon, which are timely as they directly leverage the new wealth of information. For example, SHAPE-based alignment is shown to have comparable accuracy to traditional sequence-based alignment [166].The alignment can be further improved when combining sequence information with SHAPE data. In addition, SP data-directed partition function can be used to calculate Shannon entropy, which in turn is useful in discovering well-defined RNA structures [39]. These and additional timely applications are described in a recent review [53]. Another exciting direction is the emergence of a new class of RNA structure experiments, which identify long-range and inter-molecular base-pairing interactions [192–197]. Integrating this type of information with SP data and with structure prediction algorithms is likely to pose newer challenges and trigger dedicated methods development.

The advent of SP techniques has greatly expanded our capability to understand structures of various RNAs and their functional roles. Propelled by these advances, we are standing in the era of large-scale data with increasing diversity and complexity, which in turn poses great informatics challenges in data interpretation and analysis. To maximize the potential of these data, we need to develop methods for accurate data interpretation leveraging intrinsic statistical properties of an SP protocol. Additionally, we need to better suit the methodology for comparative analysis to discover biological patterns of interest and the methodology for characterizing SP information content to more suitably feed data into structure prediction algorithms.

Acknowledgements This work was supported by the National Institutes of Health (NIH) grant [HG006860]. We thank Chun Kit Kwok and Aviran lab members - Mirko Ledda, Sana Vaziri, Hua Li and Rob Gysel - for insightful comments during the preparation of this manuscript.

References

1. P. A. Sharp. The centrality of RNA. *Cell*, 136(4):577–580, 2009.
2. S. A. Mortimer, M. A. Kidwell, and J. A. Doudna. Insights into RNA structure and function from genome-wide studies. *Nature reviews Genetics*, 15(7):469–479, 2014.
3. L. He and G. J. Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, 5(7):522–531, 2004.
4. T. R. Mercer, M. E. Dinger, and J. S. Mattick. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*, 10(3):155–159, 2009.
5. E. J. Strobel, K. E. Watters, D. Loughrey, and J. B. Lucks. RNA systems biology: uniting functional discoveries and structural tools to understand global roles of RNAs. *Current opinion in biotechnology*, 39:182–191, 2016.
6. H. M. Al-Hashimi. Structural biology: Aerial view of the HIV genome. *Nature*, 460(7256):696–698, 2009.
7. R. R. Gutell, J. C. Lee, and J. J. Cannone. The accuracy of ribosomal RNA comparative structure models. *Current opinion in structural biology*, 12(3):301–310, 2002.
8. I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.
9. D. H. Mathews, W. N. Moss, and D. H. Turner. Folding and finding RNA secondary structure. *Cold Spring Harbor perspectives in biology*, 2(12):a003665, 2010.
10. C. Ehresmann, F. Baudin, M. Mougel, P. Romby, J.-P. Ebel, and B. Ehresmann. Probing the structure of RNAs in solution. *Nucleic acids research*, 15(22):9109–9128, 1987.
11. K. M. Weeks. Advances in RNA structure analysis by chemical probing. *Current opinion in structural biology*, 20(3):295–304, 2010.
12. P. Tijerina, S. Mohr, and R. Russell. DMS footprinting of structured RNAs and RNA–protein complexes. *Nature protocols*, 2(10):2608–2623, 2007.
13. D. A. Brow and H. F. Noller. Protection of ribosomal RNA from kethoxal in polyribosomes: Implication of specific sites in ribosome function. *Journal of molecular biology*, 163(1):27–46, 1983.
14. T. D. Tullius and J. A. Greenbaum. Mapping nucleic acid structure by hydroxyl radical cleavage. *Current opinion in chemical biology*, 9(2):127–134, 2005.
15. B. Singer. All oxygens in nucleic acids react with carcinogenic ethylating agents. *Nature*, 264(5584):333–339, 1976.
16. J. J. Fritz, A. Lewin, W. Hauswirth, A. Agarwal, M. Grant, and L. Shaw. Development of hammerhead ribozymes to modulate endogenous gene expression for functional studies. *Methods*, 28(2):276–285, 2002.
17. M. Lindell, P. Romby, and E. G. H. Wagner. Lead (II) as a probe for investigating RNA structure in vivo. *RNA*, 8(04):534–541, 2002.
18. M. Lindell, M. Brännvall, E. G. H. WAGNER, and L. A. Kirsebom. Lead (II) cleavage analysis of RNase P RNA in vivo. *Rna*, 11(9):1348–1354, 2005.
19. G. Knapp. [16] Enzymatic approaches to probing of RNA secondary and tertiary structure. *Methods in enzymology*, 180:192–212, 1989.
20. K. A. Wilkinson, E. J. Merino, and K. M. Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature protocols*, 1(3):1610–1616, 2006.
21. M. Zubradt, P. Gupta, S. Persad, A. M. Lambowitz, J. S. Weissman, and S. Rouskin. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nature Methods*, 2016.
22. M. J. Smola, G. M. Rice, S. Busan, N. A. Siegfried, and K. M. Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nature protocols*, 10(11):1643–1669, 2015.
23. K. E. Watters, M. Y. Angela, E. J. Strobel, A. H. Settle, and J. B. Lucks. Characterizing RNA structures in vitro and in vivo with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Methods*, 2016.

