Title

Accurate, Automated, and Scalable Identification of RNA Structure Motifs in Structurome Profiling Data

Author

Radecki, Pierce

Publication Date

2021

Peer reviewed|Thesis/dissertation

Accurate, Automated, and Scalable Identification of RNA Structure Motifs in Structurome Profiling Data

By

Pierce Radecki

Dissertation

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Biomedical Engineering

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

Sharon Aviran, Chair

Volkmar Heinrich

Vivek Srinivasan

Committee in Charge

2021

To my parents and siblings.

Contents

List of Figures

List of Tables

Abstract

Accurate, Automated, and Scalable Identification of RNA Structure Motifs in Structurome Profiling Data

RNA is a key biopolymer that mechanistically drives many cellular processes. As a single-stranded molecule with a flexible sugar-phosphate backbone, it can fold into intricate structural conformations. The functions, interactions, and regulations of RNA are often directly attributable to these structures; as such, understanding structure is crucial to deciphering the mechanisms of RNA function and dysfunction. High quality structure models can be obtained with nuclear magnetic resonance (NMR) and X-ray crystallography. However, these methods are low-throughput, encumbered by technological limitations, and lack applicability *in vivo*. In recent decades, structure profiling (SP) experiments have emerged as a practical and scalable approach to measure the structures of RNA transcripts in their *in vivo* contexts. These methods work by exposing transcripts to chemical reagents that induce covalent modifications in a structure-dependent manner. Modifications can be mapped by high-throughput sequencing, resulting in nucleotide-wise measurements of stereochemical characteristics. SP experiments have now scaled to the level of the entire transcriptome, enabling structure studies with unprecedented scope and depth. However, the data from these experiments have been largely underutilized due to a lack of computational tools capable of readily processing their massive scale when linking structure to function.

This dissertation focuses on the development of methods to interpret transcriptomic structure profiling data. I devise a novel statistical model of SP data and couple it to a data-driven structure recognition algorithm, yielding an accurate, automated, and scalable tool for identifying structures and structure-function relationships. Application of the method to diverse datasets demonstrates its utility in several domains. Specifically, it reveals novel insights on mRNA structure dynamics, characterizes structures within viral RNA genomes, profiles the RNA-protein interactome, and links specific structure motifs to post-transcriptional regulation. The method is adaptable for future types of profiling experiments and readily scales to the evolving scope of structure studies. Altogether, these results have helped further our understanding of *in vivo* RNA dynamics and provide the RNA community with a versatile tool to assess the transcriptomic structural landscape.

Acknowledgements

I am grateful to my Ph.D. mentor, Dr. Sharon Aviran, without whom none of this would have been possible. Thank you for welcoming me into your lab, guiding me through graduate research, challenging me to always give my best, and allowing me to explore my scientific curiosities. I would also like to express my immense gratitude to the other members of my committee—Dr. Volkmar Heinrich and Dr. Vivek Srinivasan—for taking the time to provide guidance, feedback, and support through the journey of doctoral research. From the beginning of my time at UC Davis, these individuals, among many others, have exuded an infectious passion for science, engineering, and medicine. For that, I am eternally thankful.

Next, I would like to thank all the lab mates that I have had the privilege of working with and learning from: Sana Vaziri, for long discussions and camaraderie through all of the challenges one faces in graduate life. Krishna Choudhary, for scientific leadership, guidance, and friendship. Thank you for sharing numerous lessons learned and fostering the growth of everyone around you. Mirko Ledda, for far more than I can state here. You patiently introduced me to RNA structure methods, taught me the technical foundation, and inspired a research space to explore and grow with. Thank you for the many caffeine-fueled discussions on just about everything, for the supportive friendship, and for having such a great time working together.

Thank you to my previous mentors, whose guidance and support enabled my pursuit of a graduate degree. To Dr. Michael Savageau, thank you for supporting my research at UC Davis and guiding the start of my journey through graduate school. To Dr. Jeffrey Cohen and Dr. Wei Bu at the National Institute of Allergy and Infectious Diseases, thank you for taking a chance on me and for giving me an incredible experience as an intern in your lab. To Kelvin Tan and Dan Wilkin from BrainScope, Inc., thank you for giving me an opportunity to learn more about software development and for supporting my pursuits in graduate studies. To Dr. Benjamin Shapiro at the University of Maryland, thank you for taking the time to introduce a young undergraduate the ways of independent research.

Thank you to all of my close friends who have shaped my journey: Sina, Ned, Dara, Austin, Ardalon, Jay, Mitesh, Pooja, Richard, Josh, Mark, the Brekkes, and everybody else.

Finally, I would like to thank my entire family. Mom and Dad, thank you for your everlasting support of my endeavors and for inspiring me through every challenge I encounter. Tyler and Leanne, I could not have asked for better role models to grow up with. Your encouragement gives me strength. Without all of you, none of this would be possible.

Chapter 1

Introduction

The diversity of life on Earth is exceptional [129], yet all known organisms follow a remarkably repetitive recipe of biochemistry as described in the Central Dogma of Biology [26]. The various shapes, sizes, and forms of life are comprised of cells that control the genetic flow of information via three central macromolecules: DNA (deoxyribonucleic acid), RNA (ribonucleic acid), and proteins (polypeptides). In this standard summarization, genetic information is stored as nucleic acid sequences in DNA, which are then transcribed into RNA, and finally translated into proteins. The universality of this dogma as it applies to all living organisms underpins the relevance and power associated with understanding these three fundamental biopolymers and the relationships between them.

The standard perspective of the central dogma viewed RNA as an intermediary molecule that carries genetic information from DNA sequences into their protein products. Since the 1960s, however, our understanding that RNA can serve phenotypic roles in biology has blossomed significantly. At first, it became apparent that the role of RNA was far more involved than just carrying genetic information as messenger RNA (mRNA), as non-coding RNA transcripts with the capability of catalyzing reactions were identified [87]. In more recent years, RNA has been discovered to serve mechanistic roles in nearly every domain of the cellular program—it regulates genes, controls splicing, dictates post-transcriptional modifications, catalyzes reactions, maintains chromosomal structures, and much more. Recent studies quantifying the total pool of cellular RNAs estimated that mRNA rarely constitutes more than 15% of cellular RNA mass [141]. Typically, ribosomal RNAs (rRNA), nucleic acid biomachines that are a central component of the translation machinery, comprise over 50% of a cell's RNA, with most of the remaining 35% comprised mostly of non-coding RNAs (ncRNAs). This large fraction further emphasizes the functional and complex roles of RNA and underscores its funda-

mental evolutionary origins. Pioneering work by Alexander Rich arrived at the discovery of the RNA double helix, the discovery of polyribosomes, and eventually the hypothesis of an ancestral "RNA world," [157, 143] where RNA served as the original biomolecule responsible on its own for orchestrating all cellular functions, including reproduction, metabolism, and adaptation. On-going research has refined the description of this hypothesis, expanded the scope of RNA's functional landscape, and provided theoretical mechanisms for the formation RNA nucleotides in primordial conditions [63]. Although the precise role of RNA in the formation of life on Earth is still not well-understood, the centrality of this biopolymer is difficult to overstate.

1.1 RNA Structure

At the mechanistic core of RNA's diverse functions is its ability to fold into and interchange between specific structural conformations. RNA structure is hierarchical and begins with a molecule's primary structure: its nucleotide sequence. Each nucleotide is comprised by a nitrogenous base, ribose sugar moiety, and phosphate group (backbone). There are four nitrogenous bases in RNA: cytosine (C), uracil (U, corresponding to DNA's thymine, T), guanine (G), and adenosine (A). Unlike DNA, RNA is exceptionally flexible, meaning that the directional strand can fold in intricate ways. As such, its secondary structure is driven by the complementarity of nucleotide bases within transcripts that allow for hydrogen bonding (i.e., base-pairing) between part of the molecule (see Figure 1.1). Similarly to the DNA "Watson-Crick" base pairs—G-C and A-T—RNA nucleotides can also interact with their complementary partners. For RNA, which utilizes uracil instead of thymine, G-C and A-U are complementary, and there's also a third possible base pair, G-U, that is referred to as the "Wobble" base pair. Intramolecular base-pairing enables the formation of stable local structure elements, such as hairpin loops and helices (see Figure 1.2). These domains can further interact with each other in the form of tertiary interactions, and they can also interact with other molecules (other RNAs, proteins, DNA, ligands, etc.), described as quaternary interactions.

In recent decades, the connection between RNA structure and function has been demonstrated in a large number of contexts. For example, thermosensors use their structure to respond to changes in temperature [133, 158], riboswitches enable gene control changing their structure in response to specific ligands [11, 124, 203], splice sites in RNA transcripts utilize structure to control alternative splicing [44, 35, 161], long non-coding

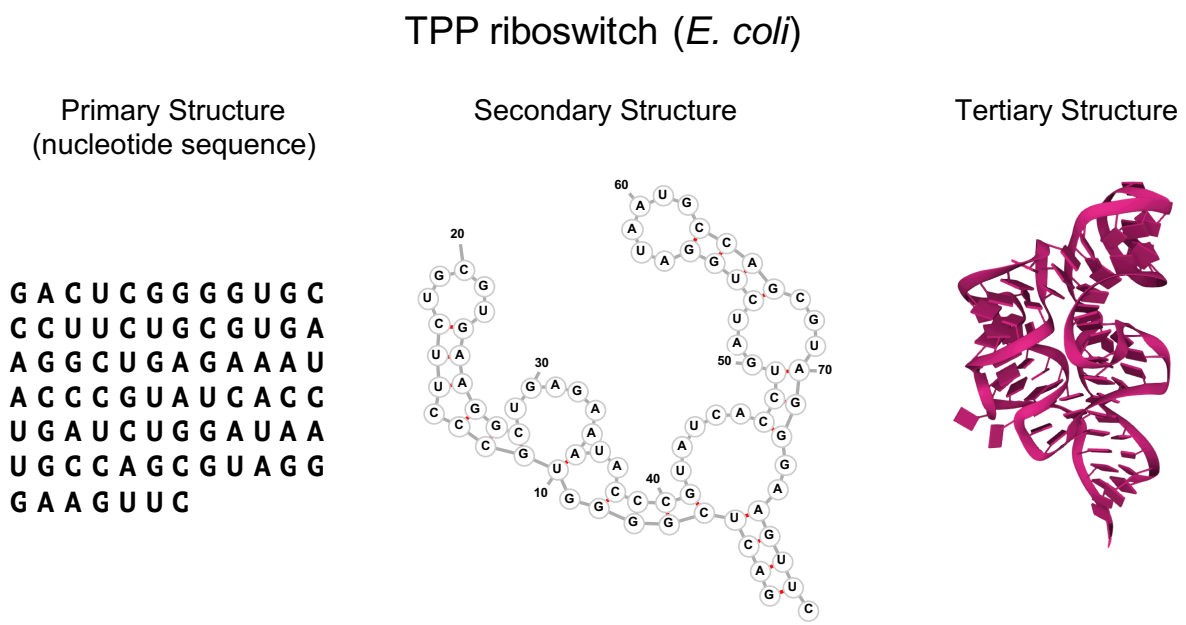


Figure 1.1: Illustration of the RNA structure hierarchy for an example transcript, the TPP riboswitch. The primary structure of any RNA is simply its linear sequence of nucleotide bases (left). The secondary structure (center) describes conformations as a set of base pairs between complementary sequence partners (i, j) within the molecule. Tertiary structure (right) is described by the three-dimensional arrangement of stems, loops, and helices within the molecule, as well as the interactions between such local domains. (PDB: 4NYA) [196]

RNAs like Xist regulate chromosomal silencing [17], mRNAs use structures to facilitate the RNA-protein interactome [24], viral genomes use structure for packaging, transport, and self-regulation [113, 155], and much more [168]. As such, understanding RNA structure will continue to help elucidate the biological functions of RNA transcripts.

1.1.1 Measuring RNA Structure

The biological relevance of RNA and the importance of accurate structure models have yielded extensive work studying them. Before discussing methods for RNA structure prediction, it is important to describe the process by which a structure is accurately determined experimentally. For atomic-resolution structure models, three primary procedures are used. The first of these is X-ray crystallography [68], which uses crystallized samples and X-ray diffraction to infer the structural conformation. The second procedure is nuclear magnetic resonance (NMR), during which magnetic fields are utilized to probe atomic nucleic by perturbing them [99, 46]. Lastly, cryo-electron microscopy (cryo-EM) enables atomic resolution structure models by combining microscopy with cryogenic temperatures and software processing to determining structures without the need for

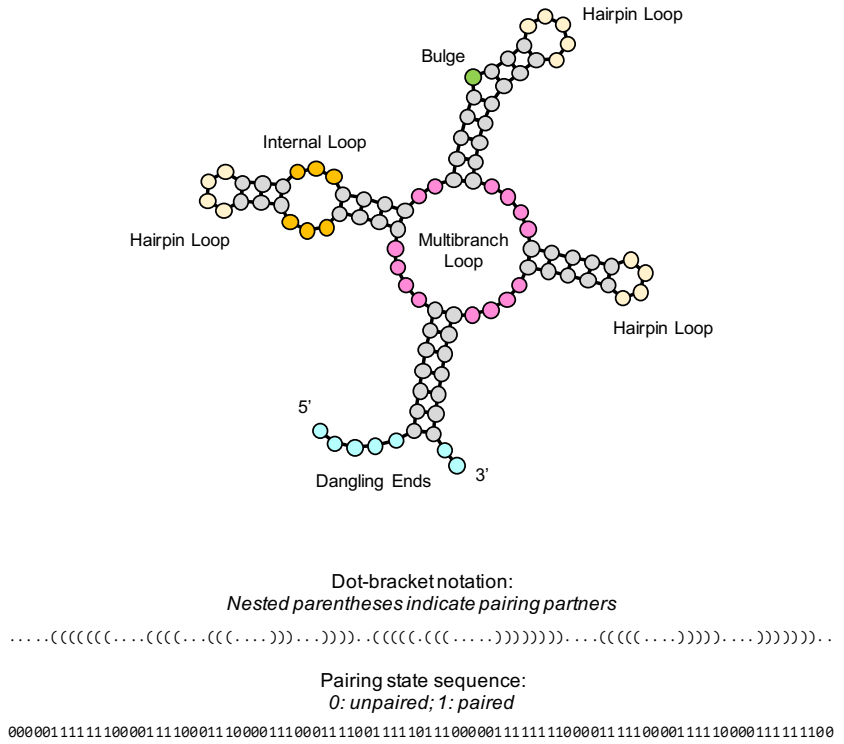


Figure 1.2: Illustration of some common RNA secondary structure motifs. Shown is an example structure with paired bases indicated in grey, hairpin loops indicated in ivory, internal loops indicated in orange, bulges indicated in green, multibranch loops (or junctions) indicated in pink, and dangling ends (also referred to as exterior loops or "single-stranded regions") indicated in cyan. The dot-bracket representation, which is often how secondary structures are encoded, is provided; nested parentheses match together to indicate the base-pairing arrangement. Structures can also be represented in a binary pairing state sequence, which simply indicates the pairing state of each nucleotide without information on the pairing partner. These notations are sometimes combined to represent structures with ambiguity in their known structure or interactions like RNA-protein binding or tertiary contacts.

crystallization [44, 211]. Predicting the structure of an RNA usually amounts to predicting the specific arrangement of base pairs in the molecule. Despite the direct and robust information acquired via these methods, they come with some critical faults. Namely, they are difficult, labor intensive, low-throughput, and restricted to short (e.g., 10-300 nt) and stable transcripts. They do not work effectively for RNAs that adopt multiple conformations. Perhaps most importantly, thought, these methods cannot provide structural information on RNA molecules within the native *in vivo* cellular environment. As a consequence, alternative methods of structure profiling have been developed which are less direct yet offer high-throughput scalability as well as measurements inside of in living cells. Such methods are central to this dissertation and therefore given special treatment in Section 1.1.3.

1.1.2 Predicting RNA Structure

The difficulties associated with directly measuring RNA structure yielded a variety of computational approaches to make predictions without the need for low-throughput experiments. After the 1980s, two primary avenues for structure prediction emerged: (1) sequence-based thermodynamic structure prediction and (2) comparative sequence analysis. In this section, we give a brief introduction to both.

Thermodynamic Structure Prediction

Sequence-based thermodynamic structure prediction is founded on the objective of accurately determining an RNA's structure from its sequence alone. Indeed, as the set of possible base-pairing arrangements in an RNA stems directly from the base-pairing compatibility of its nucleotides, the nucleotide sequence is the foundation of secondary structure formation. As such, sequence-based structure prediction is able to achieve modest accuracy in many contexts; however, as we will see in Section 1.1.3, nucleotide sequence alone is insufficient for making accurate predictions in a generalized context [48].

Thermodynamic structure prediction algorithms attempt to model the energetic stability of RNA conformations in terms of overall Gibbs free energy (ΔG) [39]. Free energy is assessed via a nearest-neighbor thermodynamic model (NNTM), which describes the free energy changes associated with RNA substructures such as hairpin loops, stacked bases, and dangling ends [186]. These models are informed via optical melting experiments in idealized conditions. Conformations which are more stable have lower overall free energy, and vice versa. That said, there are a large number of possible conformations for any feasible RNA sequence (in fact, the number of possible conformations grows exponentially with the length of an RNA). In practice, the distribution of conformations for an RNA given an NNTM follows a Boltzmann distribution, which is expressed as

$$p_i = \frac{e^{-\Delta G_i/kT}}{Z}, \text{ where } Z = \sum_i e^{-\Delta G_i/kT} \tag{1.1}$$

In this equation, p_i is the probability of observing an RNA in conformation i, ΔG_i is the Gibbs free energy of conformation i, k is the Boltzmann constant, and T is the temperature of the system. Z is referred to as the partition function. The equation naturally arrives at the conclusion that, for a specific molecule, as the Gibbs free energy of a structure increases, its probability in the Boltzmann ensemble decreases. Moreover,

the conformation with the lowest free energy will be the most prevalent. Thermodynamic structure prediction algorithms leverage this interpretation by aiming to identify the specific conformation that has the lowest free energy. This structure is referred to as the minimum free energy (MFE) structure.

In theory, the identification of the MFE structure would be straightforward if the free energy of all possible conformations was exhaustively known—simply sort the structures by free energy and select the one with the lowest ΔG_i. In practice, however, this problem is more challenging. At the center of this challenge is the computational intractability associated with exhaustively enumerating and modeling all possible conformations. As the number of possible conformations grows exponentially with the length of an RNA, there may be more conformations to consider than atoms in the universe. More sophisticated computational approaches that do not exhaustively model each possible conformation were warranted.

This difficult problem was solved by dynamic programming [137, 217]. In short, the dynamic programming solution works by recursively finding the most stable local structures in sub-fragments of the original strand and then integrating these assessments at progressively longer scales until the complete RNA structure model is computed. Rather than scale exponentially, *in silico* determination of the MFE structure was demonstrated to scale as $\mathcal{O}(L^3)$ with the length of a transcript. This cubic scaling is still not ideal (and still yields a generally computational limitation to transcripts less than 5000 nt in length), but was a profound improvement over an exponential algorithm associated with considering every individual possible conformation.

Although MFE structures provide insight on the structural conformation of RNA transcripts in biological systems, this approach is generally insufficient in accurately pinpointing biologically relevant structures or in elucidating biological structure dynamics [67, 33, 31]. This is especially true for transcripts that do not see a dominant unique structure in the Boltzmann ensemble. For RNA, which is now understood to often adopt and interchange between multiple relevant conformations, it quickly became clear that MFE predictions alone were fundamentally inadequate when modeling biologically relevant structures. Therefore, there was an emergence of structure prediction methods that explore structures in the suboptimal Boltzmann landscape [206]. Usually, such methods function by sampling suboptimal structures according to their Boltzmann probability. As such, this enables the generation of statistically representative structure ensembles in which alternative conformations and structural dynamics can be more robustly assessed

[34, 217]. Note, however, that these computations depend on a more intensive calculation of the partition function, Z, than MFE-folding alone. Despite the computational overhead, this is currently the state-of-the-art approach used to characterize the structural tendencies of an RNA sequence.

NNTM-based structure prediction methods yielded modest accuracies when applied to short and stable RNAs. However, the methods are generally insufficient when applied in biological contexts. Specifically, the methods are founded on thermodynamic parameters obtained via *in vitro* experiments, limiting their applicability *in vivo*. Beyond that, they do not naturally account for biologically relevant interactions like RNA-protein binding or interactions with other ligands. Moreover, they are poor at resolving long-range intramolecular interactions, such as tertiary contacts, cyclization bonds, or pseudoknots. Finally, their computational paradigm results in prohibitively long computations for longer RNAs, which are prevalent in biology. As such, their most meaningful applications are restricted to low or medium-throughput analyses in an idealized *in vitro* context [48].

Comparative Sequence Analysis

Comparative sequence analysis predicts structures based on sequence homologies of an RNA across various species [140, 58, 49]. Many methods exist which differ in their algorithmic implementation and objectives, but they all follow a common rationale that functional RNA structures (i.e., base pairs) should be evolutionarily conserved across a wide range or organisms. In other words, nucleotides involved in base pairs should see reduced mutation rates compared to surrounding nucleotides under the assumption that mutations which affect the structure would yield reduced fitness or non-viability. In practice, this evolutionary biases can be detected in sequence homologies and computationally analyzed to arrive at biologically relevant structure predictions.

In all types of comparative sequence analysis, phylogenetic information from multiple sequences is utilized to construct a structure model. In a simple workflow, homologous sequences are aligned and mutual information from mutation rates is utilized as an indicator of base-pairing. While this method is successful in arriving at an accurate structure model, this strategy suffers from a need of a sequence homology both sufficiently large and homologous to reliably make sequence alignments, yet diverse enough to robustly extract phylogenetic information from evolutionary changes [85, 9, 66]. Such data are not typically available when performing targeted studies on individual RNA transcripts. Even

with such data, the alignment step requires manual oversight, rendering the approach relatively low-throughput. That said, comparative sequence analyses extract information that is directly biologically relevant (in that evolutionarily conserved base pairs reflect a functional biological role). In the context of NNTM-based structure prediction, this feature cannot be overlooked.

More sophisticated comparative sequence analyses have also been developed, including a method that simultaneously aligns both RNA structures and sequences when determining a structure model [163]. This method is particularly fruitful when aligning disparate sequences that would otherwise be impossible to accurate align by their sequences alone. However, the method depends on the knowledge of at least some reference structures of the studied RNA. Despite decades of research and the universal need of accurate structure models, such references remain rare due to the experimental difficulty associated with obtaining them.

1.1.3 Structure Profiling Experiments

The low-throughput and specialized nature of RNA structure prediction methods have yielded an entirely new field of experiments referred to as structure profiling (SP) experiments [84, 184]. The succinct objective of these methods is to experimentally measure the structural characteristics of RNA molecules at nucleotide resolution by using chemical or enzymatic reagents. The result of such methods is typically termed a reactivity profile; more details are provided in the following paragraphs.

Figure 1.3 demonstrates the general schematic behind SP experiments in arriving at reactivity profiles. The process involves a sequence of several distinct steps which are referred to here as (1) structure probing, (2) reverse transcription, (3) library preparation and sequencing, (4) read mapping and stop counting, and (5) reactivity calculation.

In the first step, RNAs are exposed to the structure-sensitive reagent (sometimes referred to as the probe); see right pathway of Figure 1.3. Although both chemical and enzymatic probes have been utilized, chemical probes have seen dominant popularity in recent years [20, 205, 204, 195, 117]. The reagent modifies parts of the RNA in a structure-dependent manner; typically, unpaired or accessible bases react more strongly than paired or constrained residues. Nucleotides where the reaction occurs, however, see the formation of a covalent modification referred to as an "adduct." As such, unpaired nucleotides tend to be more likely to experience adduct formation when exposed to the structure-sensitive reagent. Adducts are subsequently detected via reverse transcription. In short, reverse

transcriptase (RT) is utilized to synthesize complementary DNA (cDNA) fragments to the probed RNA molecules. Locations on the RNA which saw adduct formation induce reverse transcription termination, meaning that cDNA fragments will stop at nucleotides modified by the reagent (note that in more modern approaches, adducts induce mutations in cDNA at these sites instead of truncations, creating a more robust signal across entire molecules). The resulting cDNA fragments are then extracted, processed, and sequenced, and the number of cDNA fragments ending at each nucleotide is counted (RT stop counts). These counts are compared to an untreated control experiment (see left pathway of Figure 1.3) in order to detect nucleotide modifications that impacted RT drop off more than is seen naturally. Excess counts when modifying RNA are indicative of structure-sensitive modification. The difference in stop counts between the probed and untreated samples are utilized to finally compute reactivity profiles, which represent the accessibility of molecule segments at the nucleotide-level.

These experiments offer specific advantages within the context of RNA structure prediction methods described earlier in this chapter. Namely, they can be conducted at small (targeted analysis of few RNAs) or large scales (tens of thousands of transcripts, or more). Perhaps most importantly, they can be applied directly in living cells. This enables quantitative assessment of *in vivo* structural dynamics, something for which NNTM-based structure prediction was shown to be insufficient. As of the time of writing, SP experiments are currently the most practical way to obtain a direct structural snapshot of RNA transcripts in their natural environments.

Precise experimental details and protocols depend on the probe selected, the structure probing context (i.e., *in vitro* or *in vivo*), their scale and scope, as well as the biological species [20]. Generally speaking, all protocols share common principles but differ in specific chemical mechanisms utilized as well as the specific stereochemistry measured. Commonly used protocols include FragSeq [188], PARS [81], and a family of chemical methodologies such as DMS footprinting [184, 216, 185] and SHAPE (2'-hydroxyl acylation analyzed by primer extension) [126, 205, 200, 177]. As a consequence of this diverse set of SP methodologies, several strategies also exist for the processing of sequencing data and computation of reactivities [20]. As such, normalization is a key problem in the field, as reactivities between experiments have disparate statistical properties and even span disjoint intervals, sometimes even for the same RNA, occluding fair comparison of data between studies.

Reactivity profiles contain rich information on the structural landscape of RNAs in-

cluded in the experiment. As described, higher reactivities are associated with a higher likelihood of being unpaired, and vice versa for paired bases. Although reactivity data do not typically indicate the base-pairing arrangement of paired nucleotides, their information was shown to be extremely impactful when integrated with NNTM-based structure prediction methods [29, 59, 109]. In order to integrate SP data with thermodynamic models, the idea of a pseudo-energy transformation of reactivities was developed. In this model reactivities are converted into a pseudo-energy term via a log-linear formulation. For example:

$$\Delta G_i^p = m \log(1 + y_i) + b, \tag{1.2}$$

where y_i is the reactivity at nucleotide i and m and b are the parameters of the log-linear model (typically set to $m = 1.8$ kcal/mol and $b = -0.6$ kcal/mol for SHAPE data [59]). Pseudo-energies are then included as penalties when using the dynamic programming algorithm with NNTM to identify the most favorable structures; in other words, high reactivities yield a large penalty for paired bases, and vice versa for low reactivities. These terms help coerce the folding algorithm to the specific conformation as measured in the experiment. Note, however, that reactivities provide a snapshot on the structural ensemble (i.e., the collection of all conformations for an RNA in the experiment), which is often not dominated by a single structure. Reactivities are typically viewed as a weighted average over the ensemble, and as such, incorporation of reactivity in folding algorithms needs to be considered carefully. In recent years, methods dedicated to ensemble dynamics have sought to disentangle the structural ensemble from single reactivity profiles with demonstrated success in some applications [106].

Importantly, the log-linear transformation of reactivity into pseudo-energies is itself parameterized based on a jack-knife approach that optimized structural predictions when using SHAPE data to predict a set of highly-structured RNAs. Although it works well in this case (often yielding structure models with >90% accuracy), it is not generalizable to all types of SP data. In order to reliably utilize the method in processing reactivities from different experiments, a computationally intensive calibration step is necessary. In this step, parameters m and b are re-optimized for the obtained data using reference structures. Circumventing the calibration step can result in incorrect and unstable predictions. In general, the disparate statistical properties of different SP datasets render their analysis difficult, and there is a lack of tools capable of automatically making structural

interpretations from them. This problem was partially solved with a likelihood-based model of pseudo-energy transformation [39], however this approach still requires its own form of calibration in the form of a statistical model.

Lastly, it is worth noting here that structure profiling experiments do not circumvent or alleviate the computational limitations associated with NNTM-based folding. SP data greatly improve NNTM-based predictions [29, 59], but the computational overhead associated with folding large numbers of long RNAs with NNTM remains burdensome. Now that SP experiments have scaled to the level of the human transcriptome (i.e., probing tens of thousands of transcripts *in vivo* simultaneously), experimentalists are generating massive amounts of rich structural data that are reinventing how we perceive the RNA structurome [132, 187, 161, 209, 35] and interactome [25]. NNTM-based folding routines are not designed to scale to data of this size, requiring weeks of compute time in order to analyze a single transcriptome-wide dataset. The field has typically resorted to *ad hoc* local folding schemes to circumvent this limitation, but this can drastically limit the impact of detected structure trends at both local and global levels. As such, methods capable of rapidly and automatically assessing structure in massive SP datasets have been warranted.

1.2 Dissertation Overview

The dissertation contains three core chapters which all relate to the development and application of a computational tool, *patteRNA*, which aims to circumvent the computational challenges associated with folding transcriptome-wide data by opting for an pattern recognition approach that mines SP data for specific structural elements. Chapter 2 describes the formulation of several automation-related improvements to the method, which facilitate comparative and integrative analysis of *patteRNA*'s predictions across different datasets and different motifs. Chapter 3 further improves the precision and speed of the algorithm by devising a novel unsupervised training scheme based on discretized reactivities. This chapter also introduces a new application of *patteRNA*'s predictions in the form of a metric quantifying structuredness of transcripts at the nucleotide-level. Chapter 4 focuses on augmenting and optimizing the method's scoring routines by incorporating NNTM-based sequence information in the form of a machine learning classifier. Lastly, we conclude in Chapter 5 with a brief summary, remarks on the outlook of RNA structure profiling data analysis, and outlooks on future method development.

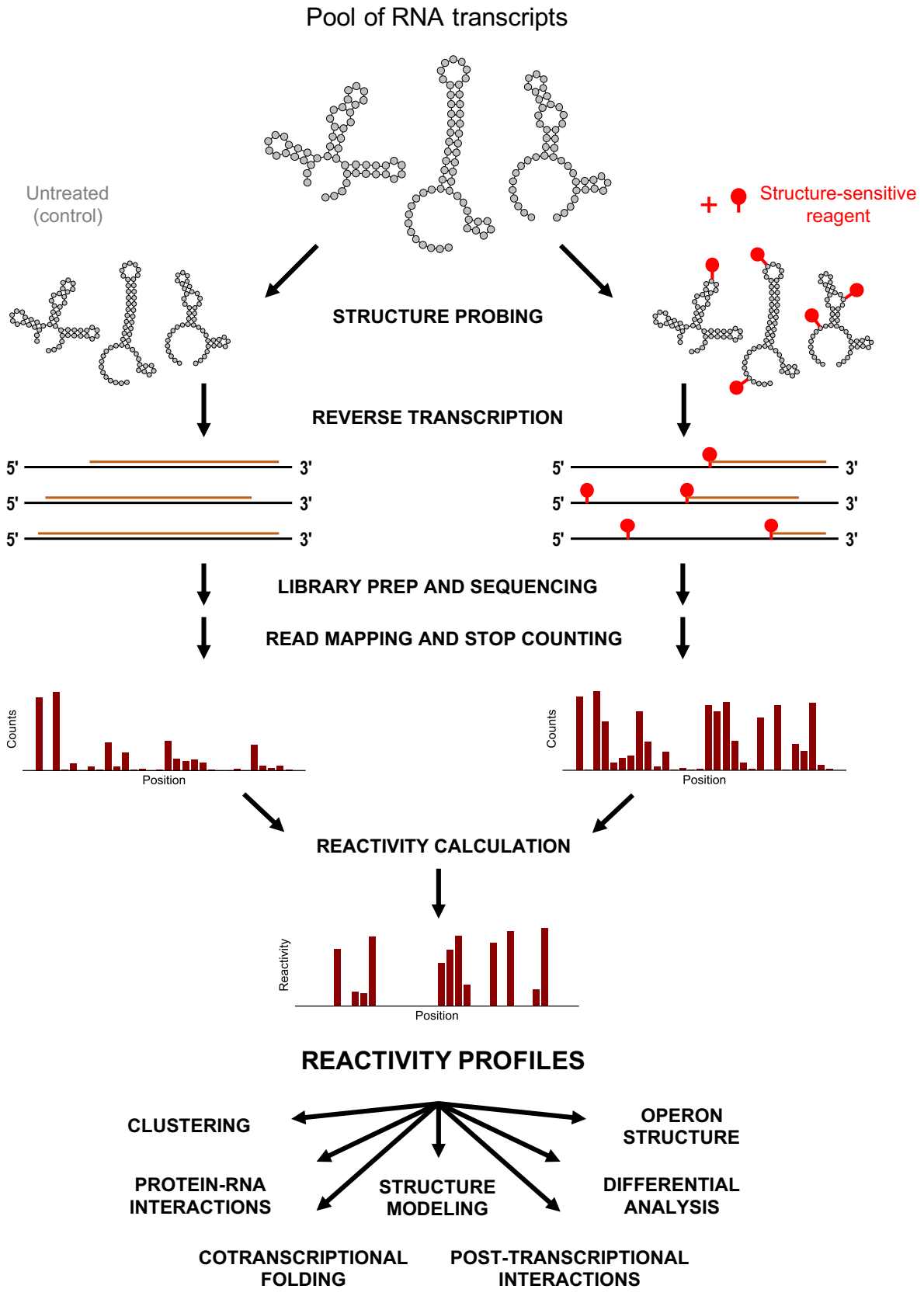


Figure caption on following page.

Figure 1.3: Overview of standard structure profiling (SP) experiment workflows. Experiments begin with a pool of RNA transcripts from which one seeks to obtain structural information. In traditional workflows, the RNAs are exposed to a structure-sensitive reagent that preferentially interacts with unpaired segments of the molecule. At sites which chemically interact with the reagent, adducts—covalent modifications—form. After exposing RNA to the reagent, transcripts are then processed for sequencing. Reverse transcriptase (RT) is used to reverse transcribe complementary DNA (cDNA) fragments along the RNA transcripts; sites which saw adduct formation either stop reverse transcription or induce mutations at the modified site. The number of cDNA fragments which end at each nucleotide (e.g., "counts") in the transcript are calculated and then compared to an untreated control experiment lacking the structure-sensitive reagent in order to compute final reactivity profiles. In short, nucleotides which saw more RT stop counts or mutations when compared to the untreated control are deemed reactive. These profiles quantify the accessibility of RNA segments at the nucleotide-level and provide rich information on the structural conformation of RNAs. Reactivity profiles can subsequently be utilized in a plethora of ways depending on the objective of the study.

Chapter 2

Automated recognition of RNA structure motifs by their SHAPE data signatures

Acknowledgement: *This chapter is reproduced from a published article in the journal Genes (Genes (2018) 9(6), doi: https://doi.org/10.3390/genes9060033, [149]). Pierce Radecki, along with Mirko Ledda, was lead author on this manuscript. Mirko Ledda was a Ph.D. candidate in the Aviran Lab. Author contributions are listed at the end of the capture. Reprinted in accordance with terms of the Creative Commons Attribution 4.0 International License.*

2.1 Introduction

RNA is one of the most important molecules for the formation, evolution, and regulation of life [38, 63]. Although it is known that RNA serves important roles at nearly all levels of cellular function, the fundamental role of RNA in biological systems has remained constant: to encode genetic information, regulate genes and serve as a catalyst of biochemical reactions [45, 47, 38, 168]. Within these contexts, the ability of RNAs to fold into specific structures is critical. For instance, the functions of thermosensors, riboswitches, aptamers, G-quadruplexes, and protein–RNA complexes all depend on the formation of intricate secondary and tertiary structures [130, 95]. The continued discovery of such functional elements has necessitated the development of methods to obtain accurate structure predictions at high-resolution. To this end, X-ray crystallography and nuclear magnetic resonance are currently the ideal RNA structure characterization

methods. However, their cost, labor requirements, and limited applicability render them low-throughput. More recently, structure profiling (SP) experiments have received considerable attention as an alternative approach for probing RNA structure that is more affordable and suitable for high-throughput applications. By providing a snapshot of the structural states of an RNA transcript at nucleotide resolution, SP experiments aim to elucidate the role of RNA structure in biologically relevant contexts [130, 93, 88, 91].

Structure profiling experiments utilize chemical or enzymatic reagents that modify or cleave nucleotides in a structure-dependent manner. Modifying reagents are sensitive to the local stereochemistry of the RNA, meaning regions which are flexible are more likely to be accessible to the reagent. As a result, accessible regions are modified more frequently compared to regions that are rigid, internalized, or obstructed. Sites of modification lead to transcription terminations or to mutations, which are then detected by sequencing. The degree of modification, termed *reactivity*, is then quantified, providing nucleotide-resolution information on a transcript's structure. The sequence of reactivities over a transcript is termed a *structure profile*. Structure profiling experiments were recently scaled to transcriptome-wide levels with the advent of next-generation sequencing. These advances have revolutionized our ability to study RNA structure at the scale of the entire transcriptome and in the complex context of a living cell, with new applications and methods continuing to emerge [95, 91, 216, 43].

Despite the recent breadth and scale of SP datasets, universal and efficient tools for their interpretation and analysis are generally lacking. There are several reasons for this, one being the difficulty in integrating nucleotide-resolution measurements to the level of biologically relevant structural elements [20]. This is critical because RNA function is typically driven by structural elements that span at least a few and often tens of nucleotides. Examples of functional elements with available consensus structures that are impacted by cellular conditions include aptamers and riboswitches, which respond to ligands [11, 124, 203, 60, 212], thermosensors that respond to temperature [72], G-quadruplexes [189, 95, 57, 160], as well as several non-coding RNAs [207, 202]. Additionally, RNA modifications, which are prevalent and dynamic, can modulate structures [62, 104]. Traditional approaches to study such elements often rely on secondary structure prediction via thermodynamic models and dynamic programming algorithms, fused with SP data [118, 156, 108, 173, 109]. While powerful, these methods do not scale well to transcriptome-level analyses [101] and are often inaccurate for long RNAs [49]. More importantly, they are based on modelling assumptions that fail to capture the full

complexity of the cellular environment [159], in particular inter-molecular interactions and varying cellular conditions. In addition, RNA structures are dynamic as illustrated by co-transcriptional folding pathways [198, 73]. In these contexts, it is valuable to be able to rapidly glean structural information from SP data alone. However, the diversity of available reagents, signal enrichment strategies, modification detection methods, and analysis pipelines results in disparate statistical properties of SP datasets. Consequently, existing SP-based methods are often specialized to the properties of the data at hand and to the study's biological objectives [194, 39, 88, 174, 91].

To address these needs, we previously developed *patteRNA*, a machine learning algorithm for mining RNA structures from SP data directly [101]. Leveraging a simplified representation of RNA structures as chains of paired and unpaired nucleotides, *patteRNA* learns the statistical properties of two components that are fundamental to all SP datasets. The first is RNA structure. Here, *patteRNA* learns how paired and unpaired nucleotides come together to form commonly observed structural motifs, such as hairpins. This is accomplished by training a Hidden Markov Model (HMM) to capture the probability of adjacent nucleotides transitioning between paired and unpaired states, and vice versa. The second feature is the SP signal, irrespective of the SP strategy employed. In this context, *patteRNA* learns which reactivity values are expected for paired nucleotides and which values are expected for unpaired ones [180, 30]. These expectations are formulated in a Gaussian Mixture Model (GMM) of reactivity values. When fused together, these two features give rise to a GMM-HMM framework [148], which allows *patteRNA* to bridge between the resolution of reactivity measurements (i.e., single nucleotide) and that of the sought-after structural elements (i.e., reactivity patterns over local regions). To implement this, the GMM-HMM statistically links every structure to every possible data pattern to assess their consistency. Equipped with these statistical modeling capabilities, *patteRNA* rapidly scans local data patterns in massive datasets, in search of regions where the data indicates that a target motif is likely to occur.

While we demonstrated *patteRNA*'s utility as a tool for automated mining of patterns in SP data, our previous work focused on detecting highly pronounced structural changes, and results derived from several datasets were generally considered in isolation [101]. To improve the algorithm's robustness and ease-of-use, we present updates to its training routine, which is now fully automated. To extend the repertoire of *patteRNA*'s applications, we introduce improvements to its scoring pipeline, which now utilizes a normalization strategy to facilitate integration and direct comparison of search results conducted

with different target motifs and datasets. Using the revised pipeline, we demonstrate our algorithm's refined capabilities of pattern recognition. Specifically, using the human immunodeficiency virus type 1 (HIV-1) Rev response element (RRE) as an example, we show that *patteRNA* can discriminate highly similar structure profiles, identify the precise location of RRE with high confidence in a whole-genome profile, and capture changes to ensemble composition in simulated data. Overall, our results suggest that data-driven models are a promising route for the discovery of functional RNA elements. Our findings also serve as further validation of *patteRNA* and its capabilities as an automated and broadly applicable RNA structure mining engine.

2.2 Materials and Methods

2.2.1 Overview of Structure Profiling Experiments

Structure profiling experiments aim at querying all RNA structures in a sample at nucleotide resolution. Chemical reagents or enzymes are used to modify the RNA in a structure-dependent manner, i.e., flexible or unpaired nucleotides are more accessible to the chemical/enzyme and are modified more frequently [173]. A common approach using chemical reagents is SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension), where modifications involve the formation of chemical adducts on hydroxyl residues of the RNA backbone. Commonly used SHAPE reagents include 1-methyl-6-nitroisatoic anhydride (1M6), 1-methyl-7-nitroisatoic anhydride (1M7), *N*-methylisatoic anhydride (NMIA), and 2-methylnicotinic acid imidazolide (NAI) [200, 177]. Chemical adducts interfere with reverse transcription, leading to either complementary DNA (cDNA) transcription terminations or mutations, which are then read out by DNA sequencing. Using two experimental conditions, one with the reagent (treated sample) and one without it (control sample), one can infer from sequencing reads a rate of modification, called reactivity, at each nucleotide [4, 3, 170, 183, 165, 105, 12]. High and low reactivities are generally indicative of unpaired (less constrained) and paired (more constrained) nucleotides, respectively. Consequently, a structure profile correlates with the underlying assayed secondary structure.

2.2.2 Improvements to *patteRNA*'s Training Routine

Building the Training Set Using Kullback–Leibler Divergence

To minimize the size of the training set, we start by compiling a histogram of all observed reactivities. The binning interval is determined automatically using the *auto* mode in the *histogram* function from the Python package *numpy* [139]. Next, transcripts are sorted in descending order of their data density, i.e., the proportion of observed values that are neither zero nor missing. Then, to build the training set, we sequentially add transcripts to our training set until its properties capture the distribution of the entire dataset. This agreement is quantified using Kullback–Leibler (KL) divergence [90] between the histogram of the entire dataset (P) and the one from the training set (Q). Note that both histograms are built using the same binning intervals to obtain probability density vectors of identical size. Formally, KL divergence ($D_{KL}(P||Q)$) is defined as:

$$D_{KL}(P||Q) = \sum_{\forall i} P_i \log \frac{P_i}{Q_i}. \tag{2.1}$$

Transcripts are added until $D_{KL}(P||Q)$ becomes smaller than a pre-set criterion, by default 0.01. Note that a drastic reduction in training runtime is expected as the computational overhead associated with the computation of the KL-divergence is eclipsed by the training phase completing significantly faster when using a subset of the data instead of the full dataset.

Determining an Optimal Number of Gaussian Components

To determine an optimal number of Gaussian components (K) per pairing state, we start by training the model with a single Gaussian per state ($K = 1$). We then compute the model's Bayesian Information Criterion (BIC), based on the number data points (n), the number of free parameters (ν) and the log-likelihood ($\log \mathcal{L}$) of the model, which is defined as:

$$\text{BIC} = -2 \log \mathcal{L} + \nu \log n. \tag{2.2}$$

Note that ν, the number of free parameters, is essentially an indicator of the model's "complexity". The BIC summarizes a model's performance penalized for its complexity (the $\nu \log n$ term) into a single metric and is commonly used in model selection [164]. The same procedure is then repeated with $K + 1$ components until an increase in BIC is observed. Such increase indicates that the currently tested model is less appropriate

than the previous, simpler, model and therefore an optimal K was found. The trained model derived from this K is then utilized for scoring.

Parameter Initialization

Parameters can be initialized either in a supervised or unsupervised manner. For supervised initialization, we use known reference structures to compute both the HMM and GMM parameters deriving from them. Specifically, for the HMM, we set the initial and transition probabilities for each pairing state equal to the frequencies observed in the reference structures. For the GMM, we start by partitioning reactivities based on the known pairing states of the reference structures, resulting in two data distributions, one for paired and one for unpaired nucleotides [180, 30]. We then fit a standard GMM, as implemented in the Python package *scikit-learn* [142], with a single Gaussian component ($\kappa = 1$) to each state-specific distribution. Next, the BIC is computed for each fitted distribution and summed into a single metric describing the performance of the fit for the two pairing states. We then increment κ by 1, repeat this procedure, and stop when the summed BIC increases. Once an optimal κ is found, we use the resulting means, variances and weights for each component and pairing state as initial parameters.

For unsupervised initialization, the default initial parameters are listed in the Appendix. Note that both means and variances depend on the input dataset. Specifically, under the initial assumption that the proportion of paired and unpaired nucleotides are identical, we can space Gaussians evenly across the data distribution using the percentiles of the reactivities distribution as shown in Figure 2.1. For variances, we initialize them as the variance of the entire data distribution.

