Computational Approaches for RNA Structure Ensemble Deconvolution from Structure Probing Data

Sharon Aviran¹ and Danny Incarnato²

¹ - Biomedical Engineering Department and Genome Center, University of California, Davis, CA, USA
² - Department of Molecular Genetics, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, Groningen, the Netherlands

Correspondence to Sharon Aviran and Danny Incarnato: saviran@ucdavis.edu (S. Aviran), d.incarnato@rug.nl (D. Incarnato), @AviranLab (S. Aviran), @danny_incarnato (D. Incarnato)

https://doi.org/10.1016/j.jmb.2022.167635

Edited by Philip C Bevilacqua

Abstract

RNA structure probing experiments have emerged over the last decade as a straightforward way to determine the structure of RNA molecules in a number of different contexts. Although powerful, the ability of RNA to dynamically interconvert between, and to simultaneously populate, alternative structural configurations, poses a nontrivial challenge to the interpretation of data derived from these experiments. Recent efforts aimed at developing computational methods for the reconstruction of coexisting alternative RNA conformations from structure probing data are paving the way to the study of RNA structure ensembles, even in the context of living cells. In this review, we critically discuss these methods, their limitations and possible future improvements.

Introduction

RNA structure is instrumental to RNA function. RNA can fold into stable and functional secondary structures, mediated by both direct base-pairing between complementary nucleobases and by base stacking. These are further compacted into tertiary structures mediated by a number of factors, including divalent cations, non-canonical base-pair interactions and interactions with proteins. However, the conformational space of an RNA molecule is huge. An n-nucleotide long RNA can theoretically fold into up to nearly 2ⁿ different secondary structures.¹ This scenario is further complicated when considering that each RNA nucleotide has 8 degrees of freedom and that the native biologically-active structure is not always thermodynamically strongly favored over competing structures.²

Determining the structure of RNA molecules is crucial to understanding their function, but performing this task starting solely from their primary sequence is a nontrivial challenge. Computational approaches based on free energy minimization aim at predicting the minimum free energy (MFE) secondary structure, that is, the most thermodynamically stable secondary structure, using a set of parameters known as nearest neighbor model (or Turner rules).³–¹¹ The accuracy of MFE structure prediction is however limited in practice, with only ~ 66% of predicted base-pairs found in reference structures.¹² The reliability of these predictions can be significantly improved by incorporating constraints from RNA structure probing experiments. These experiments are based on the use of specific probes, mostly chemicals, to query the structural state of RNA residues. Sites modified by these probes can be directly...
<table>
<thead>
<tr>
<th>Method name</th>
<th>Dependency on thermodynamics</th>
<th>Type of input data</th>
<th>Type of output</th>
<th>Size of deconvolved ensemble</th>
<th>Quantification of relative abundances</th>
<th>Maximum RNA length analyzed to date</th>
<th>Applied to</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample and cluster</td>
<td>Dependent</td>
<td>Reactivity data</td>
<td>Secondary structure ensemble</td>
<td>Depends on choice of clustering method</td>
<td>Yes</td>
<td>554 nt (SERPINA1 5' UTR) 65</td>
<td>A broad range of data types</td>
<td>-</td>
</tr>
<tr>
<td>Rsample</td>
<td>Dependent</td>
<td>Reactivity data</td>
<td>Secondary structure ensemble</td>
<td>User-defined (1,000 by default)</td>
<td>Yes</td>
<td>232 nt (HIV-1 RRE) SHAPE-seq, DMS-MaP</td>
<td>SHAPE-seq, DMS-MaP</td>
<td><a href="https://rna.urmc.rochester.edu/RNAstructure.html">https://rna.urmc.rochester.edu/RNAstructure.html</a></td>
</tr>
<tr>
<td>SLEQ</td>
<td>Dependent</td>
<td>Sequencing reads</td>
<td>Structure ensemble</td>
<td>Automatically determined</td>
<td>Yes</td>
<td>112 nt (V. vulnificus add riboswitch)</td>
<td>SHAPE-seq</td>
<td><a href="https://github.com/AviranLab/SLEQ">https://github.com/AviranLab/SLEQ</a></td>
</tr>
<tr>
<td>R2D2</td>
<td>Dependent</td>
<td>Reactivity data</td>
<td>Secondary structure ensemble</td>
<td>User-defined (100 by default)</td>
<td>No</td>
<td>114 nt (E. coli SRP RNA) SHAPE-seq</td>
<td>SHAPE-seq</td>
<td><a href="https://github.com/LucksLab/R2D2">https://github.com/LucksLab/R2D2</a></td>
</tr>
<tr>
<td>REEFFIT</td>
<td>Dependent</td>
<td>Reactivity data</td>
<td>Secondary structure ensemble</td>
<td>Same as input size</td>
<td>Yes</td>
<td>112 nt (V. vulnificus add riboswitch)</td>
<td>Mutate-and-map (M²)</td>
<td><a href="https://github.com/ribokit/REEFFIT">https://github.com/ribokit/REEFFIT</a></td>
</tr>
<tr>
<td>IRIS</td>
<td>Dependent</td>
<td>Sequencing reads</td>
<td>Secondary structure ensemble specific to wt and to SNVs</td>
<td>User-defined</td>
<td>Yes</td>
<td>330 nt (snoRNA U85) PARIS</td>
<td>PARIS</td>
<td><a href="https://github.com/JY-Zhou/IRIS">https://github.com/JY-Zhou/IRIS</a></td>
</tr>
<tr>
<td>RING-MaP</td>
<td>Independent</td>
<td>Sequencing reads</td>
<td>Deconvolved reactivity profiles</td>
<td>Automatically determined</td>
<td>Yes</td>
<td>265 nt (Bacillus stearotherophilus RNase P catalytic domain)</td>
<td>DMS-MaP</td>
<td><a href="https://webshare.oasis.unc.edu/weeksgrp/ring-map_software.zip">https://webshare.oasis.unc.edu/weeksgrp/ring-map_software.zip</a></td>
</tr>
<tr>
<td>DREEM</td>
<td>Independent</td>
<td>Sequencing reads</td>
<td>Deconvolved reactivity profiles</td>
<td>Automatically determined, up to a user-defined number ²</td>
<td>Yes</td>
<td>29,882 ³ (SARS-CoV-2 genome) in manually-tiled windows</td>
<td>DMS-MaP</td>
<td><a href="https://codeocean.com/capsule/6175523/tree/v1">https://codeocean.com/capsule/6175523/tree/v1</a></td>
</tr>
<tr>
<td>DRACO</td>
<td>Independent</td>
<td>Sequencing reads</td>
<td>Deconvolved reactivity profiles</td>
<td>Automatically determined</td>
<td>Yes</td>
<td>29,903 ⁴ (SARS-CoV-2 genome) in automatically slit and merged</td>
<td>DMS-MaP</td>
<td><a href="https://github.com/dincarnato/draco">https://github.com/dincarnato/draco</a></td>
</tr>
</tbody>
</table>
detected by reverse transcription, yielding an RNA reactivity profile that provides a measure of the likelihood of each base in the transcript of being unpaired (highly reactive) or base-paired (lowly reactive). These reactivity profiles can then be used to restrain RNA structure prediction tools, typically yielding a single structure model that better fits the experimentally-observed reactivities as compared to the MFE structure.\textsuperscript{13–16}

The main limitation of this kind of approach is that it only produces a single structure for a given RNA. In reality, despite their stability, RNA structures are highly dynamic and heterogeneous. For a given RNA, alternative mutually-exclusive conformations can coexist as part of a heterogeneous ensemble. The ability of RNA molecules to dynamically redistribute the relative abundance of specific conformations within the ensemble in response to environmental cues is crucial to their regulatory functions. RNA thermometers and riboswitches, RNA elements able to regulate translation by dynamically switching between mutually-exclusive structural conformations in response to temperature changes or to the presence of specific metabolites, are two of the better characterized examples. Several additional examples have been reported for both cellular RNAs and viral RNA genomes. For example, the mutually-exclusive interaction of HNRNPL or of the GAIT complex with the 3’ UTR of VEGFA drives the structural switch from a stem loop to a three-way junction conformation, respectively stimulating or repressing translation.\textsuperscript{17} The ensemble of the HIV-1 Rev Response Element (RRE) is populated by two alternative structures, a four-way and a five-way junction, with the latter promoting higher infection rates by increasing the Rev-mediated nuclear export of viral RNAs.\textsuperscript{18} In spite of its ascertained importance, the study of RNA structural heterogeneity, particularly in the context of living cells, has largely remained elusive, mostly because of the lack of suitable methods.