24. L. D. Poulsen, L. J. Kielpinski, S. R. Salama, A. Krogh, and J. Vinther. SHAPE Selection (SHAPES) enrich for RNA structure signal in SHAPE sequencing-based probing data. *RNA*, 21(5):1042–1052, 2015.
25. R. D. Hector, E. Burlacu, S. Aitken, T. Le Bihan, M. Tuijtel, A. Zaplatina, A. G. Cook, and S. Granneman. Snapshots of pre-rRNA structural flexibility reveal eukaryotic 40S assembly dynamics at nucleotide resolution. *Nucleic acids research*, page gku815, 2014.
26. S. Rouskin, M. Zubradt, S. Washietl, M. Kellis, and J. S. Weissman. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, 505(7485):701–705, 2014.
27. C. K. Kwok, Y. Ding, Y. Tang, S. M. Assmann, and P. C. Bevilacqua. Determination of in vivo RNA structure in low-abundance transcripts. *Nature communications*, 4, 2013.
28. Y. Ding, Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua, and S. M. Assmann. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, 505(7485):696–700, 2014.
29. Y. Ding, C. K. Kwok, Y. Tang, P. C. Bevilacqua, and S. M. Assmann. Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. *Nature protocols*, 10(7):1050–1066, 2015.
30. M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107, 2010.
31. J. G. Underwood, A. V. Uzilov, S. Katzman, C. S. Onodera, J. E. Mainzer, D. H. Mathews, T. M. Lowe, S. R. Salama, and D. Haussler. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Meth*, 7(12):995–1001, 2010.
32. J. B. Lucks, S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences*, 108(27):11063–11068, 2011.
33. D. Loughrey, K. E. Watters, A. H. Settle, and J. B. Lucks. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic acids research*, 42(21):000–000, 2014.
34. Y. Wan, K. Qu, Z. Ouyang, and H. Y. Chang. Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. *Nature protocols*, 8(5):849–869, 2013.
35. J. Talkish, G. May, Y. Lin, J. L. Woolford, and C. J. McManus. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA*, 20(5):713–720, 2014.
36. D. Incarnato, F. Neri, F. Anselmi, and S. Oliviero. Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. *Genome biology*, 15(10):1, 2014.
37. L. J. Kielpinski and J. Vinther. Massive parallel-sequencing-based hydroxyl radical probing of RNA accessibility. *Nucleic acids research*, page gku167, 2014.
38. M. G. Seetin, W. Kladwang, J. P. Bida, and R. Das. Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol. *RNA Folding: Methods and Protocols*, pages 95–117, 2014.
39. N. A. Siegfried, S. Busan, G. M. Rice, J. A. Nelson, and K. M. Weeks. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature methods*, 11(9):959–965, 2014.
40. R. C. Spitale, R. A. Flynn, Q. C. Zhang, P. Crisalli, B. Lee, J.-W. Jung, H. Y. Kuchelmeister, P. J. Batista, E. A. Torre, E. T. Kool, et al. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, 2015.
41. C. K. Kwok, A. B. Sahakyan, and S. Balasubramanian. Structural Analysis using SHALiPE to Reveal RNA G-Quadruplex Formation in Human Precursor MicroRNA. *Angewandte Chemie International Edition*, 55(31):8958–8961, 2016.
42. C. K. Kwok, G. Marsico, A. B. Sahakyan, V. S. Chambers, and S. Balasubramanian. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nature Methods*, 2016.

43. C. K. Kwok, Y. Tang, S. M. Assmann, and P. C. Bevilacqua. The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends in biochemical sciences*, 40(4):221–232, 2015.
44. Z. Lu and H. Y. Chang. Decoding the RNA structurome. *Current opinion in structural biology*, 36:142–148, 2016.
45. C. K. Kwok. Dawn of the in vivo RNA structurome and interactome. *Biochemical Society Transactions*, 44(5):1395–1410, 2016.
46. M. Kubota, D. Chan, and R. C. Spitale. RNA structure: Merging chemistry and genomics for a holistic perspective. *BioEssays*, 37(10):1129–1138, 2015.
47. J. T. Low and K. M. Weeks. SHAPE-directed RNA secondary structure prediction. *Methods*, 52(2):150–158, 2010.
48. R. Lorenz, D. Luntzer, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger. SHAPE directed RNA folding. *Bioinformatics*, page btv523, 2015.
49. E. J. Merino, K. A. Wilkinson, J. L. Coughlan, and K. M. Weeks. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *Journal of the American Chemical Society*, 127(12):4223–4231, 2005.
50. R. Lavery and A. Pullman. A new theoretical index of biochemical reactivity combining steric and electrostatic factors: An application to yeast tRNAPhe. *Biophysical chemistry*, 19(2):171–181, 1984.
51. J. L. McGinnis, J. A. Dunkle, J. H. Cate, and K. M. Weeks. The mechanisms of RNA SHAPE chemistry. *Journal of the American Chemical Society*, 134(15):6617–6624, 2012.
52. S. R. Eddy. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annual Review of Biophysics*, 43:433–456, 2014.
53. K. M. Kutchko and A. Laederach. Transcending the prediction paradigm: novel applications of SHAPE to RNA function and evolution. *Wiley Interdisciplinary Reviews: RNA*, 2016.
54. S. Aviran and L. Pachter. Rational experiment design for sequencing-based RNA structure mapping. *RNA*, 20(12):1864–1877, 2014.
55. Y. Wan, K. Qu, Q. C. Zhang, R. A. Flynn, O. Manor, Z. Ouyang, J. Zhang, R. C. Spitale, M. P. Snyder, E. Segal, et al. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, 505(7485):706–709, 2014.
56. J. Ritz, J. S. Martin, and A. Laederach. Evaluating our ability to predict the structural disruption of RNA by SNPs. *BMC genomics*, 13(4):1, 2012.
57. K. E. Watters, T. R. Abbott, and J. B. Lucks. Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq. *Nucleic acids research*, 44(2):e12–e12, 2016.
58. Y. Bai, A. Tambe, K. Zhou, and J. A. Doudna. RNA-guided assembly of Rev-RRE nuclear export complexes. *Elife*, 3:e03656, 2014.
59. K. Choudhary, N. P. Shih, F. Deng, M. Ledda, B. Li, and S. Aviran. Metrics for rapid quality control in RNA structure probing experiments. *Bioinformatics*, 32(23):3575–3583, 2016.
60. S. Aviran, J. B. Lucks, and L. Pachter. RNA structure characterization from chemical mapping experiments. *The 49th Annual Allerton Conference on Communication, Control, and Computing*, pages 1743–1750, 2011.
61. Y. Wan, M. Kertesz, R. C. Spitale, E. Segal, and H. Y. Chang. Understanding the transcriptome through RNA structure. *Nature Reviews Genetics*, 12(9):641–655, 2011.
62. J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.
63. Y. Ding and C. E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic acids research*, 31(24):7280–7301, 2003.
64. E. Rogers and C. Heitsch. New insights from cluster analysis methods for RNA secondary structure prediction. *Wiley Interdisciplinary Reviews: RNA*, 7(3):278–294, 2016.
65. S. Quarrier, J. S. Martin, L. Davis-Neulander, A. Beauregard, and A. Laederach. Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *RNA*, 16(6):1108–1117, 2010.