2.2.3 Computing Raw *patteRNA* Scores

Using a trained model, *patteRNA* rapidly scores sites in the data for consistency with a target motif. Scoring consists of quantifying the nucleotide-wise agreement between the target motif and the considered site, using a probabilistic framework [101]. At each nucleotide in a scored site, we compute the probability ratio of the target path, T, over the inverse-target path, T'. The inverse-target path is simply the opposite state sequence of the target. Because we only consider two pairing states (paired and unpaired), there exists only a unique T' for any given T. The probability ratio is derived from the GMM-HMM with the GMM capturing the likelihood of the target path given emission probabilities of reactivity values in the scored site, while the HMM captures the likelihood of the target

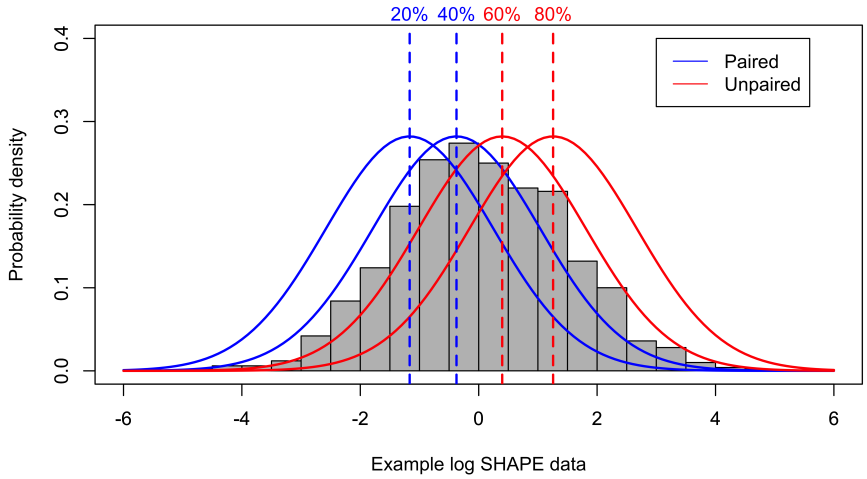


Figure 2.1: Initialization of four Gaussian components using data percentiles. Grey histograms represent the distribution of example data. In this case, the parameter $K = 2$ (i.e., two components per pairing state) and each Gaussian component is represented by a solid line with blue indicating the two components used to model paired nucleotides, and red, unpaired ones. Gaussian means are spaced at regular percentile intervals, in this case at 20%, 40%, 60% and 80% of the data distribution density, respectively.

given its state sequence as transition probabilities. It is subsequently log-transformed to handle nucleotides in the data where one pairing state is highly preferred over the other. At these nucleotides, the probability ratio would otherwise explode or collapse to exceedingly large or small values, leading to numerical overflow. The sum of log probability ratios is computed over nucleotides in the target site to produce a total raw score. More formally, we define a raw score as:

$$\text{score}(\text{target} = T, \text{site} = S, \text{model} = \theta) = \sum_{\forall i} \log \frac{\text{Prob}(S_i = T_i|\theta)}{\text{Prob}(S_i = T_i'|\theta)} \tag{2.3}$$

In practice, posterior probabilities at each nucleotide and for each pairing state are computed during training, hence scoring is a rapid process that simply involves log-transformation and summation of pre-computed values.

2.2.4 Sequence-Based Constraints

An important consideration when using *patteRNA* is the option to use sequence-based constraints. Simply put, sequence constraints are a set of rules describing which pairs of nucleotides are allowed to form base pairs. We follow the canonical set of valid base pairs when enforcing sequence constraints. Base pairs considered valid are G–C and

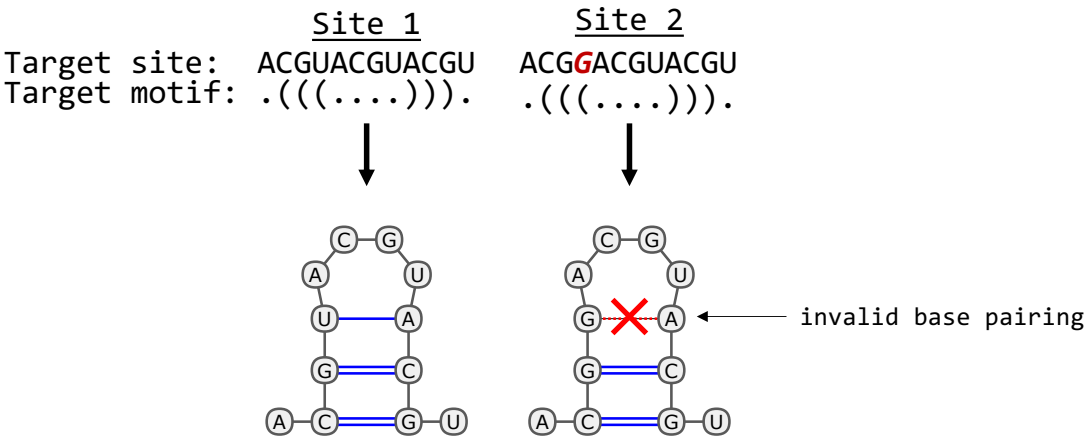


Figure 2.2: Illustration of sequence constraints. When comparing the target motif to the nucleotide sequence in Site 1, all base pairings follow the canonical rules (G–C, A–U, G–U allowed). This site consequently "passes" sequence constraints. On the contrary, the nucleotide sequence in Site 2 gives rise to non-canonical base pairings. Specifically, a G–A pairing is deemed invalid. As such, this site violates sequence constraints

A–U (Watson–Crick), as well as G–U (wobble). Note that, when enforcing sequence constraints, we do not output scores at sites whose sequence violates the constraints implied by the target structure. Visual examples of sequence-structure comparisons that pass or violate sequence constraints are summarized in Figure 2.2.

2.2.5 Comparative Motif Scoring

patteRNA normalizes raw scores by comparing them to the distribution of raw scores under the null hypothesis (H_0, defined as sites that do not harbor the target). To build the null distribution, we randomly sample raw scores from sites violating sequence constraints (see Section "*Sequence-Based Constraints*" in Materials and Methods). To do so, we scan all transcript sequences in a rolling window of the same length as the target path to create a pool of regions, from which we sample up to 5000 null raw scores (or as many as possible, if fewer than 5000 sites in the data violate sequence constraints). If sequence constraints are not enforced, we sample up to 5000 raw scores across the entire dataset.

Once null scores are compiled, we fit null distributions for each target motif using a skew-logistic (also known as a generalized logistic) probability density function (PDF). Optimal parameters are determined by maximum likelihood estimation using the implementation in SciPy [139]. We then normalize each raw score with respect to the target motif's null distribution by determining the probability of observing a raw score greater

or equal to it, also known as the survival function. This probability is log-transformed to output a c-score, which we write:

$$c\text{-score} = -\log_{10}(1 - F(\text{score}; \alpha, \beta, \gamma)), \tag{2.4}$$

where $F(\text{score})$ is the cumulative PDF of the fitted null for a target motif and $\{\alpha, \beta, \gamma\}$, the shape, shift (location) and scale parameters, respectively. By definition, c-scores are always positive and not upper-bounded. Higher c-scores indicate that the considered site is more likely to harbor the target motif. Importantly, the log-transformation serves to convert the $1 - F(\text{score}; \alpha, \beta, \gamma)$ term, which is diminishingly small for sites likely harboring the target motif, to an easily interpreted normalized score. Null distributions with fewer than 100 samples are discarded, and normalized scores for the associated target motifs are not produced. In such cases, *patteRNA* outputs a warning to the user indicating the normalized scores are not computed because the null distribution cannot be estimated reliably.

2.2.6 Benchmarking *patteRNA* Scores

To benchmark *patteRNA*'s normalization procedure against real data, we compiled a collection of 21 reference RNAs, referred to as the Weeks set [101]. This dataset was used to produce Figures 2.5 and 2.6. In Figure 2.5, we scored the Weeks set for three target motif kernels: (1) 70% paired (state path: `01111111000011111110`); (2) 50% paired (state path: `00011111000011111000`); and (3) 30% paired (state path: `00000111000011100000`). To investigate the effects of motif length, scores were also generated for each kernel when repeated two, three, and four times, as indicated by the 2×, 3×, and 4× labels in Figure 2.5A. Post-processing, statistical analysis, and figure generation were performed using in-house Python scripts. Training on the Weeks set used log-transformed reactivities and completed in 13 EM-iterations and in 5 s. Scoring of all benchmarking motifs was completed in 29 s.

To demonstrate our normalization pipeline, as shown in Figure 2.6, the following three target motifs were used: (1) hairpin, stem length 3 and loop length 8 (dot-bracket: `(((........)))`); (2) hairpin, stem length 4 and loop length 4 (dot-bracket: `((((....))))`); and (3) hairpin-internal loop composite (dot-bracket: `.....((..((.....)).))).....`). Targets were scored using the same trained model obtained with the Weeks set, as described above in this section. Scoring and normalization to c-scores were com-

pleted in 8 s.

2.2.7 HIV Rev Response Element Mutant Analysis

Previous work by Sherpa et al. [169] on the structure of the RRE in HIV-1 resulted in SHAPE profiles for seven variants of RRE. Collectively, these seven SHAPE profiles are referred to here as the Sherpa set. Two of these variants correspond to isolated isomers of RRE separated via native polyacrylamide gel electrophoresis (PAGE); they are denoted 5SL (five stem-loop) isomer and 4SL (four stem-loop) isomer. The other five profiles were generated from five RRE mutants (Mutants A–E) designed to stabilize or disrupt the two native forms. The seven RRE SHAPE profiles in the Sherpa set, each 232 nucleotides in length, were used collectively to train *patteRNA*. It is noted that the predicted structures of 5SL and 4SL are identical to Mutants A and B, respectively, hence the Sherpa set is comprised by seven SHAPE profiles with six unique nucleotide sequences predicted to give rise to five unique secondary structures. The full RRE structures are shown in Figure 2.3. *patteRNA* was then used to score the seven profiles for both their full-length predicted structures (232-nt) and the SL III/IV region (59-nt). Thus, each profile received five full-length scores as well as five scores at each possible 59-nt window, or a total of $5 + 5 \times (232 - 59) = 1326$ scores. Data were log-transformed prior to *patteRNA*'s run, hence the `--log` argument was not used. Analysis was performed twice, with and without sequence constraints enforced. Training converged in 61 iterations and 30 s. Scoring and normalization was completed in 3 s.

2.2.8 Searching the HIV Genome for Rev Response Element Motifs

RRE motifs were searched in four whole-genome structure profiles of HIV-1, three of which were generated by Siegfried et al., who employed high-throughput mutational profiling in conjunction with 1M7, 1M6, and NMIA SHAPE reagents (SHAPE-MaP) [170]. The fourth profile, generated by Watts et al. [199], was obtained with the 1M7 SHAPE reagent and capillary-based cDNA quantification. *patteRNA* was trained on each profile independently using log-transformed reactivities. The trained model for each HIV genome was subsequently used to score sites in the data for similarity to all five full-length structures of RRE from the Sherpa set as described in HIV RRE Mutant Analysis (see Figure 2.7). When scoring, sequence constraints were not enforced, thereby generat-

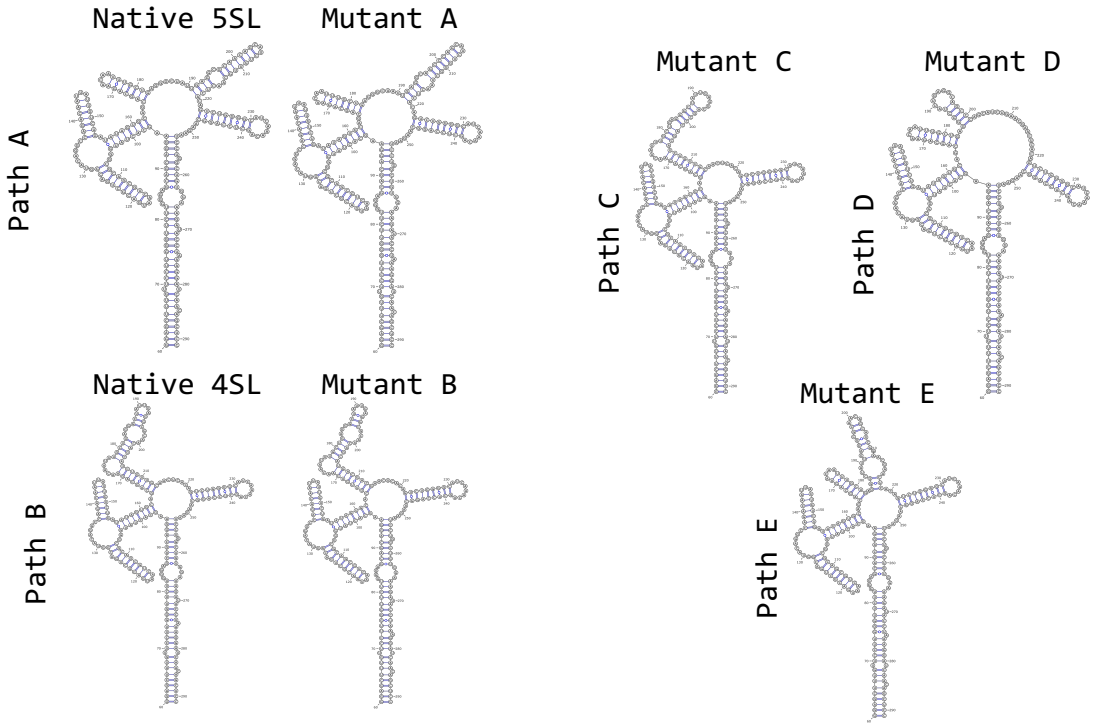


Figure 2.3: Secondary structures of the *in vitro* RREs (nt 60-291), as predicted by Sherpa et al.

ing five scores for every possible 232-nt window. Sequence constraints were not enforced because we sought to assess how scores compared between the known site of RRE and other sites in the genome that violate sequence constraints. Training converged in under 100 iterations and 3 min for all profiles. Scoring was completed for all profiles in under 90 s, for a total runtime per genome of approximately 2–4 min.

Each profile was then scored for the presence of the 59-nt SL III/SL IV region as represented in Figure 2.4. Scoring was performed with and without sequence constraints. With sequence constraints, the search space was consequently reduced to only the exact location of SL III/SL IV in the genome (nt 7409–7467) (i.e., no other sites in the genome satisfied sequence constraints). Furthermore, only Paths A, B, and E satisfied the sequence constraints at this site, so only scores from these paths are reported. Without sequence constraints, scores were generated at every possible 59-nt window within the HIV-1 genome. Using the associated trained model, scoring was completed for each profile in under 30 s.

To compare c-scores directly between searches in the HIV-1 genome and a larger dataset, we utilized publicly available *in vivo* transcriptome-wide PARS data (reference GM12878) from Wan et al. [194]. The data were processed as described previously, and

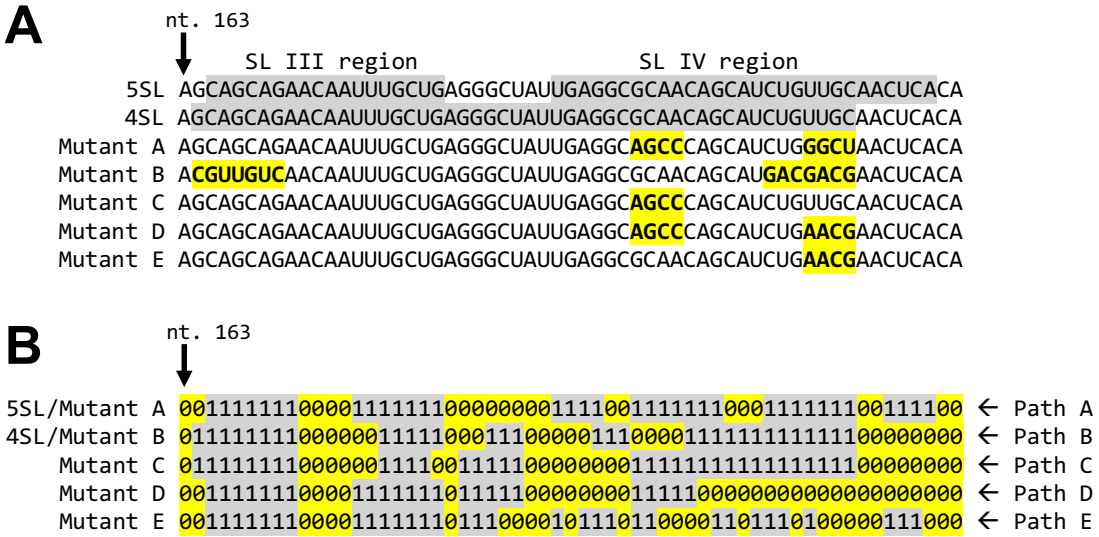


Figure 2.4: Sequences and pairing state paths of the SL III/SL IV region for RRE variants in the Sherpa set. **(A)**: Nucleotide sequences for the SL III/SL IV region (nt 163-221) in RRE included in the Sherpa set. In the 5SL structure, SL III and SL IV fold into distinct stem-loops (indicated in grey). In the 4SL structure, these two stem-loops rearrange and merge to form a single larger stem-loop known as SL III/IV. Mutations are highlighted in yellow with bold text. **(B)**: Binary pairing state representation of the native isomers and mutants of RRE within SL III/SL IV. Unpaired and paired nucleotides are represented by 0 and 1, respectively. Secondary structures related to these sequences are illustrated in Figure 2.3

the same trained model was used [101]. Using the revised pipeline, we scored the full-length 5SL and 4SL RRE conformations at 1,114,957 possible sites on 649 transcripts with at least 75% data density (i.e., $\leq$25% missing values) from the PARS dataset. Searches were conducted without sequence constraints and scoring was completed in about 8 min. We then ranked *c*-scores obtained at the location of the RRE in all HIV-1 SHAPE profiles from the Siegfried and Watts sets directly against *c*-scores obtained with searches in the PARS data.

2.2.9 *In Silico* SHAPE Mixtures of HIV-1 Structure Variants

SHAPE profiles were created in silico to emulate mixtures of pure 4SL and 5SL conformations, as isolated by Sherpa et al. Synthetic mixture profiles were created in 10% increments from 100% 5SL to 100% 4SL by taking a weighted average of the 4SL and 5SL reactivities at each nucleotide. Each mixture was then scored against the 5SL and 4SL 59-nt target paths of the SL III/SL IV region (see Figure 2.7 and Figure 2.4), using a model trained from the seven profiles in the Sherpa set.

2.3 Results

2.3.1 Overview of *patteRNA* Workflow

patteRNA first reads a dataset to train its HMM-GMM model. After training, the model can be used to mine for user-specified structures (referred to as target motifs). During this phase, which we call *scoring*, *patteRNA* attributes a score to each considered region in the input RNAs, which we call a *site*. The score is computed as the log ratio of the probability of the target motif over the probability of the target motif's inverse (see Section "*Computing Raw patteRNA Scores*" in Materials and Methods). A higher score indicates that a site is more likely to harbor the target motif. Central to our method is a simplified representation of secondary structures (target motifs) as a sequence of nucleotides in one of two pairing states, namely, paired (denoted by 1) or unpaired (denoted by 0). We hereby use the term *path* to refer to a sequence of consecutive nucleotide pairing states as represented in *patteRNA*. Note that this is a simplification of the conventional representation of secondary structures, where the requirement to specify pairing partners is eliminated, as these are not revealed by SP data.

2.3.2 Score Normalization for Comparative and Integrative Analyses

When scoring a dataset against a single target motif, it is straightforward to parse which scores correspond to sites where the motif is more likely to occur: simply rank sites by their scores and look for top-scoring ones. However, when scoring a dataset against multiple target motifs and collectively considering the results of these searches, rank-based analysis is insufficient. At the root of this issue is our observation that scores can be biased due to properties of the target motif. Each target motif produces a distribution of scores that might vary greatly in its statistical properties and dynamic range. Such discrepancies pose a challenge to both integrative and comparative analyses of *patteRNA*'s outputs, as they render scores incomparable between distinct searches. This is particularly relevant when conducting searches for functional elements that can fold into several plausible conformations, when comparing a motif and its sub-motifs components, or for comparative analysis across varying experimental conditions [69, 23, 20, 176, 106]. For example, if scores for motif A span a different range than scores for motif B, a rank-based analysis of scores between A and B is not appropriate as these scores originate from different distributions. To illustrate this point, consider scores for three 20-nt motifs with paths "`01111111000011111110`" (70% of nucleotides are paired), "`00011111000011111000`" (50% paired), and "`00000111000011100000`" (30% paired). Figure 2.5A shows raw score distributions for these target motifs when searching across a reference set of 21 *in vitro* SHAPE profiles, which we call the Weeks set [29, 59, 100, 30, 101]. Score distributions when two (2×), three (3×), and four (4×) repeats of the state-sequences for these motifs are concatenated and searched are also included in these plots to assess the effects of a target motif's length, without affecting state composition (i.e., the proportion of paired to unpaired states). In this context, the original 20-nt paths (1×) are denoted as the "kernels" of the concatenated forms (2×, 3×, and 4×). Immediately apparent is a drastic difference in the mean and skew of score distributions associated with each motif in Figure 2.5A. There are two main issues with this. First, one cannot merge and then rank scores from multiple searches to infer which sites are likely to harbor any of the sought targets, as certain searches might dominate the top of the list. This, in turn, warrants separate analysis of each search. Second, scoring a site of interest against two alternative targets might not reveal which target is more likely to be present.

The statistical properties of score distributions were found to primarily depend on the length of the target, its state composition, and the proportion of predicted paired/unpaired

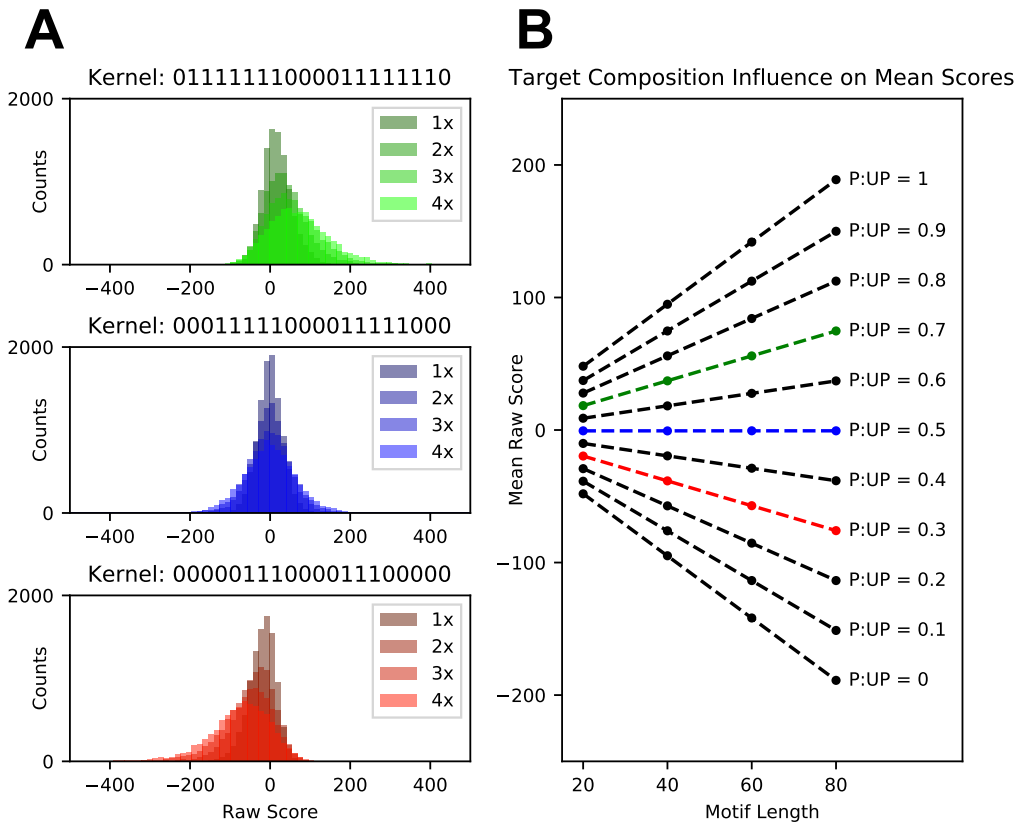


Figure 2.5: Distributions of raw scores associated with three target motif kernels. (**A**) Raw score distributions for three motif kernels of equal length and composed of 70% paired (top), balanced (middle), and 30% paired nucleotides (bottom). Overlaid are distributions for longer motifs obtained by concatenating kernels two (2×), three (3×), and four times (4×). (**B**) Mean raw scores for each of the distributions (red, blue, and green). The x-axis represents motif lengths corresponding to kernels repeated 1× (20-nt), 2× (40-nt), 3× (60-nt) and 4× (80-nt). Mean raw scores for distributions obtained when searching the Weeks set for additional kernels spanning a range of possible pairing state compositions are denoted in black. Dashed lines indicate a linear regression of the mean scores observed when extending a kernel's length.

nucleotides in the data. Firstly, longer targets generally give rise to score distributions with larger variances. This is because scores are constructed as a sum of log ratios of probabilities at each nucleotide in the scored region (see Equation (2.3) in Materials and Methods) [101]. Consequently, scores for longer targets involve summation over a larger number of terms, each with their own variance, thereby leading to overall increased spread. This bias can be seen in Figure 2.5A, where distributions of scores expand as progressively longer motifs are scored. Secondly, shifts to the mean of a score distribution are driven by an imbalance in the state composition (i.e., paired/unpaired ratio) of the target motif. To illustrate this point, the means of the score distributions for the target

motifs described earlier are shown in Figure 2.5B, where each kernel (green, blue, and red) has a unique composition. Results show that means are influenced by the target's length (x-axis) and state composition (individual curves). Note that we also observed that the magnitude of the shift in the mean is proportional to the composition of predicted pairing states across all nucleotides in the data (data not shown). Additionally, although the state composition of the target and the state composition of the data both influence the mean, we found that if the state composition of the target is balanced (i.e., 50/50 paired/unpaired), the mean of the distribution will be at zero regardless of the imbalance in the data.

To allow meaningful comparisons of *patteRNA* scores across datasets and target motifs, we developed a normalization strategy that, given a target, accounts for the statistical properties of its scores in a given dataset. The normalization step results in a comparative score, termed c-score, which quantifies the statistical significance of a site's raw score, given an estimated null distribution of raw scores associated with the target. Hereafter, the term "raw score" refers to *patteRNA* scores as described previously [101], while the term c-score refers to normalized scores. To determine the significance of a raw score, we require a distribution of raw scores (null distribution) at sites that do not harbor the motif (our null hypothesis, H_0). In practice, we do not know with absolute certainty where a motif will not occur. However, by using nucleotide sequence information, we can identify sites that are highly unlikely to harbor a motif because non-canonical base pairings would be required to give rise to the target motif. Specifically, sites where the nucleotide sequence allows for the formation of the motif via Watson–Crick or wobble base pairs are considered as putative positives. Conversely, sites that preclude motif formation are classified as falling under the null hypothesis. This filtering process is hereby called "sequence-based constraints." By applying sequence constraints and randomly sampling null sites, we can approximate the score distribution under the null hypothesis. Given the null distribution, a c-score for a given raw score, r, is the $-\log_{10}$ of the probability of observing raw scores that exceed r (in other words, the area under the null distribution above r). Note that logarithmic transformation is applied to increase the separation between diminishingly small values which are strongly indicative of the presence of the target motif, similarly to common practices in genome-wide association studies [123]. As such, c-scores are always positive and not upper bounded, and a larger c-score is indicative of a stronger match between targets and scored sites. The null distribution is then fitted using a skew-logistic PDF. The rationale for a parametric description of the

null is that it allows inferences in situations where the considered raw score falls outside the range of the null raw scores. The choice of a skew-logistic PDF was motivated by our observation that null distributions are generally non-Gaussian and often skewed (see Figure 2.5A). Note that, if sequence constraints are not enforced, the null distribution is instead constructed using scores from all sites in the data (see Methods). Under these circumstances, the null distribution will be biased, as it includes scores from true positive sites. Nevertheless, if the target motif is not widespread in the data, which is commonly the case, then this bias will only marginally affect *c*-scores as the null distribution will still contain a vast majority of negative sites.

An illustration of how our normalization framework converts raw scores at putative positive sites to final *c*-scores is shown in Figure 2.6. We illustrate the normalization process for three target motifs, namely, a short stem/long loop hairpin, a long stem/short loop hairpin and a hairpin-internal loop composite (Figure 2.6A). First, *patteRNA* computes raw scores at sites precluding formation of the target motif. These scores are used to approximate the true distribution of scores under the null (H_0) hypothesis (Figure 2.6B, left panels). A skew-logistic density function is then fitted to these null scores (black curve) and used to quantify significance of raw scores at putative sites (Figure 2.6B, right panels, where the overlaid dashed curve is the fitted null distribution). Finally, raw scores at target sites are converted into *c*-scores (Figure 2.6C) using the null distribution.

Differences in null distributions, distributions of raw scores from putative sites, and *c*-score distributions convey important observations from our normalization pipeline. First, null distributions for the three target motifs in Figure 2.6A differ in their statistical properties, and the skew-logistic PDF models observed data with high fidelity. Secondly, the distribution of raw scores from putative sites are different compared to their associated nulls. Namely, putative scores from a 4-nt stem/loop hairpin (Figure 2.6, middle) are noticeably shifted toward positive values compared to the null. In line with expectations, this shows that sites satisfying sequence constraints are more likely to emit SHAPE data in agreement with the presence of that hairpin. Although a similar shift exists for the hairpin with a shorter stem and a longer loop (Figure 2.6, top), the distinction is less dramatic. We presume that this is due to the longer loop destabilizing the hairpin more frequently compared to a hairpin with a shorter loop, which is generally assumed to be more stable. In addition, this can also be driven by sequence constraints not filtering out sites where a hairpin with a shorter loop would be feasible. In other words, while we considered a hairpin with a long loop at a site, it is more likely that the

site harbors a hairpin with a shorter loop and a longer stem if the sequence permits it, as this would be energetically favorable. The distribution of scores at putative sites for the third motif, a short hairpin containing an internal loop, closely follows the null distribution, suggesting that this target is not commonly present in the data and sequence constraints alone are a weak indicator of the motif's presence. Finally, the distribution of *c*-scores reflect these relative differences. Namely, there is an enrichment of *c*-scores greater than 1 for the short-loop hairpin that is more pronounced compared to the long-loop hairpin (Figure 2.6C, top and center panels). Comparatively, this enrichment is absent for the third motif (Figure 2.6, bottom panel).

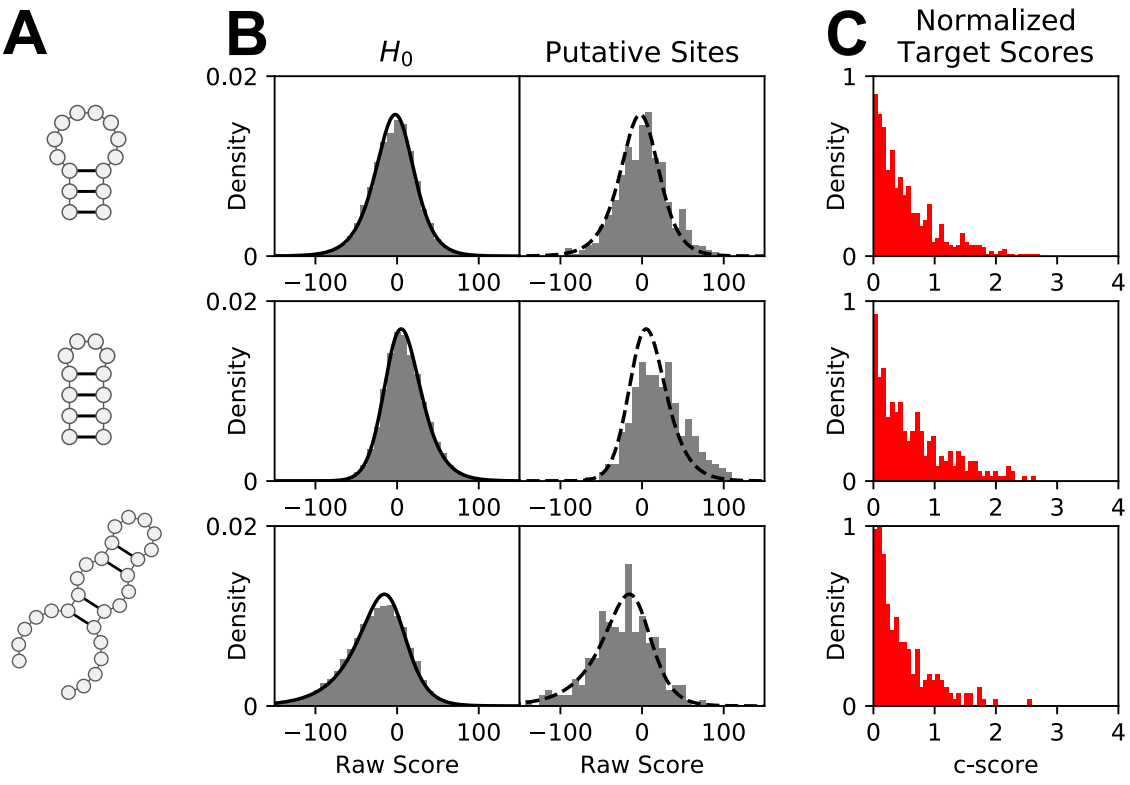


Figure 2.6: Normalization of *patteRNA* raw scores to *c*-scores. (**A**) Secondary structure of the target motifs. (**B**) Raw scores at null sites (H_0, left) and raw scores at putative sites satisfying sequence constraints (right). Null sites refer to sites where the RNA sequence precludes formation of the target motif. The solid black curves correspond to a skew-logistic density function fitted on the null scores. On the right panels, the same fitted density is superimposed (dashed curve) and is used to normalize target scores. (**C**) Distributions of normalized putative scores, i.e., *c*-scores.

In summary, we have demonstrated that *patteRNA*'s raw scoring scheme is subject to biases arising from a target motif's paired/unpaired composition as well as its length. Moreover, we observed an additional bias due to the proportion of paired/unpaired nucleotides in the dataset. As we highlighted, these biases preclude a direct comparative

analysis of different target motifs across datasets. To improve our algorithm's ability to assess relative significance of target scores, we developed a normalization pipeline that produces *c*-scores, which provide a more meaningful metric with which to interpret results from distinct searches.

2.3.3 Targeted Search of Alternative Motifs in HIV-1

Essential to viral replication and RNA trafficking in HIV is the Rev-RRE regulatory system [146]. The RRE is an RNA element present in all unspliced and partially spliced viral mRNA transcripts from an HIV-infected host cell [155]. The viral protein Rev localizes to the nucleus and binds to RRE in a cooperative manner, forming the Rev-RRE complex. Next, Crm1 and other host proteins are recruited by the Rev-RRE complex, which is then exported to the cytoplasm along with its attached mRNA transcript. Due to its highly-structured nature and implications in HIV replication, RRE has been subject to extensive structural analysis. Its structure has been characterized by crystallography [32, 75], small-angle X-ray scattering [42, 6], probing experiments [83, 18, 103, 199, 170, 6], and other methods [28, 76]. Even with the wealth of data collected, the secondary structure of RRE has remained controversial. Studies have arrived at either a 4SL [18, 115, 103] or a 5SL [28, 83, 199] structure, although slightly deviant structures have also been suggested [6]. The two principal structures, 4SL and 5SL, are shown in Figure 2.7A,B. These competing conformations are largely identical. Both predict the formation of a central loop, from which a number of stem-loops fold. The specific region of RRE that has remained controversial is the SL III/SL IV region (nt 163-221, see dashed frames in Figure 2.7A,B). It is believed that SL III and SL IV either exist as two separate stem-loops (5SL structure) or combine to form a larger stem-loop, denoted SL III/IV (4SL structure). Notably, although the mesoscale structural arrangements of these two conformations are quite different, their pairing state paths are highly similar (see Figure 2.4). As such, this presents an important challenge for analysis by *patteRNA*, which is blind to information on pairing partners and only considers the pairing state of each nucleotide when mining SP data for target structures.

To understand the role of SL III/SL IV in Rev binding, Sherpa et al. isolated two co-existing structural isomers of wild-type HIV-1 pNL4-3 RRE and performed SHAPE. From SHAPE-directed predictions, the authors concluded that each isomer corresponded to the canonical 5SL and 4SL structures [169]. They further produced RRE mutants intended to strengthen or disrupt specific base pairings in the SL III/SL IV region. Their

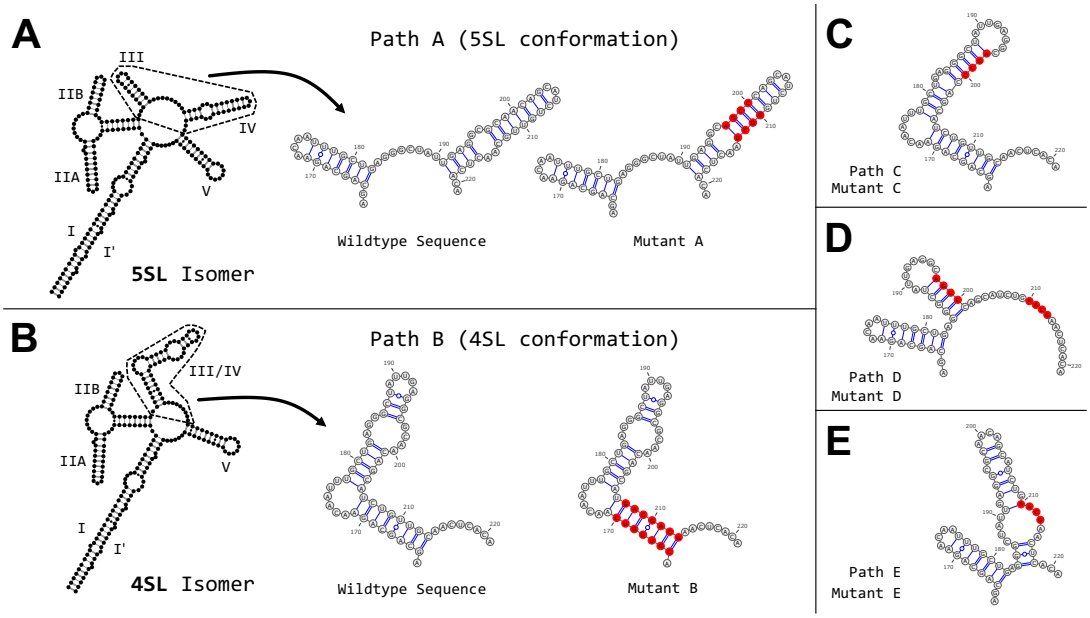


Figure 2.7: Predicted secondary structure of the Rev response element (RRE). (**A**) Full-length predicted structure of the five stem-loop (5SL) isomer of RRE. Stem-loops are indicated by their numeral. The region of interest (SL III/SL IV) is indicated with dashed lines, and expanded to show base pairings. The structure of RRE Mutant A, a variant of RRE that prefers the 5SL conformation, is also shown. (**B**) Full-length predicted structure of the four stem-loop (4SL) isomer of RRE along with a similar comparison as made in (**A**). Shown to the right is the structure of Mutant B, an RRE variant preferring the 4SL conformation. (**C**–**E**) Predicted secondary structure of the SL III/IV region for three additional RRE mutants. All mutants were produced by Sherpa et al. [169] with induced mutations highlighted in red.

experiments resulted in seven RRE transcripts with SHAPE profiles: two corresponding to the 5SL and 4SL wild-type isomers and five corresponding to mutants denoted A to E. The secondary structures within the SL III/SL IV region (as predicted by Sherpa et al. using data-directed minimum free energy models) of these seven transcripts are shown in Figure 2.7A–E, with the induced mutations highlighted in red. The binary pairing state paths for each mutant are shown in Figure 2.4B. Note that Mutants A and B share identical secondary structures with 5SL and 4SL, respectively, as they were designed to stabilize the two wild-type conformations. Moreover, while seven transcripts are considered, the native 4SL and 5SL isomers share the same underlying nucleotide sequence. Hence, this dataset, hereby called the Sherpa set, contains seven SHAPE profiles built from six unique sequences that give rise to five distinct predicted structures. As *patteRNA* represents these structures as pairing-state paths, they are denoted Paths A–E in our subsequent analyses (see Figure 2.4).

To determine *patteRNA*'s ability to distinguish between highly similar paths, we

searched for each of the structures illustrated in Figure 2.7 in all seven SHAPE profiles. Note that no sequence constraints were enforced. Our results indicate that, for all but one RRE mutant profile, our algorithm assigns the highest *c*-score to its corresponding predicted path (Figure 2.8A–E). The exception is Mutant D, where the highest score was given narrowly to Path C over its expected path, Path D (Figure 2.8D). This misclassification is driven by high reactivities at locations in Path D where nucleotides are expected to be paired (nt 170, 181, 182, and 188). As our algorithm depends solely on data to infer pairing states, Path D is demoted because of direct contradictions between the predicted path and the observed data.

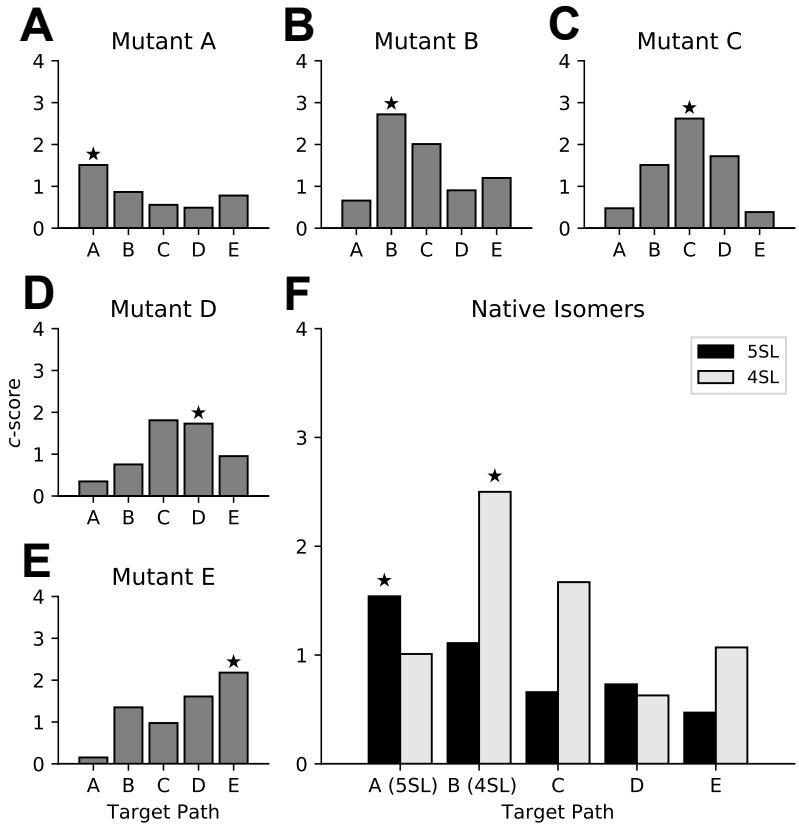


Figure 2.8: *patteRNA* scores on the Sherpa set of RRE SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension) profiles. (**A**–**E**) Each panel corresponds to a SHAPE profile for an RRE mutant. Grey bars indicate *patteRNA*'s *c*-scores for the five Paths A–E. Highlighted with a star is the score for the predicted path in the tested profile. (**F**) *c*-scores for the two native 5SL and 4SL isomers. Bars correspond to scores for Paths A–E on the 5SL (black) and 4SL (grey) profiles. Similar to the other panels, stars highlight scores for the predicted path in each profile, namely Path A for 5SL and Path B for 4SL. All scores correspond to the SL III/SL IV region (nt 163–221).

Having demonstrated that *patteRNA* can discriminate between highly similar structural RRE variants, we proceeded to investigate scores for the 5SL and 4SL native isomers. Our results show that the 5SL profile scores highest for Path A, and the 4SL profile scores

highest for Path B (Figure 2.8F). These results are in perfect agreement with Sherpa et al., as Paths A and B correspond to the sequence of pairing states for the predicted 5SL and 4SL native structures, respectively. In summary, our results support the conclusion that the two native isomers are in fact folding into the 5SL and 4SL conformations. Of the two isomers, the 4SL motif appears to be more readily detected by *patteRNA*. This is evidenced by the higher c-score when scoring the 4SL profile for its predicted state path, Path B, than when scoring the 5SL profile for its predicted state path, Path A. This difference in c-score magnitude indicates that SP data are in stronger agreement with the 4SL isomer predicted structure, compared to the 5SL isomer. This originates from reactivity values in the 5SL profile that contradict the pairing state sequence of Path A. Specifically, nucleotides 169 and 176 are observed to emit high reactivities, despite having been predicted to be in paired states within SL III. Nucleotides 195, 196, 215, 216 comprise an unpaired internal loop within SL IV, however these nucleotides emit very low or zero reactivity. A likely explanation stems from the predicted structure for 5SL having been obtained using a data-driven thermodynamic-based algorithm (RNAstructure) [29]. Such algorithms [118, 108, 156] consider the possible base-pairing arrangements (not just paired/unpaired states) in the context of the entire RNA and can subsequently consider situations in which a stem-loop is likely to fold, despite the underlying sequence necessitating an internal loop, which may or may not be reactive in SHAPE experiments. As such, it is not surprising to observe deviations between a SHAPE profile and a predicted structure. Given prior knowledge on the structure of RRE, it is possible that the 5SL conformation harbors tertiary interactions altering reactivities at nucleotides in the SL III/SL IV region. We speculate that the 5SL conformation (Path A) could leave the end of stem-loops SL III and SL IV more exposed, subsequently causing heightened reactivities for paired nucleotides. Conversely, low reactivities in the SL IV internal loop may also be explained by the rigidity of the stem-loop. Alternatively, tertiary interactions from other regions of RRE could prevent the internal loop from behaving as unpaired nucleotides in SP experiments.