In this review, we will focus on the challenges connected with the study of RNA structural ensembles and how recent combined experimental and computational efforts are making it possible to begin dissecting the heterogeneity of RNA structural ensembles.

### High-throughput sequencing-based methods for RNA structure interrogation

Over the past decade, a wide variety of methods relying on high-throughput sequencing (HTS) has been introduced to study RNA structures.\textsuperscript{19–20} The main strength of these methods is their capability to concurrently query thousands of RNAs, hence allowing them to interrogate the entire transcriptome in a single experiment. Among these, RNA structure probing experiments are probably the

<table>
<thead>
<tr>
<th>Method name</th>
<th>Dependency on thermodynamics</th>
<th>Maximum RNA length analyzed to date</th>
<th>Size of deconvolved ensemble</th>
<th>Type of output</th>
<th>Type of input data</th>
<th>Quantification of relative abundances</th>
<th>Size of deconvolved ensemble</th>
<th>Applied to</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>DANCE-MaP</td>
<td>Independent</td>
<td>Sequencing reads</td>
<td>Automatically determined, up to a user-defined number\textsuperscript{2}</td>
<td>Deconvolved reactivity profiles</td>
<td>DMS-MaP data used with DANCE-MaP has been generated under experimental conditions that promote DMS to react also with GU bases, besides A/C bases.</td>
<td>Yes</td>
<td>332 nt (Homo sapiens 7SK RNA)</td>
<td>Automatically determined, up to a user-defined number\textsuperscript{2}</td>
<td>DMS-MaP data used with DANCE-MaP has been generated under experimental conditions that promote DMS to react also with GU bases, besides A/C bases.</td>
</tr>
</tbody>
</table>

\textsuperscript{1} Only a subset of representative examples is included.

\textsuperscript{2} The upper limit to the maximum number of identified conformations is enforced by the user to avoid overestimation issues (default: 2 for DREEM, 5 for DANCE-MaP).

\textsuperscript{3} Analysis is performed in short windows. DREEM is not capable of automatically slide windows, nor of merging the information deconvolved from overlapping windows, therefore this operation has to be done manually.

\textsuperscript{4} Analysis is automatically performed in sliding windows (length: 90% the median read length, offset: 5% the median read length), and the information from consecutive overlapping windows is further automatically integrated.

\textsuperscript{5} DMS-MaP data used with DANCE-MaP has been generated under experimental conditions that promote DMS to react also with GU bases, besides A/C bases.
most popular because of their simplicity and wide applicability. HTS-based probing approaches share a similar experimental design: modification of RNA residues in a structure-dependent fashion, detection of the modification sites via reverse transcription (RT), sequencing of the resulting cDNA products, and computational reconstruction of an RNA reactivity profile (Figure 1, left). While a wide variety of methods have been developed, there are two main distinctive features: the type of probe used to interrogate the structure of the RNA and the RT readout strategy.

Probes are either enzymatic or chemical. Chemical probes have rapidly taken over enzymatic probes for a number of reasons, mainly their higher resolution and their suitability for in vivo RNA structure analyses. Additionally, contrary to enzymatic probes, chemical probes can interrogate RNA structures at many different levels. Nucleobase-specific probes, such as dimethyl sulfate (DMS) and 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC), can inform on the base-pairing respectively of A/C and G/U bases. Selective 2'-hydroxyl acylation ana-
lyzed by primer extension (SHAPE) reagents, such as 2-methylnicotinonic acid imidazolide (NAI), 2-aminopyridine-3-carboxylic acid imidazolide (2A3) and 5-nitroisatoic anhydride (SNIA), can inform on the local RNA backbone flexibility. Recently developed bifunctional probes such as trans-bis-isatoic anhydride (TBlA) and SHARC (Spatial 2′-Hydroxyl Acylation Reversible Crosslinking) reagents, form covalent crosslinks between structurally-proximal bases, hence directly informing on the tertiary structure of RNA molecules.

Sites of enzymatic cleavage or chemical modification are mapped via RT. Traditional readout of RNA structure probing experiments was based on the detection of RT drop-off events, either caused by the premature end of the RNA molecule due to enzymatic cleavage, or by the inability of the reverse transcriptase to read through sites of chemical modification. More recently, a readout strategy dubbed mutational profiling (MaP) has been introduced. In MaP, special reverse transcriptases or RT conditions are used to favor the read-through on chemically-modified RNA bases. In doing so, the reverse transcriptase incorporates a random nucleotide, resulting in a mutation in the cDNA. While in RT drop-off-based methods each cDNA only provides information on a single base per transcript, in MaP-based methods multiple sites of modification, corresponding to bases being simultaneously in the same structural state (i.e. unpaired, or structurally-flexible) within the same RNA molecule, are concurrently recorded within the same cDNA.

Aside from traditional structure probing, a separate class of methods has been developed to enable the direct capture of RNA-RNA interactions (Figure 1, right). These methods are again characterized by a very similar experimental design: stabilization of RNA-RNA interactions via crosslinking, partial digestion of RNA duplexes, intramolecular ligation of the two strands, sequencing of the result chimeras and computational identification of paired RNA regions. While these techniques are theoretically more informative than RNA structure probing experiments when it comes to mapping base-pairing interactions, the percentage of informative chimeric reads in these experiments is quite low, usually below 10%, hence making the interrogation of the entire transcriptome very challenging. Nonetheless, only direct RNA-RNA interaction mapping approaches have the potential to identify potential pseudoknots in RNA structures.

Incorporating experimental data into RNA secondary structure prediction: The state-of-the-art

Data derived from the aforementioned analyses can directly inform the prediction and modeling of RNA structures. For RNA structure probing experiments, this typically involves three main steps: calculating raw per-base reactivities, normalizing reactivities, and incorporating reactivities as pseudo-energy constraints during free energy minimization, although alternative approaches have also been proposed for the latter step. Since these steps have been reviewed in detail elsewhere, in what follows we briefly summarize common practices and emphasize the aspects most relevant to the material covered in this review. A number of tools are available to automate these steps. Different approaches for calculating raw reactivities have been proposed, but usually they involve subtracting the background signal, obtained by performing RT in the absence of any chemical or enzymatic RNA treatment, from the signal obtained from the probed sample. In some instances, a denatured control is also included, obtained by probing the RNA under denaturing conditions. As both the probing and modification detection efficiencies are non-uniform along the RNA, because of strong sequence context-dependent biases, the inclusion of a denatured control allows estimating the upper limit of the reactivity theoretically measurable for each base in an RNA molecule. Raw reactivities are then normalized. The aim of the normalization is to enable the direct comparison of RNA reactivity profiles across different experimental conditions. Different normalization schemes have been proposed, typically yielding reactivity values ranging from 0 to ≥ 1. Reactivity values are then used to constrain RNA secondary structure prediction. Older approaches involved empirically defining a reactivity threshold to impose a hard constraint on a base, forcing it to be unpaired in the predicted structure. Nowadays, hard constraints have been almost completely replaced by soft constraints (also commonly referred to as restraints). Restraining a structure prediction involves converting base reactivities into pseudo-free-energy change terms, which are free-energy change contributions used to adjust the scores assigned by the nearest neighbor energy model to either reward or penalize the formation of base-pairs. In other words, base-pairing should be favored for bases having low reactivity and disfavored for bases having high reactivity. The approach devised by Deigan and colleagues is the first and probably the most popular, and it has been implemented in all the major programs for structure prediction by free energy minimization.
method assumes a linear relationship between the logarithm of the reactivity and the corresponding pseudo-free-energy change term. This linear relationship is described by a line with unknown slope and intercept. Typically, optimal slope and intercept values are determined by grid search, a process during which the structure of a reference RNA with a known structure is iteratively predicted with different slope-intercept value pairs to find pairs that maximize the prediction accuracy. Alternative methods that model the information in structure probing experiments have also been proposed. Nonetheless, to date, no method appears to dramatically outperform the others. These methods will not be discussed here as they have been thoroughly reviewed elsewhere.