66. J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11(1):1, 2010.
67. J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.
68. J. U. Guo and D. P. Bartel. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science*, 353(6306):aaf5371, 2016.
69. M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.
70. S. Anders and W. Huber. Differential expression of RNA-Seq data at the gene level–the DESeq package. *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)*, 2012.
71. C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2):1, 2014.
72. K. A. Leamy, S. M. Assmann, D. H. Mathews, and P. C. Bevilacqua. Bridging the gap between in vitro and in vivo RNA folding. *Quarterly Reviews of Biophysics*, 49, 2016.
73. X. Hu, Y. Wu, Z. J. Lu, and K. Y. Yip. Analysis of sequencing data for probing RNA secondary structures and protein–RNA binding in studying posttranscriptional regulations. *Briefings in bioinformatics*, page bbv106, 2015.
74. P. Cordero, W. Kladwang, C. C. VanLang, and R. Das. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry*, 51(36):7037–7039, 2012.
75. B. Lee, R. A. Flynn, A. Kadina, J. K. Guo, E. T. Kool, and H. Y. Chang. Comparison of SHAPE Reagents for Mapping RNA Structures Inside Living Cells. *RNA*, 2016.
76. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008.
77. K. Sorefan, H. Pais, A. E. Hall, A. Kozomara, S. Griffiths-Jones, V. Moulton, and T. Dalmay. Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence*, 3(1):1, 2012.
78. A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology*, 12(3):1, 2011.
79. B. Li, A. Tambe, S. Aviran, and L. Pachter. Prober: A general toolkit for analyzing sequencing-based'toeprinting'assays. *bioRxiv*, page 063107, 2016.
80. S. Aviran, C. Trapnell, J. B. Lucks, S. A. Mortimer, S. Luo, G. P. Schroth, J. A. Doudna, A. P. Arkin, and L. Pachter. Modeling and automation of sequencing-based characterization of RNA structure. *Proceedings of the National Academy of Sciences*, 108(27):11069–11074, 2011.
81. A. Selega, C. Sirocchi, I. Iosub, S. Granneman, and G. Sanguinetti. Robust statistical modeling improves sensitivity of high-throughput RNA structure probing experiments. *Nature Methods*, 2016.
82. K. E. Deigan, T. W. Li, D. H. Mathews, and K. M. Weeks. Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences*, 106(1):97–102, 2009.
83. M. F. Sloma and D. H. Mathews. Chapter four-Improving RNA secondary structure prediction with structure mapping data. *Methods in enzymology*, 553:91–114, 2015.
84. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.
85. B. Li and C. N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1):1, 2011.
86. M. J. Smola, J. M. Calabrese, and K. M. Weeks. Detection of RNA–protein interactions in living cells with SHAPE. *Biochemistry*, 54(46):6867–6875, 2015.
87. M. J. Smola, T. W. Christy, K. Inoue, C. O. Nicholson, M. Friedersdorf, J. D. Keene, D. M. Lee, J. M. Calabrese, and K. M. Weeks. SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proceedings of the National Academy of Sciences*, page 201600008, 2016.