Overall, these results demonstrate *patteRNA*'s ability to discern structures in SP data, even when trained on relatively small datasets and when tasked with highly similar motifs in terms of their nucleotide pairing states. Although the algorithm's performance in this situation is not impeccable (i.e., Mutant D is narrowly misclassified as Mutant C), our results are promising given the inherent limitations of our framework, which uses SP data alone and is therefore blind to pairing partners. Scores shown here are specific to the SL

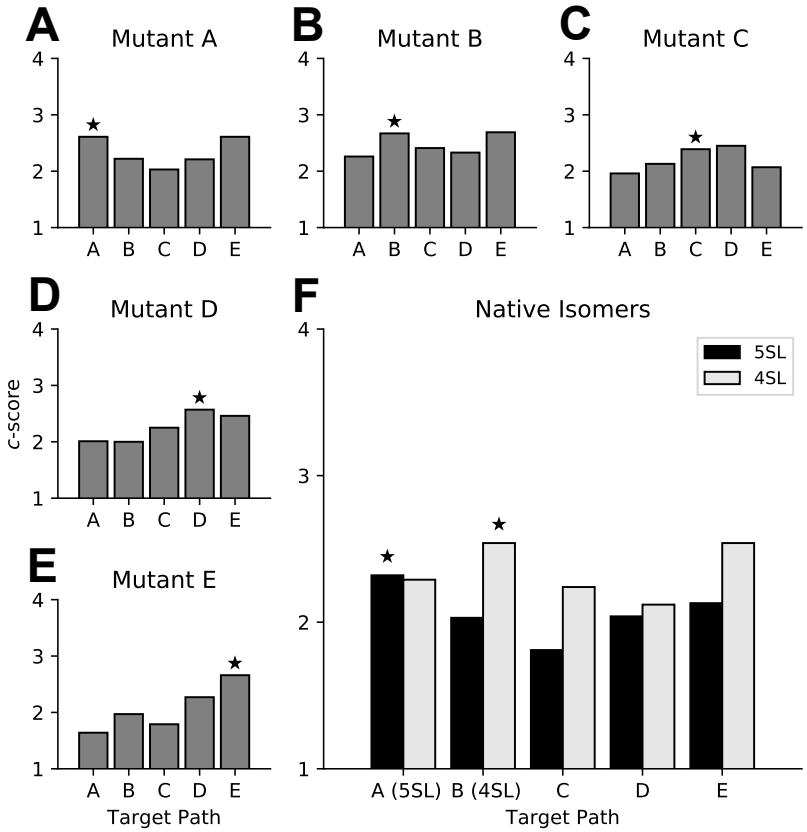


Figure 2.9: *patteRNA* scores on the Sherpa set of RRE SHAPE profiles when searching full-length RRE paths. **(A-E)** Each panel corresponds to a SHAPE profile for an RRE mutant. Grey bars indicate *patteRNA*'s *c*-scores for the five paths A-E. Highlighted with a star is the score for the predicted path in the tested profile. **(F)** *c*-scores for the two native 5SL and 4SL isomers. Bars correspond to scores for paths A-E on the 5SL (black) and 4SL (grey) profiles. Similar to the other panels, stars highlight scores for the predicted path in each profile, namely path A for 5SL and path B for 4SL. Note that *y*-axes start at 1 to better highlight differences in *c*-scores between paths, which relate primarily to differences in 59 out of 232 nucleotides when searching the full-length path.

III/SL IV region (nt 163–221), however the performance of the algorithm when searching for the full-length versions of Paths A–E convey the same conclusions (Figure 2.9).

Having observed *patteRNA*'s ability to resolve similar variants of RRE from different SHAPE profiles, we set out to investigate how well it can recognize RRE in the entire HIV-1 genome. At first, this task might seem less challenging in comparison to previous analyses we performed on human transcriptomes [101] due to the relatively small size of the HIV-1 genome. However, the data analyzed in [101] contained mRNAs that are believed to be predominantly unstructured whereas the HIV-1 genome comprises numerous highly structured elements. The latter scenario thus poses a greater challenge in discriminating between signal and background.

We utilized two HIV-1 pNL4-3 SHAPE datasets from the Weeks Lab. The first one, by Watts et al. [199], was obtained using the 1M7 reagent and capillary-based cDNA quantification. The second dataset, by Siegfried et al. [170], comprises three SHAPE-MaP profiles probed using 1M6, 1M7 and NMIA reagents. This resulted in a total of four whole-genome profiles, in which we searched for the presence of the five full-length RRE structures (Paths A–E) included in our analysis of the Sherpa set (Figure 2.3). Note that *patteRNA* training and scoring were performed on each profile independently.

Our results show that our algorithm successfully identified the exact location of the RRE structure in all four profiles (Figure 2.10, see Figure 2.11 for complete scoring results). This is demonstrated by highest *c*-scores falling precisely at the expected start location of RRE (nucleotide 7306, Table 2.1). Table 2.1 contains the highest scoring site in the whole-genome profiles for each of the five paths, A–E. Interestingly, top scores at this site are given to either the 4SL or 5SL native structures in all profiles. This is expected, as Paths C–E correspond to RRE mutants whose mutations were created artificially to render native conformations unfeasible. Note, however, that Paths C–E are still detected because we searched for the full-length RRE motif, while induced mutations are understood to drive structural rearrangements only within the SL III/SL IV region. In other words, all targets have identical structures outside of SL III/SL IV, meaning that differences in scores primarily relate to reactivity differences in only 59 out of the 232 nucleotides in RRE.

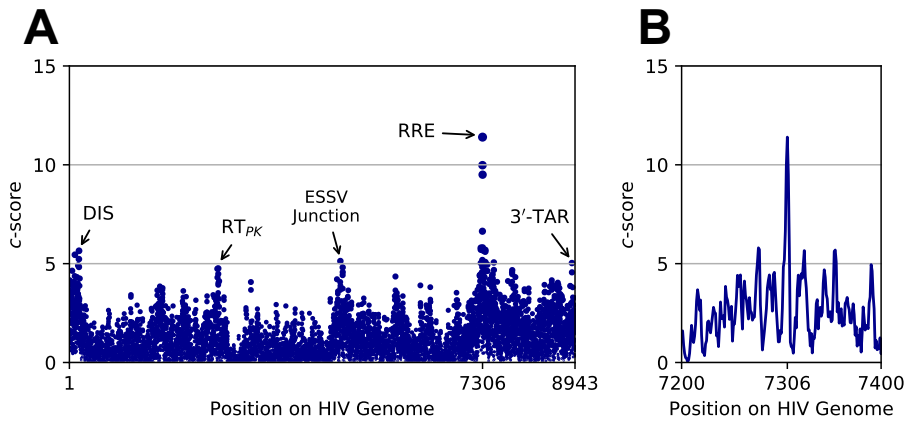


Figure 2.10: *patteRNA* scores when searching for the 4SL native structure of RRE across human immunodeficiency virus (HIV) genome profiles. (**A**) *c*-scores across the entire HIV-1 RNA genome as probed with *N*-methylisatoic anhydride (NMIA) by Siegfried et al. The peak at nucleotide 7306 corresponds to the known start site of the RRE. Other labeled peaks correspond to known structured elements in HIV-1. Scores end at nucleotide 8943 as this is the last location in the 9174-nt genome able to accept the 232-nt target paths. (**B**) Inset of *c*-scores around the RRE start site.

While the true site of RRE is consistently assigned the highest c-score over all sites in each genome, we also observed signals at other structured regions of HIV. For example, the dimerization initiation site (DIS), reverse transcriptase pseudo-knot (RT_{PK}), exonic splicing silencer ESSV junction, and 3′-TAR all give rise to detectable c-score peaks (Figure 2.10A). Because the searched RRE motif is highly structured ($> 65\%$ paired states), it is not surprising to observe heightened scores at other highly structured regions. Interestingly, the structure of the ESSV junction is not well characterized, however, recent studies have identified this region as structurally conserved across HIV and simian immunodeficiency virus (SIV) [100]. Our observations suggest that this region, readily known to influence transcription and replication [10, 80], may harbor an intricate structure related

Table 2.1: Highest *patteRNA* scores when searching Rev response element (RRE) motifs across four whole-genome human immunodeficiency virus type 1 (HIV-1) SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension) profiles. Genomes were searched for the five RRE structures reported in the Sherpa set. All top c-scores occur at the known site of RRE in the HIV-1 pNL4-3 genome (nt 7306–7537).

Dataset	**Reagent**	**Search Target**	**Top c-Score**
Siegfried Set	NMIA	Path A (5SL)	10.6
		Path B (4SL)	11.4
		Path C	11.0
		Path D	10.6
		Path E	11.4
	1M6	Path A (5SL)	12.2
		Path B (4SL)	12.4
		Path C	11.5
		Path D	12.4
		Path E	12.4
	1M7	Path A (5SL)	11.5
		Path B (4SL)	13.0
		Path C	12.2
		Path D	11.8
		Path E	11.9
Watts Set	1M7	Path A (5SL)	12.9
		Path B (4SL)	13.2
		Path C	12.7
		Path D	11.9
		Path E	12.6

NMIA: N-methylisatoic anhydride
1M6: 1-methyl-6-nitroisatoic anhydride
1M7: 1-methyl-7-nitroisatoic anhydride

to its roles in splicing.

Large fluctuations in *c*-scores were also observed in the vicinity of the known location of RRE (Figure 2.10B). These are due to pairing state agreements and contradictions when sliding the target motif's path around the true site of RRE. Because RRE is comprised by stretches of paired and unpaired nucleotides, the overlap between pairing states of the target path and those of the underlying structure of RRE will vary greatly as the target path is considered near the true site. Finally, we observe that the 4SL structure consistently ranks as the top scorer, indicating that it may be the dominant conformation in the HIV-1 genomes probed in these studies.

In addition, to place these results in the context of searches in larger datasets, we conducted a search for the two native conformations of the full-length RRE (5SL/Path A and 4SL/Path B) in a subset of highly data-dense transcripts from a human transcriptome-wide PARS dataset [194]. Searches were conducted without sequence constraints. To establish the theoretical rank that the RRE would be assigned if present in human data, *c*-scores obtained at the location of the RRE in all HIV-1 SHAPE profiles (see Table 2.1 for details) were ranked against the *c*-scores from the PARS searches for both 5SL and 4SL. Our results indicate that both conformations would rank first out of 1,114,957 sites in the PARS dataset for all HIV-1 genomes (see Figure 2.12). This suggests that the RRE would be easily identified even at a scale much larger than a 9kb viral genome. In addition, note that the distribution of *c*-scores is skewed towards high values for searches in the HIV-1 genome, as viral genomes are understood to be generally much more structured compared to mRNAs. Such statistical differences underscore the additional challenge faced by searches in relatively structured domains.

Having observed that *patteRNA* can find the RRE motif in whole-genome SHAPE profiles, we performed a similar search, but instead, we considered the SL III/SL IV region exclusively. This search provides a more specific measure of the structural nature of SL III/SL IV without overwhelming signal from the rest of the full-length RRE. The results of this search (Table 2.2) are in line with our full-length search results. Depending on the reagent used, *c*-scores indicate the either the 4SL conformation is dominating (Siegfried set, NMIA and 1M7) or both the 5SL and 4SL conformations co-exist (Siegfried set, 1M6; Watts set, 1M7). When searching the smaller 59-nt motif, less information on the target results in a reduced signal at the true positive site of SL III/SL IV (nt 7409–7467) compared to the rest of the genome. Subsequently, this site is not assigned the highest *c*-score across all sites. As such, we include in Table 2.2 the percentile of *c*-scores to

highlight the ranking of the site relative to the rest of the data. We also repeated this search with sequence constraints enforced, to prune scores corresponding to sites with nucleotide sequence precluding formation of the target path. Despite the shorter length of the SL III/SL IV target paths, the true site of RRE is nevertheless the only site in the data satisfying sequence constraints. Paths A, B, and E satisfy sequence constraints at this site only, while Paths C and D are rendered invalid because of base pairings deriving from induced mutations. Moreover, Path E, which satisfies sequence constraints at the true site in the data, is assigned a relatively low *c*-score compared to the native isomers 5SL/4SL. Taken together, these results support previous work concluding that 5SL and 4SL are the true underlying structures of RRE.

Table 2.2: *patteRNA* scoring of the SL III/SL IV RRE region (nt. 7409-7467) in genomic SHAPE data against the candidate paths A–E described in the Sherpa set. *c*-scores with and without sequence constraints are included. Paths C and D violate sequence constraints and are not reported. *c*-score percentiles are included to gauge the significance of the reported values. Percentiles correspond to the percentage of scored sites falling below the reported *c*-score.

Dataset	**Reagent**	**Search Target**	***c*-score (Percentile)** *no sequence constraints*	***c*-score** *sequence constraints*
Siegfried Set	NMIA	5SL/Path A	0.75 (81st)	0.75
		4SL/Path B	1.58 (97th)	1.52
		Path C	1.18 (93rd)	*invalid*
		Path D	0.23 (41st)	*invalid*
		Path E	0.83 (84th)	0.78
	1M6	5SL/Path A	1.39 (95th)	1.39
		4SL/Path B	1.22 (94th)	1.22
		Path C	0.55 (71st)	*invalid*
		Path D	0.50 (67th)	*invalid*
		Path E	0.65 (77th)	0.65
	1M7	5SL/Path A	1.42 (96th)	1.48
		4SL/Path B	2.77 (99th)	2.75
		Path C	2.02 (99th)	*invalid*
		Path D	0.81 (84th)	*invalid*
		Path E	0.98 (89th)	0.99
Watts Set	1M7	5SL/Path A	2.60 (99th)	2.60
		4SL/Path B	2.50 (99th)	2.50
		Path C	2.18 (99th)	*invalid*
		Path D	0.16 (31st)	*invalid*
		Path E	0.45 (63rd)	0.45

Having established that *patteRNA* distinguishes between 5SL and 4SL structures, we investigated its ability to resolve them from profiles of heterogeneous samples where they co-exist. Sherpa et al. concluded that RRE could exist as a mixture of these two structures and demonstrated that they are not functionally equivalent. More generally, the ability of structural elements to assume more than one conformation is often critical for regulatory flexibility and sensitivity. Detecting changes in the relative abundances of alternative structures is therefore an important, yet challenging problem in biology.

To explore *patteRNA* scoring of ensembles of RRE, we simulated SHAPE profiles for mixtures of the 5SL and 4SL isomers ranging from 100% 5SL to 100% 4SL, in 10% increments. Each mixture was generated by summing the desired weight fraction of the Sherpa set profiles at each nucleotide.Because we generated mixtures from the pure isomer profiles, these data essentially emulate SHAPE profiles for ensembles comprised of varying proportions of 5SL/4SL structures. Mixtures were then scored against the 5SL and 4SL motifs, similar to the analysis performed in Figure 2.8. Here, results are considered as c-score ratios between the 5SL and 4SL targets ($c_{5\mathrm{SL}}/c_{4\mathrm{SL}}$). This ratio is indicative of the relative likelihood of the two targets given their respective c-scores. Starting with 100% 5SL (Figure 2.13), our results reveal that scores evolve monotonically from favoring 5SL until 30% of the profile is comprised of the 4SL SHAPE data, at which point scores favor 4SL, as indicated by ratios below 1. This demonstrates that c-scores reflect the gradient of mixture composition underlying the simulated data.

Although *patteRNA* was not developed to decipher ensemble dynamics, our results suggest that it can readily detect composition changes in simulated data. This also hints at further applications to statistically quantify changes in structural ensembles over a time series or differing experimental conditions. Importantly, while our results suggest *patteRNA* could be utilized as a tool to detect relative changes in ensemble composition, the exact estimation of underlying population fractions remains a challenge currently beyond the algorithm's capabilities. We therefore recommend the use of other data-directed methods designed to determine ensemble compositions when approaching this problem [23, 176, 106]. Nevertheless, the utility of *patteRNA* in differential analyses of ensemble composition is promising.

In summary, we have demonstrated that *patteRNA* is capable of discerning subtle structural variation directly from SHAPE data. At the same time, it can also detect structural motifs in the larger context of a whole viral genome. Beyond basic structure characterization and search, we have shown that *patteRNA* can glean quantitative insights

on changes in ensemble compositions of native RRE isomers. Notably, all results were obtained using a standard laptop (Intel 3.1 GHz i7 CPU and 16 GB of RAM) and completed in just a few minutes, even for whole-genome HIV datasets.

2.3.4 Automating *patteRNA*'s Training Routine

Expectation-Maximization training algorithms sometimes suffer from slow convergence. To reduce runtime, a trivial solution is to reduce the size of the training set to its bare minimum. We previously noted that *patteRNA* can be trained on a subset of the data as long as it is statistically representative of the full dataset [101]. However, we did not provide specific guidelines regarding the exact number of data points, or transcripts, to be used for training, as those could vary widely based on the probing technique, sequencing approach, and data quality [22]. To circumvent this issue, we implemented an automated procedure to build the training set based on a KL divergence criterion. Briefly, transcripts are added sequentially to the training subset and the KL divergence computed. We stop adding transcripts when the reactivity distribution of the training set is sufficiently close to the distribution of the entire dataset, as indicated by a small KL divergence. The resulting training set is generally much smaller compared to the entire dataset, thereby reducing computational requirements considerably, while arriving at a trained model still representative of the whole data as demonstrated in Figure 2.14.

Next, a central parameter in *patteRNA* is the number of Gaussian components (K) used by the GMM to link reactivities to pairing states. K controls the smoothness of the model and if K is too small for the considered data, then the model will not capture all the statistical characteristics of the data, thereby leading to prediction inaccuracies. On the other hand, as K increases, the model requires more computational resources, both in runtime and in memory. In the original implementation of the algorithm, K was a user-defined parameter, as an optimal K depends on the considered data and can thus vary greatly. However, determining an optimal K is difficult as a grid-like search over selected K values is required, followed by a manual inspection of the model's goodness-of-fit. To alleviate these issues, we implemented an automated detection of K which utilizes BIC to select among models with increasing K (see Methods).

Finally, we implemented the option to initialize the model in a supervised manner. Briefly, if reference structures are provided, we utilize them to find the best estimates of the initial parameters of the GMM-HMM model. In the more common scenario of unsupervised training, we modified parameter initialization so that both the GMM and

HMM part of *patteRNA* are more representative of a biologically meaningful solution. For this, we use initial HMM parameters derived from the Weeks set that we described previously [101]. For the GMM, we use the percentiles of the reactivity distribution to gauge the initial location (i.e., the mean) of the Gaussian components (see *Parameter Initialization* in Materials and Methods).

2.4 Discussion

The emergence of high-throughput SP experiments offers new opportunities to study the functional roles of RNA structures at the transcriptome level and in both *in vitro* and *in vivo* conditions. This new wealth of data has given rise to a critical need for methods that facilitate rapid data-driven inference of structures in large datasets. *patteRNA* is a first step toward closing this gap. We envision our algorithm's utility to be twofold. First, it provides a novel approach to identifying RNA elements, which scales well with both RNA length and the number of analyzed transcripts. This means that any functional RNA element with a known or predicted secondary structure can be mined and studied quantitatively in the context of the entire transcriptome. Specific applications include identifying de novo sites harboring a functional motif (for example, identifying regulatory elements such as splice sites, riboswitches, or thermosensors), quantifying that motif's prevalence globally or its enrichment in defined RNA regions (e.g., 5′-UTR or 3′-UTR), and quantitatively studying the impact of varying experimental conditions on the presence of said motif. When considered in the context of other structure analysis tools, our algorithm could be useful in maximizing insights derived from these tools. Specifically, for large datasets, *patteRNA* can be used to select a small list of candidate regions that can then undergo more careful characterization with targeted probing, thermodynamic modeling, or phylogenetic analysis. Additionally, while we believe these capabilities are valuable for basic research, they may prove useful for the design of RNA-based therapeutics [191, 190, 1] by identifying putative target and off-target sites of drugs designed to bind RNA or by identifying RNA designs that bind a target molecule. Second, our probabilistic framework can be applied universally across SP methods and probes, bridging their differences and standardizing the interpretation of SP data as probability estimates of nucleotide pairing states. Importantly, we do not foresee *patteRNA* replacing existing methods for secondary structure prediction [118, 108, 156] or for functional RNA element mining from homologous sequences. We rather intend it to complement these tools. Of

particular relevance are covariance models, which are built from a stochastic context-free grammar framework trained on phylogenetic sequence information [135, 134]. Previously trained covariance models of RNA families can additionally be utilized to search for similar structures within genomic datasets. This approach is analogous to *patteRNA*, in that a statistical model is trained and subsequently utilized to search and score possible matches in large datasets. While *patteRNA* aims to capture structural information from SP data alone, covariance models aim to capture it from just nucleotide sequence. As such, these approaches capture structural information at distinct levels. The prospect of a unified probabilistic framework capturing information from both sequence and probing data therefore presents an intriguing challenge for the field.

A central requirement for comparing results across distinct structural motifs and datasets is the assurance that scores convey the same information about the presence or absence of a target motif at a site, regardless of the chosen target and dataset. A robust training routine is the first means by which to ensure fair comparison of results between datasets. This can be achieved by defining strict procedures on the determination of initial model parameters that require no user inputs and promote convergence of the model to biologically relevant solutions. To this end, we implemented several improvements and routines that fully automate *patteRNA*, enhancing its ease-of-use and training robustness. We next aimed to assure that inherent properties of the target motif would not bias results. In practice, this requirement was not always met, as scores displayed dependencies on motif length and paired/unpaired composition. Moreover, the proportion of paired/unpaired nucleotides in the considered dataset might also impact scores, furthering the discrepancies between searches. Put simply, scores might be incomparable between distinct searches. To alleviate this issue, we implemented a normalization routine that converts raw scores into *c*-scores. To this end, we used sequence information to classify scored sites as either putative positives or not harboring the target motif (null sites). Scores at null sites are then used to build a null distribution of scores, against which we referenced scores for putative positive sites. Intuitively speaking, a *c*-score is simply the $-\log_{10}$ of a *p*-value, thereby converting raw scores to a generalized measure of significance. While our strategy remedied the biases inherent to searching different motifs, it should be noted that the presence of missing reactivities in the data might nevertheless lead to additional biases. This arises from the bias inherent to motif length. Specifically, missing values are treated as "no information" and hence marginally contribute to raw scores. Scoring a region with sparse data is thus analogous to scoring a shorter motif. Con-

sequently, sites containing missing reactivities tend to span a narrower range of scores, compared to sites with complete observations. This in turn reduces their likelihood to emerge as top candidates in searches. This issue should be kept in mind when using *patteRNA* to search sparse datasets or poorly covered transcripts [22]. Furthermore, note that this bias cannot be easily corrected for as this necessitates considering all possible missing data patterns, a combinatorial problem that is computationally prohibitive.

Another consideration when mining large datasets for a motif is the vast number of negative sites which can often lead to signal depletion and obfuscate correct classification of true positives. As such, the pronounced signal for RRE detected by *patteRNA*is promising even in the context transcriptome-wide searches. It should be noted that sequence constraints can significantly enhance accuracy and precision during these searches because a significant number of negative sites will be pruned, henceforth enriching the scored data for true positives. This filtering is often helpful in the context of high-throughput SP data, where the overwhelming majority of sites are expected to be true negatives. However, sequence constraints might not always be relevant.

This mainly stems from *patteRNA*'s broad applicability to search target structures that are not necessarily nested, or may not be captured by the traditional prediction paradigm of secondary structure. For example, a data signature could capture not only canonical base-pairing interactions but also inter-/intra-molecular interactions [94]. In such cases, the sequence of pairing states in the target would be representative of unreactive/reactive nucleotides rather than paired/unpaired ones. Additional situations where sequence constraints are not applicable include searches for regions that are highly accessible (e.g., loops) [107] or highly structured [170].

Finally, we considered several conformations of the RRE element in HIV-1 to assess the discriminatory power of our revised pipeline. RRE is essential for viral replication and its structure has been extensively studied, providing a well-characterized element by which to benchmark our algorithm. We showed that *patteRNA* successfully identifies RRE structure variants—a non-trivial task considering the high similarity between their pairing state sequences (Figure 2.7, Figure 2.4). These results also indicate that high-quality SHAPE data alone could suffice to resolve alternative target motifs at a site, even when the targets share many similarities. We then used simulations to demonstrate the capability to discern changes in ensemble composition as such analyses do not directly depend on a precise determination of the proportion of each underlying conformation. Next, we searched for RRE across whole HIV-1 profiled genomes and demonstrated that *patteRNA*

easily and consistently finds its known location across four independent SHAPE profiles. Importantly, different SHAPE reagents were used and analyzed separately, thereby highlighting the algorithm's robustness. Of note is the remarkable signal-to-noise ratio between the location of the RRE compared to the rest of the HIV genome. This is likely because the RRE is quite large (232-nt), such that sufficient conclusive information can be gleaned to confidently discriminate between its presence and absence. Interestingly, we detected additional signals at other well-characterized and highly structured HIV-1 elements, such as the DIS, reverse-transcriptase pseudoknot (RT_{PK}), and 5′-TAR. Moreover, our search revealed a highly structured context around the exonic splicing silencer ESSV, which to our knowledge has not previously been subjected to targeted structure probing outside of whole-genome studies. Taken together, our results highlight that future application of data-driven methods to other RNA viruses [197, 92] and whole-transcriptomes has the potential to detect novel structured elements and changes to them.

2.5 Appendix

2.5.1 Author Contributions

P.R., M.L. and S.A. developed the method, analyzed the data and wrote the manuscript.

2.5.2 Deposited Resources

Data and analysis scripts supporting the conclusions of this article are freely available at `https://doi.org/10.5281/zenodo.1256866` [150].

2.5.3 Initial Parameters of *patteRNA*

Initial parameters of *patteRNA*'s Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) when no reference structures are provided, i.e., under unsupervised initialization:

- Number of Gaussian components per pairing state (K): Auto-detected using Bayesian Information Criterion (BIC)
- Transition probabilities (derived from the Weeks set):

$$\begin{array}{c} \\ \text{Unpaired} \\ \text{Paired} \end{array} \begin{array}{c} \begin{array}{cc} \text{Unpaired} & \text{Paired} \end{array} \\ \begin{bmatrix} 0.71020019 & 0.28979981 \\ 0.19677996 & 0.80322004 \end{bmatrix} \end{array}$$

- Initial probabilities:

$$\begin{array}{cc} \text{Unpaired} & \text{Paired} \\ \left[\begin{array}{cc} 0.5 & 0.5 \end{array}\right] \end{array}$$

- Gaussian means: Based on data percentiles
- Gaussian variances: Equal to the variance of the data
- Gaussian weights: $\frac{1}{K}$

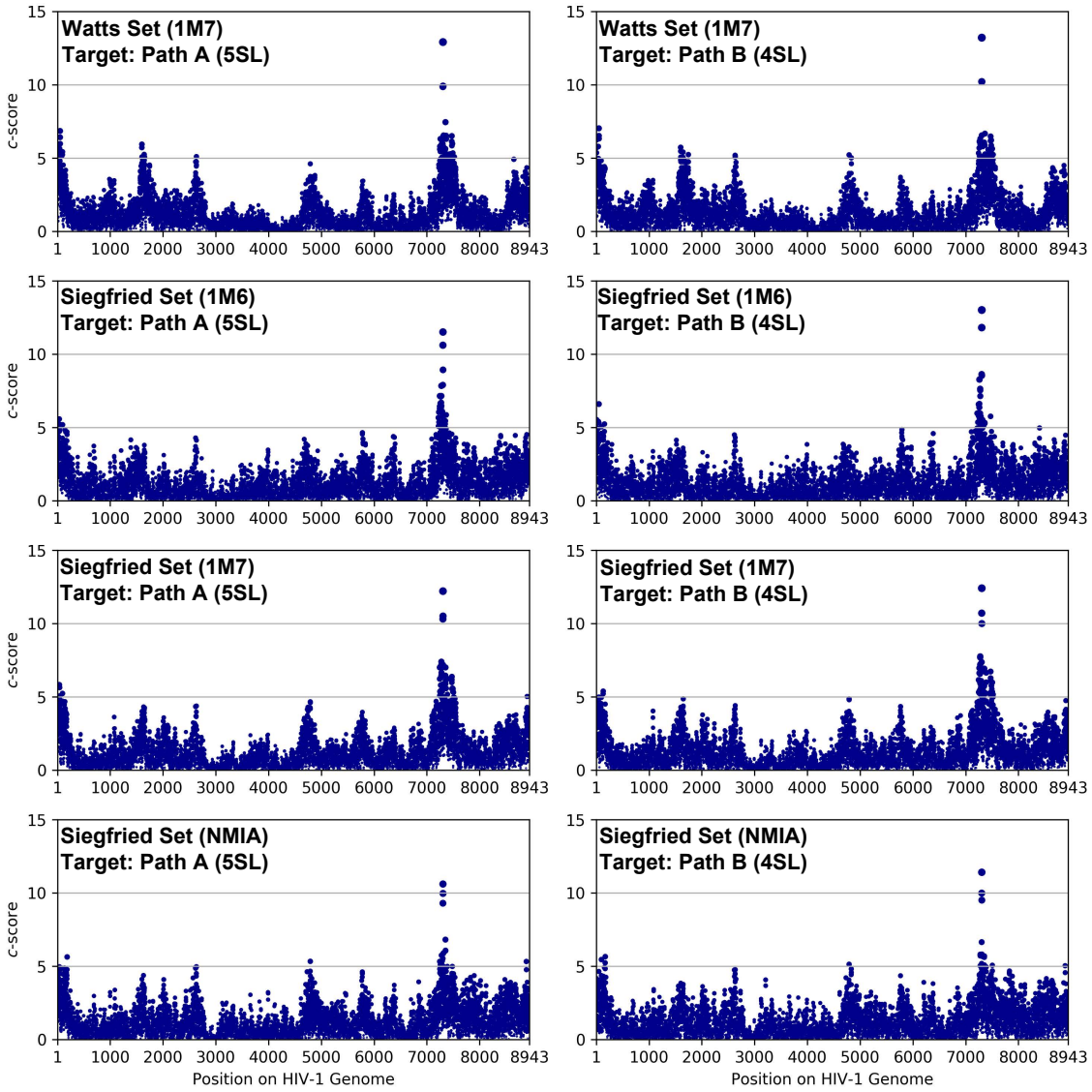


Figure 2.11: *patteRNA* scores for RRE motifs across four whole-genome HIV-1 structure profiles. *c*-scores for full-length paths A (5SL structure, left panels) and B (4SL structure, right panels) across all sites in the HIV-1 genome. Dataset and modifying reagents used are indicated in each panel and include the Watts set (SHAPE assayed with 1M7) and three profiles from the Siegfried set (SHAPE-MaP assayed with 1M6, 1M7, and NMIA, respectively). Peaks at nucleotide 7306 correspond to the known start location of the RRE.

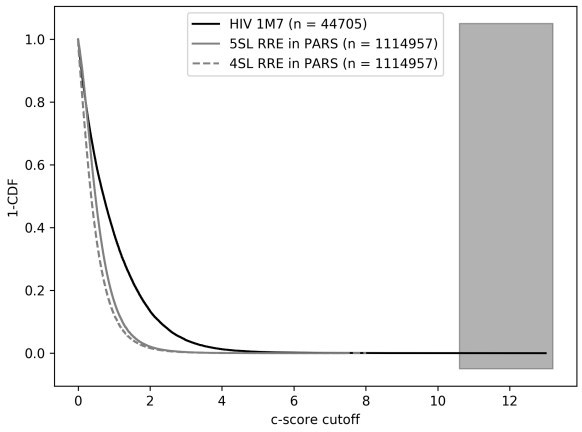


Figure 2.12: Survival functions of c-scores for the 5SL and 4SL native structure of RRE across human transcriptome-wide PARS and HIV1 SHAPE datasets. We report c-scores for searches conducted across 649 transcripts in the PARS set with data density above 75% (i.e. $\leq 25\%$ missing data), as well as c-scores from the entire HIV-1 RNA genome as probed with 1M7 by Siegfried et al. The y-axis represents the proportion of data points with c-scores above the cutoff reported on the x-axis, i.e. the survival function defined as $1 - \mathrm{CDF}(c)$, where $\mathrm{CDF}(c)$ is the cumulative distribution function. The grey rectangle highlights the dynamic range of c-scores (10.6 to 13.2) obtained at the location of the RRE for all considered RRE paths and HIV-1 SHAPE profiles (see Table 2.1 for details).

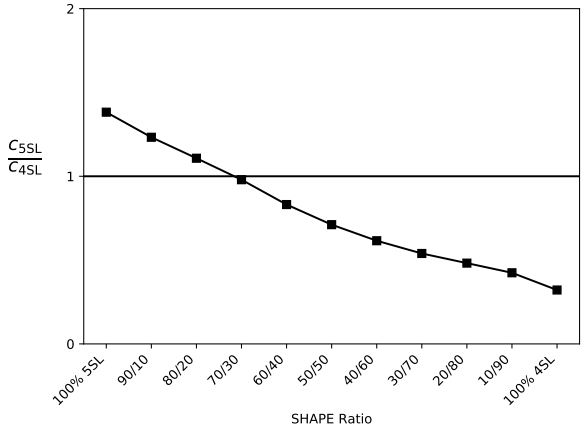


Figure 2.13: *patteRNA* score ratios (5SL/4SL) for mixtures of the 5SL and 4SL native isomers of the RRE. The x-axis corresponds to SHAPE profiles emulating various mixtures of the 5SL/4SL conformations. The y-axis corresponds to c-score ratios between the 5SL and the 4SL paths ($c_{5\mathrm{SL}}/c_{4\mathrm{SL}}$). Results indicate a stable progression of -score ratios initially favoring the 5SL structure until the SHAPE data is comprised by 30% 4SL, at which point the 4SL structure receives higher scores.

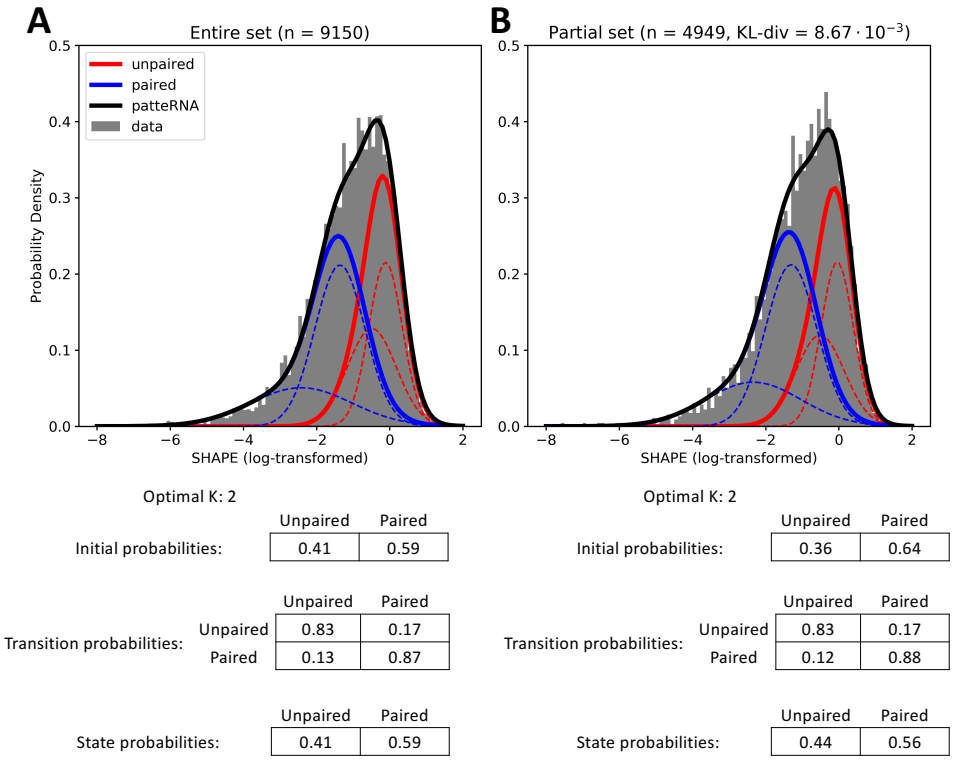


Figure 2.14: Comparison of trained models using an entire dataset and a reduced training subset. The input data are based on the HIV-1 genome probed with 1M7 from the Siegfried set and partitioned into 100 bp fragments to mimic multiple transcripts. Gaussian Mixture Models (black lines) learned by *patteRNA* as well as Hidden Markov Model parameters for **(A)** the entire dataset and **(B)** a training subset determined using KL-divergence. Grey histograms represent the distribution of the SHAPE data. Distributions associated with paired and unpaired nucleotides are shown in blue and red solid lines, respectively(solid colored lines). Individual Gaussian components are highlighted by dashed colored lines (two for each pairing state as the optimal $K = 2$ for this dataset).

Chapter 3

Rapid structure-function insights via hairpin-centric analysis of big RNA structure probing datasets

Acknowledgement: *This chapter is reproduced from an article in peer-review for publication in the journal NAR Genomics and Bioinformatics (Radecki P., Uppuluri R., and Aviran S. 2021) [151]. Pierce Radecki was lead author on this manuscript. Rahul Uppuluri was an undergraduate volunteer in the Aviran Lab. Author contributions are listed at the end of the capture. Reprinted in accordance with terms of the Creative Commons Attribution 4.0 International License.*

3.1 Introduction

RNA structure is driven primarily by the complementarity of nucleotide bases comprising it, which allows for hydrogen bonding between various segments of the molecule. Intramolecular base pairing, combined with the flexible and single-stranded nature of the molecule's backbone, allows for intricate secondary and tertiary structural elements. These structures, as well as their ability to dynamically change between relevant configurations, are known to play central roles in almost every facet of cellular regulation [48, 131, 44, 27, 166, 41]. Understanding the structures of RNA is therefore important, which has led to an explosion of methods which probe [177, 175, 195, 111, 185, 214, 53, 19, 68, 46, 211], computationally predict [59, 65, 156, 172, 127, 147, 106, 109, 40, 37, 100], and interpret them in various contexts [145, 48, 132, 86, 25, 114, 138, 166, 61]. Structure profiling (SP) experiments currently provide the most practical approach for measur-

ing RNA structures in their natural environment. These experiments work by exposing RNA to chemicals, enzymes, or photons which react differentially with parts of the molecule depending on their structural context (for example, paired/unpaired nucleotides or ds/ssRNA) [177, 175, 195, 111, 213, 204, 214, 185]. Specific protocols vary, but typically the probing reaction induces changes to the RNA bases or backbone which are detected via sequencing or electrophoresis as mutations or truncations [200, 112]. The rate of mutation or truncation at a particular nucleotide is used to summarize that nucleotide's reactivity with the probe [4]. These data contain critical information on the structural conformation of an RNA, and incorporating them as soft constraints within thermodynamics-based folding algorithms greatly improves their accuracy [29, 59, 109].

Next-generation sequencing has allowed SP experiments to scale to the level of the whole cell (i.e., transcriptome-wide). Exploration of these data have typically begun with straightforward global-level quantifications and simple comparisons [188, 81, 178, 35, 194, 111]. More recent studies expanded the intricacy of structural analysis to disentangle the dynamic functional roles of RNA structure in fundamental cellular processes [20]. For example, Saha et al. compared reactivity profiles in the vicinity of spliced introns and retained introns, and found evidence of increased structure upstream and decreased structure downstream of retained introns [161]. Yang et al. characterized structural impacts on miRNA-mediated mRNA cleaving by computing mean reactivity and mean base-pairing probability profiles around miRNA target sites, which illuminated a strong connection between transcript cleavage and unpaired bases immediately downstream of the miRNA target site [209]. Work by Mustoe et al. [132] and Mauger et al. [122] have linked changes in gene expression within E. coli and human cells to the structural dynamics within coding sequences and UTRs as quantified by local median reactivities. A slew of recent works have investigated the role of RNA structure within the interplay between RNA helicases and transcription termination, alternative splicing, translation initiation, and translation efficiency [96, 56, 193, 102]. Twittenhoff et al. [187] performed structure probing of *Y. pseudotuberculosis* at different temperatures and used averaged reactivity scores to highlight differential structure changes due to temperature in 5'UTRs versus coding regions in addition to using condition-wise reactivity differences to identify temperature-sensitive genes.

A common theme to such studies is the quantification of local "structuredness" and comparisons of it at global scales. To this end, measures of structure are typically founded on basic statistical summarization of reactivities, sometimes combined with data-directed

thermodynamics-based folding algorithms to quantify base-pairing probabilities. Current state-of-the-art algorithms for predicting base-pairing probabilities (and specific RNA structures) are founded on dynamic programming strategies and a nearest neighbor thermodynamic model (NNTM) [137, 119]. Although relatively efficient, these scale as $\mathcal{O}(L^3)$ with the length of an RNA, meaning that complete folding analyses of long RNA transcripts are often computationally infeasible. NNTM-based processing (i.e., RNA folding and computation of base-pairing probabilities) of the massive data associated with recent studies is thus challenging. As a consequence, transcriptome-wide studies have typically utilized ad-hoc folding strategies which attempt to strike a balance between computational overhead and prediction quality by locally folding pre-screened candidate regions or rolling windows of long transcripts. Even with such compromises, *in silico* analyses can take days to complete, depending on the scale of the experiment. The process itself is also susceptible to high error rates especially in molecules with multiple stable conformations [49]. It is worth noting that some of the aforementioned experiments relied solely on simple reactivity summarization; nevertheless, even in such situations, detections are typically limited to the most pronounced effects. More sophisticated analysis which accounts for structure in addition to reactivity has the potential to refine such findings and expand on them [21, 116]. This highlights a need for methods capable of rapidly extracting pertinent structural information from reactivity data.

Motivated by this need, we harnessed *patteRNA*, an NNTM-free method we previously introduced for rapidly mining structural motifs [101, 149], to quantify global trends in RNA structure dynamics from SP data. Briefly, the method works in two phases: training and scoring. The training phase learns a hidden Markov model (HMM) of secondary structure and a Gaussian mixture model (GMM) of the reactivity distributions of paired and unpaired nucleotides. The learned distributions are used to score sites for their likelihood to harbor any target structural motif (see Figure 3.1A). *patteRNA* can automatically process data from any type of SP experiment. Although we previously demonstrated that *patteRNA* accurately detects structural motifs in diverse datasets, we found that there was nevertheless room for significant improvement. Namely, there was a need for improved precision of motif detection, particularly pertaining to the vast search space encountered in transcriptome-wide experiments. Additionally, we found that our method, although suitable for comparative analysis of motifs [149], did not provide a clear quantitative framework for making practical and direct structural inferences in large datasets.

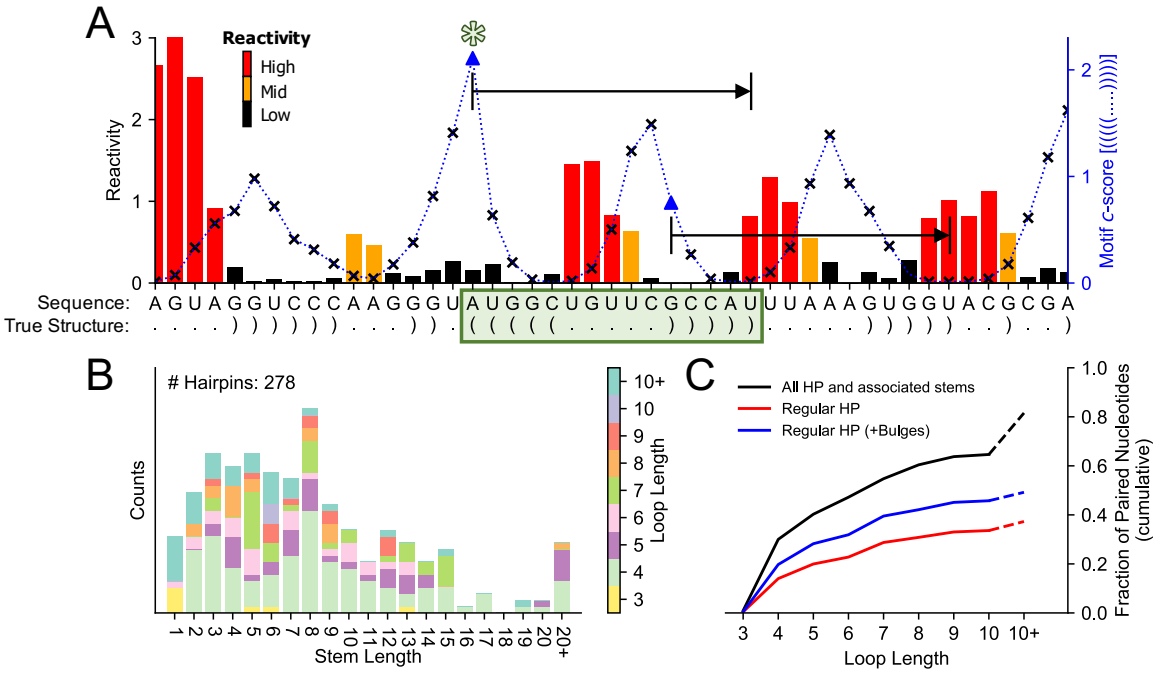


Figure 3.1: Identification of structural motifs in probing data and representation of hairpins in structures. **(A)** Schematic illustrating reactivity profile (black, yellow, red) for a region against the corresponding *patteRNA* *c*-score profile (blue) when mining for a hairpin with loop length 5 and stem length 5 (dot-bracket: "`(((((.....)))))`"). The score profile represents the likelihood of the target motif occurring at the site corresponding to using the current nucleotide as the start (left side) of a sliding window. This profile achieves a maximum at the true positive site of the mined hairpin (score indicated with star, site indicated as green box). Locations which satisfy sequence constraints necessary for the base pairs of the motif are denoted by triangle-shaped markers on the score profile, and vice versa for x-shaped markers (thus, only sites denoted with triangles are considered by *patteRNA* when scoring). The precise bounds of the sites which satisfy the sequence constraints of the motif are also indicated with black arrows. . Data shown are SHAPE-Seq reactivities from the 23S rRNA of E. coli (nt 2531-2576) [29]. Reactivities are color coded according to their magnitude (high: > 0.7; mid: > 0.3 and ≤ 0.7; low: ≤ 0.3). **(B)** Distribution of hairpin stem and loop lengths in a diverse set of structured RNAs (referred to as the Weeks set; see Methods). The vast majority of hairpins have stem lengths shorter than 15 nt and loop lengths between 3 and 10 nt. **(C)** Fraction of paired nucleotides in the Weeks set which can be represented as belonging to a regular hairpin (red), a regular hairpin with up to one or two bulges of length 1–5 nt (blue), or any/all type of hairpin and associated stems (black).

In this article, we expand and improve the capabilities of *patteRNA* and demonstrate that motif detection can be used to rapidly quantify RNA structuredness in SP datasets. As a first step, we investigate the properties of hairpin elements in RNA structures and their prevalence among all structural elements, revealing that hairpins readily detectable by *patteRNA* (hairpins without bulges) constitute over 30% of paired nucleotides. We then present an improved unsupervised training approach which yields more accurate motif detection, especially for hairpins, and benchmark it against diverse types of data. Next, we describe a novel measure, the hairpin-derived structure level (HDSL), which uses *patteRNA*'s detected hairpins to quantify the local structure context around nucleotides. We apply HDSL to three recent large-scale SP datasets to demonstrate that our hairpin-driven analysis is 1) capable of recapitulating, strengthening, and expanding on previously detected structural effects and 2) orders of magnitude faster than comparable NNTM-based routines. Simply put, our method bridges the gap between quick but naïve data summarization and intensive but more sophisticated folding-based analysis to provide rapid structure-aware interpretations. Overall, the results of our work also serve to further our understanding of the ways in which diverse SP datasets can be automatically quantified and interpreted without dependence on the assumptions driving NNTM predictions and the complexities associated with them.