Computational methods to deconvolve RNA structure ensembles

The main limitation of approaches based on free energy minimization is that nucleotide reactivities are interpreted as the signal generated by a single predominant RNA structural conformation. Rather, many RNAs likely exist as an ensemble of many alternative structures, coexisting in thermodynamic equilibrium. Therefore, reactivity profiles derived from probing experiments likely represent an average measurement over all the coexisting conformations for a given RNA sequence. The ensemble is not a static entity: within the ensemble, RNA molecules can interconvert between each possible structural state, at a rate that depends on the energetic barrier separating each conformation pair.

The dynamic, multi-state nature of the RNA folding landscape has been traditionally understood through the lens of statistical mechanics and modeled by the Turner nearest neighbor thermodynamic framework. For a given sequence, the Turner model assigns each secondary structure a free-energy change score reflecting its thermodynamic stability, from which the probability that the sequence folds into this structural state can be computed by the Gibb-Boltzmann equation. This quantitative framework results in a statistical characterization of the landscape as a probability distribution over all secondary structures (or folded states), called the Boltzmann equilibrium distribution. Despite the theoretical elegance, calculating such probabilities in practice requires determining a normalizing constant, called a partition function, which is the sum of a score-dependent function over all the structures. In the case of RNA folding, the exponential growth in the number of structures with increasing RNA length prohibits calculating this constant by structure enumeration. However, pioneering work by McCaskill enabled efficient computation of the partition function and the probabilities of observing each potential base pair in the ensemble under the thermodynamic model, thereby greatly expanding the sophistication and scope of computational structure analysis.

Methods that leverage the thermodynamic model to obtain more detailed representations of the ensemble in the form of explicit suboptimal secondary structures were also developed. For example, it is possible to compute all the suboptimal secondary structures within a (small) energy increment from the MFE. The main limitation of this approach is that it explores only the low-energy tail of the Boltzmann distribution, which for many systems, might provide a biased and incomplete description of the landscape. To overcome this limitation, Ding and Lawrence extended McCaskill’s approach to equilibrium partition function calculation into a method that samples suboptimal secondary structures from the Boltzmann distribution, known as stochastic sampling or Boltzmann sampling. Boltzmann sampling is motivated by the premise that sufficiently large samples closely approximate the Boltzmann distribution to allow accurate estimation and detection of structural properties whose exact calculation is computationally intensive or infeasible. Examples of popular applications include estimating the probability that a stretch of nucleotides adopts a single-stranded form (also termed region accessibility) or the probabilities of observing a certain helix, other structural features, combinations of features, or groups of structures. Moreover, one of the most powerful applications is to detect key structural features de novo from such samples and evaluate their relative abundance. Boltzmann sampling is widely used in RNA analysis and design, especially in the context of multiple functional structures, and implemented in the Sfold software package and other major secondary structure prediction software packages. A typical sample size is in the range of 1,000–10,000 structures, as it is generally believed to be sufficient to attain reproducibility of substructural signals between samples. However, there is no guarantee such size provides sufficient sampling depth to capture all relevant structural features. The ideal size largely depends on the RNA length, the sequence, and the targeted signal resolution (e.g., base pair vs structural features vs complete structures). Unfortunately, it can be challenging to determine, and as transcripts increase in length, the required sampling depth often becomes prohibitive. Importantly, even for typical sample sizes, extracting the key structural information and visualizing the data are highly non-trivial tasks that warrant dedicated methods. We will elaborate more on such analyses and available tools in the next subsection.

The success and widespread adoption of methods for data-directed MFE structure prediction have inspired the development of
methods that harness probing data to inform the prediction of more complex structure landscapes. In such analyses, it is particularly important to capture the key differences and similarities between structures in the underlying ensemble and to quantify the relative abundance of individual structures or of populations or similar structures. Computational methods for the analysis of RNA structure ensembles can be categorized into two main classes: thermodynamics-dependent and thermodynamics-independent (Figure 2 and Table 1). We wish to point out that dependence on or independence of thermodynamics only relates to the way these approaches tackle the problem of deconvolving the structure probing data. Thermodynamics-dependent methods make use of thermodynamics-based structure modeling to deconvolve the data, while thermodynamics-independent methods deconvolve the data by their direct clustering. However, structure modeling still relies on thermodynamics also for the latter class. This point will be addressed in greater detail in what follows.

Thermodynamics-dependent methods

Thermodynamics-dependent methods use both the thermodynamic model and the data to identify a set of explicit structure models that are consistent with the data. A shared element is the generation of a multitude of alternative structures based on the thermodynamic model, typically by sampling the Boltzmann distribution (Figure 3). However, methods differ greatly in how they
process the information embedded in these structures and how they model the relationship between structures and probing data. Additional differences include the stage in the analysis at which the data are considered, their integration with thermodynamics, modeling assumptions, model fitting techniques, and the output ensemble description. Some methods have been designed for specific experiments or data types, while others have a broader scope but nonetheless were optimized for certain experiments. In what follows, we review the main methods developed to date. For each method, we first outline its key principles, followed by further details that aim to highlight the aforementioned points and key differences.

**Sample and cluster. Overview.** The earliest and simplest strategy to data-directed ensemble analysis, which we herein refer to as ‘sample and cluster’, entails two steps (Figure 3). First, a stochastic sample is generated by data-directed thermodynamic modeling. Second, the sample is mined via cluster analysis to extract key structural information. One could also optionally reduce the dimensionality of the sampled structures for visualization purposes, either before or after clustering. Probing data are integrated in the first step, where they are used to adjust the scores assigned by the thermodynamic model. This is achieved in a manner akin to data-directed free energy minimization, that is, by converting data into soft pseudo-free-energy constraints. However, in
this case, adjustments are made to the partition function calculations that lie at the core of stochastic sampling. If the sample is mined for key structural variation prior to dimensionality reduction, one could use a number of methods developed for cluster analysis of RNA secondary structures. These methods are not discussed here as they have been thoroughly reviewed elsewhere.71 Briefly, they group structures based on common or similar structural features, where major differences between them lie in how they represent structures to capture biologically important features and how they define similarity between structures. Additionally, a representative structure is constructed for each cluster to highlight the structural similarity and difference between clusters. Population fractions are determined directly from cluster sizes. When including dimensionality reduction, structures are first represented as n-dimensional vectors, and a pairwise distance (or dissimilarity) metric is defined in that space, based on which vectors are projected to 2D or 3D space. Numerous popular techniques have been used, including Principal Component Analysis (PCA), multidimensional scaling (MDS), t-SNE, and t-SNE.