88. A. C. Solem, M. Halvorsen, S. B. Ramos, and A. Laederach. The potential of the riboSNitch in personalized medicine. *Wiley Interdisciplinary Reviews: RNA*, 6(5):517–532, 2015.
89. Y. Wan, K. Qu, Z. Ouyang, M. Kertesz, J. Li, R. Tibshirani, D. L. Makino, R. C. Nutter, E. Segal, and H. Y. Chang. Genome-wide measurement of RNA folding energies. *Molecular cell*, 48(2):169–181, 2012.
90. F. Righetti, A. M. Nuss, C. Twittenhoff, S. Beele, K. Urban, S. Will, S. H. Bernhart, P. F. Stadler, P. Dersch, and F. Narberhaus. Temperature-responsive in vitro RNA structurome of Yersinia pseudotuberculosis. *Proceedings of the National Academy of Sciences*, 113(26):7237–7242, 2016.
91. M. Corley, A. Solem, K. Qu, H. Y. Chang, and A. Laederach. Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. *Nucleic acids research*, page gkv010, 2015.
92. M. B. Abdullah. On a robust correlation coefficient. *The Statistician*, pages 455–460, 1990.
93. L. D. Goodwin and N. L. Leech. Understanding correlation: Factors that affect the size of r. *The Journal of Experimental Education*, 74(3):249–266, 2006.
94. R. Müller and P. Büttner. A critical discussion of intraclass correlation coefficients. *Statistics in medicine*, 13(23-24):2465–2476, 1994.
95. J. L. Gastwirth. The estimation of the Lorenz curve and Gini index. *The Review of Economics and Statistics*, pages 306–316, 1972.
96. S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic acids research*, 22(11):2079–2088, 1994.
97. J.-H. Zhang, T. D. Chung, and K. R. Oldenburg. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *Journal of biomolecular screening*, 4(2):67–73, 1999.
98. E. Pollom, K. K. Dang, E. L. Potter, R. J. Gorelick, C. L. Burch, K. M. Weeks, and R. Swanstrom. Comparison of SIV and HIV-1 genomic RNA structures reveals impact of sequence evolution on conserved and non-conserved structural motifs. *PLoS Pathog*, 9(4):e1003294, 2013.
99. F. A. Cowell and M.-P. Victoria-Feser. Robustness properties of inequality measures. *Econometrica: Journal of the Econometric Society*, pages 77–101, 1996.
100. R. Liang, E. Kierzek, R. Kierzek, and D. H. Turner. Comparisons between chemical mapping and binding to isoenergetic oligonucleotide microarrays reveal unexpected patterns of binding to the Bacillus subtilis RNase P RNA specificity domain. *Biochemistry*, 49(37):8155–8168, 2010.
101. E. J. Hawkes, S. P. Hennelly, I. V. Novikova, J. A. Irwin, C. Dean, and K. Y. Sanbonmatsu. COOLAIR Antisense RNAs Form Evolutionarily Conserved Elaborate Secondary Structures. *Cell Reports*, 16(12):3087–3096, 2016.
102. Z. Xue, S. Hennelly, B. Doyle, A. A. Gulati, I. V. Novikova, K. Y. Sanbonmatsu, and L. A. Boyer. A G-Rich Motif in the lncRNA Braveheart Interacts with a Zinc-Finger Transcription Factor to Specify the Cardiovascular Lineage. *Molecular Cell*, 2016.
103. G. M. Rice, C. W. Leonard, and K. M. Weeks. RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *Rna*, 20(6):846–854, 2014.
104. Y. Wu, B. Shi, X. Ding, T. Liu, X. Hu, K. Y. Yip, Z. R. Yang, D. H. Mathews, and Z. J. Lu. Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic acids research*, 43(15):7247–7259, 2015.
105. K. Choudhary, L. Ruan, F. Deng, N. P. Shih, and S. Aviran. SEQualyzer: Interactive tool for quality control and exploratory analysis of high-throughput RNA structural profiling data (in press). *Bioinformatics*, page btw627, 2016.
106. K. Rother, M. Rother, P. Skiba, and J. M. Bujnicki. Automated modeling of RNA 3D structure. *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, pages 395–415, 2014.
107. J. E. Tabaska, R. B. Cary, H. N. Gabow, and G. D. Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14(8):691–699, 1998.

108. E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of molecular biology*, 285(5):2053–2068, 1999.
109. R. B. Lyngsø and C. N. Pedersen. RNA pseudoknot prediction in energy-based models. *Journal of computational biology*, 7(3-4):409–427, 2000.
110. J. Ruan, G. D. Stormo, and W. Zhang. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20(1):58–66, 2004.
111. J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC bioinformatics*, 5(1):1, 2004.
112. J. Ren, B. Rastegari, A. Condon, and H. H. Hoos. HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *Rna*, 11(10):1494–1504, 2005.
113. S. Cao and S.-J. Chen. Predicting RNA pseudoknot folding thermodynamics. *Nucleic acids research*, 34(9):2634–2652, 2006.
114. J. Reeder, P. Steffen, and R. Giegerich. pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic acids research*, 35(suppl 2):W320–W324, 2007.
115. K. Sato, Y. Kato, M. Hamada, T. Akutsu, and K. Asai. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, 27(13):i85–i93, 2011.
116. M. Andronescu, A. Condon, D. H. Turner, and D. H. Mathews. The determination of RNA folding nearest neighbor parameters. *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, pages 45–70, 2014.
117. T. Xia, J. SantaLucia Jr, M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner. Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson- Crick Base Pairs. *Biochemistry*, 37(42):14719–14735, 1998.
118. D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of molecular biology*, 288(5):911–940, 1999.
119. R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied mathematics*, 35(1):68–82, 1978.
120. M. S. Waterman and T. F. Smith. RNA secondary structure: A complete mathematical analysis. *Mathematical Biosciences*, 42(3-4):257–266, 1978.
121. R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences*, 77(11):6309–6313, 1980.
122. M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148, 1981.
123. M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bulletin of mathematical biology*, 46(4):591–621, 1984.
124. N. R. Markham and M. Zuker. UNAFold. *Bioinformatics: Structure, Function and Applications*, pages 3–31, 2008.
125. J. S. Reuter and D. H. Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*, 11(1):129, 2010.
126. R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):1, 2011.
127. S. R. Eddy. How do RNA folding algorithms work? *Nature biotechnology*, 22(11):1457–1458, 2004.
128. D. H. Mathews and D. H. Turner. Prediction of RNA secondary structure by free energy minimization. *Current opinion in structural biology*, 16(3):270–278, 2006.
129. B. A. Shapiro, Y. G. Yingling, W. Kasprzak, and E. Bindewald. Bridging the gap in RNA structure prediction. *Current opinion in structural biology*, 17(2):157–165, 2007.
130. Y. Bai, X. Dai, A. Harrison, C. Johnston, and M. Chen. Toward a next-generation atlas of RNA secondary structure. *Briefings in bioinformatics*, page bbv026, 2015.
131. P. Ge and S. Zhang. Computational analysis of RNA structures with chemical probing data. *Methods*, 79:60–66, 2015.