3.2 Materials and Methods

3.2.1 Data

Details about the datasets used throughout this study are compiled in Table 3.1. In short, six datasets were used. Central to the development of our method is the Weeks set, a diverse dataset of 22 non-coding RNAs with high-quality *in vitro* SHAPE data and known structures (∼10,000 nt total) [101]. We used this dataset to perform benchmarks as well as to query the structural properties of structured RNAs (i.e., the representation of hairpins within them). Reference structure models were also obtained from the RNA Secondary Structure and Statistical Analysis Database (RNA STRAND) [2] and Rfam [79] to provide a more expansive set of data by which to query hairpin representation and characteristics. The remaining three datasets are recent SP datasets on which we applied *patteRNA* to demonstrate its suitability for obtaining biologically relevant insights in various contexts. This includes transcriptomic data for mRNAs *in vitro* and *in vivo* in *E. coli* [132], *in vitro* and *in vivo* reactivities for the SARS-CoV-2 genome [114], and *in vitro*

and *in vivo* transcriptome-wide reactivities for two human cell lines, K562 and HepG2 [25]. References for the sources of each dataset are provided in Table 3.1 with accession numbers included where applicable.

Note that for RNA STRAND data, the entire collection of structure models was not utilized. STRAND houses 4666 high-quality RNA structures as determined from NMR, X-ray crystallography, or comparative sequence analysis. For our work, we heuristically pruned the number of structures significantly (to 797 structures) to account for unequal representation of RNA classes within the database (specifically, the over-representation of ribosomal RNA structures). This pruning was achieved by sampling a defined number of structures from each RNA type in the database. The total numbers of original structures within each RNA type, as well as the corresponding numbers of RNA structures sampled, are given in Table 3.2. A simple visualization of the fraction of (1) transcripts, (2) nucleotides, and (3) hairpins in the pruned data coming from each RNA class is given in Figure 3.2. The numbers used for sub-sampling were heuristically determined but were guided by the composition of pruned data as observed in visualizations like the one shown in Figure 3.2. We found that the utilized values led to a fairly balanced set of data from the perspective of transcript composition, nucleotide composition, and hairpin composition.

3.2.2 Hairpin Counting and Quantification in Known Structures

To better understand the representation of hairpins with RNA structures, we parsed sets of reference structures and denoted hairpin elements according to three schemes: (1) all hairpins (hairpins and associated stems, with or without bulges), (2) regular hairpins (hairpins without bulges or internal loops), and (3) regular hairpins with or without bulges. The specific definitions used for each scheme are as follows (see Figure 3.3 for an example structure with defined hairpin motifs indicated).

All hairpins (hairpins and associated stems, with or without bulges)

Hairpins in reference dot-bracket structures were retrieved by first identifying hairpin-loops, and then backtracking to determine the full stem length. Hairpin loops are defined as locations in the dot-bracket structures where a base pair flanks a sequence of unpaired states of any length (for example, "`(...)`" or "`(.......)`"). Once a hairpin loop is identified, the stem length is determined by walking along the structure in both directions

Dataset Name	Description	Size	References
Weeks set	22 well-studied RNAs with reference structures and high-quality SHAPE data	11,070 nt	Hajdin 2013, Deigan 2009, Ledda 2018
STRAND data	797 diverse RNAs with experimentally determined structures (via NMR, crystallography, or comparative sequence analysis) [no probing data]	276,290 nt	This work, Andronescu 2008
Rfam data	Secondary structure models information by covariance models for 3,935 RNA families [no probing data]	526,608 nt	Kalvari 2020
Manfredonia data	SARS-CoV-2 genome probed by: • *In vitro* DMS-MaP • *In vitro* SHAPE-MaP • *In vivo* SHAPE-MaP	3 x 29,903 nt	Manfredonia 2020 GSE151327
Mustoe data	194 *E. coli* mRNA transcripts probed by SHAPE across three conditions (each condition is the average of two replicates) • Cellfree (*in vitro*) • Incell (*in vivo*) • Kasugamycin (*in vivo* + 10 mg/mL kasugamycin)	3 x 442,421 nt	Mustoe 2018 PRJEB23974
Corley data	*In vivo* and *in vitro* icSHAPE data (as well as fSHAPE data, not included in the dataset size) for RNA transcripts in two human cell lines: K562 and HepG2 (each condition is the average of two replicates)	2 x 40.8 million nt (K562) 2 x 35.4 million nt (HepG2)	Corley 2020 GSE149767

Table 3.1: Summary of datasets used throughout this study.

until a branching base pair is detected (i.e., a ")" to the left of the stem-loop or a "(" to the right). At this point, the stem length is called as the number of nested base pairs before the first branching base pair on either side of the stem. As a consequence, bulges

	RNA Class	Total Transcripts in STRAND Database	Transcripts Sampled for STRAND Dataset
	5S rRNA	148	70
	16S rRNA	621	30
	23S rRNA	102	20
	Signal Recognition Particle RNA	388	60
	Transfer Messenger RNA	637	50
Small RNAs	Synthetic RNA	140	100
	Transfer RNA	48	10
	Small Nuclear RNA	4	4
Regulatory Elements	Group I Intron	139	60
	Group II Intron	42	30
	Cis-regulatory Element	41	40
	7SK	1	1
	IRES	3	3
	RNAIII	4	4
	RNase E 5'UTR	6	6
	Ciliate Telomerase RNA	18	18
	Vertebrate Telomerase RNA	6	6
Ribozymes	Hairpin Ribozyme	1	1
	Hammerhead Ribozyme	136	70
	Hepatitis Delta Virus Ribozyme	7	7
	Other Ribozyme	23	10
	Ribonuclease P RNA	455	100
Other	Viral and Phage RNA	13	13
	Y RNA	14	14
	Other rRNA	136	50
	Other	20	20

Table 3.2: Number of RNA transcripts from each class of the full STRAND database included in the STRAND dataset used in this study.

and internal loops are generally ignored, so long as they occur before a branching base pair. Loops which are involved in pseudo-knotted base-pairing are treated as unpaired loops for the purpose of hairpin identification.

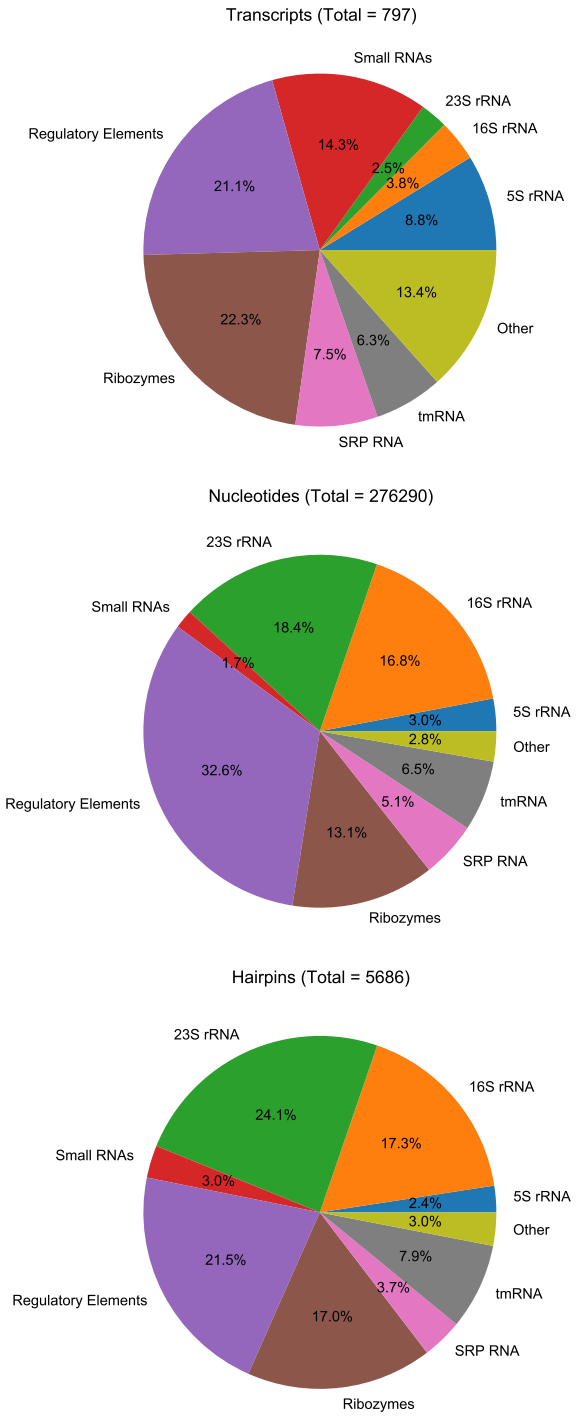


Figure 3.2: Fractional representations of transcripts, nucleotides, and hairpins for each RNA class in the STRAND data. Shown are the compositions of the data by each RNA class for transcript counts (top), nucleotide counts (middle), and hairpin counts (bottom). Hairpins are counted ignoring bulges and internal loops (see Methods: *Hairpin Counting and Quantification – All Hairpins*).

Regular hairpins (hairpins without bulges or internal loops)

We defined regular hairpins as hairpins having a stem length between 4–15 nt and loop length between 3–10 nt with no bulges or internal loops within the helix. For these motifs,

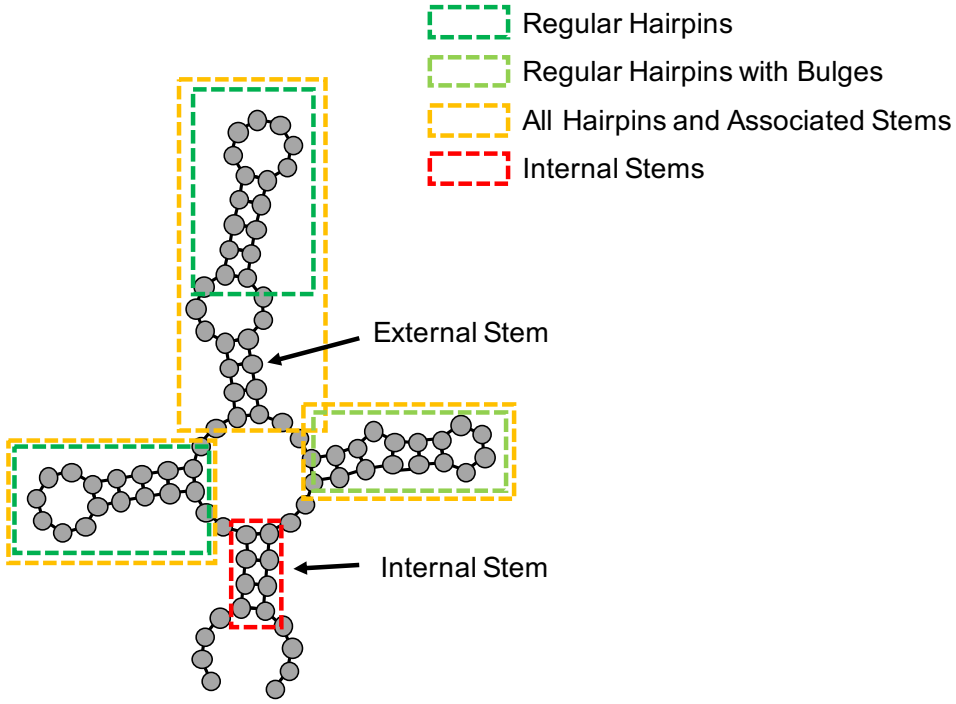


Figure 3.3: Example structure illustrating hairpin structures as defined for our analyses versus internal stems beyond the current scope of the method. Regular hairpins (dark green boxes), regular hairpins with bulges (light green boxes), and all hairpins and associated stems (yellow boxes) are indicated. External stems (helices that do not terminate in a hairpin loop but nevertheless fall within a local hairpin context outside of any bifurcations, marked within figure) are in principle within the scope of *patteRNA*'s analysis, but general searches that consider such motifs carry significant computational overhead. This is due to the combinatorial explosion of considered motifs when enumerating all combinations of internal loop sizes and positions and/or bulge sizes and positions. Internal stems (red box) do not fall into any of the defined hairpin categories and are beyond the scope of *patteRNA*'s analysis as they are underpinned by non-local base pairing.

identifying their locations amounts to simply searching the dot-bracket data for the exact dot-bracket sequence defined for each hairpin size. For example, a regular hairpin with stem length 4 and loop length 4 has dot-bracket sequence
"((((....))))";
a regular hairpin with stem length 7 and loop length 5 has dot-bracket sequence
"(((((((.....)))))))".
As before, loops which are involved in pseudo-knotted base pairing are treated as unpaired loops for the purpose of hairpin identification.

Regular hairpins with or without bulges

Identifying locations of regular hairpins that may also have one or two bulges was performed similarly to the identification procedure used for regular hairpins. However, due

to the increased flexibility of dot-bracket sequences and combinatorial explosion of qualified motifs when allowing for bulges, we used a regular expression scheme to perform the search. The regular expression has the form
"`({2,10}.{0,5}({3,10}.{3,MAXLOOP}){3,10}.{0,5}){2,10}`",
where `MAXLOOP` is the maximum loop length to include in the search. This regular expression, in order to permit flexibility for the position of bulges along the stem when identifying hairpins with bulges, also matches some motifs with stem lengths longer than 15 nt. As such, any constructed structure patterns with a stem longer than 15 nt through were discarded prior to the search. As before, loops which are involved in pseudo-knotted base pairing are treated as unpaired loops for the purpose of hairpin identification.

3.2.3 Discretized Observation Model (DOM)

The discretized observation model serves as an alternative approach for describing the probabilities of a particular state (unpaired/paired) to yield a particular reactivity value (state emission distributions). Typically, the emission distributions are modeled as continuous distributions, as is the case when *patteRNA* uses a GMM of reactivity. However, the DOM framework instead discretizes reactivities based on percentiles, then constructs probability mass functions (PMFs) over the discrete reactivity classes for the two pairing states. The state PMFs are then learned in an unsupervised fashion by coupling the emission model to an HMM and performing expectation-maximization (EM) optimization of parameters, analogously to the original GMM implementation. Also analogous to the GMM's number of Gaussian kernels, the resolution of bins used in the DOM is gradually increased until an optimal model is reached via a minimum in Bayesian information criteria (BIC) [149]. Typically, 7–10 bins are deemed optimal.

A more complete description of the mathematical formulation behind the DOM, including initialization and M-step parameter updating, is available in the Appendix of this chapter.

3.2.4 Scoring with *patteRNA*

patteRNA mines structural elements as represented in dot-bracket notation. In the context of *patteRNA*, this representation of a structure is referred to as a target motif. To mine for a motif, *patteRNA* first encodes the structure as a sequence of pairing states (states denoted as $i \in 0, 1$, where 0 is unpaired and 1 is paired), called the target path. Then, all possible locations in the data are scored for the presence of the target path.

With sequence constraints enforced, this amounts to all sites in an RNA where the nucleotide sequence permits folding of the target motif via Watson-Crick and Wobble base pairs. Sequence constraints can also be disabled, and in such situations all windows of length equal to the length of target motif are considered. (i.e., a full sliding window approach). Regardless of sequence constraints, the *patteRNA* score for a site (a window of length n beginning at nucleotide m) is defined as the log ratio of joint probabilities between the target path and its inverse path (i.e., the opposite binary sequence) [101]. More specifically,

$$\text{score}(z) = \log \frac{\Pr(y, z|\theta)}{\Pr(y, z'|\theta)}. \tag{3.1}$$

Here, y is the reactivity profile at a site, z is the target binary state path, z' is the inverse path, θ represents the parameters of a trained GMM/DOM-HMM model. The parameters of the trained model include the transition ($a_{i,j}$ for states i and j) and initial probabilities for paired and unpaired states within the Markov model, as well as an emission model (either a GMM or DOM) that described the likelihoods of paired and unpaired states to yield specific reactivity values. For a GMM [101], the emission model is parameterized by Gaussian weights, means and variances ($w_{i,k}$, $\mu_{i,k}$, and $\sigma_{i,k}$ respectively, where k corresponds to an individual Gaussian kernel in the learned mixture distributions). For a DOM, the emission model is simply parameterized by the learned discrete probability mass function of paired and unpaired nucleotides ($p_{i,k}$, where k is a bin in the discretization scheme). A trained GMM/DOM-HMM model enables computation of $b_{i,t}$ (the emission likelihood for state i and nucleotide t) as well as $\alpha_{i,t}$ and $\beta_{i,t}$ (the forward and backward probabilities for state i at nucleotide t, respectively, as computed via the forward-backward algorithm [148]). For the full formulation of emission likelihoods when using GMMs and DOMs, see the Appendix of this chapter.

The expanded score formulation as the joint probability ratio between the target path and its inverse stems naturally from the Markov model of pairing states–specifically, the probability of a particular path is the product of the probability of reaching the beginning of the path and observing upstream data ($\alpha_{i,t}$), the probabilities of transitioning between each pair of consecutive states in the path ($a_{i,j}$), the likelihoods of states in the path emitting the observed site reactivities ($b_{i,t}$), and the probability of observing downstream data given the final state of the target path ($\beta_{i,t}$). In other words,

$$\log \frac{\Pr(y, z|\theta)}{\Pr(y, z'|\theta)} = \log \left[\frac{\alpha_{z_m,m}}{\alpha_{z'_m,m}} \frac{\beta_{z_{m+n},m+n}}{\beta_{z'_{m+n},m+n}} \prod_{t=m+1}^{m+n} \frac{a_{z_{t-1},z_t} b_{z_t,t}}{a_{z'_{t-1},z'_t} b_{z'_t,t}} \right]. \tag{3.2}$$

A score of zero indicates the target path and inverse path are equally likely, and a positive score indicates the target path is more likely (and vice versa). Locations with the highest scores are subsequently deemed most likely to harbor the target motif.

To facilitate the comparative analysis of scores between different motifs and datasets, scores were further processed into *c*-scores as previously described [149] by normalizing against a null distribution of scores estimated via sampling of scores from locations which violate the sequence compatibility necessary for the motif's base pairs (and therefore can be presumed to not harbor the target motif) [149]. The resulting *c*-scores are the $-\log_1 0$ of a *p*-value, meaning they are strictly positive and theoretically have no upper bound. That said, a *c*-score above 2 is intuitively considered a strong indicator of the motif (corresponding to a *p*-value of 0.01), with *c*-scores between 0.5 and 2 providing moderate evidence in favor of the motif. Example SP data with real *patteRNA* scores superimposed is illustrated in Figure 3.1A.

3.2.5 Posterior Pairing Probabilities

patteRNA computes pairing probabilities as described [101]. Briefly, a parameterized GMM-HMM or DOM-HMM model is utilized to compute emission likelihoods for each nucleotide, followed by the forward and backward probabilities via the forward-backward algorithm. Posteriors are then computed as the product of the forward and backward probabilities and appropriately scaled such that P(paired) + P(unpaired) = 1 for each nucleotide.

3.2.6 Hairpin-Driven Structure Level (HDSL)

The hairpin-driven structure level (HDSL) is a nucleotide-wise measure quantifying the local level of structure from SP data. HDSL is initialized using posterior probabilities to be paired as computed by *patteRNA*. Then, the profile is augmented using hairpin *c*-scores calculated by *patteRNA*. For each detected hairpin with *c*-score greater than 0.5, the value $0.2 \times (c\text{-score} - 0.5)$ is added to the profile at all nucleotides covered by the hairpin. After profile augmentation, profiles are clipped to the interval [0, 1], and then profile smoothing is achieved via a 5 nt sliding-window mean followed by a 15 nt sliding-window median to give the final HDSL profile. Analogous approaches using just a sliding

mean or just a sliding median were also tested, but we found that the best results were obtained when coupling the two summary statistics together (data not shown).

The parameter values used in profile augmentation (e.g., a slope of 0.2 and a *c*-score threshold of 0.5) were determined by a grid-based optimization scheme seeking to maximize the observed difference between HDSL for nucleotides in well-folded segments of the SARS-CoV-2 genome and HDSL for nucleotides outside of these regions (see Figure 3.4). In this context, well-folded segments were defined as low SHAPE, low Shannon entropy regions as called by Manfredonia et al. [114]. The SARS-CoV-2 genome was selected for this optimization as it is distinguished from the other datasets by having both regions of high structure and un-structuredness (compared to the Weeks set, which is generally highly structured) in addition to a partially validated preliminary reference structure model (compared to the Mustoe or Corley data, which lack reliable structure models). Note that the results shown in Figure 3.4 demonstrate a large region of HDSL parameterizations which greatly improve the distinction between well-folded and less-folded segments over posteriors alone (see top left cell of each heatmap in Figure 3.4 as approximately representing the use of posteriors alone). In other words, other parameterizations arrived at similar results to the parameterization used here. Generally speaking, we observed that as the c-score threshold is increased, the slope of augmentation must also be increased in order to allow the reduced number of considered sites to sufficiently impact the final HDSL signal.

A flow chart illustrating the flow of information as handled by *patteRNA*, including the relationship between HDSL and the training and scoring phases, is included as Figure 3.5. In summary, HDSL integrates patteRNA's normalized scores (*c*-scores) for hairpins with posterior pairing probabilities to arrive at a nucleotide-wise measure of structuredness. Whereas *c*-scores (and non-normalized scores) are assigned only at specific sites in the data which satisfy the sequence base pairing requirements of a motif–e.g., a type of hairpin–HDSL is computed at all nucleotides. This is because all nucleotides are assigned a posterior pairing probability via the GMM/DOM-HMM. Hairpin scores are used to augment this profile to improve its relevance to local structure elements, but regions lacking any strong hairpins scores are still assigned pairing probabilities and as such are assigned HDSL based on those outputs.

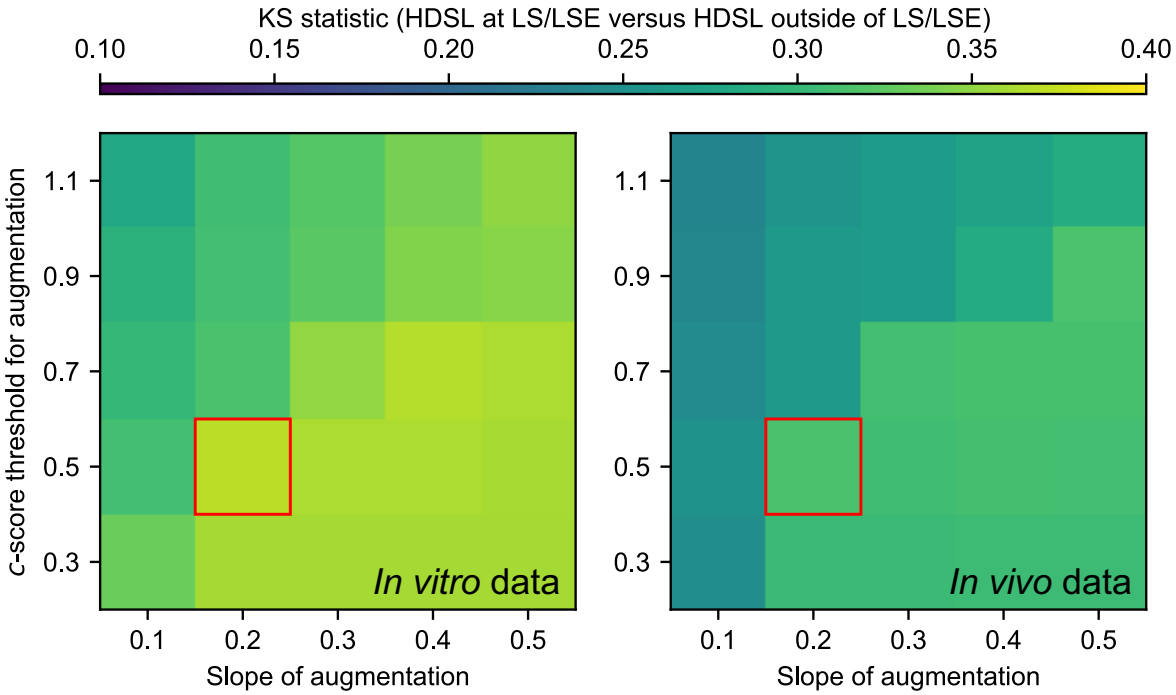


Figure 3.4: Optimization of HDSL augmentation parameterization scheme using SP data from the SARS-CoV-2 genome. Shown are Kolmogorov-Smirnov statistics between nucleotides in low SHAPE, low Shannon entropy (LS/LSE) regions and nucleotides outside of them for tested parameterizations of the HDSL augmentation scheme. Indicated in red is the parameterization selected for use in the final HDSL metric applied in the manuscript. Note that all parameterizations tested yield a statistically significant difference in HDSL between LS/LSE regions and nucleotides outside of them ($p < 10^{-200}$, 2-sample Kolmogorov-Smirnov test). This significance exists for all parameterizations as the basis of HDSL is pairing probabilities, which, on their own, are significantly different between the tested regions. As such, parameterizations of HDSL augmentations serve primarily to enhance the distinction between such regions by considering the presence of hairpin elements.

3.2.7 Computation of Statistical Performance Metrics

The accuracy of *patteRNA* to detect motifs is primarily assessed through the receiver operating characteristic (ROC) and precision-recall (PR) curves. These curves were computed by varying a theoretical *c*-score threshold between called positives and negatives and, at each threshold, computing the true-positive rate (TPR/recall), false positive rate (FPR), and precision (also referred to as positive predictive value, PPV). A site is deemed a positive if all base pairs in the target motif are also present in the corresponding location of the reference structure. These performance profiles are then visualized (ROC: FPR vs. TPR, PR: TPR vs. PPV) and summarized using the area under the curve (AUC) of the

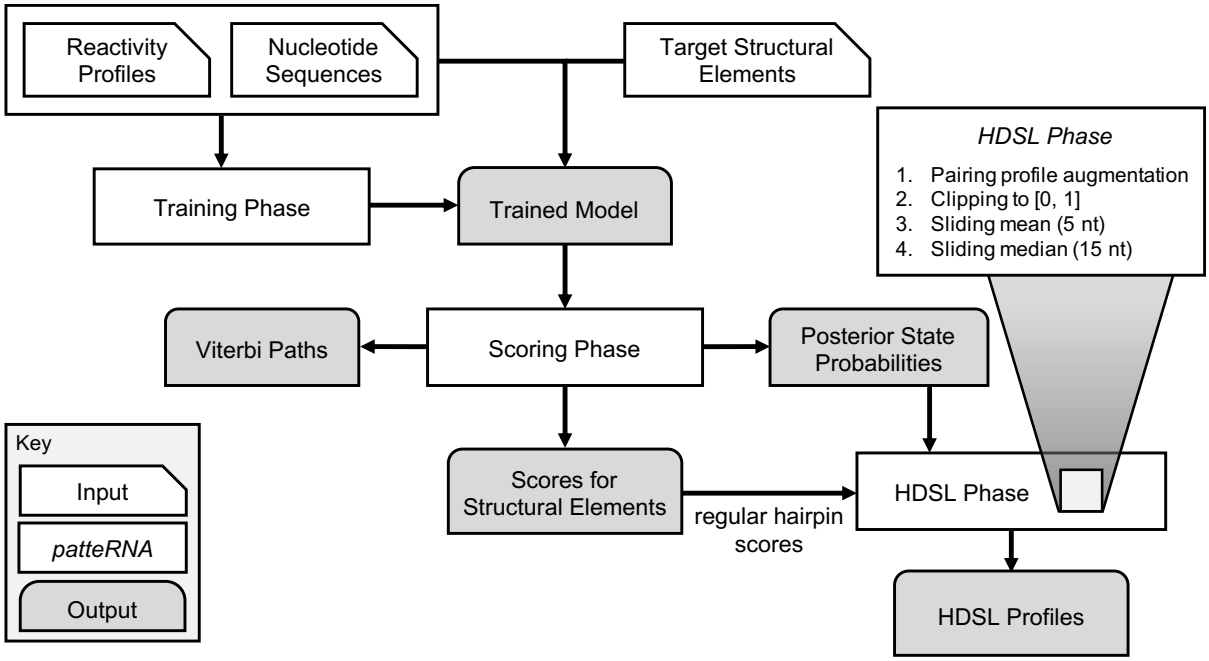


Figure 3.5: Overall flow of data and computing behind *patteRNA* and hairpin-derived structure level (HDSL). The measure is initialized as the pairing probability profiles, which are then augmented by boosting values at sites covered by highly scored hairpins (see Methods). The subsequent profile is clipped to the interval [0, 1] and local smoothing is achieved with sliding window mean and sliding window median approaches with windows of size of 5 nt and 15 nt, respectively.

ROC and average precision (AP) of the precision-recall curve. The *Scikit-learn* Python module (v0.24) was utilized to perform these computations.

3.2.8 Simulated Datasets and Benchmarks

We generated simulated data for RNAs in the Weeks set by sampling reactivities according to various state distributions schemes (see Table 3.3). 50 replicates of each scheme were generated for the performance benchmarks using in-house Python scripts. *patteRNA* was then used the train and mine the replicates for regular hairpins using the "`patteRNA $SHAPE $OUTPUT -f $FASTA [--GMM or --DOM] --hairpins`" command. The "`-l`" flag was added to use log-transformed data where applicable; training was performed independently for each replicate. Overall performance for a scheme was summarized as the mean of average precisions for the 50 replicates.

Scheme Name	Paired Distribution	Unpaired Distribution
Heitsch distributions (Sükösd 2013)	*Helix-end*: GEV(μ = 0.09, σ = 0.114, ξ = -0.821) *Stacked*: GEV(μ = 0.04, σ = 0.040, ξ = -0.763)	Exponential distribution with λ = 1.468
Gaussian / Gaussian (poor)	Gaussian distribution with μ = 0, σ = 1	Gaussian distribution with μ = 0.5, σ = 1
Gaussian / Gaussian (medium)	Gaussian distribution with μ = 0, σ = 1	Gaussian distribution with μ = 1, σ = 1
Gaussian / Gaussian (high)	Gaussian distribution with μ = 0, σ = 1	Gaussian distribution with μ = 2, σ = 1
Exponential / Gaussian	Exponential distribution with λ = 2	Gaussian distribution with μ = 2, σ = 1
Exponential / Exponential	Exponential distribution with λ = 2	Exponential distribution with λ = 1/2

Table 3.3: Parameters of state distributions used to generate artificial data on the Weeks set. GEV: generalized extreme value.

3.2.9 Averaging and Integrating HDSL over mRNA Coding Sequences

We delineated the regions surrounding the 432 genes in the Mustoe data into 4 groups: (1) start site; ±30 nt around AUG, (2) 5'UTR; -70 to -31 nt from AUG, (3) 3' UTR; +1 to +40 from STOP codon, and (4) coding sequences; +31 nt from AUG to the STOP codon. For the start site, 5'UTR, and 3'UTR, averages were taken at each aligned position as these groups each have a constant length. For situations where all regions might not exist for a gene, aligned HDSL profiles were included in the analysis as far as the nucleotide sequence allowed, and remaining positions were treated as missing values and omitted from subsequent averaging. For instance, if the 5'UTR was 50 nt (i.e., less than 70 nt), those 50 nt were aligned with the corresponding locations and

the missing 20 nt upstream were treated as missing values. For coding sequences (which inherently have a non-constant distribution of lengths), the profiles were interpolated to a vector of length 300 to allow for aligned averaging relative to the beginning and end of the window. 99% confidence intervals were computed using the Wald formulation (mean HDSL $\pm$ 2.576 $\times$ SE).

3.2.10 Hairpin Mining Performance of NNTM Partition Function Approach

We benchmarked the performance of partition function approaches to detect hairpins in the Weeks set by using the "`RNAsubopt`" command from ViennaRNA to generate 1000 structures for each transcript in the Weeks set, using that transcript's SHAPE data as soft constraints ("`RNAsubopt -p 1000 --shape $SHAPE_FILE < $SEQUENCE`"). For each hairpin in the generated structural ensemble, a "score" was assigned as the fraction of structures in the structural ensemble which contain the base pairs comprising that hairpin. Predicted hairpins and their scores were organized into a single list which was then processed into a receiver operating characteristic and precision-recall curve as done for *patteRNA*'s predicted hairpins (see *Computation of Statistical Performance Metrics*).

3.2.11 Local Folding Calculations

Windowed partition function calculations were performed using the "`RNAfold -p`" command from ViennaRNA [65]. Three schemes were utilized: windows of length 3000 nt, spaced 300 nt apart; windows of length 2000 nt, spaced 150 nt apart; and windows of length 150, spaced 15 nt apart. In each case, sequences within each window were parsed using custom Python scripts and then processed sequentially with RNAfold. Only the time required to run RNAfold commands was measured in timing benchmarks (no integration of windowed outputs or post-processing were accounted for). RNALfold benchmarks were performed using the default arguments of the command to process all sequences in the Corley data sequentially. All timing comparisons in this study were performed on an AMD Ryzen 9 5900X CPU running Ubuntu 20.04 LTS.

3.2.12 *patteRNA* Training and Scoring

Unless otherwise noted, all *patteRNA* analyses were performed with default training parameters (KL divergence for training set: $D_{KL} = 0.01$, convergence criterion $\epsilon = 0.0001$,

automatic determination of model complexity, k, via Bayesian information criteria) [149]. With the exception of benchmarks investigating the effect of log-transforming data, log-transformed data were always used when using a GMM and non-transformed data were used when using a DOM. Scoring for regular hairpins was achieved using the "`--hairpins`" flag and computation of HDSL profiles was achieved with the "`--hdsl`" flag.

3.3 Results

3.3.1 Overview of *patteRNA* Mining

To mine structure elements from SP data, *patteRNA* first learns the statistical properties of the data via the training phase. The purpose of this procedure is to estimate the distributions of reactivities associated with paired and unpaired nucleotides, respectively. Training is unsupervised and has been shown to accommodate diverse data distributions (see Ledda et al. [101] for a complete description). With the dataset characterized via its statistical model, *patteRNA* can then mine for structural motifs. Figure 3.1A demonstrates key concepts related to *patteRNA*'s motif mining. When mining a particular structural element (i.e., the target), sites which satisfy the sequence constraints necessary for the target's secondary structure are scored for their probing data's consistency with its pairing state sequence [101, 149]. Sites which do not satisfy sequence constraints can also be scored, however these sites are almost certainly all negatives and can therefore be discarded (the only exception being the possibility of non-canonical base pairs). Sites which harbor the target motif presumably have SP data consistent with the desired state sequence and therefore score highly. *patteRNA*'s overall objective is to identify sites harboring particular structural elements, such as hairpins, as accurately as possible.

3.3.2 Hairpins Comprise a Significant Portion of Structural Elements

To assess the plausibility of a hairpin-centric approach in making general assessments of structure, we examined a diverse dataset of 22 RNAs with known structures (∼10,000 nt) [101] to quantify the distribution of hairpins present as well as the proportion of base pairs contained within hairpins. We refer to this dataset as "the Weeks set." Analyzing the 278 distinct hairpins in the Weeks set reveals that a majority fall within a narrow range

of stem and loop lengths (Figure 3.1B). Specifically, hairpins most frequently have loop lengths between 3 and 10 nt, and stem lengths 15 nt or less. In other words, although their properties are diverse, there is a range of stem and loop sizes which represents a majority of hairpins (83%). Later in the study will we leverage these characteristic properties to focus our searches on this most representative subset of hairpins.

Our results also illustrate that hairpins comprise a large fraction of structural elements. We first focused on hairpins with no bulges or internal loops (i.e., unpaired stretches flanked by some number of base pairs), which we call regular hairpins, and found that around 35% of paired nucleotides reside in such structures (Figure 3.1C). If you also consider hairpins with up to two bulges each with length up to 5 nt, this coverage increases to over 50%. This suggests that, although hairpins are only a subset of RNA structural elements, they are indeed the most prevalent, and therefore identifying them in SP data could provide a strong quantification of general structural trends.

Understanding that the Weeks set is a small sample of structures to draw conclusions from, we repeated this hairpin counting and quantification on a diverse set of 797 reference structures from the STRAND database [2] and 3,935 reference consensus structures for RNA families in Rfam [79], representing a more complete profile of structured RNA properties. The distributions of hairpins in these datasets are shown in Figure 3.6 and recapitulate the observations from the Weeks set. The STRAND data suggests that regular hairpins specifically comprise a slightly larger fraction (40%) of structural elements than is seen in the Weeks set (35%), while the Rfam data suggest this fraction is slightly less (30%). We noted that Rfam data was slightly biased by an overrepresentation of microRNA families, typically comprised by long (¿20 nt) stem-loops. As such, Figure 3.6 also shows the representation of hairpins in Rfam when microRNAs are removed. In this case, we observe the the hairpin trends align closely to what is observed with STRAND and the Weeks set, with approximately 35 to 40% of paired nucleotides residing in regular hairpins.

One can further expand the definition of a hairpin to also include the associated stems that extend from a hairpin element up to the first nucleotide that base pairs outside of the nested context of this element (see Figure 3.3 for examples). We refer to these helices as external stems and note that such motifs are prevalent in structured RNAs. Figure 3.1C shows that relaxing the definition of a hairpin to include external stems leads to over 80% coverage of paired nucleotides, with the remaining 20% of base pairs described by longer-range interactions—e.g., internal stems (see dashed red frame in Figure 3.3) and

pseudoknots. Although external stems are nevertheless the scope of the *patteRNA*-based analysis that follows, this high coverage indicates that a large majority of RNA structure can be represented as simple motifs with local base pairing. Moreover, it's important to note that virtually all types of canonical RNA structure motifs necessarily exist in the context of hairpin elements—internal stems, multi-branch junctions, etc., only exist in the presence of hierarchical domains which all terminate in a hairpin-like fashion.

In the context of *patteRNA*, we note that there are practical limitations on the types of searches that can be performed. Specifically, although structures comprised by internal loops, bulges, and external stems are within the permitted scope of minable motifs described solely by local base-pairing, the automated identification of such motifs in SP data is computationally burdensome. This is due to the combinatorial explosion of considered motifs associated with allowing for flexibility in the position and size of internal loops and bulges. For instance, regular hairpins are comprised by 96 distinct motifs (12 stem lengths and 8 loop lengths), but regular hairpins with bulges (as defined in this work) are comprised by a set of motifs with size larger than 20,000 due to the many possible bulge locations and sizes within each regular hairpin motif. Allowing for the presence of various internal loops further increases the space of motifs by orders of magnitude. Although permitted by *patteRNA*, such more comprehensive searches scale poorly to transcriptome-wide applications. As such, the analyses that follow generally focus on mining and assessment of regular hairpins.

3.3.3 Simplified Reactivity Model Improves Accuracy of Motif Detection

In an attempt to improve *patteRNA*'s performance, we investigated alternative statistical models of reactivity and their downstream effects on scoring accuracy. While the GMM approach performs well, especially at the task of approximating the underlying state distributions, we encountered issues in motif scoring. Namely, reactivities from the tails of the overall data distribution would be strongly predicted to be paired or unpaired. This isn't an inherent problem, as the most extreme reactivities should theoretically be the best candidates for confident prediction. However, these reactivities present problems during scoring as they have the propensity to dominate the score for sites they fall into. In other words, a single extreme reactivity consistent with the target state sequence could yield a high score for a site, even if data within that site is otherwise inconsistent with the target (and vice versa). Generally speaking, for SP data such as SHAPE, the most

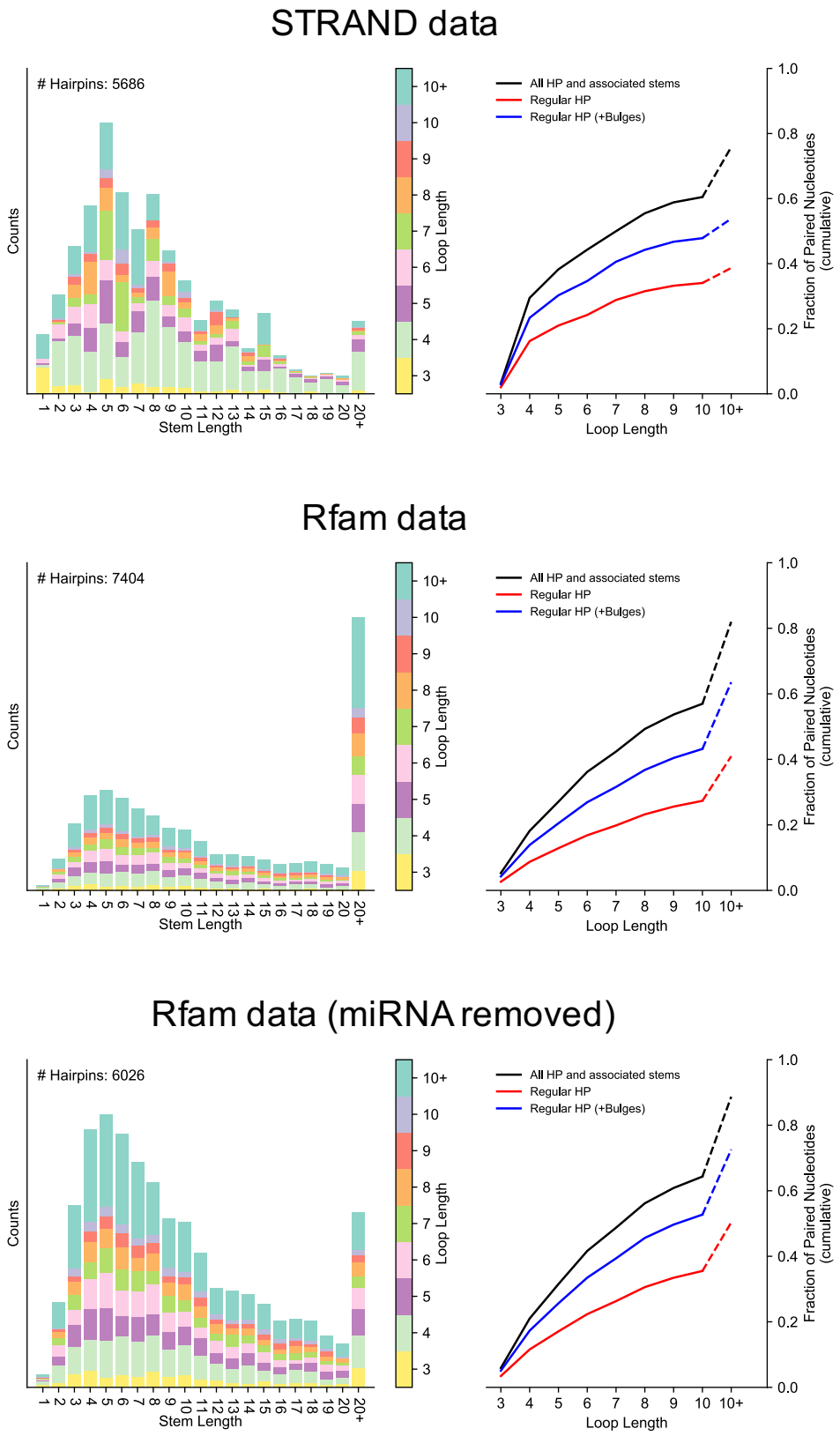


Figure 3.6: Hairpin stem and loop lengths (stacked histograms) and structural coverage of hairpins (right plots) in diverse sets of representative RNA structures obtained from STRAND and Rfam (see Methods). Results are also shown for Rfam data when omitting microRNAs.

extreme reactivities are only about 3-5 times more likely to be in one state over the other [39], yet the GMM often arrives at likelihood ratios 10 or 100 times larger than this empirical ratio. Such predictions have negative consequences on the interpretation of scores.

Motivated by these issues, we devised a simplified framework for unsupervised learning of the state reactivity distributions. It entails a discretized observation model (DOM) which substitutes for the GMM component of the statistical model (i.e., the emission probabilities), resulting in a DOM-HMM model of SP data. The DOM entails modeling reactivities as a discrete distribution where they are binned into classes based on percentiles. During training, pseudo-counts are estimated for each class (E-step) and then utilized in the M-step to infer the discrete reactivity distribution for paired and unpaired states. A schematical comparison of the GMM and DOM approaches is shown in Figure 3.7A (see Methods and Appendix for a complete mathematical formulation).

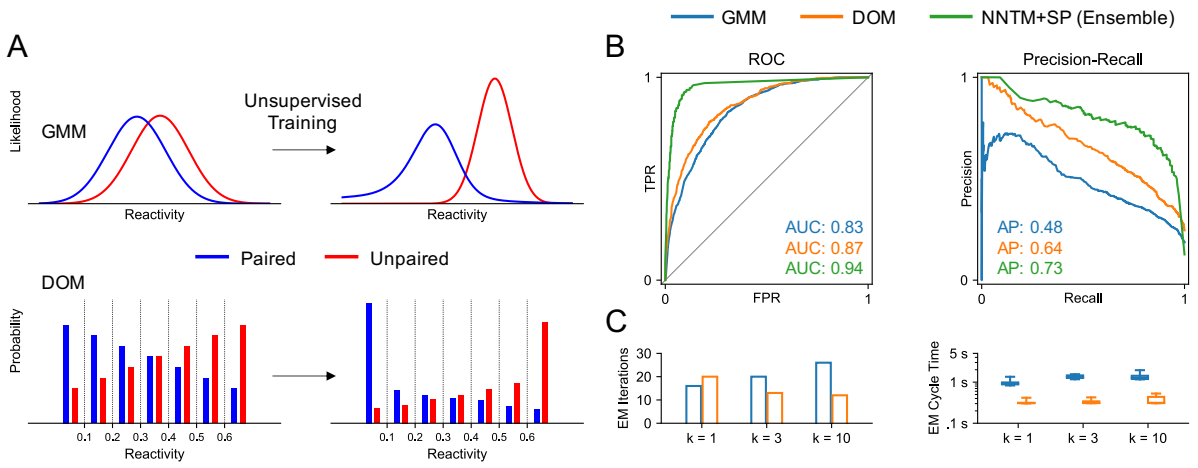


Figure 3.7: A discretized observation model (DOM) of reactivity improves hairpin detection precision when compared to a Gaussian mixture model (GMM). **(A)** Schematic illustration of GMM and DOM approaches in the content of *patteRNA*'s unsupervised learning scheme. The DOM is founded upon a percentile-based discretization of reactivities which yields a discrete emission probability scheme. The discretization scheme it itself optimized during training based on Bayesian information criteria (BIC) of models using progressively smaller bins. **(B)** Receiver operating characteristic curves and precision-recall curves when mining regular hairpins in a reference dataset ("the Weeks set," see text) with *patteRNA* using either GMM (blue) or DOM (orange) approaches, or when using data-driven NNTM-based folding (green). **(C)** Timing benchmarks of unsupervised training via GMM and DOM on the Weeks set. Shown are the number of EM iterations required for convergence on the Weeks set and time required for a single EM iteration. 5 repetitions were used when measuring EM cycle times.