Method details. Analysis starts with data in the form of reactivities, which are input to data-directed partition function and stochastic sampling routines implemented in popular structure prediction software packages.40–41 Since these routines take reactivities as input, ‘sample and cluster’ is widely applicable to diverse probing techniques, irrespective of the readout (i.e., RT drop-off, MaP, or cleavage) or cDNA quantification (i.e., electrophoresis or sequencing) strategies.73 Yet, it is worth noting that pseudo-free-energy parameters used by these routines are currently set by default to values determined by optimization of single-structure predictions directed by SHAPE data from in vitro experiments.38–39 Parameters may thus benefit from re-calibration in situations where the reagent or experimental conditions are different and when suitable training data are available. Furthermore, these values were determined based on predictions for structured RNAs that adopt one predominant fold and might be suboptimal for ensemble prediction. Importantly, the stochastic sampling step limits consideration to nested secondary structures, such that pseudoknots, non-canonical base pairs, or tertiary interactions are not modeled. The optional dimensionality reduction requires representing structures as n-dimensional vectors and defining a distance metric in vector space. One popular representation encodes structures into binary vectors, where 0 and 1 correspond to the absence or presence of each of the possible base pairs, respectively. It is often used in tandem with the so-called base pair distance between structures, defined as the number of base pairs in either structure but not in both.73 Other representations and distance metrics have also been used, e.g., nucleotide paired/unpaired states65,76 or counts of small structural features with/without their positions,74,77 jointly with Euclidean distance. In terms of sequence lengths to which this approach has been applied, to the best of our knowledge, the longest is 554 nt.

Newer approaches. The ‘sample and cluster’ framework was introduced shortly after Boltzmann sampling73 and about a decade before data-directed sampling methods were implemented as the direct extension of soft-constrained MFE structure prediction.40–41 Indeed, ‘sample and cluster’ is often used with traditional sampling (i.e., without data). Since it was not conceived with data-derived information in mind, its recent adaptation couples tools developed for other tasks: free-energy minimization with soft constraints and cluster analysis for secondary structure prediction. In contrast, the following methods were designed to leverage data for modeling folding ensembles. Like ‘sample and cluster’, they obtain multiple structures, but subsequently, they also interpret the data with respect to these structures based on a model of the relationship (or consistency) between the structures and the data (Figure 3). The interpretation step is method-specific, but its goal is always to learn how to assign relative abundances to structures in a manner consistent with the data. From this point onward, the first method (Rsample) differs from the others as it obtains a new, data-reweighted stochastic sample. Thus, it can be used as the data-directed sampling module in ‘sample and cluster’ analyses.77 The other methods assign abundances to previously-obtained structures. In doing so, they build on an earlier idea, termed ‘sample and select’, which was introduced to predict a single structure from sequence and data. In ‘sample and select’, many alternative structures are stochastically sampled with no input from the data. Subsequently, the structure most consistent with the data is selected based on the agreement between the base-paired/unpaired states in each structure and a discretized representation of the data as a paired/unpaired state sequence.38–39 These methods extend this idea from selection of one structure to redistribution of relative abundances over a given set of structures. Abundances can be adjusted such that only one or few structures are deemed likely and are thus selected as representatives of structural sub-populations. Alternatively, abundances can be redistributed across the entire set, in which case an additional clustering step allows one to ‘select’ a representative structure from each cluster.

Rsample. Overview. Rsample (for Restrained sample) uses probing data to guide the thermodynamic model toward sampling structures that are consistent with the data.30 The idea is to
sequentially generate two large stochastic samples of alternative secondary structures, where one learns from the first sample how to adjust the thermodynamic calculations, such that structural features in the second sample are brought to better agreement with the data. Adjustments take the form of nucleotide pseudo-free-energy change terms, where each term is computed based on the discrepancy between the reactivity observed and the reactivity predicted by considering all sampled structures. This differs from the popular pseudo-free-energy approach to MFE prediction, which interprets the data in the absence of structure models. Once a second sample is obtained, it is clustered to arrive at a simplified description of its structural variation in the form of a small number of cluster-representative structures and their relative abundances. One could also opt to predict a single representative structure. To predict reactivities based on structures in the first sample, Rsample employs a statistical empirical model of reactivity data, which is typically compiled from several thousands of reactivities observed for structured RNAs with known structures. The model captures the statistical behavior of reactivities, depending on a nucleotide’s structural context, that is, whether it is unpaired, paired at a helix end, or paired inside a helix. For each nucleotide in each structure, a value is drawn according to its structural context, and reactivities are predicted by averaging over all structures. This strategy allows Rsample to account for the high variability and ambiguity often characteristic of probing data, in particular the observation that unpaired nucleotides can often be unreactive to SHAPE or DMS.

**Method details.** Rsample takes probing data and the RNA sequence as input. It outputs the second sample of suboptimal structures (1,000 by default, but is controllable). A supplementary script implements a classic divisive hierarchical clustering algorithm, which is widely used to mine large structure samples for structurally-similar sub-populations. The number of clusters is variable and automatically determined using the Calinski-Harabasz index, which evaluates clustering quality by the compactness within clusters as well as the separation between clusters. For each cluster, a centroid structure is constructed to include all base pairs whose probability within that cluster exceeds 0.5. Population fractions are determined directly from cluster sizes. Input data take the form of reactivities, hence Rsample is applicable to diverse probing techniques. As noted, Rsample employs a reactivity model to predict reactivities from structures. Such data-driven models require substantial data collection and have proven to be reagent-specific, and in some cases, specific to experimental conditions and/or the probing technique. It is thus important to use a model that best captures the statistical properties of the analyzed data. Currently, a SHAPE model compiled from *in vitro* experiments is used by default. A DMS model compiled from *in vivo* experiments is also available, as well as the option for users to specify a model. In terms of modeled features, Rsample relies on standard partition function calculations and therefore cannot predict structures with pseudoknots, non-canonical base pairs, or tertiary interactions, nor account for RNA-ligand/RNA-protein interactions. To date, it has been applied to RNAs whose lengths range from few tens to few thousands nucleotides, although analyses of multiple conformations were performed only for sequences up to 232 nt long.

**SLEQ. Overview.** SLEQ (Structure Landscape Explorer and Quantifier) approaches ensemble modeling from a statistical inference angle, whereby a generative model of the data is used to deduce a parsimonious description of the landscape. The data are first sorted into modification patterns indicating which nucleotides were modified or co-modified, and pattern frequencies are summarized. The statistical relationship between a given set of structures and these patterns is characterized by calculating all pairwise structure-to-pattern likelihoods, that is, how likely a structure is to generate an observed pattern. In other words, this is a means to quantify consistency between the structures and the sequencing reads. To enforce the notion that the data are ensemble-averaged measurements, the likelihood of observing a modification pattern in the experiment is computed as the sum, over all structures, of the likelihood that each structure generates this pattern, weighted by its (unknown) relative abundance. Therefore, the likelihood that a given structure ensemble generates the entire data (i.e., patterns and their frequencies) is viewed as a mixture of structure-specific pattern likelihoods. The mixture is deconvolved from the observed pattern frequencies using non-negative linear least squares (NNLS) regression. NNLS optimization searches for the relative abundances that bring the predicted pattern likelihoods as close as possible to their frequencies in the data. Notably, SLEQ was designed with the goal of explaining the data with as few structures as possible, hence a parsimonious solution that assigns non-negligible weights only to a small number of structures is desired. Although NNLS is not designed to necessarily favor parsimonious selection, it has been used to obtain sparse solutions in other domains, and in the case of structure probing, prior analyses of real and simulated data for small RNAs resulted in selections of very few structures. Structure-to-pattern likelihoods are derived from a model that is general, in the sense that it is not learned from prior experiments. The model makes two simplifying and broadly applicable assumptions: (1) structurally constrained nucleotides are inaccessible to the reagent and (2) accessibility of unconstrained nucleotides may only
depend on their base identity. Modeling nucleotides as inaccessible/accessible to a reagent, as opposed to paired/unpaired, allows capturing interactions beyond canonical base pairing.