132. K. J. Doshi, J. J. Cannone, C. W. Cobaugh, and R. R. Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC bioinformatics*, 5(1):1, 2004.
133. M. Zuker et al. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52, 1989.
134. F. Deng, M. Ledda, S. Vaziri, and S. Aviran. Data-directed RNA secondary structure prediction using probabilistic modeling. *RNA*, pages 1109–1119, 2016.
135. J. L. McGinnis, Q. Liu, C. A. Lavender, A. Devaraj, S. P. McClory, K. Fredrick, and K. M. Weeks. In-cell SHAPE reveals that free 30S ribosome subunits are in the inactive state. *Proceedings of the National Academy of Sciences*, 112(8):2425–2430, 2015.
136. D. H. Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *Rna*, 10(8):1178–1190, 2004.
137. S. H. Bernhart, I. L. Hofacker, and P. F. Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–615, 2006.
138. Y. Ding, C. Y. Chan, and C. E. Lawrence. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *Rna*, 11(8):1157–1166, 2005.
139. C. B. Do, D. A. Woods, and S. Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, 2006.
140. Z. J. Lu, J. W. Gloor, and D. H. Mathews. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *Rna*, 15(10):1805–1813, 2009.
141. M. Hamada, K. Sato, and K. Asai. Prediction of RNA secondary structure by maximizing pseudo-expected accuracy. *BMC bioinformatics*, 11(1):586, 2010.
142. P. Cordero and R. Das. Rich RNA Structure Landscapes Revealed by Mutate-and-Map Analysis. *PLoS Comput Biol*, 11(11):e1004473, 2015.
143. R. R. Breaker. Riboswitches and the RNA world. *Cold Spring Harbor perspectives in biology*, 4(2):a003566, 2012.
144. J. Parsch, J. M. Braverman, and W. Stephan. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics*, 154(2):909–921, 2000.
145. P. P. Gardner and R. Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC bioinformatics*, 5(1):140, 2004.
146. J. J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Müller, et al. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC bioinformatics*, 3(1):1, 2002.
147. L. Rupert, G. Stefan, and S. Gerhard. Construct a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic acids research*, 27(21):4208–4217, 1999.
148. I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *Journal of molecular biology*, 319(5):1059–1066, 2002.
149. S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC bioinformatics*, 9(1):1, 2008.
150. B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic acids research*, 31(13):3423–3428, 2003.
151. Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjölander, R. C. Underwood, and D. Haussler. Stochastic context-free grammers for tRNA modeling. *Nucleic acids research*, 22(23):5112–5120, 1994.
152. B. Knudsen and J. Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6):446–454, 1999.
153. D. Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825, 1985.
154. J. H. Havgaard, R. B. Lyngsø, G. D. Stormo, and J. Gorodkin. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, 21(9):1815–1824, 2005.
155. D. H. Mathews and D. H. Turner. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *Journal of molecular biology*, 317(2):191–203, 2002.

156. A. O. Harmanci, G. Sharma, and D. H. Mathews. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC bioinformatics*, 8(1):1, 2007.
157. J. Gorodkin, L. J. Heyer, and G. D. Stormo. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic acids research*, 25(18):3724–3732, 1997.
158. O. Perriquet, H. Touzet, and M. Dauchet. Finding the common structure shared by two homologous RNAs. *Bioinformatics*, 19(1):108–116, 2003.
159. I. L. Hofacker, S. H. Bernhart, and P. F. Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–2227, 2004.
160. M. Hochsmann, T. Toller, R. Giegerich, and S. Kurtz. Local similarity in RNA secondary structures. In *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE*, pages 159–168. IEEE, 2003.
161. S. Siebert and R. Backofen. MARNA: A Server for Multiple Alignment of RNAs. In *In Proceedings of the German Conference on Bioinformatics*, pages 135–140, 2003.
162. C. E. Hajdin, S. Bellaousov, W. Huggins, C. W. Leonard, D. H. Mathews, and K. M. Weeks. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proceedings of the National Academy of Sciences*, 110(14):5498–5503, 2013.
163. Y. Tang, E. Bouvier, C. K. Kwok, Y. Ding, A. Nekrutenko, P. C. Bevilacqua, and S. M. Assmann. StructureFold: genome-wide RNA secondary structure mapping and reconstruction in vivo. *Bioinformatics*, page btv213, 2015.
164. J. M. Watts, K. K. Dang, R. J. Gorelick, C. W. Leonard, J. W. Bess Jr, R. Swanstrom, C. L. Burch, and K. M. Weeks. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, 460(7256):711–716, 2009.
165. S. Montaseri, M. Ganjtabesh, and F. Zare-Mirakabad. Evolutionary Algorithm for RNA Secondary Structure Prediction Based on Simulated SHAPE Data. *PLOS ONE*, 11(11):e0166965, 2016.
166. C. A. Lavender, R. Lorenz, G. Zhang, R. Tamayo, I. L. Hofacker, and K. M. Weeks. Model-free RNA sequence and structure alignment informed by SHAPE probing reveals a conserved alternate secondary structure for 16S rRNA. *PLoS Comput Biol*, 11(5):e1004126, 2015.
167. I. V. Novikova, A. Dharap, S. P. Hennelly, and K. Y. Sanbonmatsu. 3S: shotgun secondary structure determination of long non-coding RNAs. *Methods*, 63(2):170–177, 2013.
168. R. Lorenz, M. T. Wolfinger, A. Tanzer, and I. L. Hofacker. Predicting RNA secondary structures from sequence and probing data. *Methods*, 2016.
169. K. Zarringhalam, M. M. Meyer, I. Dotu, J. H. Chuang, and P. Clote. Integrating chemical footprinting data into RNA secondary structure prediction. *PLoS One*, 7(10):e45160, 2012.
170. S. Washietl, I. L. Hofacker, P. F. Stadler, and M. Kellis. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic acids research*, 40(10):4261–4272, 2012.
171. Z. Ouyang, M. P. Snyder, and H. Y. Chang. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome research*, 23(2):377–387, 2013.
172. Z. Sükösd, B. Knudsen, J. Kjems, and C. N. Pedersen. PPfold 3.0: fast RNA secondary structure prediction using phylogeny and auxiliary data. *Bioinformatics*, 28(20):2691–2692, 2012.
173. S. Sahoo, J. S. Pedersen, et al. ProbFold: A probabilistic method for integration of probing data in RNA secondary structure prediction. *Bioinformatics*, page btw175, 2016.
174. W. Kladwang, C. C. VanLang, P. Cordero, and R. Das. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nature chemistry*, 3(12):954–962, 2011.
175. Z. Sükösd, M. S. Swenson, J. Kjems, and C. E. Heitsch. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic acids research*, 41(5):2807–2816, 2013.