We benchmarked the capacity of *patteRNA* to identify regular hairpins in the Weeks set via the GMM and DOM. We assessed their discriminatory power primarily via the re-

ceiver operating characteristic (ROC) and precision-recall curve (PRC), which are shown in Figure 3.7B. Our results indicate that the DOM approach improves both the area-under-the-curve (AUC) of the ROC and the average precision (AP) of the PRC. Although the improvement to AUC appears minor, average precision was increased from 0.48 with a GMM to 0.64 with a DOM. Precision is a crucial performance metric in structure motif mining where the vast majority of scored sites are negatives (even with sequence constraints applied), so the improvements seen in the DOM are important through this perspective. Notably, precision at the highest scores is much better in the DOM compared to the GMM, which is susceptible to numerous negatives at the highest hairpin scores despite decent precision at moderate scores. This is evidenced by the large fluctuations in precision at low levels of recall for the GMM (see the top left of precision-recall plot in Figure 3.7B). The DOM approach, on the other hand, is far more reliable for returning positive hits at the highest scores. Figure 3.7B also includes a benchmark for data-directed NNTM folding algorithms which shows that *patteRNA* is, although improved via the DOM, generally unable to match the precision of RNA folding. Notably, NNTM folding was performed with an ensemble-based approach, which, although much slower, outperforms a single MFE calculation [101].

Importantly, the presented results show overall performance on the collection of all regular hairpins, which is comprised predominantly by motifs with shorter stems. Shorter stems present a challenge to *patteRNA*, as fewer base pairs render sequence constraints less effective in controlling the number of negative sites considered in the analysis. When comparing performance on individual motifs, however, we find that *patteRNA* matches the precision of NNTM-ensemble methods for longer stems. In some cases, such as hairpins with stem length 6 and loop length 7, it even surpasses the performance of the NNTM approach (see Figure 3.8). We also observe a universal trend for the DOM to outperform the GMM at the motif-level, further validating its superior performance.

Not only does the DOM improve precision, but the model itself is described by fewer parameters and trains faster than a GMM. As seen in Figure 3.7C, faster training is achieved in two distinct ways. First, the DOM generally requires fewer EM iterations to converge. Second, EM iterations are significantly faster. The latter is presumably due to the DOM's simpler M-step formulation, which reduces to simple counting as opposed to the GMM which requires multiplication and squaring to update the means and variances of each Gaussian kernel.

Given the rapidly evolving field of structure probing and disparate statistical proper-

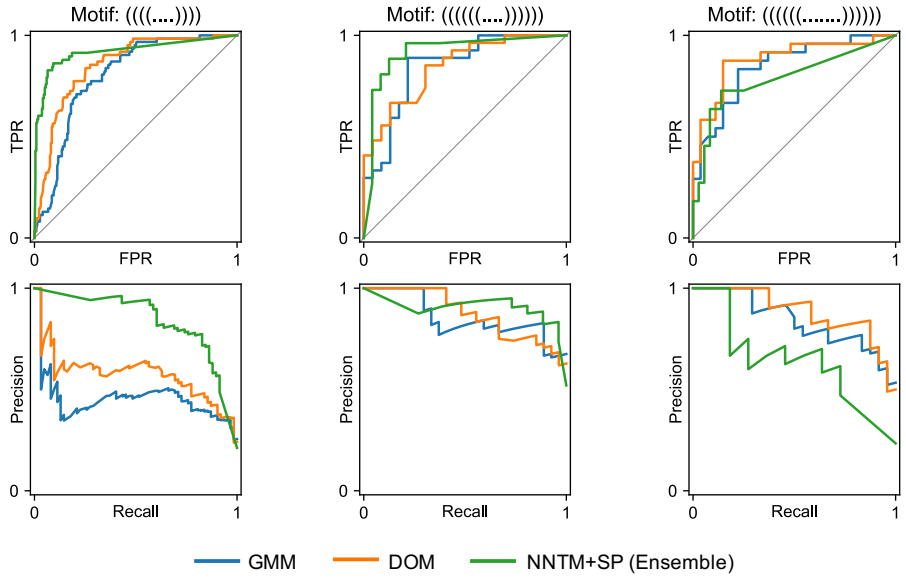


Figure 3.8: Performance of *patteRNA* (GMM), *patteRNA* (DOM), and NNTM+SP (Ensemble) approaches on identifying locations of individual motifs.

ties of SP datasets [20], we also investigated whether the benefits from the DOM generalize to other data distributions. Different probes have different quality [13, 20], different conditions yield different quality [20], and the quality of probes is constantly improving [117]; therefore, adaptability of methods is crucial. Benchmark datasets like the Weeks set are not currently available for the plethora of probes used, so we resorted to simulations. We constructed several artificial datasets and benchmarked *patteRNA*'s performance via the GMM or DOM approaches. We sampled reactivities for the underlying structures in the Weeks set according to various state distributions, including empirically-fitted distribution models from Sükösd et al. [180], referred to as the Heitsch distributions, as well as a collection of mock distributions with varying classification power (i.e., various degrees of separation between the state distributions). For each scheme, 50 replicates were created, and we benchmarked performance against both the regular and log-transformed data. We note that the fidelity of the GMM is dependent on the Gaussianity of the data, presenting a weakness of this approach as the decision to log-transform can have a major impact on scoring efficacy.

The results of the benchmarks are shown in Table 3.4. Generally speaking, the DOM matches or exceeds the performance of the GMM. Depending on the data properties, the DOM's performance gain ranges from minute to transformative. In only one of the benchmarks did the GMM outperform the DOM (poor quality Gaussian/Gaussian data), and only by a small margin. This specific outcome might be explained by the DOM's

simplification of SP data which effectively clips extreme reactivities when discretizing the data. In datasets of poor quality, the most extreme reactivities likely provide the only opportunity for reliable inference on pairing state, so it's possible that the relatively coarse discretization scheme reduces the information content of the data. Regardless, it's worth noting that data of such poor quality is uncommon, especially in light of ongoing improvements to experimental protocols and probe quality [117, 195, 167, 175]. Our results also demonstrate the adaptability of the DOM and its robustness to non-Gaussian data, which render the method broadly applicable. When using the DOM, log-transforming is largely irrelevant to model performance, as the discretization scheme is founded on data percentiles. The lone exception to this rule is when handling reactivities below zero, which are necessarily binned together if data is log-transformed.

Overall, these results demonstrate the benefit of the DOM approach in more efficiently and effectively mining structures from SP data. Note, however, that the GMM still provides a specific utility when one's objective is to arrive at continuous models of the state reactivity distributions (e.g., to use for simulations, or for data inspection). *patteRNA* includes both implementations such that the respective approach can be used depending on the intended use-case.

	Mean AP			
Data Scheme	**GMM**	**GMM (log data)**	**DOM**	**DOM (log data)**
Heitsch Distributions (Sükösd 2013)	0.43	0.58	**0.63**	**0.63**
Gaussian / Gaussian (poor)	0.32	**0.36**	0.34	0.34
Gaussian / Gaussian (medium)	0.48	0.48	**0.49**	**0.49**
Gaussian / Gaussian (high)	0.65	0.62	**0.72**	**0.72**
Exponential / Gaussian	0.58	0.55	**0.71**	**0.71**
Exponential / Exponential	0.52	**0.57**	**0.57**	**0.57**

Table 3.4: Average precisions of *patteRNA* for hairpin mining when utilizing a Gaussian mixture model (GMM) or discretized observation model (DOM) of reactivity against various artificial data schemes (see Table 3.3). For all benchmarks, average precision was averaged over 10 replicates. Bold entries highlight the best performing approaches for each scheme. AP: average precision

3.3.4 Summarizing Structuredness in RNAs from Hairpin Detection

As hairpins comprise a large fraction of structural elements, we sought to utilize *patteRNA* to quantitatively summarize local "structuredness." Due to the plethora of cellular processes affected by RNA structures, there are numerous contexts in which summarizing local structure is important. To name a few examples, one might wish to find structural domains and druggable pockets in viral genomes [121, 145, 114], quantify connections between mRNA structure and gene regulation [81, 101, 171, 161, 154, 64], identify transcriptome-wide where RNA is differentially affected by particular stimuli [25, 125], or compare structure between conditions and/or logical regions of genomes [132, 48, 50]. The most popular approach for quantifying structuredness relies on a combination of two metrics: local reactivity and local Shannon entropy. Local reactivity is generally computed via a rolling mean or median with windows ranging 25-500 nt, while local Shannon entropy derives from base pairing probabilities computed via NNTM folding routines. The combination of these two metrics yields regions which are largely unreactive (i.e., base paired) and stable (i.e., tending to adopt one conformation). We note that each metric by itself is generally insufficient in this context, as low reactivity regions sometimes include regions which see multiple competing conformations (but are nevertheless highly paired), and low Shannon entropy can also be observed for regions which are preferentially single stranded.

To integrate *patteRNA*'s results into a quantification of structuredness, we propose a nucleotide-wise measure we term the hairpin-derived structure level, or HDSL. At the highest level, HDSL combines *patteRNA*'s computed base pairing probabilities with information from hairpin searches. This allows us to consider the locations of stable hairpins in addition to the overall pairing propensity of regions, the former of which typically does not account for all structured regions (e.g., external stems, stems with numerous bulges, or stems with non-canonical base-pairing). Briefly, the posterior pairing probabilities are used as a starting point. They are then amplified at nucleotides covered by highly scored hairpins, depending on the hairpin c-score—the higher a hairpin is scored, the larger the boost. Next, the profile is clipped to [0, 1] and locally smoothed by taking a 5 nt rolling mean followed by a 15 nt rolling median (see Figure 3.5 and Methods for a complete description). In summary, HDSL integrates posterior pairing probabilities with the locations of detected regular hairpins to arrive at a nucleotide-wise measure of structuredness that is mindful of local structure elements. Whereas c-scores quantify the likelihood for

specific sites in the data to harbor a specific structure motif, HDSL is computed at all nucleotides and considers a representative collection of hairpins simultaneously. This is because all nucleotides are assigned a posterior pairing probability via the GMM/DOM-HMM, and as such, all nucleotides can be assigned HDSL. This is distinct from *c*-scores which are only assigned at sites in the data which satisfy the sequence constraints necessary for the considered targets. We explored the properties of HDSL and validated its utility as an indicator of local structure by applying it to three recent datasets that were previously used to assess local structuredness in diverse contexts.

3.3.5 Trends in Detected Hairpins Recapitulate Known mRNA Dynamics in *E. coli*

We analyzed the set of 197 mRNA transcripts (comprising 432 genes) in E. coli probed *in vitro*, *in vivo*, and *in vivo* + kasugamycin with SHAPE-MaP by Mustoe et al. [132]. In addition to Mustoe et al.'s analysis, previous studies have demonstrated that mRNAs fold differentially in cells compared to *in vitro* [171, 159, 50, 122]. *In vivo* mRNAs have been observed to be less structured than their *in vitro* counterparts, with the magnitude of structural changes correlated with translation [8, 86]. These effects have been observed most strongly in the context of the 5'UTR and CDS of highly expressed genes. Conversely, structural changes have also been observed around the 3'UTR, but evidence demonstrating both a decrease [159] and increase [8] in structures has been published in the literature, possibly correlating to the degree of post-transcriptional regulation of transcript decay [8]. We applied HDSL to Mustoe et al.'s data and investigated to what degree our measure reveals structural changes along mRNA transcripts in a prokaryotic organism like E. coli.

The results of our analysis are compiled in Figure 3.9. In Figure 3.9A, we compare averaged HDSL profiles over the 432 genes included in the study between *in vitro* and *in vivo* conditions. The averaged HDSL profiles are delineated into 3 groups: nucleotides near the start site (AUG $\pm$ 30 nt), nucleotides within the coding sequence (at least 31 nt downstream of AUG), and nucleotides in UTRs (5'UTR: 31–70 nt upstream of AUG; 3'UTR: first 40 nt after STOP). Our results demonstrate that, as expected, UTRs are generally the most structured regions of the transcripts. They also show a strong intrinsic effect for mRNA to be relatively less structured around the start codon in both conditions. Moreover, *in vivo* data show that factors in this condition work to further unfold structures around the start site, as HDSL is significantly lower around the start

codon *in vivo* than *in vitro*. Interestingly, we did not detect a strong signal for structures in coding sequences (AUG+31 nt onward) to be de-structured overall when accounting for the region around the start codon separately. It is worth noting that the reduction of HDSL around the start of coding sequences in the *in vivo* condition is only detected if the area around the start codon is delineated separately from the UTRs and CDS. Figure 3.9B shows the global HDSL trends in logical mRNA regions when (1) delineating start sites from UTRs and CDS and (2) delineating based solely on CDS/UTR boundaries. Our results indicate that HDSL is significantly different between the conditions only in the region proximal to start codons. This contrasts with the original analysis by Mustoe et al. which did not consider start sites separately (i.e., considered only CDS versus non-CDS), concluding that coding sequences are relatively less structured in cells based on a slight increase in reactivities *in vivo* versus *in vitro* for nucleotides in CDS (demonstrated via reactivity scatterplot comparison of the two conditions and a fitted linear model slope greater than 1). Our analysis suggests that global changes to reactivity profiles within CDS between conditions are not significant, yet effects specific to the start codon region are significant. These effects are likely partially responsible for previous inferences on *in vivo* structure dynamics. Notably, the specific relevance of structure around this region of mRNA transcripts has been observed and recognized as important in several other studies on organisms of varying genetic complexity [14, 35, 194, 122].

To further substantiate the effects we observed, we checked the similarity of *patteRNA*'s detected hairpins for each pairwise comparison of the three conditions included in the original study. Ideally, in the absence of significant structural remodeling between two conditions, we expect to find the same hairpins in both. On the other hand, if two conditions are substantially different, we expect to see larger differences in the hairpins detected by *patteRNA*. Searching for the aforementioned set of regular hairpins (see *Hairpins Comprise a Significant Portion of Structural Elements*) and using a *c*-score threshold of 1 to indicate a "detected" hairpin, we computed the fraction of hairpins reproducible in both conditions of each comparison (Figure 3.9C). We see that *in vivo* and *in vivo* + kasugamycin have the highest level of hairpin conservation (less than 10% of detected hairpins are not present in both conditions, meaning >90% similarity in detected hairpins). This high similarity serves as a basic quality control measure, as the *in vivo* + kasugamycin condition, although affected by changes to translation initiation, is nevertheless highly similar to the *in vivo* condition. On the contrary, comparing *in vivo* to *in vitro* data shows that 20% of detected hairpins are unique to one condition. The

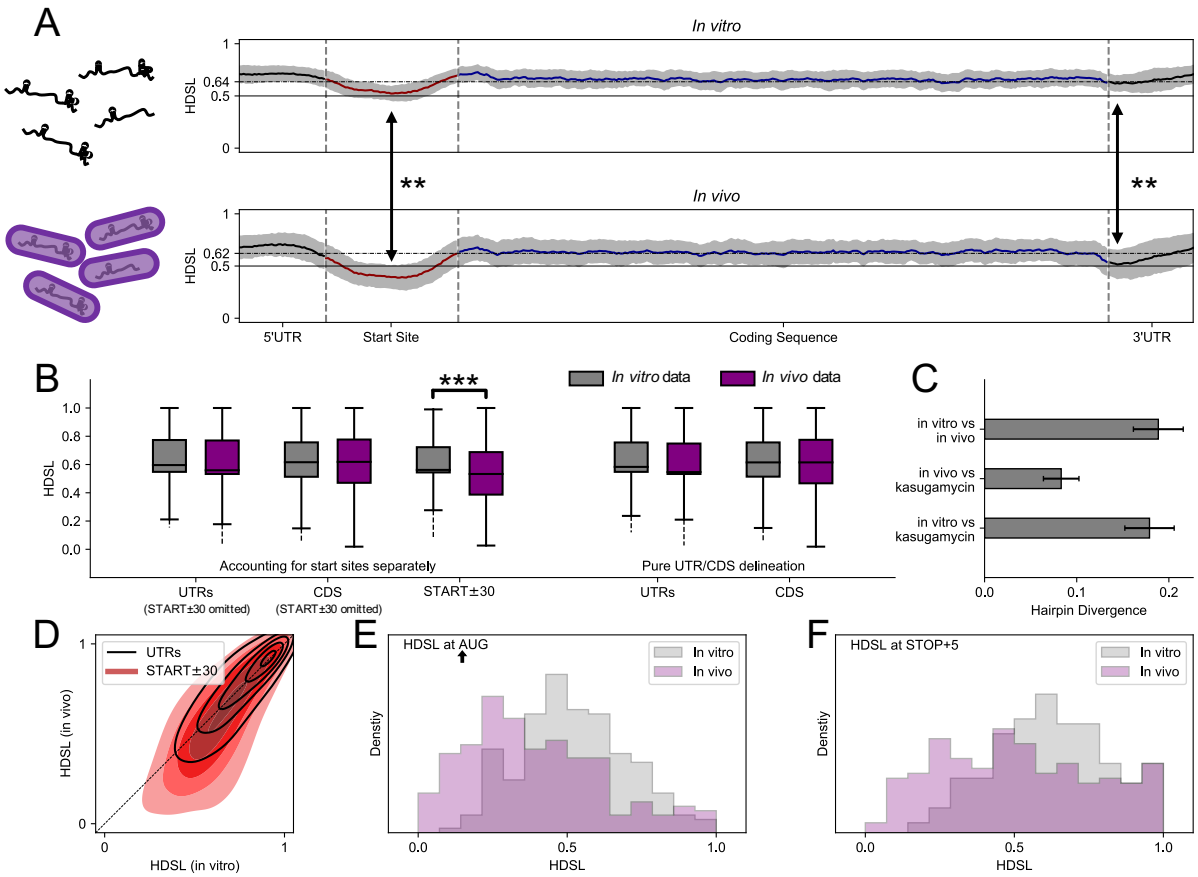


Figure 3.9: Hairpin-derived structure level (HDSL) demonstrates regional differences in structure changes between *in vivo* and *in vitro* structures for mRNA transcripts in E. coli (probed by Mustoe et al. [132]). **(A)** Averaged HDSL profiles across all genes (N = 432) for nucleotides around the start codon (±30 nt, red), within the coding sequence (AUG+31 to STOP), and 5'/3'UTRs (black). Grey area indicates the 99% CI of mean HDSL (Wald interval, see Methods). Dot-dashed lines indicate mean HDSL over all nucleotides in each condition. **(B)** HDSL trends between *in vitro* and *in vivo* conditions with delineating mRNA regions by UTRs and CDS (right) versus accounting for the region around the start codon separately (left). Delineating the region around start codons separately from CDS and UTRs reveals a signal occluded by the other delineation scheme. **(C)** Hairpin divergence (fraction of *patteRNA*-detected hairpins unique to one condition) for the three pairwise comparisons between *in vivo*, *in vitro*, and *in vivo* + kasugamycin conditions. Error bars represent the exact binomial (Clopper-Pearson) 99% CI. **(D)** 2D density plot of HDSL between the two conditions shown in (A) indicates a bias for weakly structured regions *in vitro* to become more unstructured *in vivo*. **(E)** Histograms of HDSL at the adenosine of start codons for both conditions in (A). **(F)** Histograms of HDSL at the fifth nucleotide after the STOP codon for both conditions in (A). Stars indicate Wilcoxon signed-rank tests for mean HDSL at the noted positions to be equal in both conditions. ** indicates $p < 1 \times 10^{-60}$, *** indicates $p < 1 \times 10^{-100}$.

very high level of similarity between *in vivo* and *in vivo* + kasugamycin reaffirms that the differences observed in Figure 3.9A between *in vivo* and *in vitro* reflect real differential effects, rather than the impact of biological variation or artifacts from *patteRNA*'s imperfect hairpin detection scheme.

To further investigate the differences between the conditions around start codons, we visualized the condition-wise correlation of HDSL for all nucleotides within this region (Figure 3.9D). We detected a tendency in this area for the most structured regions *in vitro* to remain structured *in vivo* (see top right of distribution, which is tightly concentrated around the diagonal). The density of HDSL in Figure 3.9D does reveal a tendency for HDSL to be reduced in the *in vivo* condition, but mostly for regions with moderate HDSL *in vitro*. Thus, the overall de-structuring effect from Figure 3.9A appears to be driven by unfolding of moderately structured regions. Figure 3.9E compares the HDSL distribution between *in vitro* and *in vivo* at the adenosine residue of the start codon. There is a noticeable reduction in HDSL in the *in vivo* condition ($p < 1 \times 10^{-60}$, Wilcoxon signed-rank test), presumably driven by translation and possibly other cellular effects destabilizing mRNA structure, as discussed above. There is also a noticeable reduction in HDSL near the start of the 3'UTR (Figure 3.9F, $p < 1 \times 10^{-60}$, Wilcoxon signed-rank test), although this effect disappears on average for nucleotides farther away from the end of the coding sequence (see Figure 3.9A). Overall, our results demonstrate that HDSL can rapidly measure local structure and gives results consistent with prior analyses.

3.3.6 HDSL Correlates Strongly with Structured Regions of SARS-CoV-2

To further explore the properties of HDSL, we applied it to the SARS-CoV-2 genome. Recently, multiple labs have independently probed the genome with SHAPE [114, 71] and dimethyl sulfate (DMS) [114, 97]. These works have resulted in a complete structure model of the genome, highlighted by the identification of structured elements across its entire length. Here, we focus on SP data generated by Manfredonia et al. [114], which contained SHAPE data both *in vitro* and *in vivo*. Other studies either had data for only one condition or relied on DMS, which only reports reactivity for A and C nucleotides.

We first characterized the consistency of *patteRNA*'s detected hairpins with the complete structure model proposed by Manfredonia et al. We took the published structure model as ground-truth, searched for all predicted regular hairpins, and quantified the accuracy of *patteRNA* via the ROC (Figure 3.10A) and PRC (Figure 3.10B). Our results

reveal good consistency between detected hairpins and hairpins in the predicted genome structure, as evidenced by AUCs around 0.88 and APs above 0.65 from analyses for both conditions.

Next, we used *patteRNA* to generate *in vivo* and *in vitro* HDSL profiles. Inspecting them in the 5'UTR reveals trends consistent with the currently accepted structure models (see Figure 3.10C) [113, 71, 215, 97, 114]. Namely, HDSL is high at known stable stem-loops, such as SL2, SL4, SL5A/C, SL7, and SL8. A weaker signal is found at SL6, which also shows differential structuredness between *in vitro* and *in vivo* data. Comparative analysis [162], *in vivo* RNA-RNA interactions [215], and multiple probing datasets [71, 114, 97] support the presence of this element. However, mutagenesis studies on a related coronavirus, murine coronavirus (MHV), demonstrated that disrupting this stem loop did not significantly affect virus viability [208]. Given that SL6 is within ORF1ab, it is possible that the element is transient in nature. That said, NMR experiments concluded SL6 stably forms and additionally measured a significantly larger internal loop than was predicted with *in silico* structure models [192]. The internal loop, also identified as a major binding site for the N protein [74], appears to be responsible for high reactivities and the observed differential structuredness of SL6 between *in vitro* and *in vivo* data. Similarly, for SL3, although comparative sequence analysis and NNTM-based folding with *in vitro* data suggest the presence of this stem-loop, *in vivo* data does not agree with its presence [114, 71, 97]. NMR investigations concluded that the stability of the element is strongly influenced by ionic conditions [192], and studies on RNA-RNA interactions suggest that this stem-loop is unfolded *in vivo* to facilitate genome cyclization, as the region is involved in a long-range interaction with the 3'UTR [215]. As such, differential structuredness between *in vitro* and *in vivo* conditions is consistent with current understandings of the stem-loop element. Finally, we observe relatively low HDSL for SL5B, and element confirmed via RNA-RNA interactions [215] and NMR [192]. NMR studies, however, suggest that the upper part of the stem is destabilized at physiological temperatures by the presence of SL5C. The presence of a bulge and high reactivities near the apical loop of SL5B subsequently result in attenuated HDSL observations around this element, as the structure scores poorly for the regular hairpin motifs considered by *patteRNA* when summarizing structuredness. Although a complete analysis of the SARS-CoV-2 genome is beyond the scope of this study, full HDSL profiles for the two conditions are included in Figure 3.11.

Generally speaking, there is a reasonable correlation between HDSL *in vitro* and

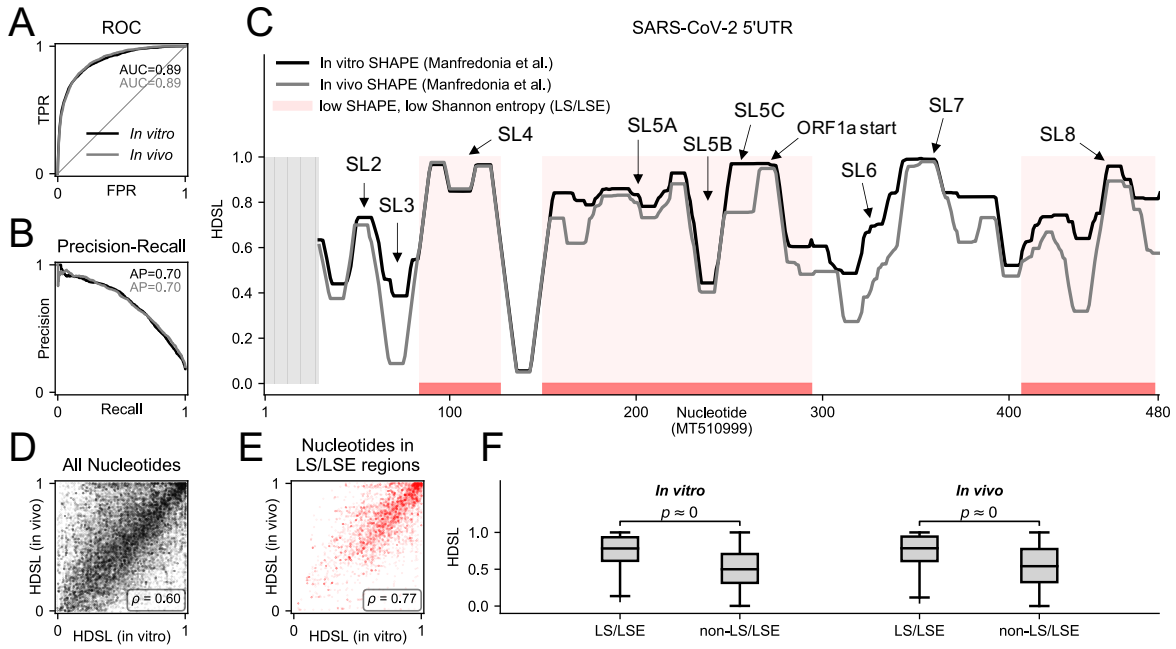


Figure 3.10: HDSL demonstrates correlated and differential structuredness between *in vitro* and *in vivo* SHAPE experiments on SARS-CoV-2 by Manfredonia et al. [114]. **(A, B)** Receiver operating characteristic curves and precision-recall curves for *patteRNA*'s detected hairpins. **(C)** HDSL profiles for the 5'UTR of SARS-CoV-2 *in vitro* and *in vivo* with low SHAPE, low Shannon entropy (LS/LSE) regions (called by Manfredonia et al.) indicated in red. Grey regions indicate no data. **(D)** Scatterplot of HDSL *in vitro* and HDSL *in vivo* for all nucleotides of the genome. **(E)** Scatterplot of HDSL *in vitro* and HDSL *in vivo* for nucleotides in LS/LSE regions. **(F)** Boxplot comparison of HDSL profiles within LS/LSE regions and outside of them for both conditions.

HDSL *in vivo* (Figure 3.10D), although some deviation is expected given that *in vivo* contexts alter RNA dynamics. We also compared the properties of HDSL within Manfredonia et al.'s called "low SHAPE, low Shannon entropy" regions (regions with locally low SHAPE and Shannon entropy). Inspecting HDSL properties within these regions confirms they are characterized by very high HDSL levels, as seen in Figure 3.10E and Figure 3.10F. We investigated this association in more detail by correlating Shannon entropy with the following: SHAPE reactivity, pairing probabilities from *patteRNA*, and HDSL (see Figure 3.12). Our results show that reactivity is loosely correlated with Shannon entropy, but pairing probabilities correlate slightly better. However, HDSL shows an even stronger correlation, suggesting that it captures structuredness better than the former measures. Lastly, our results on the SARS-CoV-2 genome indicate that HDSL profiles retain sufficient resolution to capture locations of specific structural elements (e.g., individual stem-loops in the 5'UTR), boding for the plausible use of our measure to assist in more detailed analyses of regions in addition to quantifying local structuredness.

The application of HDSL on these data allows for the unique opportunity to benchmark it against a previously characterized transcript with both structured and unstructured regions. In that context, we remark that HDSL was developed with the intention of assisting in global structure quantifications and comparisons (e.g., the analysis presented on the Mustoe data) rather than a tool for *de novo* detection of structured regions. Nevertheless, our results suggest it could also provide utility for *de novo* applications. In such cases, structured regions could be detected by defining criteria based on high HDSL that persists across long spans of nucleotides (e.g., over 50 nt). As seen in Figure 3.10, structured elements of the SARS-CoV-2 genome are typically associated with long stretches of HDSL greater than 0.8. We recommend thresholds around this value when seeking to identify structured regions. When quantifying changes in structure, however, the use of HDSL is more flexible. Depending on the specific application and degree of structure in the RNAs being studied, the magnitude of HDSL should be considered in addition to any relative changes in it across differing cellular conditions or logical transcript regions.

3.3.7 RBPs Bind RNA at Structured Regions

Corley et al. [25] devised a novel experimental procedure called fSHAPE which can detect RNA nucleotides engaging in hydrogen bonding with RNA binding proteins (RBPs). fSHAPE works by chemically probing RNA transcripts in the presence and absence of native binding factors, then quantifying the degree of modification change between the

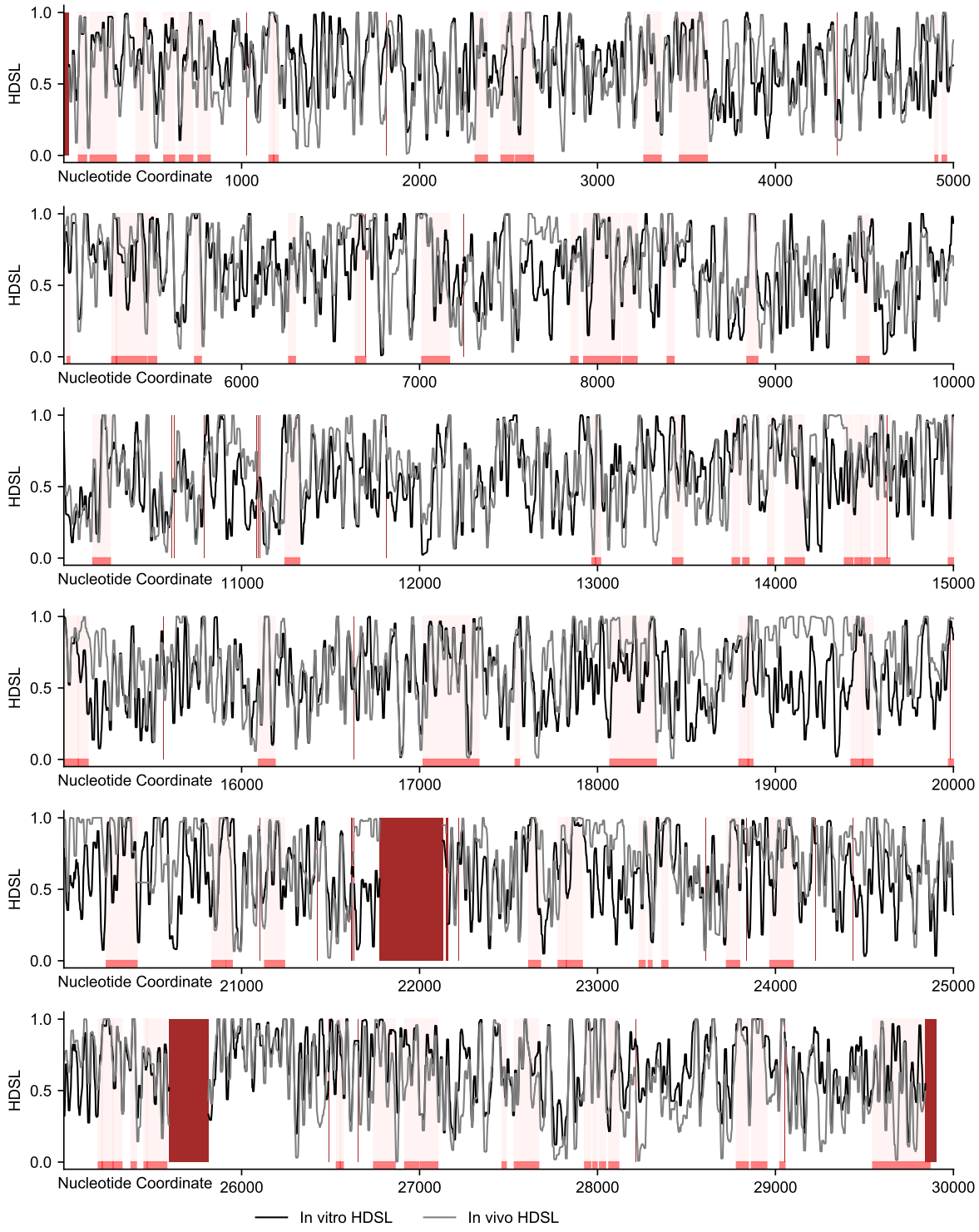


Figure 3.11: Complete HDSL profiles from *in vitro* and *in vivo* SHAPE data probing the SARS-CoV-2 genome from Manfredonia et al. Brown areas represent locations of missing data.

two conditions. Nucleotides bound by RBP would presumably be more reactive in the absence of binding factors, which translates to a high fSHAPE score. Integrating fSHAPE information with standard reactivity profiles therefore allows one to examine the struc-

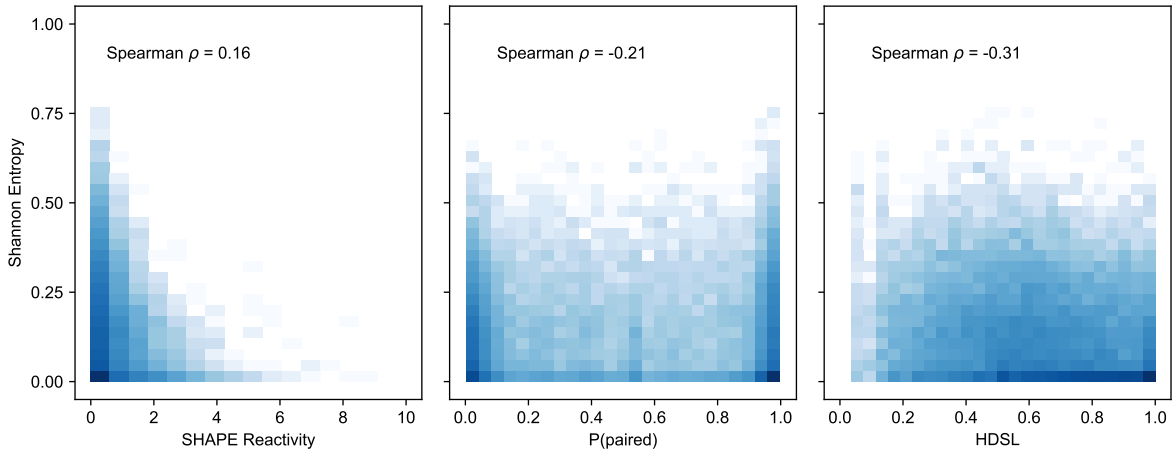


Figure 3.12: Bivariate histograms of Shannon entropy for nucleotides in the SARS-CoV-2 genome (as computed by Manfredonia et al using *in vitro* SHAPE data) against reactivity (left), posterior probability to be paired as computed by *patteRNA* under a DOM model (center), and hairpin-derived structure level (HDSL) (right).

tural context of RBP binding sites. In this regard, Corley et al. performed icSHAPE in tandem with fSHAPE to perform such analyses transcriptome-wide on human cell lines (K562, HepG2, and HeLa). Their work showed that nucleotides with high fSHAPE scores tend to fall in areas with relatively low Shannon entropy when compared to the regions flanking them, allowing them to conclude that RBP tend to associate with RNA in the general context of stable structured regions.

We sought to use HDSL to address the same question, namely, is there a structural context characteristic to RBP binding? To this end, we processed their icSHAPE data with *patteRNA*, mined for regular hairpins, and computed HDSL profiles. We first investigated what association exists, if any, between high fSHAPE nucleotides and pairing probabilities as computed by *patteRNA*'s DOM-HMM. Simply put, we found that nucleotides with high fSHAPE (fSHAPE > 2) are almost unanimously unpaired (Figure 3.13A), while nucleotides with lower fSHAPE follow a distribution encompassing both states yet biased towards paired states ($p < 10^{-307}$) for all low/high fSHAPE comparisons in Figure 3.13A, Mann-Whitney U test). The association of high fSHAPE with unpaired nucleotides recapitulates what Corley et al. demonstrated with pairing probabilities computed via partition function approaches.

However, despite the increased accessibility observed at single nucleotides with high fSHAPE, when one expands the context to the nucleotides' local neighborhood (i.e., via HDSL analysis), one observes significantly more local structure around nucleotides with high fSHAPE compared to nucleotides with low fSHAPE (Figure 3.13B). This result is

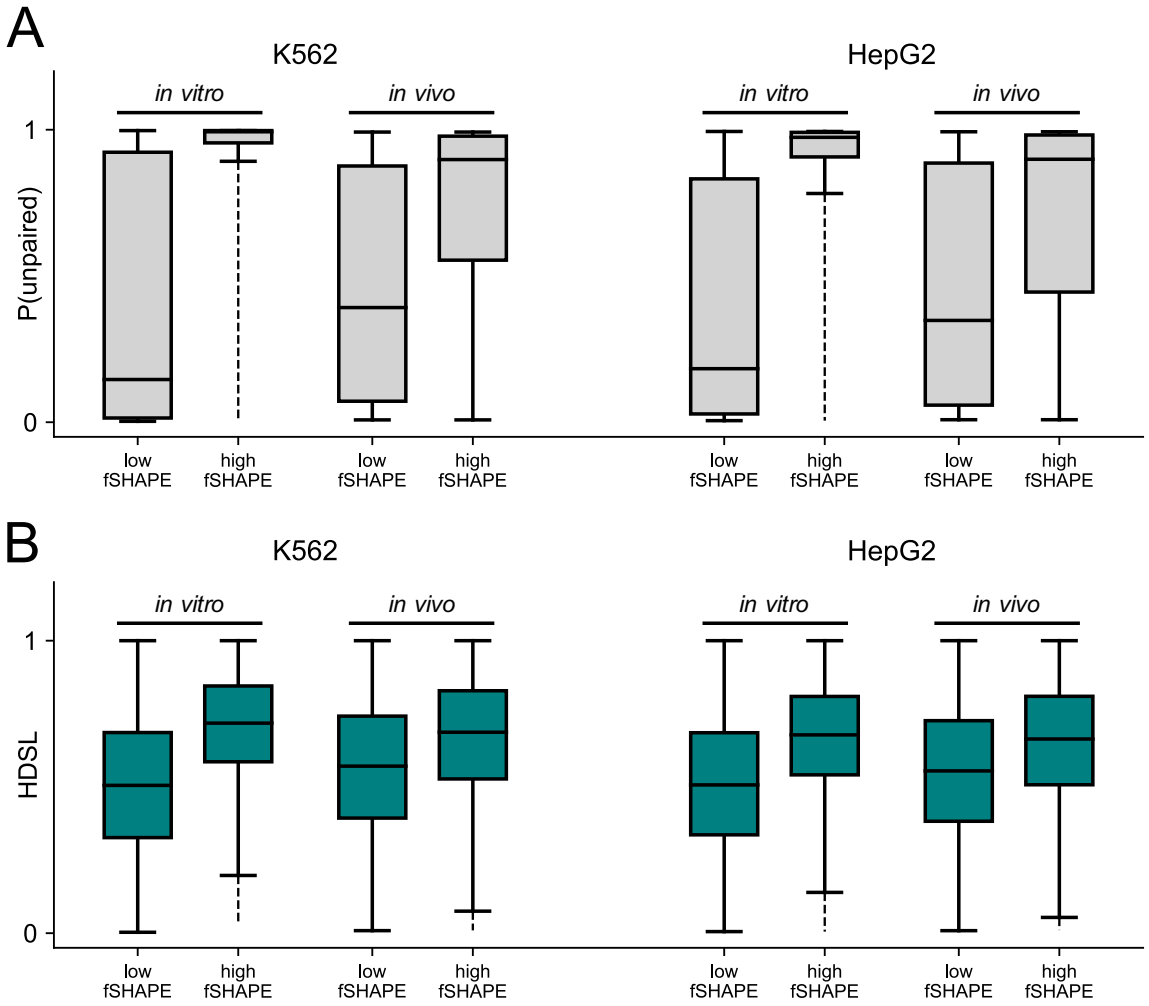


Figure 3.13: *patteRNA* demonstrates a strong association of RNA structure and RBP binding sites in human cell lines probed as by Corley et al. (Corley et al. 2020). **(A)** Unpaired probability boxplots (determined from icSHAPE reactivity via *patteRNA*'s DOM-HMM) for nucleotides with low fSHAPE (fSHAPE < 0) and high fSHAPE (fSHAPE > 2). Within each of the two cell lines, K562 and HepG2, results are presented for both *in vitro* and *in vivo* SHAPE data. **(B)** HDSL boxplots for nucleotides under the same conditions as (A). Although reactivities indicate that nucleotides likely involved in RBP binding (i.e., nucleotides with high fSHAPE) are remarkedly accessible and therefore likely unpaired, HDSL demonstrates that these reactive nucleotides more frequently occur in the general context of structured regions when compared to nucleotides with low fSHAPE. $p < 10^{-307}$ for all low/high fSHAPE comparisons in panels (A) and (B) (Mann-Whitney U test).

consistent with results from NNTM analyses performed by Corley et al., whose interpretation again depended on the computation of Shannon entropy. Our results were achieved without any folding steps and are more statistically significant ($p < 10^{-307}$) for all low/high fSHAPE comparisons in Figure 3.13B, Mann-Whitney U test) than originally demonstrated. They were also generated orders of magnitude faster than a comparable

NNTM approach, as we will show next. We note that current approaches for summarizing local structuredness from SP data alone, specifically local median reactivity, are generally insufficient for reaching this conclusion (see Figure 3.14). This highlights the capability of our method to extract more information from big SP datasets without relying on the additional assumptions and computational overhead of thermodynamic modeling.

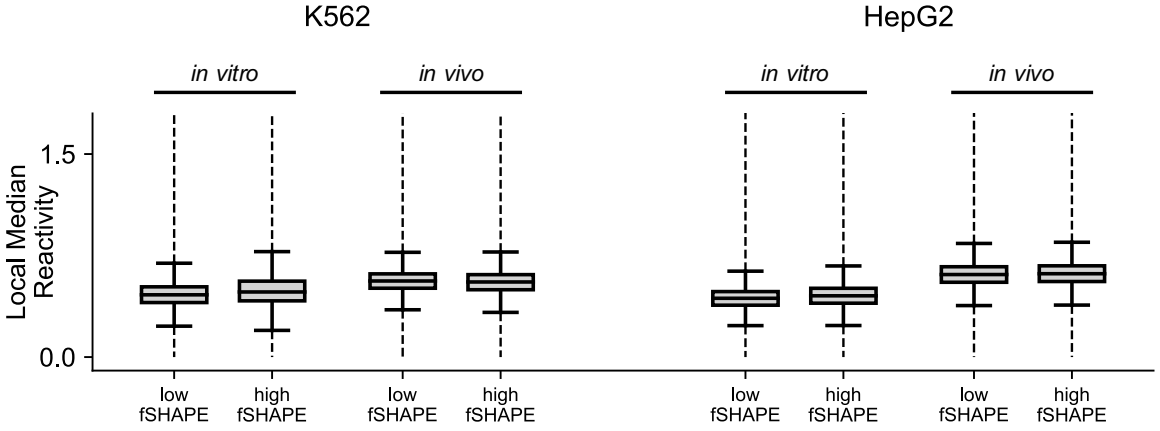


Figure 3.14: Boxplots of local median (51 nt windows) icSHAPE reactivity for nucleotides in the Corley et al. data, as classified by high fSHAPE (fSHAPE > 2) and low fSHAPE (fSHAPE < 0).

3.3.8 *patteRNA* Processes Large Data Rapidly

An especially appealing property of *patteRNA* is its ability to process big datasets rapidly. To demonstrate its speed in the context of existing methods, we timed our analyses and compared to partition function-based assessment of structure. To this end, we processed the Weeks set, SARS-CoV-2 genome, Mustoe data, and Corley data with three sliding-window partition function analyses of varying computational overhead: partition function calculations with windows of length 3000 nt, spaced 300 nt apart; windows of length 2000 nt, spaced 150 nt apart, and windows of length 150 nt, spaced 15 nt apart. The results of the benchmarks are in Figure 3.15. We observe that *patteRNA* is orders of magnitude faster than sliding-window partition function analysis for massive datasets (e.g., SP data on human transcriptomes). Specifically, *patteRNA* processed the largest dataset included in this study, the Corley data, in less than 1 hour when using a single-threaded implementation (compared to roughly 1 and 7 days for partition function calculation via 150 nt and 2000 nt windows, respectively; 3000 nt window calculations on the Corley data were not performed as they could not be completed in reasonable timeframe). Ad-

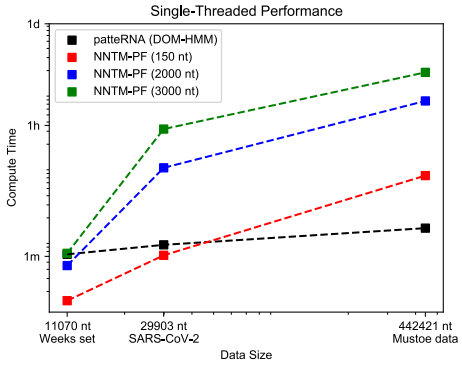
ditionally, our method is natively parallelized, and benchmarks using 12 threads allow *patteRNA* to process such data in less than 10 minutes. Analogous parallelization of partition function-based approaches on large batches of RNA transcripts is relatively simple in theory, but not natively provided "out-of-the-box" for ViennaRNA (meaning it's up to the user to program their own parallelized calls to the relevant methods). An alternative RNA folding package, RNAstructure [156], does provide scalable parallelization out-of-the-box, but the core folding implementation is about one to two orders of magnitude slower than ViennaRNA. The method was therefore not included in our comparison.