**Method details.** SLEQ takes probing data, RNA sequence, and candidate structures as input and outputs a small set of structures and their relative abundances, interpreted as representative of distinct populations of structural variants. It selects them from input candidates and estimates their abundances using its generative model. Their number is variable and automatically determined by the algorithm. Differently from most other thermodynamics-dependent methods described here, input data take the form of sequencing reads that report either mutation or drop-off events, all spanning the region of interest. Reagents supported include non-base-selective chemistries, such as SHAPE and in-line probing, as well as the base-selective DMS. SLEQ also provides flexibility with respect to candidate structures. Although commonly generated by stochastic sampling, one could alternatively curate a candidate set to reflect prior knowledge or otherwise include hypothesized structures to test their relevance. A hybrid ‘spike in’ approach, by which one merges a small, curated set with a large stochastic sample, was also taken, whereby pseudoknotted structures were added to the sample. Large samples from distinct but related sequences, such as SNVs, were also combined. Another option is to perform ‘sample and cluster’ analysis and feed cluster representatives to SLEQ to estimate their relative abundances. In terms of modeled structures, any interaction (e.g., tertiary, non-canonical, intermolecular, non-nested) known to render certain regions inaccessible to the reagent can be encoded into structures as constrained sites. However, despite various strategies to enhance the content of candidate sets, it is important to note that relying on input structures is risky as it precludes de novo prediction of structures from the data. If relevant structures are missing, predictions are bound to be biased, and moreover, such situations are difficult to detect. We discuss this limitation of ‘sample and select’ methods in more detail later. By design, SLEQ embeds a very simple structure-to-data model with few parameters to maintain generality and alleviate risks of overfitting. However, the downside is that variation in reactivities is interpreted as arising solely from shifts within the ensemble, disregarding known discrepancies between reactivity and structural context, particularly for unpaired nucleotides. To date, SLEQ has been used to analyze RNA sequences up to 112 nt long.

**R2D2.** **Overview.** R2D2 (Reconstructing RNA Dynamics from Data) was developed to guide modeling of non-equilibrium RNA cotranscriptional folding pathways using data obtained from specialized assays that probe structures of nascent RNAs at all intermediate lengths of a target sequence. It extends on ‘sample and select’ by repeating it multiple times to obtain multiple structures consistent with the data and it also quantifies consistency differently. For each intermediate transcript length, R2D2 stochastically samples 100 secondary structure sets. Each structure is assigned a consistency score with respect to the data observed for its length, and in each sample, the best-scoring structure is selected. For each intermediate transcript, this results in 100 secondary structures (not necessarily distinct), all deemed consistent with the data. One could then opt to construct a consensus structure per each length, where only base pairs observed in at least 50% of the structures are retained. Consensus structures can be further used as a starting point for molecular dynamics simulations to generate 3D models of transitions between predicted intermediate states. To evaluate consistency between structures and the data, R2D2 compares base-paired/unpaired states to reactivities using a dedicated distance function that was optimized through grid search to maximize accuracy of single-structure predictions of known RNA structures. It treats paired and unpaired states separately to account for the known tendency of some unpaired nucleotides to display low reactivity.

**Method details.** Although R2D2 was designed to model folding pathways, an application in which out-of-equilibrium folded states are of interest, its secondary structure prediction module is applicable to structure ensembles under equilibrium conditions. This is because it employs conventional stochastic sampling routines, treats each intermediate transcript independently of the others, and uses data from standard probing assays to guide predictions. R2D2 takes probing data and the RNA sequence as input. It outputs up to 100 distinct secondary structures, yet it does not estimate their relative abundances nor it determines structurally distinguishable populations. Input data take the form of reactivities, irrespective of readout or cDNA quantification choices. A couple notable departures from common sampling practices concern sample size and composition. Specifically, each sample comprises 150,000 structures, as compared to the typical 1,000–10,000. Furthermore, these samples are assembled from 3 independent 50,000-structure stochastic samples, whose generation is guided by either soft data constraints or hard data constraints or solely by thermodynamics. These unconventional choices should not negatively impact equilibrium predictions and aim to increase the likelihood that structures disfavored in equilibrium conditions are included in the samples, or more generally, to increase structural diversity. As noted above, R2D2 determines structure-to-data consistency using new custom parametric data preprocessing and distance function. However, data are
still interpreted as the signature of a single predominant structure, similarly to ‘sample and select’ and earlier work. Since Yu et al. probed structure with SHAPE-seq, parameters of the consistency evaluation pipeline were optimized to predict known structures for 6 relatively short RNAs from SHAPE-seq data. Data-directed sampling routines were similarly optimized, where values deviate from ones typically used with SHAPE data. Therefore, for different data, it may be helpful to re-optimize this pipeline when possible. When analyzing RNAs much longer than the sequence analyzed to date (117 nt), the sample size and number of samples might also need to be adjusted. Similarly to Rsample, R2D2 relies on conventional stochastic sampling, which is limited to nested secondary structures.

REEFFIT. Overview. REEFFIT (RNA Ensemble Extraction from Footprinting Insights Technique) models structure ensembles using data obtained from mutate-and-map (M²) experiments that probe the structures of a target sequence and that of a single point mutant at each possible position along the RNA. REEFFIT is motivated by the observation that certain point mutations give rise to substantially different reactivity profiles compared to the wild type sequence, and that distinct mutants often display similar alternative profiles. The method assumes that the wild type (wt) and all probed single-nucleotide variants (SNV) share a common set of coexisting alternative structures and that the profoundly altered reactivity patterns are a manifestation of significant shifts in their relative abundances. In other words, a mutation can potentially perturb the ensemble through altering equilibrium fractions, such that some structures become predominant. This concept is expressed mathematically via a linear generative model, where the reactivity profile of each sequence (i.e., wt and SNVs) is a noise-corrupted aggregate structural signature of a shared, typically large, underlying ensemble of alternative secondary structures, where each structure’s contribution amounts to its individual structural signature, weighted by its relative abundance. The sequence-specific weights (i.e., relative abundances pertaining to wt and each SNV), the individual structure signatures, and the nucleotide-specific noise levels are unknown and need to be estimated. To jointly deconvolve multiple sequence-dependent mixtures, REEFFIT employs a Bayesian statistical inference framework and an expectation–maximization (EM) algorithm to estimate the most likely values of the unknown parameters given the M² data. In Bayesian inference, one incorporates prior knowledge and/or modeling assumptions by imposing constraints on the statistical behavior of unknown parameters. These take the form of prior probability distributions and serve to direct or stabilize the estimation. In REEFFIT’s model, a structure-context-dependent reactivity model derived from existing data, similar to that employed by Rsample, serves as prior distribution on structure signatures, such that they are based on samples drawn from the model. Due to the large number of unknown parameters, Gaussian priors on the weights serve to reduce overfitting through shrinking weight magnitudes. Additional constraints on the weights are derived from thermodynamic calculations to avoid large discrepancies between estimates and thermodynamics-based predictions of SNV-to-wt weight ratios.

Method details. REEFFIT takes three inputs: structure probing data for the wt and SNVs, the probed RNA sequences and, optionally, a collection of secondary structures. For each sequence, it outputs the predicted weights for all the input structures along with the estimated uncertainties in these weights. One can subsequently mine the input structures for substructural signals using the methods we described earlier. Such analysis can also account for the output weights, for example, by first clustering structures, irrespective of them, and then, for each probed sequence, summing the weights over the structures in each cluster. While users have the flexibility to independently compile the underlying set of secondary structures, REEFFIT also provides two built-in options. The first consists of constructing a large set by first generating up to 200 structures, whose free energies are within 5% of the MFE for each sequence and by then merging all sets. The second further processes the output of option (1) into a small set via clustering and choice of cluster representatives, essentially applying the ‘sample and cluster’ pipeline prior to ensemble modeling. Input data take the form of reactivities, hence REEFFIT is broadly applicable to M² experiments employing various readout and cDNA quantification strategies. However, like Rsample, it relies on a data-driven reactivity model to predict data signatures for input structures, and such models have been shown to be reagent- and/or experimental condition-dependent. Currently, an exponentially-distributed model derived from electrophoresis-based SHAPE in vitro data is coded into REEFFIT. It is thus helpful to verify that the data to be analyzed are consistent with this model. In that context, it is worth noting that REEFFIT was recently applied to a related, yet distinctly different data type collected by a scaled version of the M² strategy, dubbed M²-seq. This sequencing-based technique relies on the MaP readout and further processes reads with multiple mutations into pairwise mutation co-occurrence counts. Prior to analysis with REEFFIT, these counts were pre-processed and subsequently quantile-normalized to bridge differences between the statistical properties of the exponential SHAPE model and the processed DMS-MaP co-mutation counts. In terms of structure models,
REEFFIT currently generates non-nested secondary structures and assumes that user-specified structures also have this property. To date, it has been applied to sequences up to 112 nt long. Since the computational burden depends on RNA length, users can choose between two inference strategies, termed soft EM and hard EM, where the latter is less computationally intensive.