176. N. D. Berkowitz, I. M. Silverman, D. M. Childress, H. Kazan, L.-S. Wang, and B. D. Gregory. A comprehensive database of high-throughput sequencing-based RNA secondary structure probing data (Structure Surfer). *BMC bioinformatics*, 17(1):1, 2016.
177. Y. Wu, R. Qu, Y. Huang, B. Shi, M. Liu, Y. Li, and Z. J. Lu. RNAex: an RNA secondary structure prediction server enhanced by high-throughput structure-probing data. *Nucleic acids research*, page gkw362, 2016.
178. M. Norris, J. Cheema, C. K. Kwok, M. Hartley, R. J. Morris, S. Aviran, and Y. Ding. FoldAtlas: a repository for genome-wide RNA structure probing data. *Bioinformatics*, page In print, 2016.
179. F. Li, Q. Zheng, L. E. Vandivier, M. R. Willmann, Y. Chen, and B. D. Gregory. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. *The Plant Cell*, 24(11):4346–4359, 2012.
180. S. A. Mortimer, C. Trapnell, S. Aviran, L. Pachter, and J. B. Lucks. SHAPE-Seq: High-Throughput RNA Structure Analysis. *Current protocols in chemical biology*, pages 275–297, 2012.
181. D. Incarnato, F. Neri, F. Anselmi, and S. Oliviero. RNA structure framework: automated transcriptome-wide reconstruction of RNA secondary structures from high-throughput structure probing data. *Bioinformatics*, page btv571, 2015.
182. J. Goecks, A. Nekrutenko, and J. Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):1, 2010.
183. J. König, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, and J. Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, 17(7):909–915, 2010.
184. E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature methods*, 13(6):508–514, 2016.
185. J. E. Squires, H. R. Patel, M. Nousch, T. Sibbritt, D. T. Humphreys, B. J. Parker, C. M. Suter, and T. Preiss. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic acids research*, page gks144, 2012.
186. D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, 485(7397):201–206, 2012.
187. K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey. Comprehensive analysis of mRNA methylation reveals enrichment in 3 UTRs and near stop codons. *Cell*, 149(7):1635–1646, 2012.
188. S. Edelheit, S. Schwartz, M. R. Mumbach, O. Wurtzel, and R. Sorek. Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m 5 C within archaeal mRNAs. *PLoS Genet*, 9(6):e1003602, 2013.
189. P. J. Batista, B. Molinie, J. Wang, K. Qu, J. Zhang, L. Li, D. M. Bouley, E. Lujan, B. Haddad, K. Daneshvar, et al. m 6 A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell stem cell*, 15(6):707–719, 2014.
190. T. M. Carlile, M. F. Rojas-Duran, B. Zinshteyn, H. Shin, K. M. Bartoli, and W. V. Gilbert. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*, 2014.
191. D. Incarnato, F. Anselmi, E. Morandi, F. Neri, M. Maldotti, S. Rapelli, C. Parlato, G. Basile, and S. Oliviero. High-throughput single-base resolution mapping of RNA 2-O-methylated residues. *Nucleic Acids Research*, page gkw810, 2016.
192. G. Kudla, S. Granneman, D. Hahn, J. D. Beggs, and D. Tollervey. Cross-linking, ligation, and sequencing of hybrids reveals RNA–RNA interactions in yeast. *Proceedings of the National Academy of Sciences*, 108(24):10010–10015, 2011.
193. V. Ramani, R. Qiu, and J. Shendure. High-throughput determination of RNA structure by proximity ligation. *Nature biotechnology*, 33(9):980–984, 2015.
194. Y. Sugimoto, A. Vigilante, E. Darbo, A. Zirra, C. Militti, A. DAmbrogio, N. M. Luscombe, and J. Ule. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. *Nature*, 519(7544):491–494, 2015.

195. E. Sharma, T. Sterne-Weiler, D. OHanlon, and B. J. Blencowe. Global Mapping of Human RNA-RNA Interactions. *Molecular cell*, 62(4):618–626, 2016.
196. Z. Lu, Q. C. Zhang, B. Lee, R. A. Flynn, M. A. Smith, J. T. Robinson, C. Davidovich, A. R. Gooding, K. J. Goodrich, J. S. Mattick, et al. RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell*, 165(5):1267–1279, 2016.
197. J. G. A. Aw, Y. Shen, A. Wilm, M. Sun, X. N. Lim, K.-L. Boon, S. Tapsin, Y.-S. Chan, C.-P. Tan, A. Y. Sim, et al. In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Molecular cell*, 62(4):603–617, 2016.
198. K. Darty, A. Denise, and Y. Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974–5, 2009.