We also compared our method to RNALfold [67], an optimized routine within the ViennaRNA package designed to rapidly scan long RNAs for locally-stable structural elements. As expected, we found that this method is capable of processing large data significantly faster than the sliding-window partition function approaches, yet it is nevertheless outpaced by *patteRNA*. Moreover, this method only returns structural elements with sufficiently low free energy ("significantly low" energies judged via an SVM) and, to the best of our knowledge, has not been well-benchmarked against reference structures. Furthermore, RNALfold does not attempt to integrate its results to summarize local structuredness, which is key to the type of comparative analyses performed in this study and a central theme of a broad range of recent SP-based studies [25, 159, 8, 20]. Nevertheless, this method arrives at a more specific and comprehensive description of local structures (i.e., it can de-novo identify stems with bulges and internal loops), whereas *patteRNA*'s analyses here focus specifically on hairpin elements. We note that the incorporation of such local folding routines would likely improve the efficacy of future methods aiming to summarize local structure in large SP datasets, and our results show promising evidence that localized folding can be incorporated without major sacrifices to computational speed.

3.4 Discussion

RNA structure probing experiments are rapidly evolving in terms of their design, scale, and quality. This evolution is accompanied by a need for versatile and scalable methods capable of extracting information from diverse and massive SP data. *patteRNA* is one such tool which was developed to rapidly extract insights from such data. Here, we have demonstrated reformulation of the *patteRNA* framework which increases its speed, adaptability, and precision, enabling it to scale well to data containing millions or billions

A



B

Compute Times for Corley et al.
K562 *in vivo* SHAPE

	patteRNA	NNTM-PF (2000 nt)	NNTM-PF (150 nt)	*RNALfold*
Single-Threaded	50m 15s	1w 1d**	1d 1h	1h 35m
Multi-Threaded (12 parallel processes)	9m 14s			

**extrapolated from data in panel A

Figure 3.15: Compute times for *patteRNA* regular hairpin mining and NNTM windowed partition function (150, 2000 and 3000 nt windows) on the datasets used in this study. **(A)** Singled-threaded measured compute times on the Weeks set, SARS-CoV-2 genome, and Mustoe data. NNTM can process the Weeks set more rapidly than larger datasets due to the relatively small size of most RNA transcripts ($>$ 1000 nt) in this set. *patteRNA*, on the other hand, scales best for larger datasets. **(B)** Estimated compute times for NNTM-PF (150 and 2000 nt) on the K562 transcripts from Corley et al. (40.8 million nt) against measured compute times for *patteRNA*. To estimate single-threaded compute time for NNTM-PF, the linear time relationship between the SARS-CoV-2 and Mustoe time points was extrapolated to 40.8 million nt. Also included in this panel is the compute time for RNALfold, an optimized approach in the ViennaRNA package for rapidly scanning long RNAs for structural elements under the constraint of a maximum base-pairing distance of 150 nt. All times were computed via the arithmetic mean of three replicates.

of nucleotides. Moreover, we have shown that RNA structure can be rapidly quantified and compared in various contexts by detecting the signatures of hairpin elements. Our work expands the repertoire of analyses which *patteRNA* is capable of and demonstrates the power of simpler schemes when interpreting reactivity information. As seen with our benchmarks using a DOM approach, relatively low-resolution discretization schemes (akin to those used to highlight low/medium/high reactivities when visualizing SP data) are valuable when quantifying and mining motifs.

In the context of RNA structure determination, we note that *patteRNA* is not envisioned as a competing method or replacement to traditional NNTM-based approaches. Rather, we view the method as a tool to be used in tandem to RNA folding. As seen in Figure 3.7, NNTM-based ensemble methods provide a far more accurate prediction of specific structures and are capable of assessing the entire structure landscape including bulges, internal loops, and internal stems. The analyses via *patteRNA* shown here, on the other hand, intentionally compromise on the type of structures considered in the analysis in order to maximize the speed and scalability of the approach. This is evidenced by the relatively lower sensitivity of our method when compared to NNTM-based

partition function analyses (Figure 3.7B). It's worth noting, however, that HDSL handles the low sensitivity of hairpin detection by utilizing posterior pairing probabilities to quantify structure in regions where no highly scored hairpins are found. In other words, structured regions which house no detected hairpins are still likely to see high HDSL assuming local reactivities are moderately low. It's also worth mentioning that, although overall sensitivity on the representative set of hairpins benchmarked was relatively lower than NNTM-based ensemble approaches, benchmarks for individual motifs (Figure 3.8) reveal that patteRNA's *c*-scores are capable of matching and outperforming partition function analyses for hairpin motifs with longer stems. In summary, although HDSL considers a partial landscape of detected hairpins as provided by *c*-scores, the formulation is driven primarily by the most confident hairpin predictions, resulting in a measure of structure significantly more correlated to Shannon entropy than local reactivities or pairing probabilities alone (Figure 3.12). Nevertheless, the sensitivity of hairpin detections underpinning the method leaves room for improvement, for example, by combining simple thermodynamic assessments of local structure [152]. As a consequence of these compromises, *patteRNA* is most useful when assessing structure properties in large-scale data. For instance, as we demonstrated, it could be utilized to quantify macroscopic structural trends related to specific regions, or it could be used to identify regions of RNA which see differential structuredness associated to some factor, which might then be followed by more intensive RNA folding approaches (e.g., partition function computation). In this way, *patteRNA* helps mitigate the computational limitations of such methods, especially for those who do not have advanced computing hardware at their disposal. Finally, although analyses in this study generally focus on using *patteRNA* to derive information on structuredness via hairpins, the method itself is fundamentally a versatile structure-mining algorithm which has been demonstrated to effectively search for putative functional motifs across in transcriptome-wide data [101].

Our analysis of the SARS-CoV-2 5'UTR is distinguished from the others by a comparison of HDSL with specific structures that have been validated in a plethora of ways, including NMR spectroscopy (Wacker et al. 2020). We remarked on a great correspondence of HDSL peaks and stable structural elements, indicating that HDSL captures more than just local structure—it retains information on specific motifs with high resolution. This observation is important in the context of our analysis of Corley et al.'s fSHAPE data. Namely, the increase in HDSL around sites with high fSHAPE (Figure 3.13B) suggests the possibility that RBP frequently associate not only in the context of stable

structured regions, but specifically in the context of hairpin-like elements. RBP which recognize sequence motifs in hairpin-loops have previously been identified [5, 77], but our results demonstrate the plausibility that the association between hairpin elements and RBP is more prevalent than previously thought. This is not entirely unexpected, as RBP are known to bind both dsRNA and ssRNA in a manner that correlates with the structure of the protein [55]. Moreover, RBP binding ssRNA are observed to associate at unpaired bases stemming from RNA helix irregularities (e.g., bulges and internal loops) [78], also placing them in the context of hairpin elements. Recent studies have further documented that structured RNAs interact with a larger number of proteins than less structured RNAs [55]. Our result further strengthens the utility of *patteRNA* in mining biologically relevant structures transcriptome-wide.

Looking ahead to future development of rapid analysis of SP data, *patteRNA* is well-suited to adapt to evolving probing technologies and datasets. That being said, its current implementation does come with several limitations. First, motif mining depends on the definition of specific secondary structures, which limits its application to situations where a specific structure or small collection of similar structures can be defined. For motifs like hairpins, this means that considering situations where a bulge or internal loop may or may not be present complicates analyses due to the combinatorial explosion of unique secondary structures needed to define all possible hairpin architectures through loop size, bulge size, and bulge position. *patteRNA* is already capable of exhaustively mining such motifs, but such analyses come at the cost of significant computational overhead, generally working against the utility of the method. A more efficient approach for motif mining which naturally considers alternative similar structures within a region could theoretically address some parts of this limitation. Secondly, although the circumvention of RNA folding enables rapid computational analyses, it also handicaps the accuracy of the approach, as the energetic favorability of sequences within stems and loops is ignored. The incorporation of an optimized local folding routine could likely assist in this regard, although the coupling of such models into a statistical model like *patteRNA* is non-trivial. Nevertheless, methods like RNALfold [67] bode for the potential incorporation of NNTM-derived information without sacrificing on speed and scalability. Regardless of these limitations, however, *patteRNA* remains a viable computational method for the rapid assessment and quantification of structural trends in the largest SP datasets.

3.5 Appendix

3.5.1 Author Contributions

P.R., R.U., and S.A. developed the method and analyzed the data. P.R. and S.A. wrote the manuscript.

3.5.2 Deposited Resources

Python scripts for generating simulated datasets, computing statistical benchmarks (e.g., ROC and PRC), and post-processing of HDSL profiles related to genes in the Mustoe data will be made available on the online version of the manuscript with which this chapter is associated [151].

3.5.3 Complete DOM Formulation

The *patteRNA* Approach and Training Algorithm

The overall objective of *patteRNA* is to infer the location of RNA structure motifs, such as hairpins, within SP data. To do this, *patteRNA* considers a simplified view of RNA structure as a sequence of hidden states, namely unpaired nucleotides (state 0) and paired nucleotides (state 1). For a particular RNA transcript in a dataset, the hidden states – denoted $\omega = \{\omega_1, \omega_2, \ldots, \omega_T\}$ – (i.e., the structure) are unknown, but SP data provide information on these states in the form of values $y = \{y_1, y_2, \ldots, y_T\}$ where T is the transcript length. It is easy to see that if y provides perfect information on the structure (e.g., if a hypothetical experiment yields 0 and 1 for unpaired and paired states, respectively), the hidden states can be directly read from the data. In practice, however, the observed data is an imperfect indicator of pairing state, necessitating statistical processing mindful of the uncertainty in how observations inform pairing state.

patteRNA applies a simple model of structural context in the form of a Hidden Markov Model (HMM) to make predictions on nucleotides' pairing state. The HMM is integrated with a probabilistic model of observations (GMM), yielding a statistical framework capable of learning the properties of SP data in an unsupervised fashion. For the specific details of this implementation, we refer the reader to the original *patteRNA* study [101]. The overall procedure follows the Expectation-Maximization (EM) algorithm (see Figure 3.16), which works to arrive at a model which best explains the observed data. The final model can then be used to quantitatively identify loci in datasets which are likely to

harbor specific RNA structures (referred to as structure motifs).

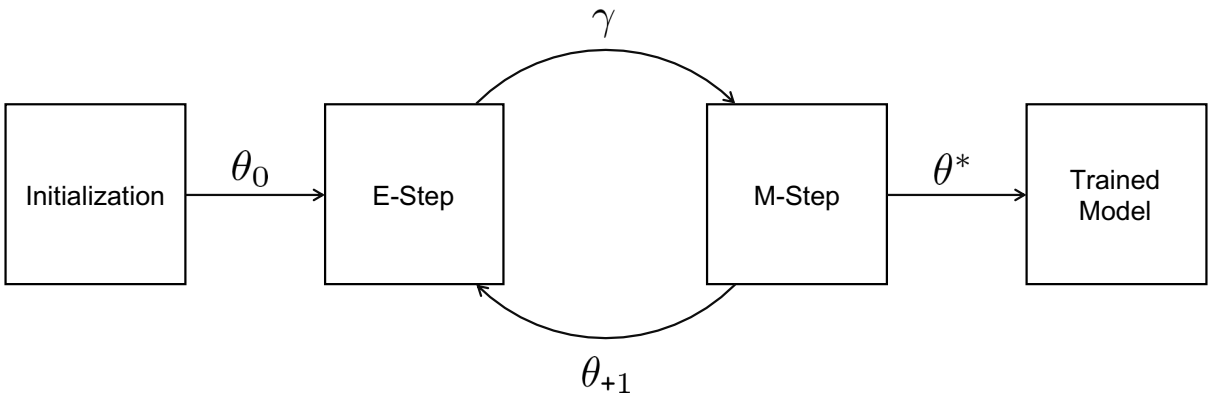


Figure 3.16: Expectation-Maximization training procedure of *patteRNA*. The process begins with an initial model, θ_0, which is constructed using a combination informed default values and key statistical properties of the dataset. The E-step is comprised by the Forward-Backward algorithm, which arrives at posterior marginals, γ, of pairing state given the model. These posteriors are then utilized by the M-step to arrive at a new model, θ_{+1}, which can subsequently be used to refine the posterior marginals in the next E-step. The EM-cycle is repeated until convergence criteria are satisfied whence the final trained model, θ^* is saved.

Modeling Emissions with a GMM

A Gaussian mixture is a continuous probability distribution constructed by summing multiple Gaussian distributions (referred to as kernels, each with their respective mean μ_k and variance σ_k^2), each with an associated weight, w_k, such that $\sum_K w_k = 1$ and therefore $\int_y \text{GMM}(y) = 1$, where $\text{GMM}(y) = \sum_K w_k \mathcal{N}(x; \mu_k, \sigma_k^2)$. K refers to the number of Gaussian kernels comprising the mixture, and as K increases, so does the flexibility of the mixture distribution to approximate continuous distributions of arbitrary shape.

Here, we describe the most relevant aspects of the GMM as needed to introduce the analogous discretized observation model forthcoming in the next section (see Ledda and Aviran [101] for a complete description of how a GMM is coupled to an HMM to yield an unsupervised learning framework). This includes the definition of the GMM likelihood (which is used in conjunction with the HMM during the E-step to compute state probabilities), how the likelihood function is augmented to account for two cases of special observations (zeros and missing values) which are given special treatment, and how the parameterization of the GMM is updated during the M-step.

The GMM primarily serves to provide a model of emission likelihoods (i.e., given that a nucleotide t has state $\omega_t = i$, what is the likelihood of emitting an observation y_t), which are then utilized in conjunction with the HMM to disentangle the statistical properties

of observations from each state ($i = 0$ for unpaired states; $i = 1$ for paired states). Importantly, this means that there are two GMM distributions within our algorithm—one for unpaired and paired states, respectively. In other words, the GMM is used to compute the state emission likelihoods, $b_{0,t}$ and $b_{1,t}$, where

$$b_{i,t} = P(y_t|\omega_t = i) \tag{3.3}$$

$$= \sum_K w_{i,k}\mathcal{N}(y_t; \mu_{i,k}, \sigma^2_{i,k}) \tag{3.4}$$

However, in order to handle the practical nature of SP data, some augmentation is required. Specifically, SP data is understood to often contain missing values (denoted $\varnothing$), which cannot be processed directly through the GMM distribution. Additionally, SP data often contain zeros. This is not a problem on its own; however, *patteRNA* is typically applied to log-transformed observations, because the log-transform has been observed to induce Gaussianity (and subsequently improve the quality of training). Because zeros (and negative values) cannot be log-transformed, they must be handled separately from the continuous GMM when log-transforming observations. To handle missing values and zeros, the following modification is made:

$$b_{i,t} = \sum_K b_{i,k,t} \tag{3.5}$$

where

$$b_{i,k,t} = \begin{cases} \phi_i w_k & \text{if } \varnothing \\ \nu_i w_k & \text{if } 0 \\ (1 - \phi_i - \nu_i) w_k \mathcal{N}(y_t; \mu_{i,k}, \sigma^2_{i,k}) & \text{otherwise} \end{cases}$$

Here, $\phi_i = P(y_t = \varnothing|\omega_t = i)$ and $\nu_i = P(y_t = 0|\omega_t = i)$. This augmentation of the emission likelihood accounts for the two additional discrete special cases while ensuring that $\int_y b_{i,t} = 1$.

Optimizing the parameters of the GMM during the M-step amounts to a pseudo-counting problem over posterior marginals γ for each kernel at nucleotides with continuous observations. Specifically,

$$\bar{\mu}_{i,k} = \frac{\sum_Q \sum_T \gamma^*_{i,k,t} y_t}{\sum_Q \sum_T \gamma^*_{i,k,t}} \tag{3.6}$$

$$\bar{\sigma}_{i,k} = \frac{\sum_Q \sum_T \gamma^*_{i,k,t}(y_t - \mu_{i,k})^2}{\sum_Q \sum_T \gamma^*_{i,k,t}} \tag{3.7}$$

$$\bar{w}_{i,k} = \frac{\sum_Q \sum_T \gamma^*_{i,k,t}}{\sum_Q \sum_{K^*} \sum_T \gamma^*_{i,k^*,t}} \tag{3.8}$$

where

$$\gamma^*_{i,k,t} = \begin{cases} \gamma_{i,k,t} & \text{if } y_t \in 0, \varnothing \\ 0 & \text{otherwise} \end{cases}$$

and Q represents the number of transcripts in the training set.

Finally, it is worth considering the number of parameters necessary to describe the GMM (and the two discrete cases). For each state, we have a GMM with K kernels, a likelihood for zero, and a likelihood for missing values. Each GMM kernel itself is described by three parameters (mean, variance, and weight), so the total number of parameters is

$$N_{\text{params}} = N_{\text{states}}(3K + 2) = 6K + 4 \tag{3.9}$$

For a typical training procedure using three kernels, this amounts to 22 parameters.

The Discretized Observation Model

The discretized observation model (DOM) serves to replace the GMM implementation by providing an alternative approach for describing emission likelihoods. The model works by discretizing the observed SP data, y, into some number of bins, each of which can then be treated as a discrete class. The emission likelihood function therefore becomes a discrete probability mass function lacking any specific architecture constraining the shape of the distribution. In other words, we define emission likelihoods not directly on y_t, but rather on a discretization transformation of the observations $\mathbb{D}(y_t)$.

$$b_{i,t} = P(y_t|\omega_t = i) = P(\mathbb{D}(y_t)|\omega_t = i) \tag{3.10}$$

Assuming the boundaries of a binning strategy with K bins are represented by $\mathbf{B} = \{B^{(1)}, B^{(2)}, \ldots, B^{(K-1)}\}$, the discretization function is defined in the following manner.

$$\mathbb{D}(y_t) = \begin{cases} 1 & \text{if } y_t < B^{(1)} \\ 2 & \text{if } B^{(1)} \leq y_t < B^{(2)} \\ 3 & \text{if } B^{(2)} \leq y_t < B^{(3)} \\ \vdots & \\ K-1 & \text{if } B^{(K-2)} \leq y_t < B^{(K-1)} \\ K & \text{if } B^{(K-1)} \leq y_t \\ K+1 & \text{if } y_t = \varnothing \end{cases} \tag{3.11}$$

or, more succinctly,

$$\mathbb{D}(y_t) = \sum_{k}^{K+1} k H^{(k)}(y_t) \tag{3.12}$$

where

$$H^{(k)}(y_t) = \begin{cases} 1 & \text{if } B^{(k-1)} \leq y_t < B^{(k)} \\ 1 & \text{if } y_t = \varnothing \text{ and } k = K+1 \\ 0 & \text{otherwise} \end{cases}$$

For convenience, we also define $B^{(0)} = -\infty$ and $B^{(K)} = \infty$. This means that for log-transformed data, observations which were originally less than or equal to zero are automatically placed into their own class ($H^{(k)}(y_t) = 0$) and processed separately, as done in the GMM.

We now have a finite number of classes, and the emission likelihoods are just discrete probabilities which can be (1) initialized in a simple and reasonable fashion (see Figure 3.17) and (2) optimized according to a very simple scheme during the M-step:

$$P(k|\omega = i) = p_{i,k} = \frac{\sum_Q \sum_T \gamma_{i,t} H^{(k)}(y_t)}{\sum_Q \sum_T \gamma_{i,t}} \tag{3.13}$$

This single equation represents the entire M-step for updating the emission likelihoods for all bins, and can computed extremely quickly due to the vectorized nature of the arithmetic. Moreover, this model tends to require fewer parameters than a comparable GMM. Two parameters are required for each of the K bins, plus two more for an additional bin to handle missing data. Thus, the number of parameters describing the DOM is given by

$$N_{\text{params}} = (K+1)N_{\text{states}} = 2(K+1) \tag{3.14}$$

For a DOM model with $K = 7$, this amounts to 16 parameters.

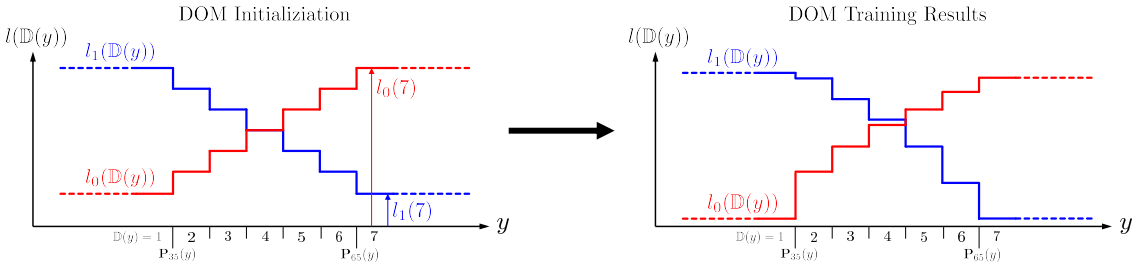


Figure 3.17: Illustration of an example initial parameterization of the DOM approach and training results.

Binning Strategies

he locations of the bins underpinning DOMs are determined based on the quantiles/percentiles of the data. In this way, data without zeros or negative values are guaranteed to be binned (and trained on) identically regardless of whether the data are log-transformed or not.

The specific bin edges for a given number of bins, K, are computed from a linear space of percentiles. For instance, if $K = 2$, a single bin edge is placed at the 50th percentile of the data. In other words, $B^{(0)} = -\infty$, $B^{(1)} = \mathbf{P}_{50}$, and $B^{(2)} = \infty$, where $\mathbf{P}_{50}$ is the 50th percentile of the data. For $K = 3$, the two finite bin edges would be $B^{(1)} = \mathbf{P}_{33.3}$ and $B^{(2)} = \mathbf{P}_{66.7}$. This linear spacing of bins logically continues for larger values of K. Data percentiles are computed via the `percentile` method of the NumPy Python module using linear interpolation of non-integer percentiles where appropriate.

When modeling SHAPE data, we found that model efficacy during the scoring phase could be further maximized by constraining the binning interval to only within the range of observations falling between the 35th and 65th percentiles. Thus, the bin edges are computed as a linear space of percentiles between 35 and 65 instead of 0 and 100. Due to the improved scoring performance observed with this approach, this is the default behavior of *patteRNA*. We recommend using this default approach when mining SHAPE data for motifs, but the bounds can be turned off by using the `--no-bounds` flag when calling the *patteRNA* command. This flag may yield better results for probing data which is statistically dissimilar from SHAPE.

Model Selection

The number of bins in the DOM (and analogously, the number of kernels of the GMM), is iteratively increased until a minimum in Bayesian information criteria (BIC) is found [149]. BIC is defined as

$$\mathrm{BIC} = -2 \log \mathcal{L} + \nu \log n \tag{3.15}$$

where $\log \mathcal{L}$ is the log-likelihood of the model, ν is the number of parameters describing the model, and n is the number of observations used during training.

Within *patteRNA*, K is initially set to be 4 bins for the DOM approach, as fewer bins were almost always found to yield higher BIC in both real and simulated SP datasets. Typically, *patteRNA* converges to a trained model comprised by between 5 and 10 bins. The number of bins can also be manually controlled by the user using the `-k` flag.

Chapter 4

Accurate detection of RNA stem-loops in structurome data reveals widespread association with protein binding sites

Acknowledgement: *This chapter is reproduced from an article in preparation for submission to the journal RNA Biology (Radecki P., Uppuluri R., Deshpande K., and Aviran S. 2021) [152]. Pierce Radecki was lead author on this manuscript. Rahul Uppuluri and Kaustubh Deshpande were undergraduate volunteers in the Aviran Lab. Author contributions are listed at the end of the capture.*

4.1 Introduction

Beyond serving as a carrier of genetic information, RNA plays key mechanistic roles in diverse cellular processes. These functions are regularly attributed to the molecule's ability to fold into specific structures [48, 131, 44, 27, 166, 41, 7]. Driven by its flexible backbone and the complementarity of nucleotide bases comprising it, the structures of RNA are intricate and dynamic [179, 48]. Although high-quality structure models of RNA transcripts are important in understanding their function and dysfunction, accurate determination of structures, especially *in vivo*, is challenging. High-resolution structure models can be obtained with experimental measurements from X-ray crystallography [68], nuclear magnetic resonance [46], and cryo-EM [44, 211], yet these methods are low-

throughput and incapable of measuring structures in living cells. Comparative sequence analyses can also glean structural information from sequence homologies, but this process depends on a sufficiently large set of related sequences, which limits the scope of their application [140, 58, 98]. The advent of nearest-neighbor thermodynamic models (NNTM) combined with efficient energy minimization algorithms were a critical step in increasing the throughput of structural predictions by enabling computational folding based on nucleotide sequences [137, 217]. Despite their popularity, however, the accuracy of predictions is generally poor, especially when applied *in vivo* or to long transcripts [49]. structure profiling (SP) experiments have emerged as a practical and high-throughput approach to measuring the structure of RNA molecules [84, 205]. Although these methods are diverse, they help inform structure models by providing nucleotide-level measurements of conformational characteristics. Importantly, they can be applied *in vivo*, and, with the advent of next-generation sequencing, are scalable.

SP experiments follow common principles [20]. Briefly, they expose RNA to chemical reagents or enzymes that react with parts of the molecule in a structure-dependent manner (e.g., when using common acylation reagents, single stranded nucleotides react more strongly than double stranded regions). This reaction induces the formation of adducts or cleavages [84, 205, 204, 117], which can then be detected during sequencing as either truncations or mutations in reverse-transcribed cDNA fragments. The rate of truncation or mutation at each nucleotide is then quantified and converted into a measure called reactivity that summarizes the nucleotide's structural context; the reactivities across a transcript are termed its reactivity profile [4]. The incorporation of these data in NNTM-based folding algorithms was shown to greatly improve their accuracy [29, 109]. In this regard, SP data have served to supplement the thermodynamic models by providing direct information on the measured conformation, which is especially relevant when predicting structures *in vivo*. That said, SP experiments have scaled massively, enabling the profiling of an entire transcriptome, termed the structurome. NNTM-based folding, however, is a computationally intensive process that scales as $\mathcal{O}(L^3)$ with the length of an RNA in most applications. For transcriptomes, which contain many tens of thousands of transcripts—each of which may be thousands of nucleotides long—the computational cost associated with folding has begun to inhibit comprehensive NNTM-based analyses of structurome data. This has warranted the development of methods designed to accommodate the growing scale of SP data in making structural assessments. Such methods are useful when seeking to inspect and quantify biologically relevant changes in the structur-

ome—for instance, to highlight structural changes between different cellular conditions [132, 187], inspect the structural context of relevant regions, such as splice sites, miRNA targets, or alternative polyadenylation sites [161, 209, 35], profile the degree of structure across different types of mRNA [35], or explore the interplay between the structurome and RNA-protein interactome [25].

We previously introduced *patteRNA* as a method to address this need for scalable analysis of SP data [101]. Rather than perform complete RNA folding, it was developed to rapidly mine local structure elements from reactivity profiles via an unsupervised, versatile, and NNTM-free approach. In short, the method couples a statistical reactivity model—e.g., a Gaussian mixture model (GMM) or discretized observation model (DOM)—to a Hidden Markov model of structure [101, 149, 151]. A parameterized model subsequently enables quantitative scanning for locations that are likely to harbor a specific structure element. Versatility is a key characteristic of the method; namely, it leverages an unsupervised training step to learn the properties of any dataset (i.e., to parameterize the reactivity-structure model) before mining it. This is crucial for the automated handling of data from diverse SP experiments that consequently have disparate statistical properties [30]. Moreover, the NNTM-free nature of *patteRNA* helps it scale to the structurome level and also confers flexibility to rapidly mine complex structural elements such as pseudoknots or self-contained tertiary interactions without any significant increase in computational complexity [101]. In short, any target that can be defined via a local reactivity pattern or base-pairing arrangement can be quickly mined.

By scanning reactivity profiles alone, *patteRNA* was able to achieve reasonable accuracy when mining canonical motifs, such as hairpins/stem-loops [101, 151]. However, there was room for improvement via integration of NNTM-derived sequence information, which we believed could likely assist in situations where SP data is inconclusive. However, effective integration of NNTM with the statistical framework underpinning our approach is itself a non-trivial problem—we sought to not only improve performance, but also to maintain speed and versatility. To address this problem, we took a data-driven approach in which a large set of reference structures guided the construction of an integrative scoring classifier which considers statistical characterization of SP data in additional to local thermodynamics. This is a deviation from the unsupervised nature of our approach; nevertheless, we aimed to ensure that the classifier maintains the method's automated adaptability in analyzing any type of SP dataset. The impact of including thermodynamics on the method's efficiency was also carefully considered, as we sought to maintain

a balance between improvements to prediction quality and the increased computational overhead triggered by thermodynamic modeling.

Our results describe the development a data-driven logistic regression classifier to more accurately identify the locations of target structural elements. It considers the thermodynamic properties of local regions in addition to reactivity profiles when making predictions, which strongly improves precision, especially for shorter structure motifs. The classifier is suitable for all types of canonical local structure motifs and maintains the versatility of *patteRNA* in handling diverse types of SP data. In this process, we also create a large-scale set of RNAs with known structures from RNA STRAND [2] and use it in conjunction with a data simulation scheme to extensively train and validate our approach. Although underpinned by simulated data, we find this resource to be more effective at training data-driven classifiers than smaller sets of real data and believe it can serve as a useful resource for method development. Moreover, we apply the classifier to an integrative transcriptomic dataset on K562 and HepG2 cells that quantifies both structure and RNA binding protein (RBP) interactions [25]. We demonstrate that stable stem-loops are almost always associated with evidence of RBP binding, and that this association exists across a diverse set of stem-loop configurations. In the context of the latest RBP studies, our results expand on previous observations of the RNA-protein interactome and refine our understanding of the roles played specifically by stem-loops. This also highlights the power stem-loop profiling, where relevant tools are lacking. Overall, our work provides a major improvement to *patteRNA* while simultaneously strengthening our understanding of the functional roles played by canonical structure elements.

4.2 Materials and Methods

4.2.1 *patteRNA* Overview

patteRNA works in two phases: training and scoring. The training phase involves the utilization of an unsupervised Expectation-Maximization (EM) scheme coupled to a Hidden Markov Model (HMM) to estimate the reactivity distributions for unpaired and paired states, respectively. With these distributions in hand, *patteRNA* searches for a target motif in SP data as previously described [101, 149]. Briefly, all subsequences (referred to as sites) which satisfy the sequence constraints underlying the base pairing arrangement of the target motif are considered. These sites are then each assigned a score, which quantifies the overall consistency of the reactivity data within the site with the pairing state

sequence of the target (a higher score indicates a better agreement between the reactivity profile and target motif). Scores are further processed into *c*-scores via a normalization scheme based on an estimated distribution of scores associated with null sites (sites that do not harbor the target motif). For details on the core *patteRNA* algorithm, see [101] and [151]; for details on score normalization, see [149].

All applications of *patteRNA* in this study used default hyperparameters unless otherwise noted. When mining hairpins, the "`--hairpin`" flag was used, which searches for all hairpins with stem length between 4 and 15 nt and loop length between 3 and 10 nt. This representative collection of motifs is referred to as regular hairpins or regular stem-loops throughout our work. When mining loops, the "`--loops`" flag was used, which searches for runs of unpaired nucleotides length 3 to 10 nt flanked by one base pair.

4.2.2 The Weeks Set

The Weeks set is a dataset of 22 diverse RNA transcripts (totaling 11,070 nt) with high-quality SHAPE data and known reference structures. We use the Weeks set in this study as a reference set to benchmark the performance of *patteRNA*'s analyses and related methods on real data. This dataset was initially introduced in [101] and contains reactivity data from [29, 59, 100], see Table 4.1 for further details on the RNA molecules in the Weeks set.

4.2.3 Classifier Training Data

In order to construct a larger set of reference data by which to develop a scoring classifier, we compiled all RNA secondary structures from RNA STRAND (4,666 transcripts). Due to the presence of highly similar sequences within the data, we used CD-HIT-EST [70] to remove redundant sequences at an 80% similarity threshold, yielding 1,191 final transcripts (totaling 706,306 nt). In order to utilize these secondary structures for *patteRNA*-related analyses, we generated artificial SP data for the transcripts according to a three-state reactivity model (0: unpaired, 1: paired, 2: helix-end) with associated state reactivity distributions devised in (Sükösd et al. 2013), which we refer to as the Heitsch distributions. The distributions are defined as follows; unpaired states: exponential distribution with λ = 1.468, paired state: generalized extreme value distribution with μ = 0.04, σ = 0.040, ξ = -0.763, helix-end state: generalized extreme value distribution with μ = 0.09, σ = 0.114, ξ = -0.821. Five replicates of SP data were produced. The Python module SciPy was used to sample reactivities from the corresponding distribu-

RNA name	Length (nt)	Reference
5S rRNA (*E. coli*)	120	Hajdin 2013
5'-domain 16S rRNA (*E. coli*)	530	Hajdin 2013
5'-domain 23S rRNA (*E. coli*)	511	Hajdin 2013
Adenine riboswitch (*V. vulnificus*)	71	Hajdin 2013
Fluoride riboswitch (*P. syringae*)	66	Hajdin 2013
Group II intron (*O. iheyensis*)	412	Hajdin 2013
Group I intron (*T. thermophilia*)	425	Hajdin 2013
Group I intron (*Azoarcus sp.*)	214	Hajdin 2013
Hepatitis C virus IRES domain	336	Hajdin 2013
Lysine riboswitch (*T. maritime*)	174	Hajdin 2013
M-Box riboswitch (*B. subtilis*)	154	Hajdin 2013
Cyclic di-GMP riboswitch (*V. cholerae*)	97	Hajdin 2013
TPP riboswitch (*E. coli*)	79	Hajdin 2013
Pre-Q1 riboswitch (*B. subtilis*)	34	Hajdin 2013
SAM I riboswitch (*T. tengcongensis*)	118	Hajdin 2013
16S rRNA (*C. difficile*)	1504	Lavender 2015
23S rRNA (*E. coli*)	2904	Deigan 2008
16S rRNA (*E. coli*)	1542	Deigan 2008
16S rRNA (*H. volcanii*)	1474	Lavender 2015
P546 domain, bI3 Group I intron	155	Deigan 2008
Asp. tRNA (*S. cerevisiae*)	75	Deigan 2008
Phe. tRNA (*E. coli*)	76	Hajdin 2013

Table 4.1: RNAs in the Weeks set.

tions. The scripts used to sample reactivities for STRAND transcripts in addition to the STRAND data itself (including the sampled reactivities used in this work) are available at `https://doi.org/10.5281/zenodo.4667909` [153].

To assist in verification and benchmarking of classifiers, additional datasets were also generated by resampling (with replacement) the empirical reactivity distributions observed in the Weeks set.

4.2.4 Feature Generation

Several features were investigated insofar as their potential to provide additional information on the presence of target motifs during scoring. After preliminary investigations, we focused on the following features, in addition to the *patteRNA* c-score: cross-entropy loss (CEL) between *patteRNA* posteriors and the target state sequence, Gini coefficient of SHAPE data in a site, the local minimum free energy (LMFE), the local constrained minimum free energy (LCMFE; the local MFE with the target motif enforced as a folding constraint), and the motif energy loss (MEL; the difference between LMFE and LCMFE). Cross-entropy loss was computed as

$$\text{CEL} = \sum_i -(y_i \log p_i + (1 - y_i) \log(1 - p_i)) \tag{4.1}$$

where y_i is the pairing state of the target motif (e.g., $y_i = 0$ for unpaired states and $y_i = 1$ for paired) and p_i if the posterior pairing probability at nucleotide i of a scored site. The Gini coefficient was computed as

$$\text{Gini} = \frac{\sum_i^n \sum_j^n |x_i - x_j|}{2n^2 \bar{x}} \tag{4.2}$$

where x_i is the reactivity at nucleotide i of a site and n is the length of the target motif. The remaining three features (LMFE, LCMFE, and MEL) all depend on the thermodynamic model employed in RNA structure prediction and were computed using the ViennaRNA package (version 2.4.17) Python interface using a local window extending 40 nt in both directions from the boundaries of the target site ($c = 40$).

4.2.5 Feature Selection

To identify the set of features which best predict the presence of a target motif, we used a combinatorial approach to test various combinations of features and their potential scoring efficacy. To do this, we used the scoring feature set generated from the Weeks set hairpins. We used the c-score as a base feature in all experiments while iterating through pairwise combinations of the other features on top of it (see 4.3). Specifically, we tested all of the 2-feature approaches underpinned by the c-score and one of the other features. We then tested all of the 3-feature approaches underpinned by the c-score and all of the pairwise combinations of the other features. To quantify scoring efficacy, a logistic classifier was trained on the feature combinations and the average precision of its fitted motif probability was used to assess scoring potential.

4.2.6 Classifier Selection

After converging on a 3-feature set of *c*-score, CEL, and MEL, we explored the capacity of various binary classifiers to precisely identify true positive sites from these features as well as their ability to generalize to other datasets and target motifs beyond hairpins. After a preliminary analysis of an initial collection of standard classifiers, we explored in more detail logistic binary classification (LBC; "`LogisticRegression`" object in *Scikit-learn*), random forest classification (RFC, "`RandomForestClassifier`" object in *Scikit-learn*), linear discriminant analysis (LDA, "`LinearDiscriminantAnalysis`" object in *Scikit-learn*), and quadratic discriminant analysis (QDA, "`QuadraticDiscriminantAnalysis`" object in *Scikit-learn*). In all cases, default parameterizations were used as provided by *Scikit-learn*.

The set of hairpins mined from the STRAND dataset was used to train each classifier, and the average precision of their trained predictions was computed. Trained classifiers were then tested against the generated feature set for Weeks set hairpins and for Weeks set loops. Lastly, the classifiers were verified against feature sets obtained when mining 5 resampled replicates of the empirical data in Weeks set (for hairpins and for loops) and when mining 5 replicates of simulated STRAND sets (for hairpins and for loops). In all cases, performance was assessed by the average precision of the classifier.

4.2.7 Final Scoring Classifier Training and Selection

To train the definitive classifier used in *patteRNA*, we utilized the 5 replicates of Heitsch-simulated reactivity data for the pruned STRAND transcripts and used each replicate to generate a scoring feature set at the sites scored when mining for hairpins. These scoring feature sets were each used to train an associated logistic classifier—i.e., 5 distinct classifiers were trained simultaneously, 1 for each replicate of simulated data. Each resulting classifier was then used to process the other four feature sets (i.e., the other simulations not used to train that classifier) as well as the feature set associated with the empirical Weeks set data. The overall performance of the classifiers was assessed as the sum of the performance on the other 4 STRAND replicates (computed as the mean average precision for hairpins across the 4 replicates) and the performance on the Weeks set (average precision for hairpin mining). The classifier with the greatest total performance via this assessment was selected as the final model to utilize for distribution in the *patteRNA* method.

4.2.8 Performance Benchmarks and Verification

The accuracy of *patteRNA* and tested binary classifiers was primarily assessed via the area-under-the-curve of the precision-recall (PR) curve, referred to as the average precision (AP) of the classifier. Precision-recall curves were computed by varying a theoretical score threshold between positives and negatives, then computing the true-positive rate (recall) and precision (PPV) at each threshold. Sites were deemed true positives if all base pairs in the target motif are also present in the corresponding location of the reference structure. The *Scikit-learn* Python module was utilized to perform these computations. Scripts that perform this quantification (and others, including the receiver-operating characteristic) are available in [151].

4.2.9 Partition Function Analysis

We benchmarked the performance of partition function approaches to detect hairpins in the Weeks set by using the "`RNAsubopt`" command from ViennaRNA to generate 1000 structures for each transcript in the Weeks set, using that transcript's SHAPE data as soft constraints ("`RNAsubopt -p 1000 --shape $SHAPE_FILE < $SEQUENCE`"). For each hairpin in the generated structural ensemble, a "score" was assigned as the fraction of structures in the structural ensemble which contain the base pairs comprising that hairpin. Predicted hairpins and their scores were organized into a single list which was then processed into a precision-recall curve as was done for *patteRNA*'s predicted hairpins.

4.2.10 Analysis of Structurome and RBP Binding Data

We used *patteRNA* with the latest logistic scoring classifier described above to mine hairpins in the *in vitro* and *in vivo* icSHAPE data from K562 and HepG2 cells published by Corley et al. [25]. *patteRNA* was trained on each dataset/condition independently (e.g., K562 *in vitro* icSHAPE, K562 *in vivo* icSHAPE, HepG2 *in vitro* icSHAPE, etc.) and then used to mine them for hairpins (referred to in this analysis as stem-loops) using the "`--hairpins`" flag and default hyperparameters. We then cross-referenced the locations of high-scoring stem-loops with the fSHAPE profiles (interpreted as RBP binding signal) obtained by Corley et al. on the same cell lines.

To combine and visualize the fSHAPE profiles from searched sites which differ in terms of their stem and loop lengths, we utilized an interpolation scheme to bring fSHAPE profiles to a common length basis. fSHAPE profiles from the left and right sides of

the stem (which vary from 4 to 15 nt in length) were each processed to a length of 10, respectively. fSHAPE profiles from the loop regions were processed to a length of 6. This processing was achieved by linearly interpolating the fSHAPE profiles to a number of equally spaced points (e.g., 10 points for stems and 6 points for loops). For example, a stem of length 6 nt would be linearly interpolated to the local coordinates (1, 1.56, 2.11, 2.67, 3.22, 3.78, 4.33, 4.89, 5.44, 6), where 1 and 6 denote beginning and end of the fSHAPE profile along the stem, respectively.

Motif scores from both conditions were then combined and used to train a perceptron classifier processing condition-wise paired scores from the LBC into a predictor of strong RBP binding signal in the loop (defined as sites where fSHAPE > 2 in the stem-loop). Only sites that received a valid score in both conditions were considered in this analysis. The multi-layer perceptron (MLP) classifier object (`MLPClassifier`) from *Scikit-learn* was utilized to construct and train the classifier; the default model parameterization was used, which is defined by a single hidden layer of 100 nodes with ReLU activation following the Adam optimization algorithm [82]. Cross-validation during perceptron training was achieved by randomly setting aside 20% of the samples and using them to terminate training when convergence was observed (this behavior was defined with the hyperparameters "`validation_fraction=0.2`" and "`early_stopping=True`" when calling `MLPClassifier`. Note that the purpose of the perceptron in this case is, essentially, to fit the joint distribution of condition-wide scores as an indicator of high loop fSHAPE. We found that this non-linear relationship of scores between conditions was best captured by a simple perceptron instead of simpler linear models like logistic regression and LDA.

patteRNA was also used to mine the icSHAPE data for hairpins with bulges, which we defined as hairpins with stem length between 5 and 15 nt with one bulge (of 1-2nt) on either side of the stem. Locations of high scoring motifs were then cross-referenced against the locations of high fSHAPE, as was done for the hairpin search.

The proportion of strong RBP binding signals (defined as fSHAPE > 2) which can be explained as occurring within the loop of a detected stem-loop was quantified. In this quantification, only fSHAPE observations which coincide with valid reactivity data were included. In other words, fSHAPE data at locations lacking reactivity information was omitted, as such regions are not processed by *patteRNA* when scoring. Three score thresholds for calling detected stem-loops were used: 0.9, 0.7, and 0.5.

4.2.11 Code Availability

The latest version of *patteRNA*, version 2.1, was used for all analyses in this study. *patteRNA* is an open-source Python 3 module and is freely available at `www.github.com/AviranLab/patteRNA` under the BSD-2 license.

4.3 Results

4.3.1 *patteRNA* Overview

The overarching objective of *patteRNA* is to accurately mine structure elements from SP data in an automated fashion. To do this, *patteRNA* follows a two-step process (see Figure 4.1). The first step is the training phase, during which reactivities are utilized to iteratively optimize a joint reactivity-structure statistical model (e.g., a GMM-HMM [148, 101] or a DOM-HMM [151]). This results in an estimate of the distributions of reactivities associated with paired and unpaired nucleotides, respectively, as well as transition probabilities between paired and unpaired nucleotides. Training is unsupervised and capable of accommodating diverse data types; see [101] for a complete description of the mathematical framework.