IRIS. **Overview.** IRIS is the only method, to the best of our knowledge, designed to deconvolve RNA structure ensembles from RNA-RNA interaction mapping data.\(^{101}\) It further extends the idea of integrating RNA-RNA interaction data in the form of hard base-pair constraints, previously proposed to direct ‘sample and cluster’ analysis of coexisting alternative structures in Zika virus.\(^{36}\) IRIS takes a 3-step approach. First, the data are used to score all the possible base pairs for the strength of their interactions, as supported by read coverage. Second, MFE predictions are generated with high-scoring base pairs enforced as hard constraints, resulting in thousands of plausible secondary structures that are deemed relatively stable and contain one or two stems supported by the data. To reduce redundancy and similarity among structures, they are clustered into 100 groups. The structure with the lowest free energy in each cluster is retained as representative of the cluster, and the 100 resulting structures are candidates for selection. Third, structure selection and relative abundance estimation are achieved by fitting a linear model that expresses the data-derived base-pair scores as the noise-corrupted mixture of candidate structural signatures. This model follows the same principles that underlie SLEQ and REEFFIT, but it captures the relationship between individual structures and the data differently. Here, structures are represented as sets of base pairs, where only base pairs contribute to structure signatures. To enforce parsimonious predictions, the model is fit to the data in two steps. Initially, least absolute shrinkage and selection operator (LASSO) regression is used to select structures. LASSO is a regularized regression method that yields sparse solutions by constraining the sum of absolute values of regression coefficients, such that coefficients are driven to zero.\(^{102}\) Therefore, IRIS first selects a subset of candidates with non-zero coefficients. Since regularization strength is controlled by a parameter, IRIS applies LASSO with decreasing values until more than \(K\) structures are selected. Next, all subsets of size \(K\) are considered as candidate ensembles. Relative abundances are estimated for each subset by fitting the model via NNLS regression. Finally, a Bayesian framework is employed to determine by enumeration the most likely ensemble given the data. IRIS integrates thermodynamic considerations into this framework by setting a prior distribution on the ensembles, which depends on partition function calculations. It further models the observed base-pair scores as exponentially distributed, depending on each ensemble’s mean base-pair score.

**Method details.** IRIS takes mapped reads and an RNA sequence as input and outputs a set of secondary structures and their estimated relative abundances. The number of structures \((K)\) must be determined by the user. Three additional parameters require user input, two of which control how data are integrated into MFE predictions in the form of forced helices. These parameters vary with the RNA length (see\(^{101}\) for recommended values). The third is the number of structure candidates, set to 100 for all RNAs analyzed to date with IRIS, whose lengths range from 79 to 330 nt. Since input data takes the form of mapped reads, IRIS is applicable to several direct RNA-RNA interaction mapping techniques. However, to date, it has been applied to and optimized for PARIS data.\(^{32}\) Particularly, its Bayesian framework employs a likelihood model derived from PARIS data.

**General limitations of thermodynamics-dependent methods.** One limitation of thermodynamics-dependent methods is rooted in their dependence on prior assumptions in the form of explicit structure models and the fact that the main means to obtain these structures is through thermodynamics-based predictions. Consequently, their performance depends on the ability to sample all the relevant structures, or, in other words, to obtain samples that are well-representative of the folding landscape. For ‘sample and select’ methods, this point is critical since predictions are restricted to structures observed in one round of sampling.\(^{13}\) If a key structural feature is missing from the sample, it cannot be predicted de novo from the data. Although this issue may be alleviated when the data are used to direct both sampling and selection,\(^{30,90,101}\) this remains a concern, especially because it is often challenging to discern situations where key landmarks of the structural landscape are missing and predictions are biased. Developing strategies to evaluate different ensemble compositions given the data (also known as model selection) may help mitigate this issue, although it does not address the root cause, i.e., reliance on misrepresentative structure samples.

Three factors make it particularly difficult to accurately represent the ensemble via Boltzmann sampling: (1) the gigantic number of putative structures and their minuscule probabilities, (2) inaccuracies in the thermodynamic model, and (3) limitations on the structural features that are modeled by conventional secondary structure prediction algorithms (e.g., they do not account for the presence of pseudoknots).\(^{103}\) Moreover, accurate modeling of folding becomes even more chal-
lenging as RNAs increase in length.\textsuperscript{104–105} Model inaccuracies and restricted features are inherent to the current biophysical modeling paradigm, and overcoming these limitations will likely require extensive experiments to refine the model.\textsuperscript{8,106} specialized algorithms,\textsuperscript{82,103} or a paradigm shift.\textsuperscript{107–108} The problem of sampling rare events, on the other hand, may be mitigated to some degree by taking advantage of an ever-expanding computing power to sample deeper. However, this problem also exacerbates when RNAs increase in length due to the combinatorial explosion of possible structures and accompanying growth in structural diversity. For this reason, strategies to increase structural diversity in samples were developed, such as combining data-directed with traditional sampling.\textsuperscript{90} combining samples from a sequence with samples from multiple SNVs of it,\textsuperscript{74,87,95} combining samples directed by distinct subsets of data-derived constraints,\textsuperscript{36,101} and iteratively adapting folding thermodynamics to steer sampling away from structures already sampled repeatedly.\textsuperscript{109}

Additional confounding factors arise from data limitations. Aside from routine data quality considerations common to many genomic assays,\textsuperscript{110–111} the information content of probing data also plays a role in method performance. One challenge is presented by the high variance and ambiguity inherent to these data, manifested as significant overlap between reactivity distributions for paired and unpaired bases observed for structured RNAs.\textsuperscript{13,77,81} Predicting pairing states directly from reactivities has thus been an ongoing challenge, especially for unpaired bases, a substantial fraction of which are unreactive. There are several possible reasons for this, including structural constraints limiting the nucleotide flexibility or the burying of these bases inside a tightly folded tertiary structure, making them inaccessible to the probing reagent. As we described earlier, some methods mitigate this issue in their modeling of the relationship between structures and reactivities. However, completely resolving such discrepancies and avoiding data misinterpretations by computational means alone remains a challenge. Another issue stems from the inability of most experimental techniques to directly detect base-pairing interactions. Rather, susceptibility to modification reveals base-pairing states (paired vs. unpaired), which makes it difficult for methods to distinguish between alternative structures containing nucleotides that participate in base-pairing interactions with distinct partners. Ultimately, the ability to deconvolve a structure mixture accurately largely depends on the degree of differences in pairing states between alternative structures as well as on how informative or accurate the data are at these particular sites.\textsuperscript{112} These in turn depend on the analyzed sequence and underlying relative abundances. Data-directed ensemble deconvolution with respect to explicit structure models is thus a challenging problem that remains open.