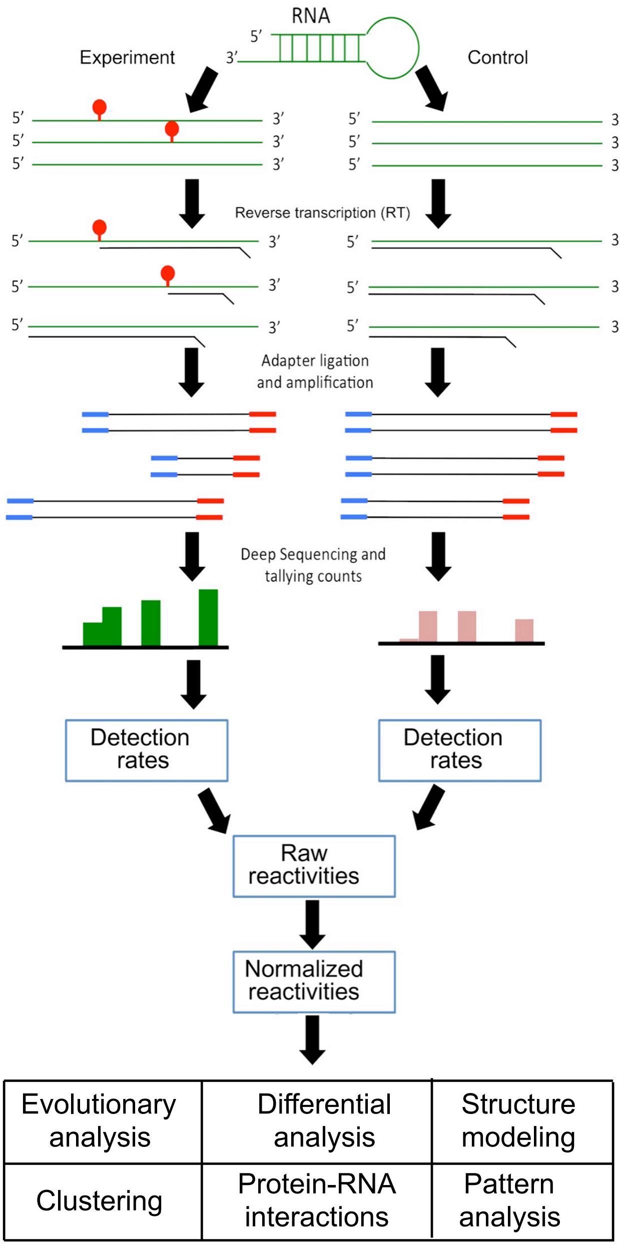

Fig. 1 Overview of structure-profiling experiments. RNA sample of interest (at the top) is probed with structure-sensitive reagent, which introduces a modification (red pins) preferentially at unpaired nucleotides. Degree of modification is read via reverse transcription and sequencing. Next, the readouts are mapped to reference sequence and normalized reactivities are calculated from counts summary of mapped reads. Reactivity profiles of probed RNAs are used for diverse downstream applications, some of which are listed.

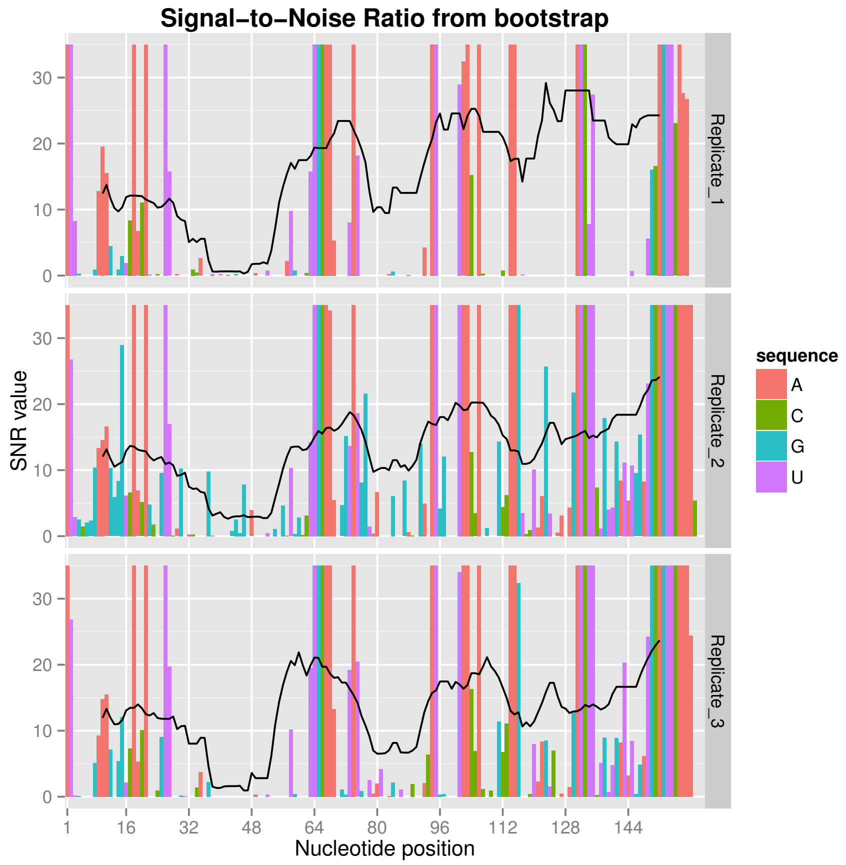

Fig. 2 **Quality screening with SEQualyzer.** Bars represent per-residue SNR and black lines represent rolling mean of per-residue SNR for windows of 20 nt. SEQualyzer estimates SNR via bootstrapping as described by Choudhary *et al.* [59] Examination of quality profiles reveals that signal quality is good for entire RNA except a short region from nucleotides 35-53 where it is poor in all replicates. For illustration purpose, we used data for P4-P6 domain of *Tetrahymena* group I intron ribozyme from Loughrey *et al.* [33]