Once the data properties have been learned, *patteRNA* mines for structural motifs in the data via a scoring step. Scoring requires the description of a specific secondary structure motif (or collection of motifs) which defines the target of *patteRNA*'s pattern recognition scheme. Typically, the user provides this motif in dot-bracket format, but *patteRNA* also has built-in routines to automatically mine some canonical motifs. For instance, *patteRNA* can automatically mine a representative set of hairpins (referred to as regular hairpins or regular stem-loops; defined as stem-loops with stem length between 4 and 15 nt, and loop length between 3 and 10 nt) via the "`--hairpins`" flag [151]. When mining a particular structural element, only loci in the provided transcripts which satisfy the sequence constraints necessary for the target's secondary structure (via Watson-Crick and Wobble base pairs) are considered during scoring; these loci are henceforth referred to as "sites." *patteRNA* scores sites by computing the log ratio of joint probabilities between the target's pairing sequence and its inverse. By default, scores are further processed into *c*-scores (comparative scores) which are a statistically-normalized measure computed by considering the significance of a score in the context of a null-score distribution constructed for each target [149]. Intuitively speaking, *c*-scores are simply

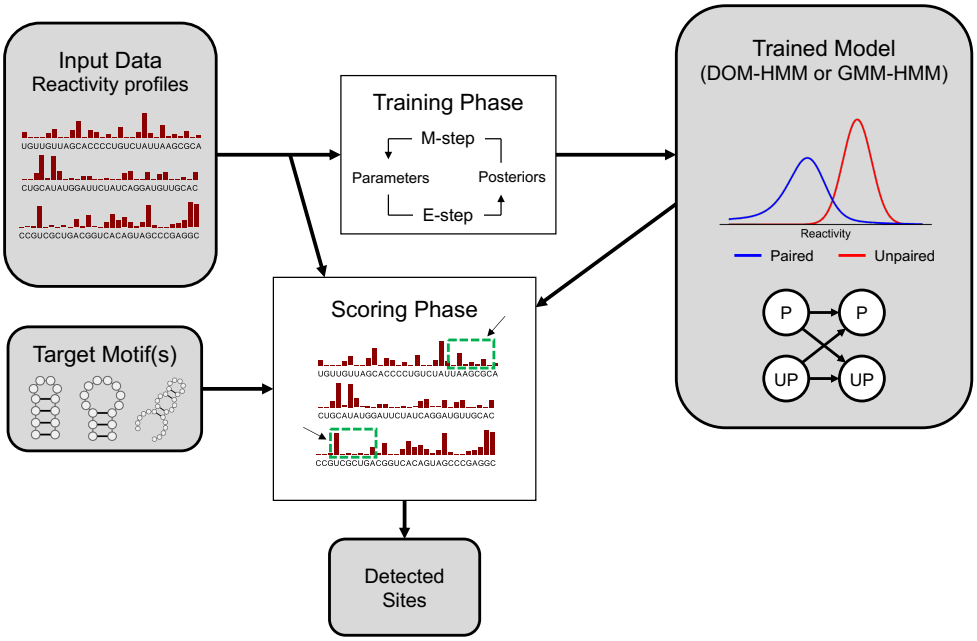


Figure 4.1: patteRNA workflow in achieving automated detection of structural elements in diverse SP data types. The statistical model used by the method is a hybrid model Reactivities are first processed in the training phase, which uses iterative optimization (EM algorithm) to learn the properties of the reactivities and structural tendencies. This process arrives as estimations of the state reactivity distributions (modeled with either a DOM or GMM) in addition to state transition probabilities and other parameters underpinning the HMM. Once a trained model has been formed, *patteRNA* can quantitatively mine the data for target motifs provided by the user. Mining is achieved by scanning the provided transcripts for loci which are compatible with the base-pairing of the motif, then scoring such sites for consistency between their reactivity profiles and the target. DOM, discretized observation model; GMM, Gaussian mixture model; HMM, Hidden Markov model.

the $-log_{10}$ of a p-value, facilitating comparative analysis of scores from different target searches. Higher scores indicate a higher likelihood of the target motif, with a c-score of 2 (corresponding to a p-value of 0.01) generally interpretable as a strong signal.

In addition to scoring, *patteRNA* can also use a trained statistical model to compute posterior pairing probabilities (i.e., for each nucleotide, the probability that it is paired or unpaired), Viterbi paths (the most likely sequence of pairing states for each transcript), and hairpin-derived structure level (HDSL) profiles (a nucleotide-wise measure of local structuredness [151]).

4.3.2 Supervised Context-Aware Scoring

patteRNA was developed as an NNTM-free method. It inspects and quantifies patterns in reactivity profiles to identify sites consistent with the presence of a sought structure motif. Sequence information is only taken into account when determining whether sites are compatible with a target motif, i.e., satisfying the sequence constraints associated with base pairs in the target. This approach has facilitated the algorithm's rapid speed when mining transcriptomic data, however, information encoded in the sequence has the potential to improve its predictions. Here, we aimed to improve *patteRNA*'s accuracy by including an assessment of information in nucleotide sequences (e.g., NNTM-based quantifications of sequence energetics). We also explored the use of additional SP data-related metrics in improving performance.

The integration of NNTM-based predictions with a statistical metric like the *c*-score is non-trivial. Therefore, we pursued the development of a data-driven scoring classifier, through which multiple features from sites would be processed in assessing the likelihood of a motif. This is a departure from the unsupervised nature of *patteRNA*. Despite this departure, we sought to maintain the broad applicability of the method to diverse data types. As such, we focused on features that we believed to generalize well across SP datasets (i.e., are data invariant).

We explored various features in conjunction with the *c*-score to underpin the classifier. Five features emerged as promising candidates and their potential was further explored in a combinatorial set of experiments. The first was the cross-entropy loss (CEL) between the target motif's pairing sequence and *patteRNA*'s computed posterior pairing probabilities (see Methods). This feature relates closely to the *c*-score but highlights the cumulative disagreement between the data and the motif more explicitly. Specifically, CEL is influenced more strongly by nucleotide-level disagreements (compared to agreements, see Figure 4.2) which may otherwise be masked by the *c*-score. As such, we speculated that this metric could assist scoring as it helps discriminate between sites that score moderately well across their entire span and sites that score strongly for some nucleotides but have strong disagreement in others. This is particularly relevant in the context of RNA structure where distinct motifs are often highly similar outside of a small number of decisive nucleotides. For example, when determining the length of a loop within a stem-loop, nucleotides near the end of the stem may inform the precise extent of the loop (e.g., a loop length of 4 nt versus 6 nt). Local disagreement can distinguish between such competing structures.

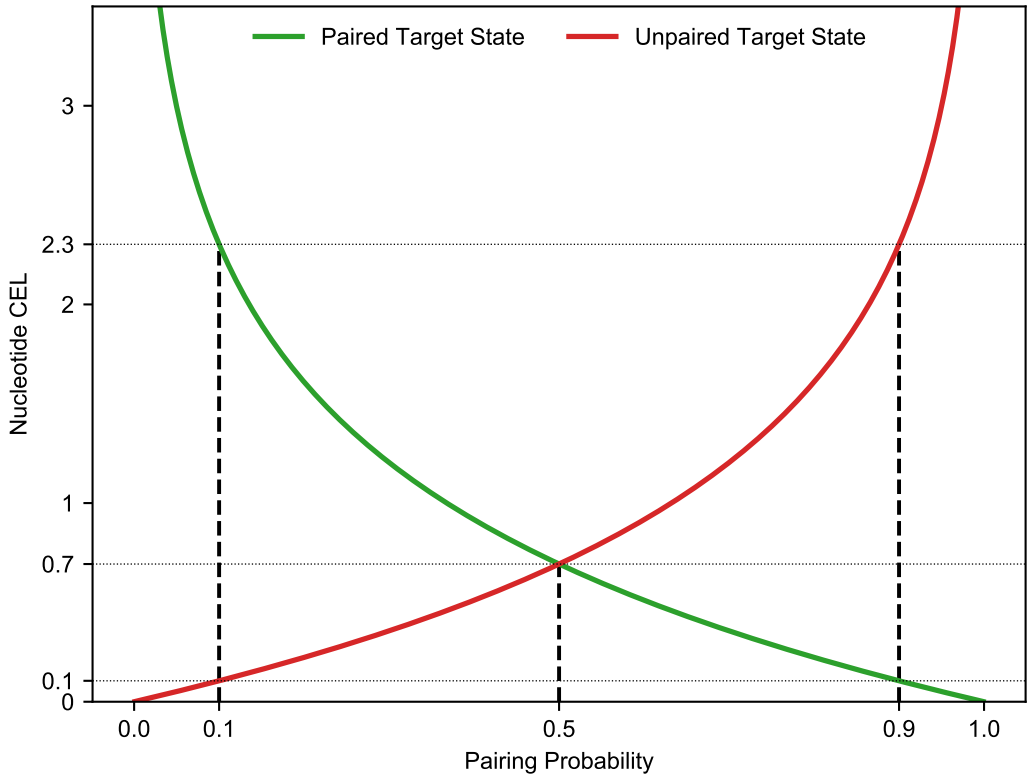


Figure 4.2: Graph of pairing probability versus cross-entropy loss. For paired target states, a pairing probability of 0.5 yields a loss of 0.7. Higher pairing probabilities reduce this loss—e.g., a pairing probability of 0.9 yields a loss of 0.1. However, due to the nonlinear nature of the loss function, lower pairing probabilities yield disproportionately larger loss—e.g., a pairing probability of 0.1 yields a loss of 2.3 (a change in loss of 1.4 from a pairing probability of 0.5, compared to the change in loss of 0.6 associated with the analogous change in loss for a pairing probability of 0.9). Thus, total CEL for a site (the sum of loss across all nucleotides in a site) is measure that is influenced most by the presence of strong disagreements within it.

The second feature was the Gini index of the reactivities at the target site, which is often used in the context of reactivity analysis [159, 20]. Gini index has previously been used to assess structural homogeneity. For instance, we expect stable conformations to yield more distinct reactivities between paired and unpaired states (high Gini index) and less stable structures or regions characterized by multiple conformations yield more intermediate values (low Gini index) [106]. As such, we speculated that Gini index could serve as a simple proxy for data quality and structural homogeneity in a site, and therefore might assist in informing where a *c*-score is more or less meaningful.

The third, fourth, and fifth features relate to the thermodynamic prediction of the local region's minimum free energy (MFE) structure. It has been shown that incorporating

thermodynamic models with SP data tends to improve predictions [29, 34, 30]. Therefore, we utilized their predictions in different ways to potentially assist as features in a scoring classifier. As such, the third feature was the local minimum free energy (LMFE) of the region around a site, where local is defined as the target site window extended in both directions by some distance (e.g., 40 nt; see Figure 4.3A). The fourth feature was the local constrained minimum free energy (LCMFE) of the region around the site, which amounts to folding with the target motif strictly enforced as a hard constraint. We thought that these two metrics, or perhaps their combination, could assist in interpreting the stability of the local region and the motif's influence on it. We also considered a fifth feature, which was the difference between LMFE and LCMFE, which we termed the motif energy loss (MEL). This measure summarizes the energetic favorability associated with the presence of the motif.

4.3.3 Feature Selection

To test the scoring potential of various feature combinations, we established a simple train-and-test pipeline for mining hairpins in a reference dataset (the Weeks set, see Methods). Various feature combinations were used to train a logistic classifier whose scoring precision was then quantified (using an 80%/20 test/train split). For each feature combination, this procedure was repeated 5 times. Mean scoring precision on the test sets was then used to assess the scoring potential of that feature combination.

We performed benchmarks in a simple combinatorial manner to investigate which features and feature combinations were most effective. The *c*-score was used as a base feature in our analysis, meaning that it was included in all combinations. The results of our preliminary feature analysis are in Figure 4.3B. In the 2-feature experiments, all candidate features except Gini index yielded a detectable improvement in precision over just using the *c*-score (which achieves an average baseline precision of 0.62), and we found that MEL yielded the strongest enhancement (to an average precision of 0.69, an 11% improvement over baseline). The 3-feature experiments were only able to incrementally improve scoring precision beyond this level. The best 3-feature combination was *c*-score, CEL, and MEL, which yielded an average precision of 0.70. Interestingly, the combination of *c*-score, LMFE and LCMFE yielded an average precision approximately equal to the observed precision with *c*-score and MEL. We also observed that none of the 4, 5, or 6-feature classifiers significantly outperformed the best 3-feature classifier on any of the benchmarks (data not shown), further validating the efficiency of the chosen scheme. We

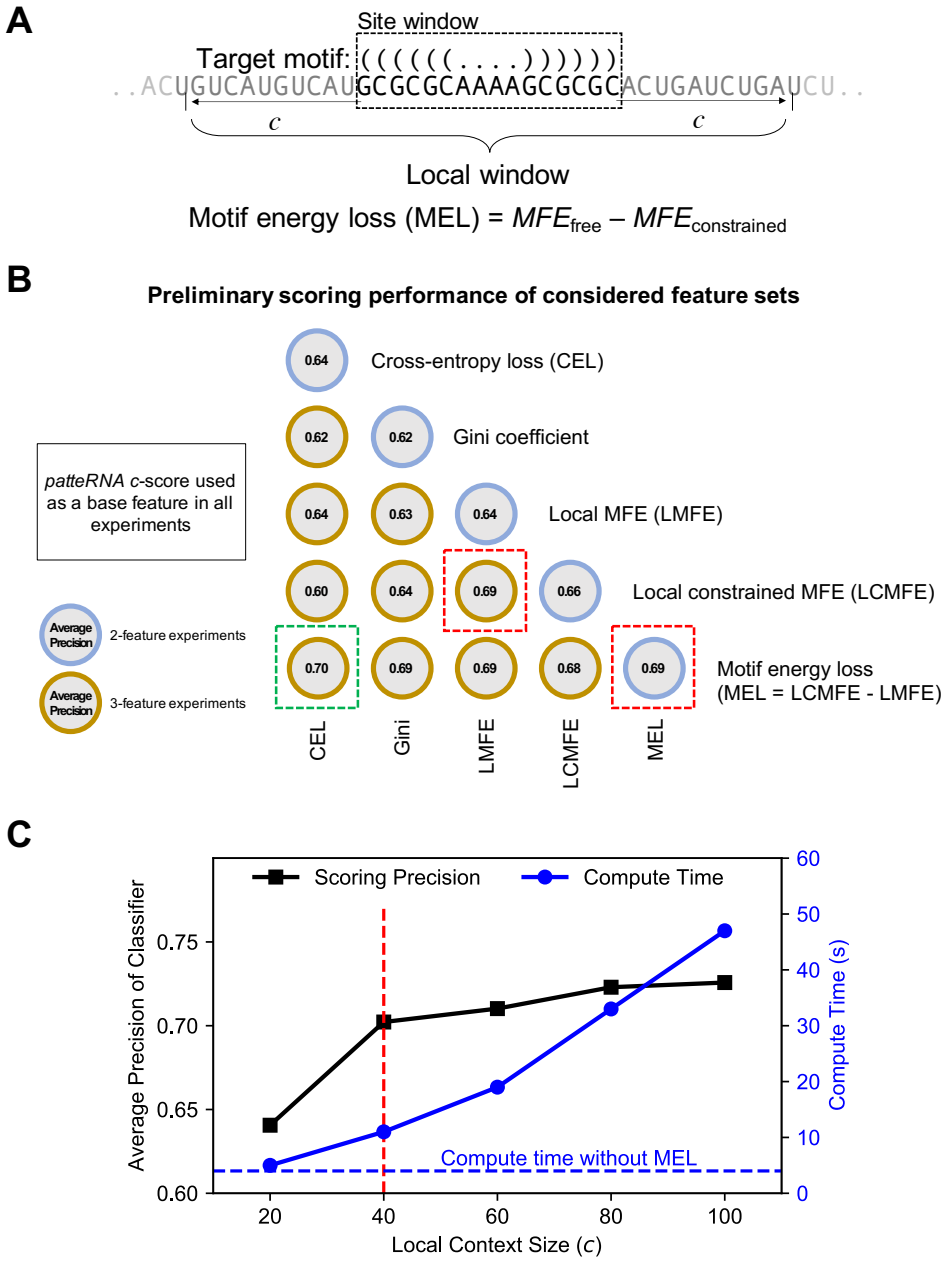


Figure 4.3: Auxiliary feature development for assisting in structure motif mining from SP dataset. **(A)** Illustration of the local window considered when computing thermodynamics-based features for scoring, such as motif energy loss (MEL). The considered window extends a distance, c (the local context size), from the boundaries of the scored site. **(B)** Preliminary scoring performance of considered feature sets. A combinatorial approach was used to test the performance of feature combinations. Features were benchmarked by using them to train a logistic classifier and then computing their average precision on a hairpin test set across five replicates; mean average precision is shown. **(C)** Determination of a suitable context size to use for MEL computations in *patteRNA*. Shown is the scoring precision when using a logistic classifier with c-score, cross-entropy loss (CEL), and MEL across various context sizes. Also shown is the measured runtime at each context length. Highlighted in red is the chosen default context size (40 nt), which strikes a balance between scoring precision and computational overhead relative to *patteRNA*'s original speed on the utilized data.

chose this set of features to use as inputs when developing and optimizing the scoring classifier to utilize in *patteRNA*'s scoring pipeline.

To determine an appropriate local context size to use for MEL, we investigated the precision of the selected feature set at regular intervals of local context length from 20 nt to 100 nt (note that the full context size used for folding is $2c + n$, where c is the context length and n is the motif length). We simultaneously measured the respective compute times. Our results, shown in Figure 4.3C, demonstrate a trade-off between feature quality and compute time as one increases the local context size. We observe that the scoring quality plateaus approximately at $c = 40$, yet the extra compute time (relative to NNTM-free scoring) rapidly grows for longer context lengths. For this reason, we decided to use $c = 40$. We note, however, that larger contexts do provide a slightly better structural interpretation. Thus, although a length of 40 nt is used for the remainder of our work in the manuscript, this parameter may to be tuned by the user.

4.3.4 Classifier Selection and Optimization

Having converged on using *c*-score, CEL, and MEL, we devised a more intensive classifier training pipeline and used it to investigate a set of standard binary classifiers for their ability to robustly model these features. Specifically, we examined logistic binary classification (LBC), random forest classification (RFC), linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA).

Our classifier training pipeline was underpinned by the use of RNA STRAND [2]. STRAND has 4,666 high-quality secondary structure models spanning a large set of RNA families, including regulatory elements, ribosomal RNA, ribozymes, synthetic RNAs, and more. After removal of highly redundant sequences with CD-HIT-EST [70], 1,191 transcripts remained, providing a much more expansive structural snapshot to use for classifier training than the Weeks set, which comprised 22 transcripts. Importantly, STRAND transcripts do not generally have SP data associated with them. Thus, we utilized simulations to generate artificial data. Reactivities were modeled as (and sampled from) the three-state model (unpaired, stacked, and helix-end) devised in [180].

Figure 4.4A demonstrates the interplay between the Weeks set data and the STRAND transcripts as used in our analysis. In short, we used simulated data on STRAND transcripts for classifier training. The Weeks set was then used to benchmark the performance of classifiers trained from STRAND simulations. We found that using STRAND transcripts (with simulated data) yielded the best results in terms of performance on the

Weeks test set benchmarks, even outperforming classifiers trained on the Weeks set directly. Verification sets were also generated by resampling additional replicates of the Weeks set and simulating additional replicates on STRAND (see Methods for details). The overall objective was to identify the best possible motif classifier for the 3 investigated features (Figure 4.4B).

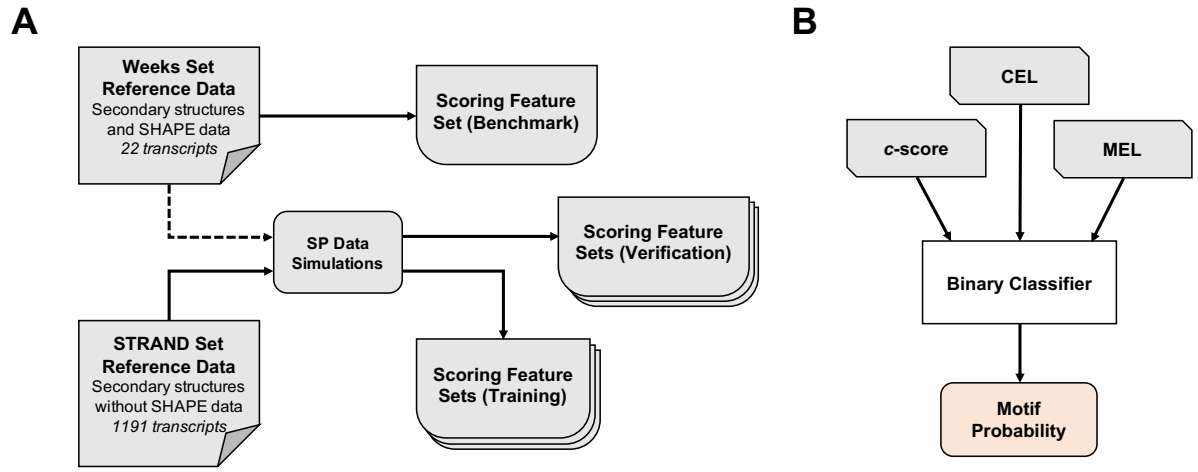


Figure 4.4: Data processing scheme for feature set generation in training, verifying, and benchmarking a binary motif classifier. **(A)** Data sources and computational flow for generating features sets used for training, benchmarking, and verification. Transcripts in RNA STRAND were used for classifier training; however, because these transcripts lack SP data, simulations were used to generate artificial reactivities on known secondary structures. The original Weeks set was used to benchmark classifiers as it contains RNAs with known structures and high-quality real-world reactivity data. The reactivities in the Weeks set were also resampled to generate additional replicates, which were also used for verification steps in addition to replicated simulation on RNA STRAND. **(B)** Schematic of the binary classification approach utilized in *patteRNA*. *c*-score, CEL, and MEL were used as the features driving assessments of motif probability.

Our results are presented in Figure 4.5. Overall, we found that the LBC provided the best results in terms of scoring consistency and translatability to verification benchmarks against other data and other motifs. Generally, we observed similar results for LBC, LDA, and QDA—-all classifiers strongly improved scoring when compared to *c*-scores on the benchmark and verification sets—yet the LBC slightly exceeded the others' performance on all tests. We also observed that the LBC was the fastest of the tested classifiers (data not shown). Interestingly, we observed that a random forest classifier was able to achieve remarkable performance on the training set but did not translate effectively to other benchmarks or validations. We presume that the classifier was overfitted due to its parameterization (described by a large number of decision trees); efforts to reduce the size and complexity of the parameterization (e.g., by reducing the number of estimators) were

unsuccessful in improving performance beyond what was observed with logistic regression. Moreover, we found the compute time in applying random forest classification to scale poorly in situations where a large number of sites (i.e., more than tens of thousands) were scored.

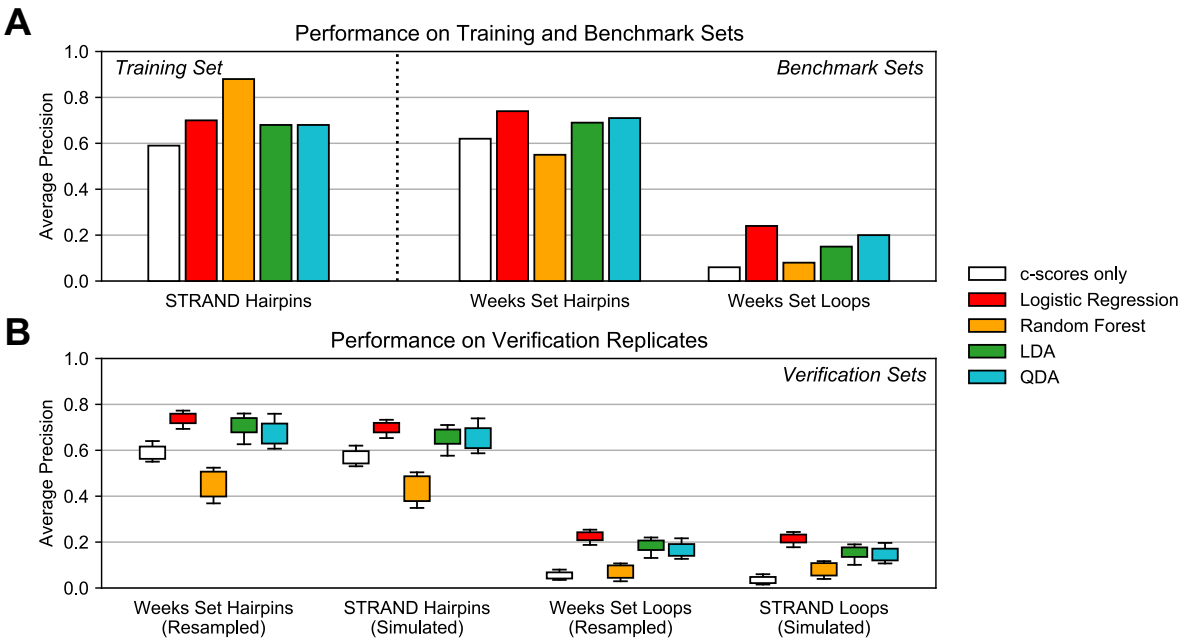


Figure 4.5: Results of experiments testing the ability of standard classifiers to fit the training set and generalize to various benchmarks and verifications. **(A)** Results on classifier performance on the training set (hairpins in RNA STRAND) and benchmark sets (hairpins and loops in the Weeks set). **(B)** Performance of trained classifiers on 5 resampled replicates of the Weeks set and 5 simulation replicates on RNA STRAND. LDA: linear discriminant analysis; QDA: quadratic discriminant analysis.

Due to these results, we decided to use an LBC trained on *c*-scores, CEL, and MEL from STRAND hairpin sites. We developed the final classifier by generating five replicates of SP data for the STRAND transcripts and using each to train a respective LBC. We benchmarked the classifiers against the Weeks set hairpins and STRAND hairpins in the other replicates and assessed their overall performance as the sum of (1) average precision on the Weeks set and (2) mean average precision on the other STRAND hairpin replicates. The classifier with the highest cumulative performance was chosen as the specific parameterization to use in *patteRNA*'s scoring, although there was little difference between the five candidates.

The final LBC performance is compiled in Figure 4.6. In short, when benchmarking on the Weeks set, we observe an increase in average precision from 0.62 with *c*-scores to 0.74, a relative improvement of almost 20%. Importantly, the precision at the highest

scores (i.e., when recall is low), is significantly increased compared to c-scores, and roughly matches the performance seen when utilizing full transcript folding (i.e., full-length transcript partition function analysis) (see Figure 4.6A, dashed box). We confirmed that the LBC yielded slightly improved scoring when using larger contexts in computing MEL, similar to that observed in 4.3C. We also utilized the entire 4,666 STRAND transcripts to benchmark *patteRNA*'s performance on various RNA classes (Figure 4.6B). As the structural properties of RNA are diverse, we observe differential performance at hairpin mining for different types of RNA. Structured transcripts defined by a high prevalence of hairpins score the best, for example, regulatory elements, small RNAs, and ribozymes. Other classes which tend to be less structured or a have large proportion of non-local base paring score relatively worse, for example, tmRNA, SRP RNA, and 5S rRNA.

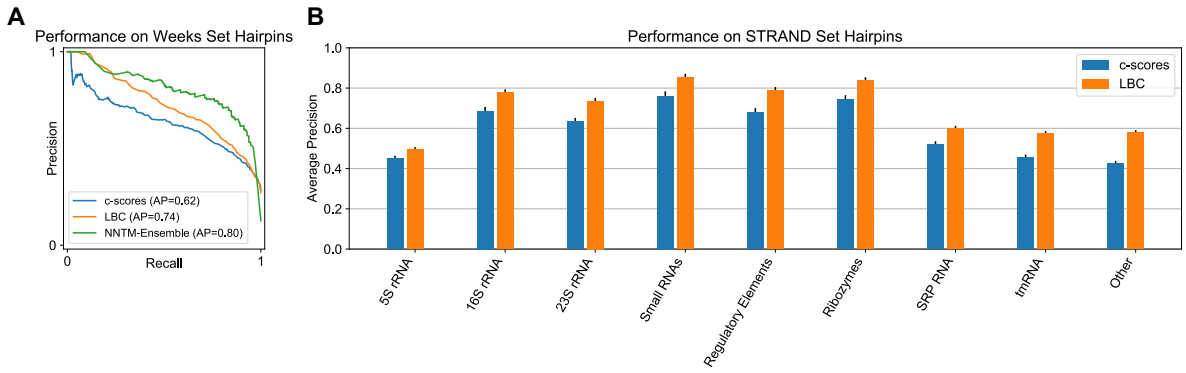


Figure 4.6: Performance of *patteRNA* when using the finalized iteration of a logistic binary classifier (LBC) natively during its scoring phase. **(A)** Precision-recall curves for hairpin detection in the Weeks set for LBC probabilities, full NNTM-Ensemble predictions, and regular c-scores absent any additional classifier processing. Dashed box indicates the region associated with the highest scores, where the LBC is able to match the precision of full-length partition function analyses. Also indicated are the performance points associated with thresholding to $c = 2$ and Prob(SL) $= 0.9$. **(B)** Average precision by RNA class when mining 5 replicated simulations on RNA STRAND transcripts. Error bars indicate standard error of the mean.

Runtime benchmarks demonstrate that our approach scales linearly and allows for transcriptome-wide mining of hairpins within an hour (see Figure 4.7). This speed is one to two orders of magnitude faster than processing the data via local partition function workflows with windows of length 150 or 3000 nt.

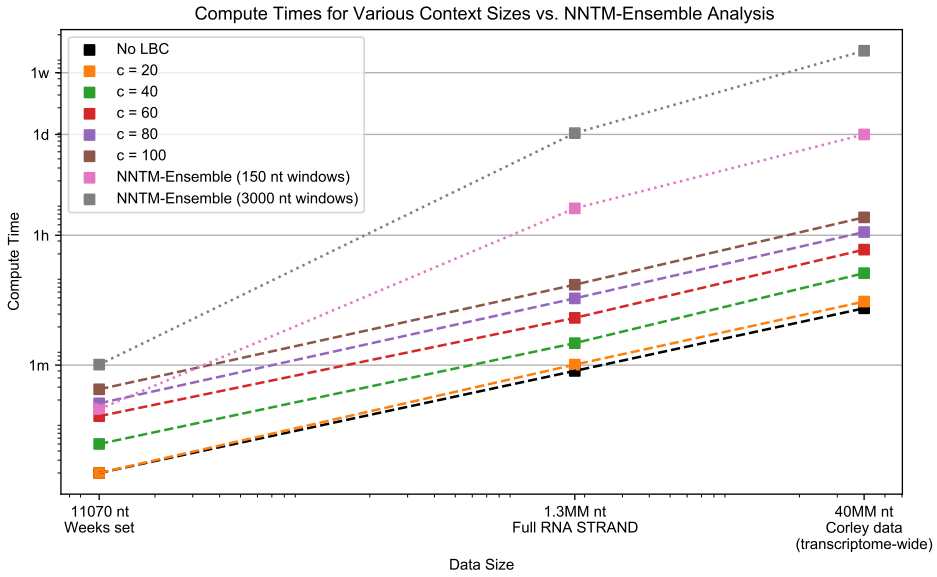


Figure 4.7: Compute timing benchmarks for *patteRNA* using a logistic binary classifier with MEL at various context lengths versus benchmarks using NNTM-Ensemble analysis.

4.3.5 Mining Structurome Data Reveals Strong Association between Stem-Loops and RBP Binding Signals

The interplay between RNA structure and RBPs has been of significant interest for several decades [24]. Such interactions are widespread, dynamic, and known to underpin important regulatory processes like splicing, trafficking, and translation [182, 51, 201, 16, 52]. Although it is believed that many RBPs prefer to associate in unstructured regions, recent *in vitro* and *in vivo* experiments have indicated that a significant portion of RBP binding occurs in structured contexts and in a structure-dependent manner [36, 136, 25, 151, 181]. That said, a mechanistic understanding of RBP binding exists only for a very small number of RBPs which have been subject to targeted research. The global trends and dynamics of RNA-protein interactions are still poorly understood, and as such, significant efforts have been directed at disentangling the complex relationships between RNA transcripts, their regulation, and the proteins which interact with them.

Corley et al. recently harnessed structure profiling to detect RBP binding sites in an experiment called fSHAPE and applied it transcriptome-wide to human cell lines [25]. The result of their work is a large set of data encompassing *in vivo* and *in vitro* icSHAPE reactivities and fSHAPE scores, the latter of which capture differential reactivity in the presence and absence of RBPs. They demonstrated that strong fSHAPE signals are highly correlated to RNA nucleotides that are unpaired and known to engage in hydrogen bonding with proteins, meaning that high fSHAPE signals are strong evidence of

RBP binding. These data enable quantitative comparisons between RNA structure (via icSHAPE reactivity profiles) and RBP binding (high fSHAPE signals).

Corley et al.'s analysis further demonstrated that strong fSHAPE signals preferentially occur in structured contexts, and our previous work harnessing *patteRNA* expanded on this result by indicating a global association between a nucleotide-wise measure of structuredness (HDSL) and high fSHAPE [151]. Both of these results, however, were obtained from "bird's-eye view" approaches in which low-resolution global trends were utilized to elucidate general properties of RNA-protein interactions. In this work, we sought to utilize *patteRNA* to associate specific structure motifs with RBP binding in a more mechanistic "bottom-up" approach. Specifically, we sought to address the questions, "to what extent does RBP binding occur in the context of stable hairpins?" and "what fraction of stable hairpins can be associated with RBP binding?"

We used the LBC to score Corley et al.'s icSHAPE data as a means of exploring the association between hairpins (also referred to as stem-loops) and RBP binding signatures. Specifically, we mined two transcriptomes (K562 and HepG2 cells) for the representative set of stem-loops introduced earlier (stem lengths between 4 and 15 nt, loop lengths between 3 and 10 nt) and cross-referenced the locations of highly scored sites with the fSHAPE data to elucidate any connections between them. The results of our analysis are compiled in Figure 4.8, where we present findings from both *in vitro* and *in vivo* icSHAPE data (K562 results shown; results for HepG2 data were very similar and are shown in Figure 4.9). We first examined the locations of a highly prevalent stem-loop motif described by a stem of 6 base pairs and a loop length of 4. Figure 4.8A depicts the mean fSHAPE signal of highly-scoring sites (black) versus poorly-scoring sites (blue). It shows that sites which scored highly for this motif (Prob(SL) > 0.9) often display a high fSHAPE signal, interpreted as evidence of RBP binding, localized to the loop region. Specifically, greater than 70% of these high-scoring sites *in vitro* displayed strong evidence of RBP binding in the loop (defined as fSHAPE > 2, the same threshold used by Corley et al.). A threshold of 0.9 was chosen as it is associated with near-perfect precision in our benchmarks on the Weeks set (see Figure 4.6A, orange dot). Interestingly, analysis of *in vivo* data arrives at a similar association, suggesting that data from one condition may suffice in determining relevant structures. A comparable signal was also detected when examining highly scored sites for a similar motif with a 6 base pair stem and a 3 nt loop (Figure 4.8B).

We expanded the scope of our analysis by inspecting the highly scored sites across

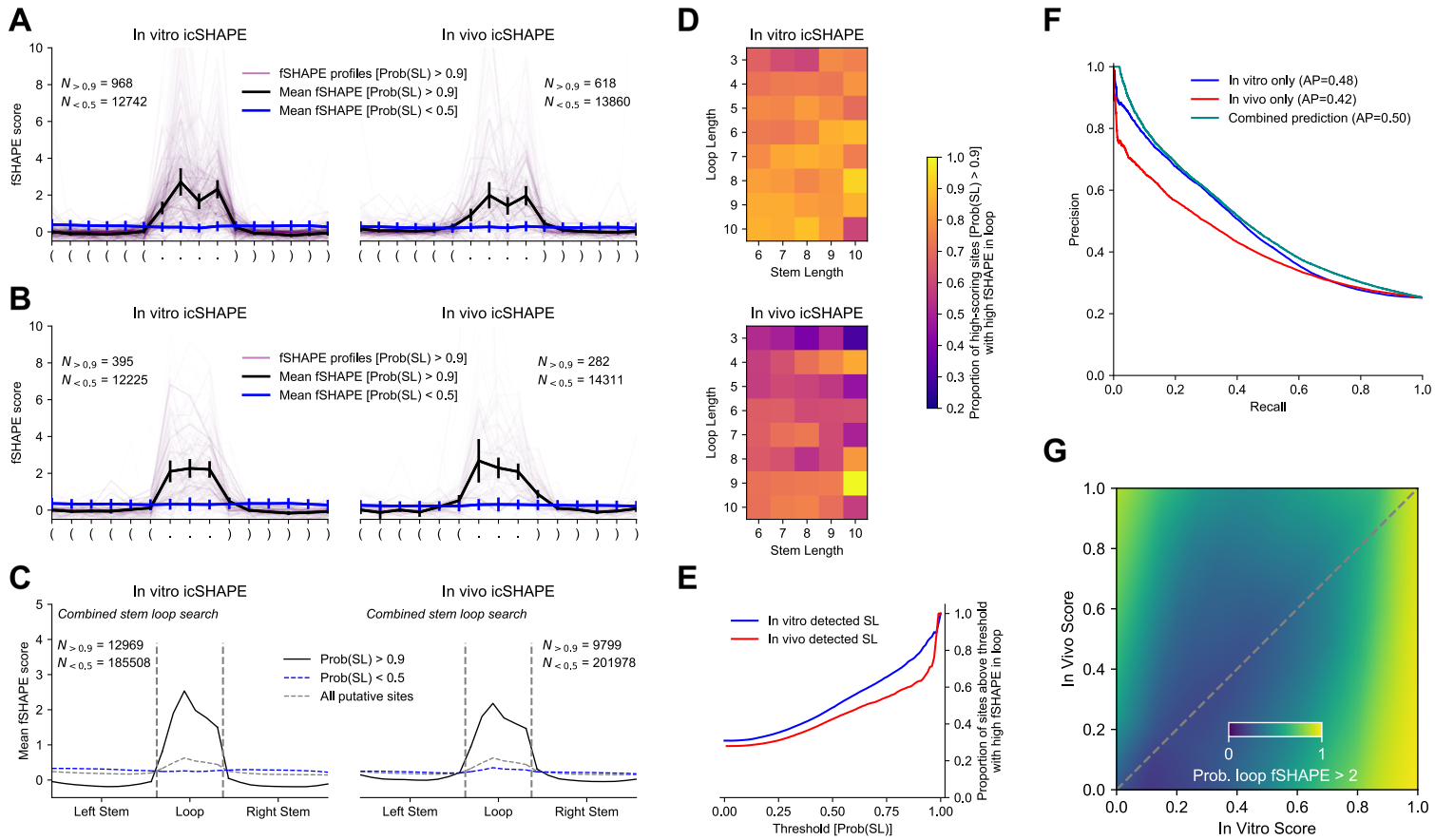


Figure 4.8: Strong association between detected stem-loops (SL) and RBP binding evidence (high fSHAPE scores) in structurome data from K562 cells. **(A)** fSHAPE profiles for sites scored highly (in vitro: left; *in vivo*: right) for a stem-loop with stem length 6 nt and loop length 4 nt. Individual fSHAPE profiles for sites with score greater than 0.9 are shown (purple) as are the mean fSHAPE profiles for sites scored above 0.9 (black) and below 0.5 (blue), respectively. **(B)** Same illustration as shown in panel (A), but for sites scored for a stem-loop with stem length 6 nt and loop length 3 nt. **(C)** Combined fSHAPE properties for sites scored when searching for a representative set of stem-loops (stem lengths 4 to 15 nt; loop lengths 3 to 10 nt; no bulges). fSHAPE profiles from scored sites were interpolated to a fixed length of 26 (10 nt left stem, 10 nt right stem, 6 nt loop; see Methods). **(D)** Proportion of high scoring sites (Prob(SL) $>$ 0.9) that have fSHAPE $>$ 2 in their predicted loop for stem-loops for each considered loop and stem length. Shown are results when mining *in vitro* icSHAPE data (top) and *in vivo* icSHAPE data (bottom). Stem-loops detected *in vitro* were more associated with evidence of RBP binding than those detected *in vivo*, but both datasets demonstrate a strong association. **(E)** Proportion of sites above indicated thresholds that have high fSHAPE in their predicted loop. **(F)** Precision-recall curves for identifying sites with high loop fSHAPE for sites that were scored in both conditions. Performance is shown for a perceptron classifier trained on condition-wise paired scores (see Methods) against the performance observed when using the condition-wise scores on their own. **(G)** Perceptron-modeled relationship between condition-wise scores and evidence of RBP binding (fSHAPE $>$ 2). The modeled distribution indicates that hairpins strongly detected *in vitro* are overwhelmingly associated with RBP binding. There is also RBP binding signal identified for hairpins only detected *in vivo* and not *in vitro* (top left).

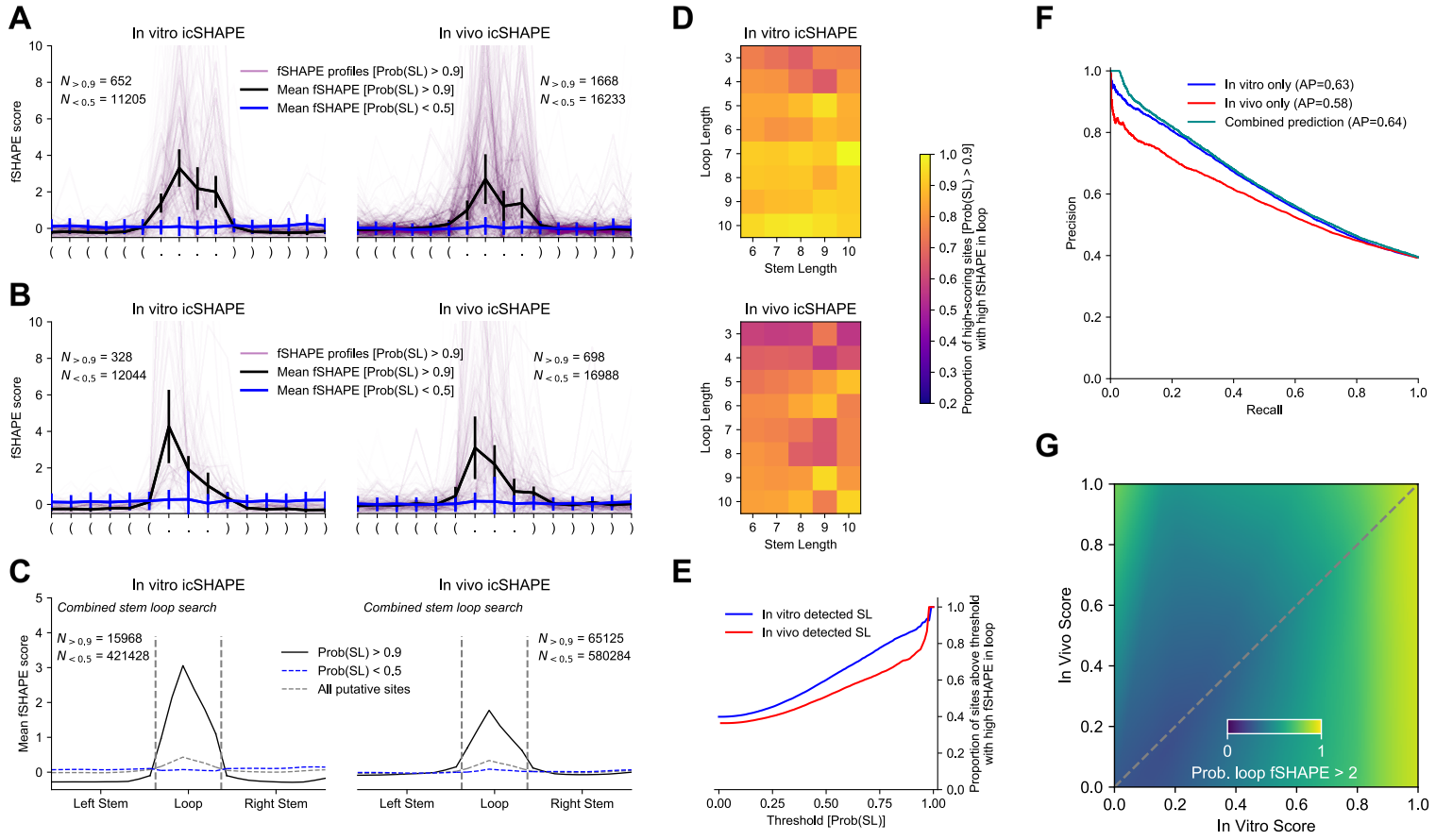


Figure 4.9: Strong association between detected stem-loops (SL) and RBP binding evidence (high fSHAPE scores) in structurome data from HepG2 cells. **(A)** fSHAPE profiles for sites scored highly (in vitro: left; *in vivo*: right) for a stem-loop with stem length 6 nt and loop length 4 nt. Individual fSHAPE profiles for sites with score greater than 0.9 are shown (purple) as are the mean fSHAPE profiles for sites scored above 0.9 (black) and below 0.5 (blue), respectively. **(B)** Same illustration as shown in panel (A), but for sites scored for a stem-loop with stem length 6 nt and loop length 3 nt. **(C)** Combined fSHAPE properties for sites scored when searching for a representative set of stem-loops (stem lengths 4 to 15 nt; loop lengths 3 to 10 nt; no bulges). fSHAPE profiles from scored sites were interpolated to a fixed length of 26 (10 nt left stem, 10 nt right stem, 6 nt loop; see Methods). **(D)** Proportion of high scoring sites (Prob(SL) $>$ 0.9) that have fSHAPE $>$ 2 in their predicted loop for stem-loops for each considered loop and stem length. Shown are results when mining *in vitro* icSHAPE data (top) and *in vivo* icSHAPE data (bottom). Stem-loops detected *in vitro* were more associated with evidence of RBP binding than those detected *in vivo*, but both datasets demonstrate a strong association. **(E)** Proportion of sites above indicated thresholds that have high fSHAPE in their predicted loop. **(F)** Precision-recall curves for identifying sites with high loop fSHAPE for sites that were scored in both conditions. Performance is shown for a perceptron classifier trained on condition-wise paired scores (see Methods) against the performance observed when using the condition-wise scores on their own. **(G)** Perceptron-modeled relationship between condition-wise scores and evidence of RBP binding (fSHAPE $>$ 2). The modeled distribution indicates that hairpins strongly detected *in vitro* are overwhelmingly associated with RBP binding. There is also RBP binding signal identified for hairpins only detected *in vivo* and not *in vitro* (top left).

all stem-loop motifs included in our search (motifs with stems of length 4 to 15 nt and loop length 3 to 10 nt) (see Figure 4.8C). For the *in vitro* icSHAPE data (left side of panel C), *patteRNA* identified 12,969 high scoring stem-loops out of 289,764 considered putative sites (i.e., those which satisfy sequence constraints associated with the searched motifs), which amounts to less than 5%. To visualize the fSHAPE data from these sites, which have different sizes, fSHAPE profiles were interpolated to a constant stem and loop length (10 nt and 6 nt, respectively; see Methods). When examining this larger representative set, we continued to observe a strong fSHAPE signal in loops and a low signal in stems of stable motifs. Moreover, an inspection of sites which score Prob(SL) < 0.5 shows complete depletion of this signal, thereby providing a negative control that strengthens the conclusions drawn from high-scoring sites.

Given the seemingly universal association between highly scored stem-loops and RBP binding signal, we sought to investigate it at the motif level for each considered target. In other words, we examined if particular stem-loops have a stronger association with RBP binding than others. Figure 4.8D shows the fraction of highly scored sites for each considered motif that also have high fSHAPE signal in their loop. Examining this association across all motifs reveals that this notable propensity of RBP binding signal within loops generally applies to all of them. Nevertheless, the association appears significantly stronger for *in vitro* than for *in vivo* scores. This is presumably due to the effect RBPs have on reactivities for unpaired nucleotides engaging in RBP binding (i.e., reduces their accessibility) and/or lower data quality *in vivo*. Adding to our previous conclusion that one condition may suffice for determining relevant structures (Figure 4.8C), our results indicate that *in vitro* structure mining is in fact preferable in some contexts when identifying motifs functionally relevant in an *in vivo* context. For example, differences between conditions are particularly stark when motif loops are short (e.g., top three rows of *in vivo* heatmap). We speculate that this difference is due to RBP occlusion of loop reactivities which is more detrimental to *patteRNA*'s scoring when loops are short. The differences between the conditions are further illustrated as a function of the threshold by which stem-loops are declared stable by *patteRNA* (Figure 4.8E). Notably, the observed associations were even stronger in HepG2 cells (Figure 4.9).