Thermodynamics-independent methods

Thermodynamics-independent methods do not make any prior assumption on RNA structure and thermodynamics. Rather they attempt to define clusters of raw sequencing reads, belonging to different structural conformations, in a thermodynamics-independent fashion (Figure 4). To easily understand the underlying principle of these algorithms, we can picture the ensemble as a $n$-dimensional abstract space, where $n$ is the length of the RNA in nucleotides. Each read, corresponding to a single original RNA molecule, is described by a 1-dimensional vector of length $n$. Mutated bases, corresponding to bases of the RNA that were modified by the chemical probe, are represented as ones, while the remaining bases are represented as zeros. In this scenario, reads originating from the same structural conformation will exhibit correlated co-mutation patterns and thus will tend to cluster together. Following the deconvolution of reads, the reconstructed reactivity profiles for the individual structural conformations are typically used to constrain thermodynamics-based RNA structure prediction algorithms (Figure 2). Below we provide an overview of the main methods developed to date, in chronological order.

RING-MaP. Method details. RNA INteraction Groups by Mutational Profiling (RING-MaP) represents the first attempt to deconvolve the reactivity profiles of multiple coexisting structural conformations for the same RNA.\textsuperscript{113} The first step in RING-MaP analysis consists of identifying the number of clusters (or conformations) for a given RNA by spectral clustering. The analyzed RNA is represented as a graph, where each vertex is a base. Vertices are connected by edges, whose weight is proportional to the number of reads in which the two bases were observed to co-mutate. Spectral analysis is then applied to the normalized Laplacian matrix derived from the graph’s adjacency matrix. The aim of this step is to find a partition of the graph into $K$ subsets of nodes, in such a way that the sum of the weights of the cut edges is minimized and the sum of the weights of the uncut edges is maximized. In other words, the algorithm attempts to define clusters of vertices (bases) that are frequently observed to co-mutate, as these likely describe bases that were simultaneously unpaired in a specific structural conformation. This is achieved by looking at the distance between consecutive eigenvalues, the eigengaps, following eigen-decomposition of the normalized Laplacian matrix. In simple terms, each eigenvalue provides a measure of the goodness of a certain cut of the graph at partitioning “structurally-related” bases from the unrelated ones: the smaller the value, the better the partitioning. For an RNA forming $K$ struc-
tures, the eigengap between eigenvalues $K$ and $K + 1$ will be significantly larger than the previous. After having determined the optimal $K$, individual reads are assigned to the individual structural conformations by K-means clustering, further allowing to reconstruct the individual conformation reactivity profiles, and to estimate the relative abundance of each structure.

Method limitations. RING-MaP analysis has three main limitations. First, the number of clusters $K$ is determined by setting an arbitrary threshold of 0.03 on the magnitude of the eigengaps. Although this threshold appears to work reasonably well for the small set of structures RING-MaP has been applied to, it is currently impossible to predict whether, under different experimental conditions, it might lead to under- or over-estimating the number of clusters. Second, reads are hard-clustered by K-means, meaning that each read can only be assigned to a single cluster. However, it is likely that multiple conformations will share a subset of unpaired

---

**Figure 4. Encoding of alternative RNA conformations in cDNA molecules via mutational profiling (MaP).**

Schematic representation of the reads (cDNAs) generated by MaP experiments, when either one or two different structures populate the ensemble. (top) The RNA folds into a single structure. In the depicted structure, positions 11–18-22 are all simultaneously unpaired (or structurally-flexible), therefore they can all be simultaneously probed. After reverse transcription, the modifications are encoded as mutations in the resulting cDNAs, theoretically generating any possible combination of co-mutation patterns of the three positions. (bottom) The RNA folds into two alternative structures. In structure A, positions 11–18 are unpaired (reactive to probing), while position 22 is base-paired (unreactive to probing). In structure B, positions 11–22 are unpaired, while position 18 is base-paired. In this case, not all possible combinations of co-mutation patterns of the three positions will be observed. For instance, as the base-pairing of positions 22 and 18 differentiates structure A from structure B, these two positions will never be observed to co-mutate.
bases, hence possibly originating a subset of reads with similar co-mutation patterns. Third, the method relies on the full coverage of the target RNA by the sequencing reads. Therefore, only shorter RNAs, or small chunks of longer transcripts, can be analyzed by RING-MaP.

**DREEM. Method details.** Detection of RNA folding Ensembles using Expectation-Maximization (DREEM) models the ensemble as a Bernoulli mixture model. Consequently, each read belonging to one of multiple coexisting structural conformations, is assumed to represent a random draw from a Bernoulli mixture model. DREEM uses an iterative expectation–maximization (EM) approach, in which the model parameters are initially randomly initialized and $K$ clusters ($K = 2$ by default) are assumed to be present. During the expectation phase, the algorithm probabilistically assigns the reads to the clusters and it then computes the likelihood of observing the data, given the model parameters. In the Maximization step, the model parameters are re-estimated by maximizing the expected value of the likelihood. In other words, the algorithm attempts to determine the best way to cluster the sequencing reads, by looking for the set of parameters that results in the best fit for the joint probability of the data points (in this case, the reads). In an attempt to avoid overfitting, DREEM applies a stringent Bayesian Information Criteria (BIC) test, to determine the appropriate dimensionality of the model. Basically, after the EM algorithm has converged, the BIC test evaluates whether the data justifies the presence of $K$ clusters. If the test is passed, the number of clusters is increased by one and the EM algorithm starts over. These steps are repeated until the BIC test fails, or $K$ exceeds a user-defined threshold.

**Method limitations.** DREEM suffers from three main limitations. First, as for RING-MaP, in its current implementation the method can only be applied to the analysis of short transcripts or transcript windows that are fully covered by the sequencing reads. Second, by assuming that the sequencing reads are Bernoulli-distributed, the method implicitly assumes that the probability of each base of being chemically modified by the probing reagent is independent from that of any other base in the RNA. However, it is likely that both the introduction of a positive charge on the target base by DMS, or of a bulky adduct on the ribose moiety by SHAPE, might affect the local structural context, hence altering the modification probability of surrounding bases. Third, the maximum number of conformations the algorithm looks for has to be manually defined by the user (two by default) and it is limited to four to reduce the computational time.

**DRACO. Method details.** Deconvolution of RNA Alternative Conformations (DRACO) is inspired by the RING-MaP approach and it attempts to mitigate some of the limitations of RING-MaP and DREEM. Analogously to RING-MaP, spectral analysis is performed to determine the number of coexisting structural conformations in the sample. However, instead of imposing a hard threshold on the minimum magnitude of the eigengaps, DRACO builds a null model by generating multiple perturbed versions of the initial data, by shuffling the mutations within the reads. Comparison of the original eigengaps to the distribution of eigengaps derived from the null model allows identifying the eigengaps having a higher magnitude than expected by chance. Once the number of clusters has been determined, DRACO groups together identical reads (reads presenting the same set of mutated bases) and it then calculates the weight of each base in the read for each cluster. From these weights, an assignment score is calculated and the read group is assigned to each cluster, proportionally to the assignment score. By exploiting this fuzzy clustering approach, if two or more clusters have the same assignment score, implying that the corresponding structures share the same co-mutation pattern, the read group will be equally divided across all of them. Additionally, if the estimated abundance of a cluster is below a user-defined threshold (5% by default), the number of clusters is decreased by one and the fuzzy clustering/weighting step is repeated. To further overcome the limits imposed by short read sequencing and to enable the analysis of long transcripts, DRACO implements a sliding window approach. Briefly, a window of size $\leq$ to the median read length in the experiment is slid along the RNA, selecting at each step only the reads fully covering the window. Spectral analysis is then performed for each window independently and consecutive windows for which the same number of clusters has been detected are merged, before reassigning the sequencing reads. This windowed strategy has made it possible to analyze the entire SARS-CoV-2 genome (~30 kb).

**Method limitations.** The main limitation of DRACO remains the impossibility to determine the structural relationships between distal non-overlapping windows in long transcripts. For instance, if a long transcript would present two short windows, located at opposite ends of the RNA, each one folding into two alternative conformations, it would be impossible for DRACO to determine which combination(s) of structures simultaneously occurred in the ensemble.