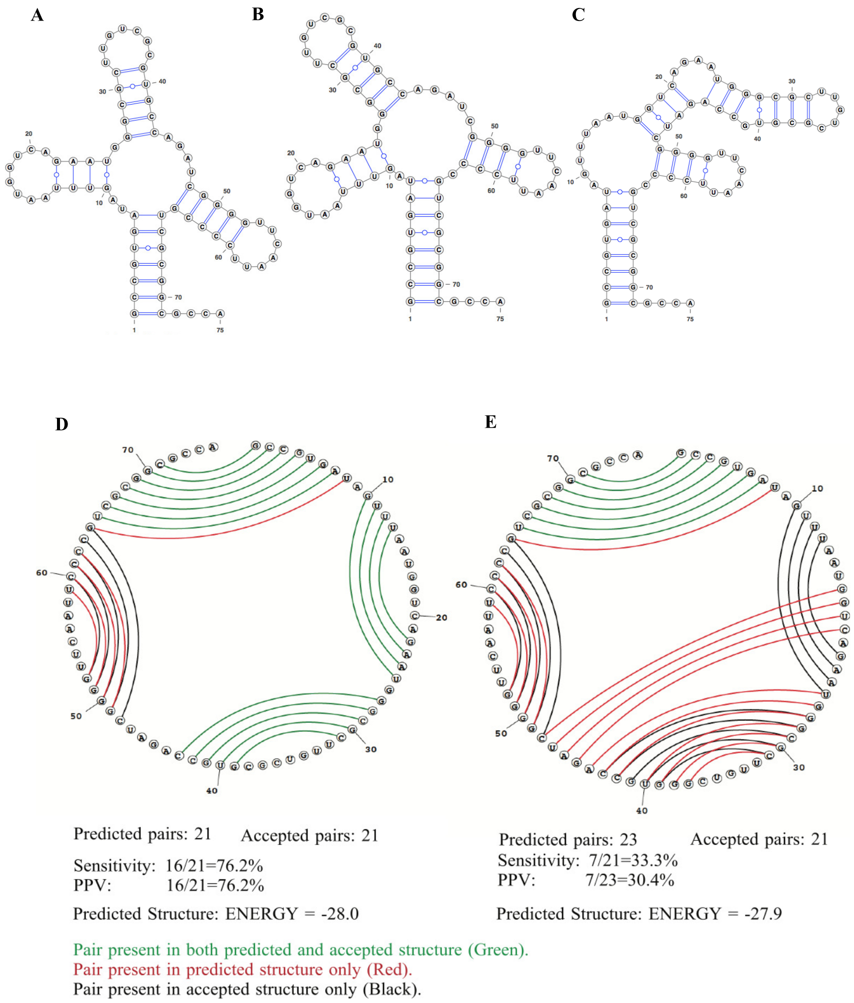
A
B
C
D
E
Predicted pairs: 21
Accepted pairs: 21
Sensitivity: 16/21=76.2%
PPV: 16/21=76.2%
Predicted Structure: ENERGY = -28.0
Predicted pairs: 23
Accepted pairs: 21
Sensitivity: 7/21=33.3%
PPV: 7/23=30.4%
Predicted Structure: ENERGY = -27.9
Pair present in both predicted and accepted structure (Green).
Pair present in predicted structure only (Red).
Pair present in accepted structure only (Black).

Fig. 3 Comparison between MFE secondary structure and one of the suboptimal secondary structures for tRNA(asp), *yeast*. A: Reference (accepted) structure. B: MFE structure. C: Suboptimal structure. D: Circular plot comparing the MFE structure in B to the reference structure in A. E: Circular plot comparing the suboptimal structure in C to the reference structure in A. Structures are predicted using the *Fold* program in RNAstructure package [125] with default parameters. Plots A, B and C are prepared with VARNA [198]. Circular plots D and E are prepared with the *CircleCompare* program in RNAstructure. In D and E, base pairs are indicated by lines. Pairs present in both the predicted and reference structures are green; pairs which are present only in the predicted structure are in red; and pairs which are present only in the reference structure are in black.

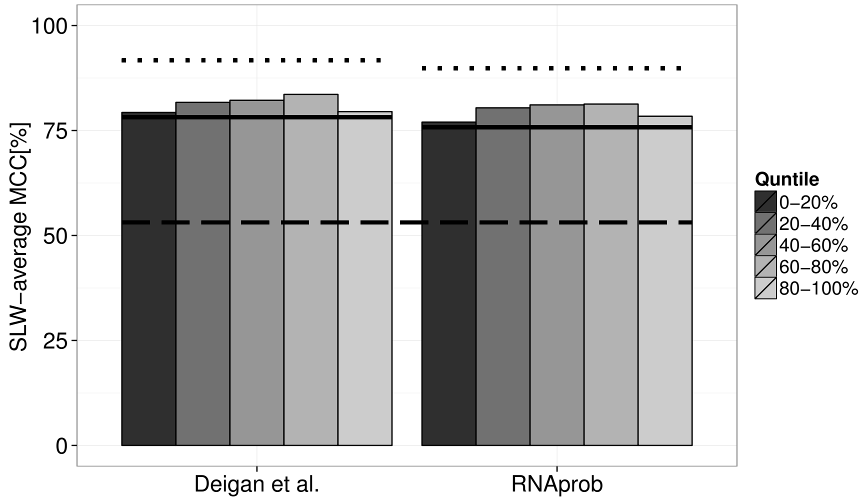

Fig. 4 Information content of SHAPE data with perfect information. Two data-directed structure prediction methods, i.e., Deigan's approach [82] and RNAprob [134], are tested on a set of 23 RNAs, as used in [134]. For RNAprob, the variant with two structural contexts and empirical decoder is used. Bars represent SLW-average MCC values of quintiles with perfect information. Upper dashed lines represent the performance with the entire SP set to perfect information. Solid lines indicate the performance with the original SP data and the bottom dashed line corresponds to the no-SHAPE control.