**DANCE-MaP. Method details.** Deconvolution and Annotation of riboNucleic Conformational Ensembles measured by Mutational Profiling (DANCE-MaP), has been recently proposed. The approach adopted by this method is largely similar to the one used by DREEM, with two notable exceptions. First, in order to prevent individual posi-
tions from dominating the clustering phase, RNA bases are split into active and inactive, depending on whether their background-subtracted mutation frequency exceeds 0.2% or not. The reactivity data for the active positions is deconvolved first and cluster proportions are determined, followed by a second deconvolution of the reactivities for the inactive positions. Second, a number of model validity tests are performed to discard situations in which the algorithm converges to non-physical solutions. Particularly, thresholds are defined for the minimum conformation abundance, the range of accepted mutation frequencies for the reconstructed reactivity profiles and the minimum difference between the reactivity profiles for different conformations.

Method limitations. DANCE-MaP shares the same limitations of the DREEM method.

General limitations of thermodynamics-independent methods. The main limitation of thermodynamics-independent methods is the incompatibility with data derived from RT drop-off-based experiments, as they rely on the analysis of co-mutation patterns in reads harboring two or more mutations, obtained from MaP experiments, to define the clusters. Additionally, although these methods are thermodynamics-independent per se, the interpretation of the deconvolved reactivity profiles still relies on free energy minimization and, as such, its accuracy depends on how effectively the contribution of nucleotide reactivities to the thermodynamics model is determined. Further limitations of these algorithms are tightly intertwined with those of MaP experiments. First of all, MaP experiments are typically characterized by low efficiencies. These are the consequence of both low RNA modification efficiency by chemical probing and low read-through efficiency at RT. For example, upon treatment with DMS, one of the most efficient chemical probes known to date, an unpaired A or C base will typically show a modification frequency of just 2–10%. It becomes intuitively evident that the efficiency of the chemical probe at MaP plays a major role in determining the efficiency and accuracy of thermodynamics-independent methods. On the one hand, a high MaP efficiency ensures a high signal-to-noise ratio, making the chemical probing signal sufficiently distinguishable from the background. On the other hand, it increases the fraction of reads carrying two or more mutations, as these provide information on the structural relationship between the bases of the RNA in analysis. Among the user-controlled experimental parameters that can help increase the efficiency and accuracy of these analyses, three are of particular relevance: the sequencing depth, the read length and the chemical modification frequency. Increasing the sequencing depth can increase the number of reads carrying two or more mutations, whilst also increasing the confidence that the observed mutations represent genuine chemical modification-induced mutations rather than arising from stochastic PCR or sequencing errors. Similarly, increasing the read length can also increase the number of reads carrying more than one mutation, as well as aid read mapping, but it will increase the risk of accumulating sequencing errors that might be misinterpreted as structural signals. Lastly, increasing the concentration or incubation time of the chemical probe can help increase the modification frequency, but it could lead to artifacts such as overmodification-induced RNA unfolding.

It is also worth pointing out that the deconvolution performed by thermodynamics-independent algorithms most likely yields only a coarse-grained representation of the actual ensemble. Indeed, these methods can only efficiently detect major structural rearrangements, involving a significant number of bases in the RNA, while they cannot pick up minor structural differences, as well as transient lowly populated conformations.

To date, all of the aforementioned methods have only been used in conjunction with DMS-MaP data. This choice is mostly dictated by two factors. First, most SHAPE reagents exhibit modification rates significantly lower than those observed for DMS. Second, the background mutation rate of SuperScript II, the reverse transcriptase typically used in SHAPE-MaP experiments, is nearly one order of magnitude higher than that of TGIRT-III, the reverse transcriptase typically used for DMS-MaP experiments.

Future directions

The ability to deconvolve RNA structure ensembles is poised to dramatically improve our understanding of RNA structure dynamics in living cells. We have here discussed a number of methods, some of which we have dubbed thermodynamics-independent because of their ability to reconstruct the individual reactivity profiles of the conformations making up the ensemble by direct clustering of the sequencing reads. Nonetheless, interpretation of the deconvolved reactivity profiles still relies on the ability of available tools for free energy minimization to accurately incorporate these experimental constraints into RNA structure models. Importantly, all these tools lack the ability to coherently incorporate multiple sources of experimental data (i.e. SHAPE, DMS, direct RNA-RNA interaction mapping, etc.) into a single predicted RNA structure model. A recently proposed approach, IPANEMAP, attempts to fill this gap by randomly sampling a combination of the Boltzmann ensembles generated by using multiple experimental datasets as constraints, followed by structure clustering and selection of the repre-
sentative structures. The method showed good performances on a small set of test RNAs. Nonetheless, this method suffers from the same limitations of the thermodynamics-dependent approaches we described earlier, as its accuracy strictly depends on the ability to effectively sample the Boltzmann ensemble. In other words, the native structure for the RNA under analysis might never be sampled from the ensemble. To potentially mitigate this issue, the RNAxplorer method has been recently introduced to enable a more efficient exploration of the RNA conformational space and to more efficiently sample underrepresented structures from the ensemble.\textsuperscript{109}

To date, thermodynamics-independent methods have solely been used in conjunction with DMS probing data. While the use of DMS has a considerable advantage in terms of signal-to-noise ratio with respect to traditional SHAPE reagents, it can only provide information for roughly 50% of the bases of the transcriptome, as it can only probe A and C bases under standard conditions. As a consequence, small yet highly dynamic regions in RNA molecules can be overlooked because of the lack of information. The authors of DANCE-MaP partially addressed this issue by performing probing at pH 8.0. Under these conditions G and U bases are transiently deprotonated, hence allowing them to be partially modified by DMS, although with 10-fold lower efficiency as compared to A and C bases.\textsuperscript{83} Hopefully, the recent introduction of novel SHAPE reagents with superior signal-to-noise ratios, such as 2A3,\textsuperscript{26} will provide the means to deconvolve RNA structure ensembles by fully taking advantage of the structural information on all 4 bases.

Irrespective of the employed method, one of the biggest challenges when it comes to deconvolving RNA structure ensembles is the efficient reconstruction of the structural landscape for long RNAs. As previously mentioned in the context of DRACO,\textsuperscript{116} thermodynamics-independent methods are largely limited by the maximum read length that could be achieved on Illumina-based instruments, hence making it impossible to determine the relationship between two structurally heterogeneous regions located far apart. The rapid improvement of long-read sequencing technologies, such as Oxford Nanopore, will likely help overcome this limitation,\textsuperscript{119} provided that the base-calling accuracy would be improved as well. Indeed, the high background mutation frequencies typical of Nanopore-based experiments would introduce significant noise, hence negatively affecting the performances of the thermodynamics-independent methods we previously described.

It is also worth mentioning that, as previously commented, thermodynamics-independent methods are not sensitive enough to detect small structural differences and, as such, the set of reactivity profiles deconvolved by these methods likely represents an aggregate of the reactivity profiles of many similar, yet distinct, structures. In this regard, the combination of thermodynamics-independent and thermodynamics-dependent methods might help refining the deconvolved structures, hence providing a more in-depth understanding of RNA structure ensembles. A first step in this direction has been taken by the authors of DANCE-MaP, who used the reactivity profiles deconvolved by their thermodynamics-independent method to calculate base-pairing probabilities for the identified structural conformations, rather than simply interpreting the data as a single structure.\textsuperscript{117}

Acknowledgements

D.I. was supported by the Dutch Research Council (NWO) as part of the research programme NWO Open Competitie ENW - XS with project number OCENW.XS22.1.015 and by the Groningen Biomolecular Sciences and Biotechnology Institute (GBB, University of Groningen).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received 14 March 2022; Accepted 5 May 2022; Available online 17 May 2022

Keywords: thermodynamics; partition function; clustering; SHAPE; RNA structure ensemble

References


genomewide or targeted RNA structure probing in vivo. Nat. Methods 14, 75–82. https://doi.org/10.1038/nmeth.4057.