Although both *in vitro* and *in vivo* detected stem-loops strongly associated with high fSHAPE signals, there were some differences between the conditions. As such, we attempted to fuse both scores into a single, more powerful predictor of stem-loops with RBP binding signals. To this end, we fitted a simple perceptron model to predict from a

site's *in vitro* and *in vivo* LBC scores whether or not the site has high fSHAPE (fSHAPE > 2) in the loop (see Methods). Using the perceptron to predict motifs with high loop fSHAPE resulted in a slightly stronger association (as quantified by average precision for indicating sites with high loop fSHAPE) between its predictions than using the *in vitro* or *in vivo* scores alone (Figure 4.8F), suggesting that changes between the two conditions can offer additional insight into RBP-motif interactions.

We attempted to interpret the perceptron's model to gain insights into scoring patterns associated with RBP binding. Its predictions are seen in Figure 4.8G and reveal two distinct patterns. The first pattern is a high *in vitro* score (irrespective of *in vivo* score, yellow region on right side of heatmap), which recapitulates key results from Figure 5A–E. However, the second pattern is associated with sites that score poorly *in vitro* but strongly *in vivo* (top left corner). These sites appear to fold into stem-loops only in the *in vivo* condition. We speculate that this pattern reflects motifs that are functional (i.e., engage in RBP binding) but only fold or become stabilized in the *in vivo* cellular context. Note, however, that this pattern is far rarer than the former. Whereas over 9900 sites fall into the first pattern (Prob(SL) > 0.9 *in vitro* with high loop fSHAPE), only 66 sites were found in the second (Prob(SL) > 0.7 *in vivo*, Prob(SL) < 0.2 *in vitro*, with high loop fSHAPE). More work is needed to investigate the association of these sites with RBP binding. Note, however, that while the second pattern appeared in our analogous analysis of HepG2 data, it was not as pronounced (Figure 4.9).

Next, we expanded the scope of our motif search to include stem-loops with bulges of 1 or 2 nt on either side of the stem. This greatly broadens the space of considered motifs and therefore increases the required computational overhead, as searching for regular stem-loops mines for 96 targets but allowing for bulges increases this number to 2,640. Overall, approximately 7.2 million sites were considered as satisfying sequence constraints for a searched motif (either a regular stem-loop or stem-loop with a bulge), 27,769 of which received a score greater than 0.9. We compiled the fSHAPE profiles of high scoring stem-loops with bulges and quantified their properties in a manner similar to our hairpin analysis. However, in addition to distinguishing loops from stems, we also distinguished bulges into their own group when quantifying fSHAPE tendencies. Our results are given in Figure 4.10 and demonstrate a similar enrichment of high fSHAPE in apical loops of stem-loops with bulges to that which was observed for stem-loops without bulges. Moreover, we also detected a marked fSHAPE increase within bulge nucleotides, also implicating them in RBP interactions. These results expand the context of our

demonstrated association between structure motifs and RBP binding signal.

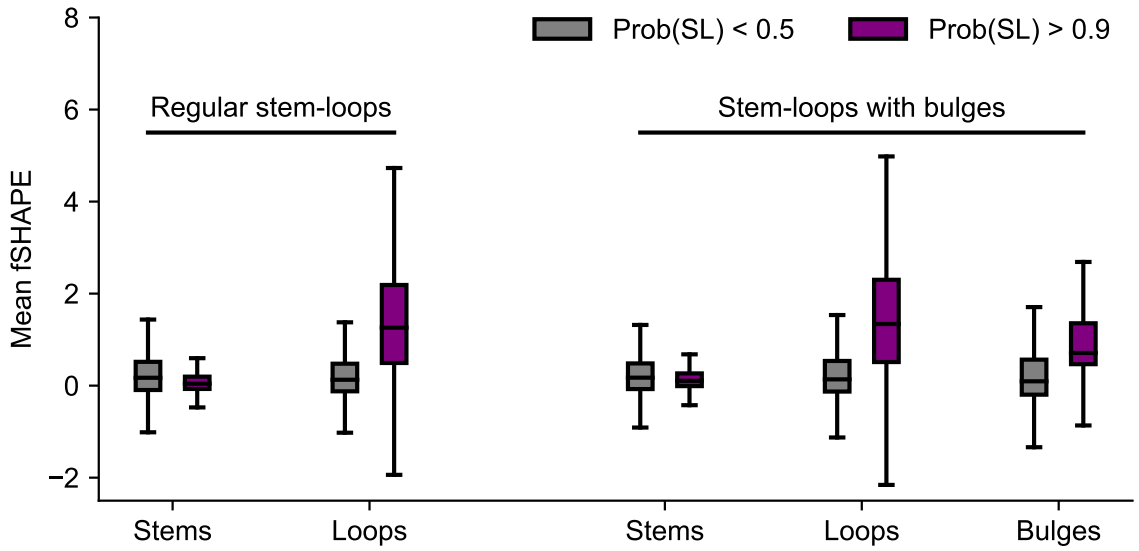


Figure 4.10: Association between RBP binding and structure motifs persists when considering stem-loop motifs with bulges in their stems. Hairpins with bulges were defined as hairpins with stem length 4 nt to 15 nt, loop length 3 nt to 10 nt, and one bulge of 1 or 2 nt on either side of the stem.

Our analysis suggests that a significant majority of stable stem-loops are likely to interact with RBPs. This naturally raised the question of what fraction of RBP binding sites can be explained as occurring in the context of a stem-loop. We estimated this fraction by computing the proportion of nucleotides with fSHAPE > 2 which occur in the loop segment of a highly scored SL motif in the *in vitro* data. The results are given in Table 4.2, showing that, of the fSHAPE data that was included in our motif searches, 19% of nucleotides with fSHAPE > 2 fall within a stem-loop motif scored Prob(SL) $>$ 0.9. Upon relaxing the threshold to 0.7, this proportion increases to 33%. Interestingly, this result is comparable to previous estimates of the proportion of RBPs interacting with stem-loop motifs versus linear motifs [77]. Similar results were observed *in vivo* and in HepG2 cells (see Table 4.3) and when using NNTM-free *patteRNA* *c*-scores (see Table 4.4). Nevertheless, the scope of our search remains somewhat limited. For example, we did not exhaustively consider all feasible bulge types (e.g., bulges larger than 2 nt or stem-loops with bulges on both sides of the stem), nor did we consider internal loops. Both types of motifs have been previously associated with RBPs [120, 24]. Despite the computational overhead associated with mining such complicated motifs, their consideration is likely to significantly increase the proportion of high fSHAPE observations explainable as occurring in a structured element.

Fraction of high fSHAPE nucleotides explained by motif sites	Motif Probability Threshold = 0.9	Motif Probability Threshold = 0.7	Motif Probability Threshold = 0.5
Hairpins without bulges	16%	30%	49%
Hairpins with or without bulges	19%	33%	55%

Table 4.2: Fraction of high fSHAPE sites in K562 cells accounted for by hairpins (without bulges) and hairpins (with or without bulges) as detected in *in vitro* icSHAPE data. Although the identified sites do not account for a majority of the high fSHAPE data, the results demonstrate that a sizable portion of RBP binding can be associated with canonical hairpin motifs.

We further investigated the association between stable stem-loops and RBP binding signals within logical regions of mRNA transcripts—5' UTRs, CDS, and 3' UTRs (noncoding RNAs were treated as their own group). Interestingly, we observed that the association is remarkably consistent between regions (Figure 4.11). Across all considered regions, approximately 75-80% of detected stem-loops had a loop which coincided with strong RBP binding signal (values indicated for K562 data; percentages were approximately 80-85% for HepG2 data). Nevertheless, we did observe large differences in the density of detected stem-loops between these regions. In all cell lines and conditions, 3' UTRs have a significantly higher rate of stable stem-loops than other regions (see Table 4.5). For instance, in K562 *in vivo* icSHAPE data, *patteRNA* identified 9.57 stem-loops per 1000 nt in 3' UTRs, compared to 4.86 and 4.07 in 5' UTRs and CDS, respectively. In the context of post-transcriptional regulation, stem-loops are known to be mechanistically involved with polyadenylation and degradation [89, 144, 54]; however, this is the first stem-loop profile of a human structurome that detects the effect at a global level.

4.4 Discussion

The evolution and growing scale of RNA structure profiling experiments has warranted methods well-suited to the analysis of millions to billions of nucleotides. *patteRNA* is one such tool which was developed with the specific aim of rapidly extracting biologically relevant insights from such data. For genome-wide analyses, high precision is often a specific objective yet challenging to achieve due to the large number of negative sites

	Fraction of high fSHAPE nucleotides explained by motif sites	Motif Probability Threshold = 0.9	Motif Probability Threshold = 0.7	Motif Probability Threshold = 0.5
K562 cells *In vitro* icSHAPE	Hairpins without bulges	16%	30%	49%
	Hairpins with or without bulges	19%	33%	55%
K562 cells *In vivo* icSHAPE	Hairpins without bulges	8%	35%	60%
	Hairpins with or without bulges	12%	37%	61%
HepG2 cells *In vitro* icSHAPE	Hairpins without bulges	10%	33%	60%
	Hairpins with or without bulges	15%	36%	62%
HepG2 cells *In vivo* icSHAPE	Hairpins without bulges	6%	28%	58%
	Hairpins with or without bulges	12%	31%	62%

Table 4.3: Fraction of high fSHAPE sites accounted for by hairpins (without bulges) and hairpins (with or without bulges) for the datasets analyzed in this study. Although the identified sites do not account for a majority of the high fSHAPE data, the results demonstrate that a sizable portion of RBP binding can be associated with canonical hairpin motifs.

considered [39]. In this work, we took a machine learning approach to improve the scoring precision by developing a classifier that accounts for local sequence energetics in addition to *patteRNA*'s statistical characterization of reactivities. To ensure broad applicability, we created a high-quality, non-redundant, and large-scale set of transcripts with known structures from RNA STRAND and used it in conjunction with a data simulation scheme to extensively train and validate our approach. Our work indicates this simulated data provide a strong suite of structural information by which to develop methods, which can augment real datasets that are currently much smaller in size. We believe this resource will be useful for others seeking to develop data-driven methods. Application of the classifier transcriptome-wide revealed that stable stem-loops are strongly associated with

	Fraction of high fSHAPE nucleotides explained by motif sites	*c*-score Threshold = 2	*c*-score Threshold = 1
K562 cells *In vitro* icSHAPE	Hairpins without bulges	16%	47%
	Hairpins with or without bulges	24%	55%
K562 cells *In vivo* icSHAPE	Hairpins without bulges	8%	40%
	Hairpins with or without bulges	20%	54%
HepG2 cells *In vitro* icSHAPE	Hairpins without bulges	8%	39%
	Hairpins with or without bulges	14%	45%
HepG2 cells *In vivo* icSHAPE	Hairpins without bulges	5%	33%
	Hairpins with or without bulges	14%	43%

Table 4.4: Fraction of high fSHAPE sites accounted for by hairpins (without bulges) and hairpins (with or without bulges) for the datasets analyzed in this study when using an NNTM-free scoring approach (*c*-scores only).

fSHAPE RBP binding signals across cell lines. This association has been previously documented for individual RBPs [36, 77, 120], however the ubiquitous nature of stem-loops to interact with RBPs *in vivo* has not been previously shown. Not only does this implicate common and canonical structural elements with RBPs, it also reinforces the notion that mining local structure elements can provide biologically relevant insights.

The results of our perceptron analysis of condition-wise paired scores demonstrated

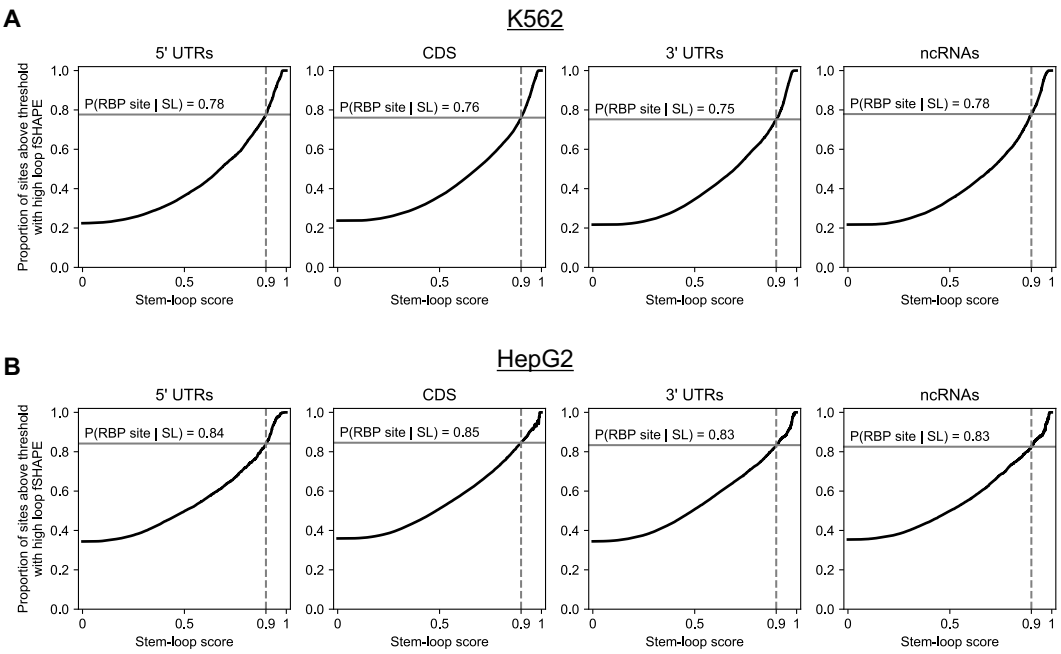


Figure 4.11: Association of RBP binding signal to detected stem-loops within logical mRNA regions. **(A)** Associations for K562 *in vitro* detected stem-loops. **(B)**Associations for HepG2 *in vitro* detected stem-loops. Proportions of sites above a score threshold of Prob(SL) > 0.9 with RBP binding signal (fSHAPE > 2 in loop) are indicated for each graph.

	Region	**K562**	**HepG2**
In vitro icSHAPE	5' UTRs	7.33	4.84
	CDS	8.71	5.71
	3' UTRs	11.63	8.61
	ncRNAs	9.78	6.31
In vivo icSHAPE	5' UTRs	4.86	4.52
	CDS	4.07	3.99
	3' UTRs	9.57	6.98
	ncRNAs	6.74	4.71

Table 4.5: Density of stem-loop detections in logical regions of mRNA transcripts from *in vitro* and *in vivo* icSHAPE data. Values are given as stem-loops per 1000 nt.

that some patterns could be leveraged to identify functional stem-loops beyond inspecting each condition independently. We found that stem-loops detected *in vitro* explain a significant (greater than 30%) fraction of RBP binding signals in Corley et al.'s data.

Another pattern that emerged was the presence of stem-loops that score poorly *in vitro* (Prob(SL) < 0.2), but highly *in vivo* (Prob(SL) > 0.8), although the prevalence of such sites was much lower than sites associated with the former pattern. We note that the perceptron analysis was primarily performed to assist in the interpretation of score changes between conditions (e.g., Figure 4.8G), and that analogous statistical analysis (i.e., via bivariate data fitting) could arrive at similar conclusions. Lastly, it is possible that a more advanced perceptron approach could better disentangle the relationship between the two conditions. For instance, a perceptron or deep neural network trained on the underlying features from each site in each condition (i.e., *c*-score, MEL, CEL, etc.) might yield more precise predictions on the identification of structure motifs with RBP binding signal.

The LBC developed in this work was demonstrated to be significantly more accurate than *patteRNA*'s *c*-scores alone. Nevertheless, we were curious to what degree using *c*-scores (i.e., an NNTM-free approach) could recapitulate the stem-loop/RBP results obtained with the LBC. We re-analyzed the data, but used a threshold of $c > 2$ to determine stable stem-loops instead of Prob(SL) > 0.9 (see Figure 4.12). As indicated on Figure 4.6A, this threshold is roughly comparable to an LBC threshold of 0.9, although it yields slightly lower precision and recall. We found that the use of *c*-scores arrived at similar conclusions to those which were obtained with the LBC, but the observed association was slightly weaker. Specifically, we observed that the association was significantly weaker for stem-loops with shorter stems (6 or 7 nt) and longer loops (5 nt or longer), especially for the *in vivo* data (Supplementary Figure 4.12F). We believe that such motifs benefit most from the thermodynamic information contained in MEL, as sequence constraints are less effective in pruning the number of negative sites considered during scoring. Nevertheless, this result recapitulates that *patteRNA*'s NNTM-free implementation provides accurate detections, especially for high-quality data. We believe that the LBC assists most in situations where motifs are short, or data quality is low.

patteRNA was developed with a specific aim of addressing the need for universal and efficient tools for analysis of a growing breadth, scale, and diversity of SP experiments. Universality is important because different experiments yield reactivities with disparate statistical properties, meaning one-size-fits-all approaches are generally suboptimal. As such, the versatility of *patteRNA* is a central characteristic of the method. In the development of a data-driven scoring approach, we sought to maintain this trait. We found that the *c*-score naturally lends itself to ensuring an automatically adaptable classifier,

as it provides a normalized measure of a site's consistency with the target motif. Serving as a measure of statistical significance against a null distribution that captures data-level and motif-level biases, this metric can be considered largely data-invariant. Moreover, the MEL feature only depends on the local nucleotide sequence, meaning that it is invariant to different reactivity distributions. The decoupling of MEL from the SP data also enables insight on the contributions of NNTM to SP data interpretation. For example, although we observed the LBC improves precision across a range of motifs, the largest relative improvement was observed for motifs with few base pairs, such as loops flanked by single base pair (see Figure 4.5). This trend was also observed in our analysis of the Corley data, where the largest differences between using *c*-scores (NNTM-free) and the LBC (NNTM-dependent) was observed for the shortest stems. Our results suggest that, when searching for motifs harboring many base pairs, folding with NNTM may not provide a significant benefit over using *c*-scores alone.

From its initial development, *patteRNA* was not envisioned as a replacement or competing method to traditional NNTM-based approaches typically used in RNA structure analyses. Rather, it was developed as a tool to be used in tandem to NNTM-based approaches. For example, it can be used to identify candidate sites for a motif of interest (e.g., broadly defined motifs, such as stem-loops, or specific structural elements, such as iron response elements), which could then be subject to more intensive structural analysis with NNTM and targeted SP experiments. In any case, the advantages of *patteRNA* emerge when analyzing large-scale data. By focusing specifically on sites that satisfy the sequence constraints for a target motif and performing minimal local MFE calculations for the LBC, our method arrives at structuromic insights orders of magnitude faster than partition-function based analyses. This speed helps mitigate the computational overhead associated with partition-function analysis of massive SP datasets, especially for those without access to cutting-edge computational hardware. In considering the future development of a method like *patteRNA*, we believe more work remains to be done, despite *patteRNA*'s demonstrated capabilities and strengths relative to transcript folding or partition function analysis. The primary limitation of our method is the dependence on the definition of specific local secondary structure motifs to use for mining. This dependence enables rapid scans in large datasets but limits the scope of the method's analysis to elements with a previously known or suspected structure. One may specify a large set of related motifs to circumvent this limitation, but this comes at an increased computational cost. The current implementation is capable of mining thousands of dis-

tinct structures in a human transcriptome within several hours (e.g., mining stem-loops with bulges), however searches with increased flexibility (e.g., accounting for more diverse bulges, longer loops, and internal loops) result in a combinatorial explosion of considered motifs to counts larger than 10,000 or 100,000. This renders such searches impractical. Nevertheless, when specifically focused on canonical local motifs, for example stem-loops or stem-loops with bulges, *patteRNA* provides rapid, accurate, and biologically relevant motif mining capabilities on structurome data.

4.5 Appendix

4.5.1 Author Contributions

P.R., R.U., K.D., and S.A. developed the method. P.R. and S.A. analyzed the data and wrote the manuscript.

4.5.2 Deposited Resources

Data and analysis scripts supporting the conclusions of this article are freely available at `https://doi.org/10.5281/zenodo.4667910` [153].

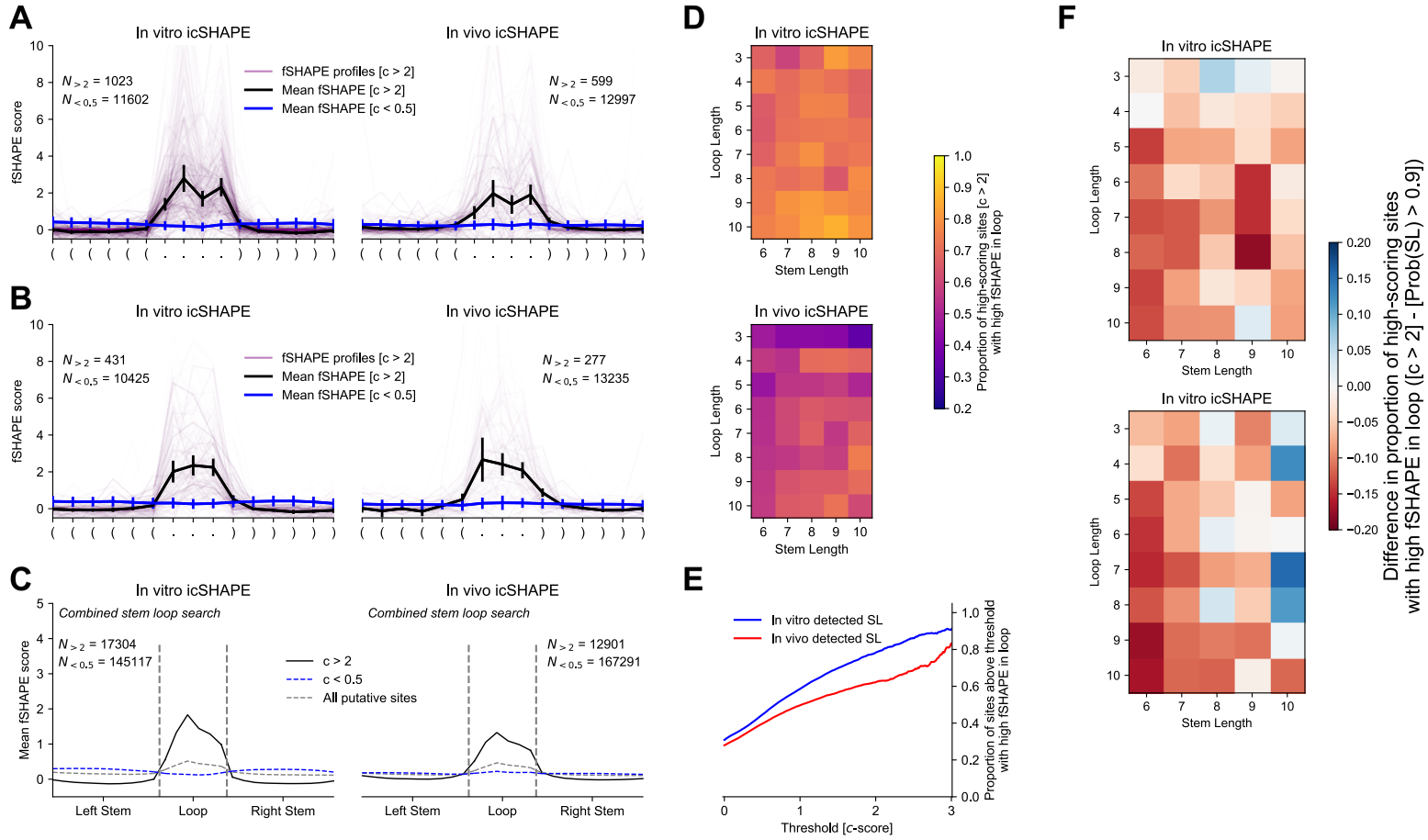


Figure 4.12: Strong association between detected stem-loops (SL) and RBP binding evidence (high fSHAPE scores) in structurome data from K562 cells when using c-scores to determine the locations of stem-loops. **(A)** fSHAPE profiles for sites scored highly (in vitro: left; *in vivo*: right) for a stem-loop with stem length 6 nt and loop length 4 nt. Individual fSHAPE profiles for sites with c-score greater than 2 are shown (purple) as are the mean fSHAPE profiles for sites scored above 2 (black) and below 0.5 (blue), respectively. **(B)** Same illustration as shown in panel (A), but for sites scored for a stem-loop with stem length 6 nt and loop length 3 nt. **(C)** Combined fSHAPE properties for sites scored when searching for a representative set of stem-loops (stem lengths 4 to 15 nt; loop lengths 3 to 10 nt; no bulges). fSHAPE profiles from scored sites were interpolated to a fixed length of 26 (10 nt left stem, 10 nt right stem, 6 nt loop; see Methods). **(D)** Proportion of high scoring sites ($c > 2$) that have fSHAPE > 2 in their predicted loop for stem-loops for each considered loop and stem length. Shown are results when mining *in vitro* icSHAPE data (top) and *in vivo* icSHAPE data (bottom). Stem-loops detected *in vitro* were more associated with evidence of RBP binding than those detected *in vivo*, but both datasets demonstrate a strong association. **(E)** Proportion of sites above indicated thresholds that have high fSHAPE in their predicted loop. **(F)** Difference in the proportion of high scoring sites that were found associated with high loop fSHAPE when using c-scores versus LBC probabilities. c-scores generally yield a slight decrease in the measured proportions. The effect is most pronounced on motifs with shorter stem lengths.

Chapter 5

Conclusion

5.1 Dissertation Summary

The central role of RNA in biology is difficult to overstate, and the functionalities of the molecule often stem directly from its ability to adopt and interchange between relevant structures. The advent of next-generation sequencing combined with structure profiling (SP) experiments have yielded an explosion in transcriptome-wide SP studies which are reinventing our understanding of the RNA structurome and interactome. New methodologies are continuing to emerge at a rapid rate, warranting the development of methods capable of rapidly and automatically assessing structure in massive and diverse SP datasets. This dissertation addressed the need for such methods by describing improvements to a statistical pattern recognition algorithm, *patteRNA*, that rapidly identifies target structural elements in probing data. Importantly, the work presented here improved nearly every facet of the algorithm. Initially, we facilitated the comparative analysis of structural predictions by developing the *c*-score and considering in more depth the versatility of the algorithm to diverse datasets. Then, we re-formulated the training phase of the algorithm using a discrete approach to be faster, more reliable, and more accurate than before. Lastly, we utilized a machine learning approach to greatly improve the precision of the method's scoring phase and target motif predictions. Throughout all this work, we regularly demonstrated the tool by using it to analyze state-of-the-art SP datasets. In this regard, the predictions from *patteRNA* were used to disentangle mRNA structure dynamics, characterize viral RNA genome structures, and elucidate the interplay between the RNA structurome and RNA-protein interactome.

First, we described the automated recognition of structure motifs by their SHAPE data using *patteRNA* [149]. This work built upon the initial description of *patteRNA*

[101] in several ways. The most significant result was the introduction of a novel metric, termed the *c*-score, that uses a comparison of a score against a measured null distribution to quantify its statistical significance. This work also resulted in numerous improvements to the method's automation routines. The most impactful of these improvements was the use of KL-divergence [90] when constructing the training set in order to more rapidly train on a minimally-representative set of transcripts. The result of this work was accurate discrimination of competing conformations of the Rev response element (RRE) in HIV-1 as well as improved accuracy on transcriptome-wide benchmarks.

Next, we reformulated the unsupervised training routine of our method with a discretized observation model (DOM) [151]. The implementation of this approach resulted in significantly faster training, more accurate estimations of the underlying state distributions, and more accurate predictions on the locations of target structural elements. The improvements to the precision of hairpin mining specifically enabled the development of a novel measure, the hairpin-derived structure level (HDSL), which integrates hairpin predictions with local reactivity trends to quantify structuredness as the nucleotide-level. Application of this measure to diverse datasets recapitulated, expanded on, and strengthened results previously obtained either with reactivity summarizations alone or NNTM-assisted quantifications.

Lastly, we devised a data-driven classifier using a machine learning approach to more precisely identify target structures. In addition to utilizing *patteRNA*'s statistical characterization of reactivity profiles, the classifier also incorporated local NNTM-based energetic information. This augmentation again improved scoring significantly without sacrificing computational speed. Our classifier development also included the construction of a large set of reference RNA structures with artificial reactivity data that we believe will be useful to others in the field seeking to develop data-driven methods of RNA structure interpretation. Applying the latest version of *patteRNA* to transcriptome-wide data revealed a marked association between stem-loops and RNA binding protein (RBP) binding sites. Not only did this serve as a general validation of our method, but it also highlighted the powerful biological relevance of identifying local structure motifs like stem-loops. We believe that the insights and methods developed in our work will facilitate the characterization and discovery of novel structural elements in high-throughput SP studies.

5.2 Future Work and Research Directions

The results of the work contained in this dissertation naturally yield promising avenues for the future of structure analysis in both smaller-scale and transcriptome-wide data. The current formulation of *patteRNA* is highly optimized, thus any future improvements to it and related methods will likely stem from fundamental reformulations of the statistical model and pattern recognition schemes instead of iterative improvements to its subroutines.

5.2.1 Statistical Extensions of *patteRNA*

patteRNA uses a simplified model of reactivities in which pairing partners are ignored during training; thus, two states are considered: paired and unpaired. Although this facilitates rapid unsupervised training, it also limits the sophistication of the statistical disentangling of the state reactivity distributions. That said, the core optimization implementation in *patteRNA* (i.e., the EM algorithm), is versatile and naturally suited to couple with virtually any parameterizable statistical model. As such, the utilization of a more intricate model, such as a stochastic context-free grammar (SCFG), is a natural extension to the method that would likely yield significant improvements. SCFGs have previously been applied in RNA structure prediction and alignment with impressive results [135, 37, 40]. Such models more naturally allow for RNA to be represented as meaningful components, such as stems, hairpin loops, and bulges. Although there is some computational overhead associated with their implementation when compared to a simplistic model like *patteRNA*'s 2-state approach, such impacts are unlikely to inhibit the analysis of transcriptome-wide data. Moreover, although these models require more data to train (due to the much larger set of parameters), the scale of high-throughput SP data should be more than sufficient to arrive at robust models. When processing smaller sets of data, smart initialization of parameters (i.e., to values observed on standard datasets) would also help alleviate any issues associated with a high parameter count to data size ratio.

5.2.2 Thrashing Conventional Sequence Constraints for Faster and More Comprehensive Searches

A key aspect of *patteRNA* is the requirement for the definition of target secondary structure motifs when mining data. Typically, targets take the form of either an individual

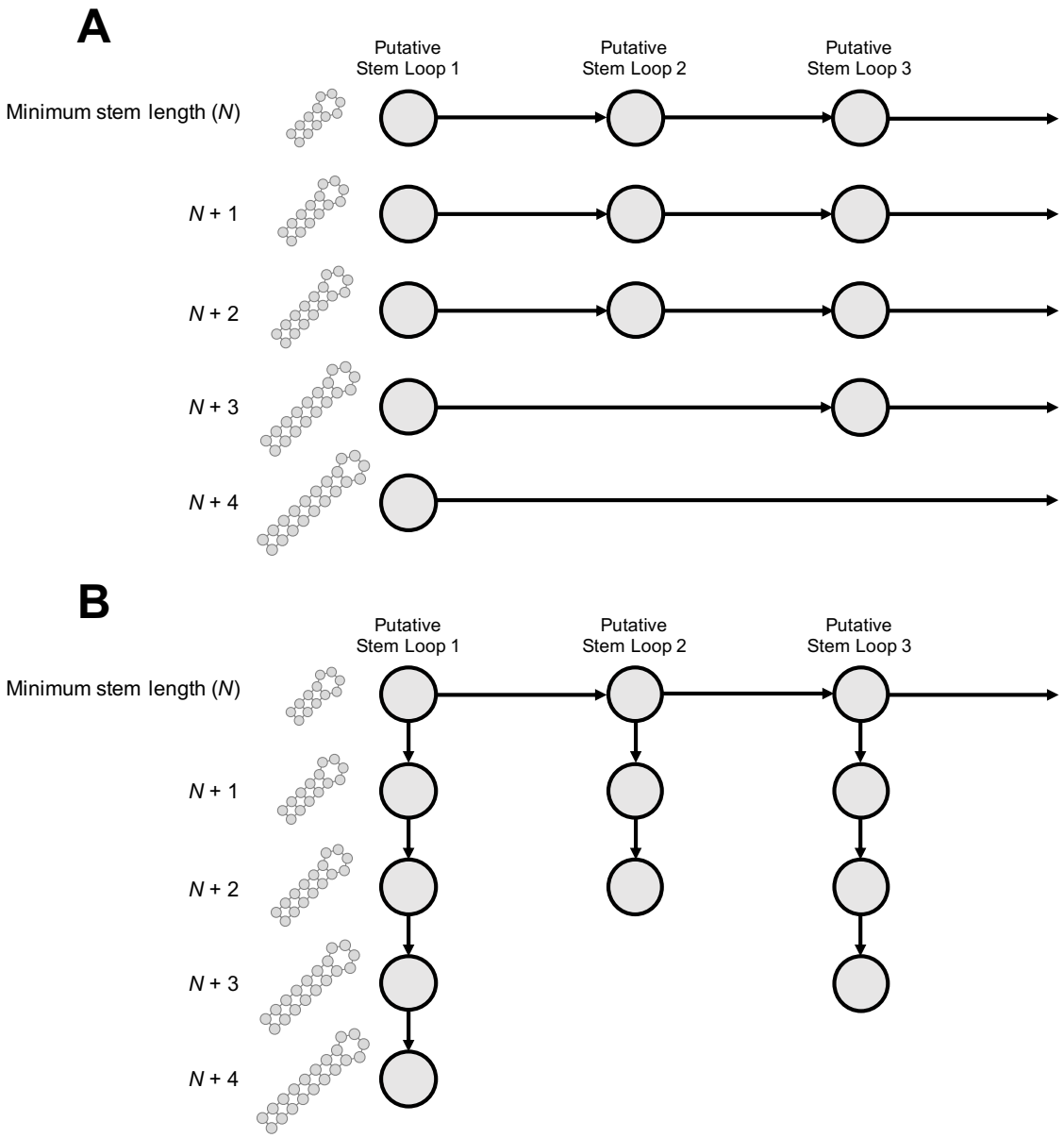


Figure 5.1: Approaches for more rapidly checking sequence constraints. **(A)** In the *patteRNA* implementation, each motif is processed separately of the others. Thus, for a given RNA, each motif is exhaustively checked against every sub-window of the transcript. **(B)** In a more optimized approach, results from nested motifs are utilized to omit the checking of almost all sites. Sites which satisfy sequence constraints for the shortest stem-loop are identified first, then each putative site is used to inform the sites to be checked for the next-longer stem. For that motif, only the next flanking base pair needs checked.

motif (e.g., a specific structure known to interact with a ligand) or a collection of motifs (e.g., a collection of representative hairpins). That said, most analyses presented in this dissertation utilized a collection of motifs. When mining for a collection of motifs, *patteRNA* follows a straightforward process for enumerating and checking the sequence constraints of all loci in the data. The approach is abstractly illustrated in Figure 5.1A.

Essentially, the method uses a brute-force strategy. Sequence constraints for individual motifs are checked independently. For a given secondary structure motif and RNA sequence, all windows in the RNA with length equal to the length of the motif are considered. For each window (i.e., loci), the sequence constraints of the motif on the local sequence are checked. If the sequence is compatible with the base pairs of the motif, that window is deemed a putative site and saved for scoring. If not, it is discarded. As such, each motif is assigned a set of sites for each RNA indicating where the sequence is compatible its formation. When searching a collection of stem-loops with loop length 4 nt and stem length 4 through 8 nt (Figure 5.1A), each motif is checked against every corresponding loci on every RNA.

Note that because the structures of some motifs exist as a substructure within another motif, redundant work is being done when checking each motif individually. For example, when checking a stem-loop with loop length 4 and stem length 5 ($N + 1$), you could automatically omit any site that failed to satisfy the sequence constraints for a stem-loop with loop length 4 and stem length 4 (N). Thus, you skip a majority of sites without having to check a single base pair. The same logic can be applied to the next stem of the collection, a stem length of 6 ($N + 2$). Only sites that satisfied the sequence constraints for the stem length of 5 ($N + 1$) need to be checked. Moreover, at those sites, only one additional base pair needs to inspected.

A slightly more sophisticated scheme is illustrated in Figure 5.1B. Instead of exhaustively considering all possible windows for each motif, we start with the stem-loop with the shortest stem length. For this motif, we check all windows on each RNA to identify putative sites. Then, for motifs with the longer stems, we walk through the set of sites identified for the shorter stem. At each site, we check if the flanking bases are also compatible with an additional base pair. If so, that site is added as a putative site for the stem-loop with stem length $N + 1$. This process is repeatedly for subsequently longer stem lengths. The entire process would be repeated for different loop lengths, although further optimizations could be enabled over the space of loop lengths by considering that even- and odd-sized loops have some redundant information for each other (e.g., a stem-loop with loop 4 nt and stem 5 nt has all of the base pairs in a stem-loop with loop 6 nt and stem 4 nt, plus one additional base pair at the top of the stem; a stem-loop with loop 5 nt and stem 5 nt has all of the base pairs in a stem-loop with loop 7 nt and stem 4 nt, plus one additional base pair at the top of the stem).

Although the process of checking sequence constraints is not very computationally

intensive, this schematic would likely lead to a detectable improvement in *patteRNA*'s runtime. The relative improvement would depend on the precise collection of motifs considered as well as the size of the data. For the standard set of hairpins searched via the "`--hairpins`" flag on the Weeks set benchmarks, the time to check sequence constraints would be cut by over 95%, as over 99% of windows are simply ignored for the stems longer than 7 nt. That said, checking sequence constraints only amounts to a small portion of the total scoring phase. Actually computing scores at sites is responsible for most of the runtime. For this reason, although the described methodology would solve the sequence constraints problem in far more scalable manner, the impact on overall compute time would not be transformative.

The transformative benefits of such a method for checking sequence constraints do not come with the improvement in compute time, but rather with the opportunity to define an entirely new search process that doesn't depend on the definition of specific motifs. For instance, with a minimum stem length defined, one can arbitrary continue to check longer and longer stems with virtually no additional computation cost. In this sense, it does not make sense to define an upper limit to stem-loop length, as one can simply extend each putative stem loop by checking additional flanking base pairs until either an invalid base pair is encountered or the end of the transcript is reached. Additionally, this approach can be naturally extended to allow for bulges when checking base pairs to extend the stem of a motif (see Figure 5.2). When checking additional base pairs, one could check not only the next flanking base pair $(i,\ j)$, but also the base pairs associated with a left $(i-1,\ j)$ and right bulge $(i,\ j+1)$. If either bulge base pair is satisfied, those motifs spawn their own branches of continued base pair checking conditioned on the presence of their bulge. Some rules would need to be enforced, such as a maximum number of allowed bulges before exploring "perfect" stem continuations only, bulge size considerations, and the allowance of internal loops. The space of considered motifs would be as expansive as the search rules permit it to grow. More restrictive rules would yield faster scoring but fewer putative sites; less restrictive rules would require more time but yield a more comprehensive assessment of local structure elements.

With a relational map of sites with motifs as related to the substructures they contain, the door would also be opened to reformulate the scoring computations themselves. Scoring time could be saved by reusing information from sub-sites. For instance, the working sum of log emission probability ratios (the core of the *patteRNA* score) can be tracked at each level, such that scores can be computed simply by modulating this sum with the

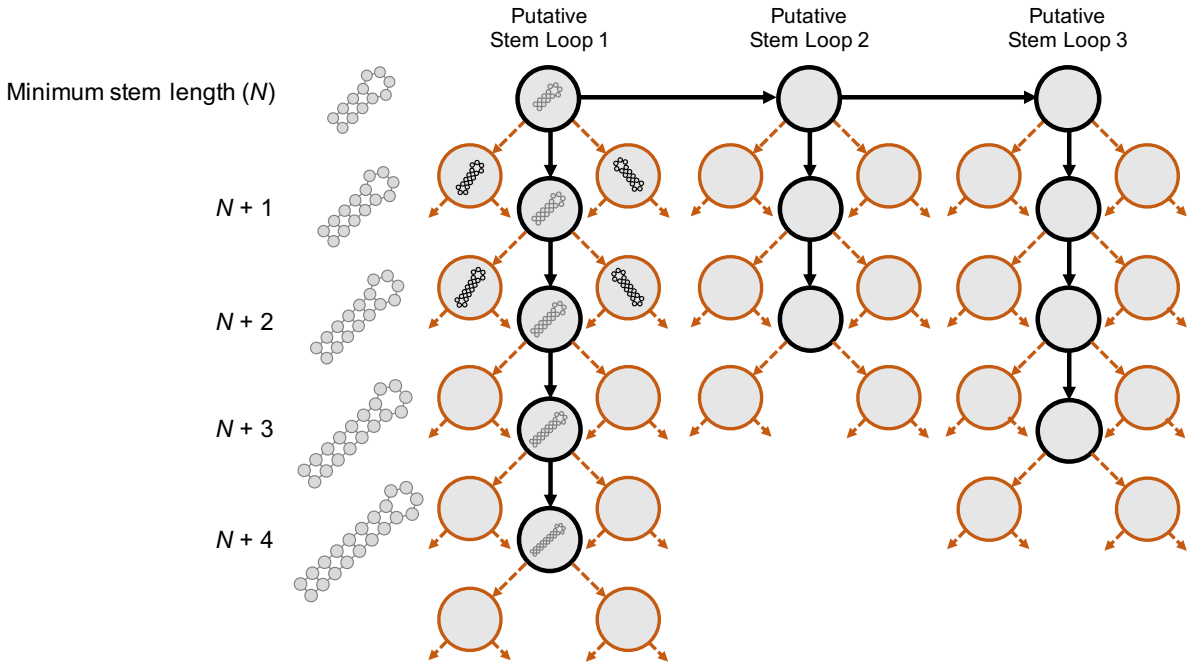


Figure 5.2: Illustration of a novel method to search for putative elements. Rather than beginning with a list of structures to search for, this approach only requires the definition of a stem-loop starting point (e.g., 4 base pairs with some loop length) then uses the sequence to guide the structural search. For each motif, the next flanking base pair (i, j) is considered. If it's compatible with the motif, the stem length is extended and the process repeats. A scheme could also allow for the presence of bulges when checking each additional stem length; for instance, one could check not only the next flanking base pair, but also the base pairs associated with a left $(i-1, j)$ and right bulge $(i, j+1)$. If either bulge base pair is satisfied, those motifs spawn their own branches (indicated in orange) of continued base pair checking conditioned on the presence of their bulge. This process is iteratively repeated until some stop condition is met.

log emission probability ratios of the flanking nucleotides contained in each longer motif, as well as the new forward and backward log probability ratios. Such an implementation could be transformative in terms of its speed and comprehensiveness. One caveat to this approach is the question of how to compute c-scores, given that your search may arrive at thousands of different secondary structure arrangements, each of which require a sampled null distribution. This problem could be solved by deriving the parameters of the null distribution according to the properties of the data and target motif instead of sampling them.

Recent work by Cao et al. [15] was able to identify characteristic SHAPE profiles associated to specific loop topologies, and they developed a novel method, SHAPELoop, that improves structure prediction by refining secondary structure models based on the locations of detected loop elements. The characteristic SHAPE profiles they identified could presumably be used within the schematic described in this section; when first

identifying putative loops with the shortest stem length, reactivities in the loop could be inspected for their consistency with the quantified reactivity pattern for the sought loop length to reinforce or prune the search space. Sites that have conflicting SHAPE data in the loop could be discarded from further analysis. Alternatively, the specific SHAPE patterns associated with certain loops could be utilized during the scoring phase to more accurate score putative sites. Regardless of the implementation, however, the versatility of the method warrants more consideration. The work by Cao et al. is specific to SHAPE data, and it is unclear how well their detected profiles would translate to other probes and SP experiments. It might be possible to reformulate their patterns into a likelihood model, but it is also likely that different probes—which measure different types of stereochemistry—would have fundamentally different biases in their expected loop profiles. The proper incorporation of probe-specific biases is a non-trivial problem.

5.2.3 Deep Learning

Recent studies have demonstrated that deep learning can accurately identify diverse structural patterns in SP data [210, 181]. Given the complexities associated with RNA structure dynamics, highly parameterized models like deep neural networks are likely to assist in making accurate assessments of structure. Indeed, in this work, even a simple single-layer perceptron was able to help discern stem-loops that engage in RBP binding from stem-loops that don't engage in RBP binding by just considering their differential scores between *in vitro* and *in vivo* conditions. Despite their popularity across many fields of science, deep learning methods in RNA structure prediction have remained relatively naïve. This can be partially attributed to a lack of high-quality reference data (e.g., structures) to use for training. Fewer than 10,000 non-redundant reference structure exist, which severely limits the scale and scope of deep learning applications. For transcriptome-wide datasets specifically, perhaps the most impactful use of SP, virtually no reference structure information exists. Additionally, there are difficulties associated with developing a neural architecture well-suited for versatile RNA sequence and structure processing. Convolutional and recurrent networks have seen the most use in published literature [210, 181, 110], but our understanding of the ideal architectures by which to process nucleotide sequence data will likely evolve in coming decades. Another issue with their general use is the interpretability of their predictions; in addition to accurate predictions, it is often desirable to understand why and how a model arrived as a specific conclusion. When using deep neural networks, interpretability is a challenge and an issue often en-

countered with ample literature describing potential solutions (see [128]). For example, Sun et al. [181] used their deep learned model of RNA-protein interactions to create saliency maps over any input data, which allowed for some interpretations on the details of specific interactions. The impact of future deep learning methods would be enhanced if they can also arrive at an interpretable system model. At any rate, careful work is needed to consider the specific predictions addressable by deep learning models in RNA structure biology and the best architectures by which to construct them.

5.3 Closing Remarks

The future of RNA structure research is almost certainly going to depend on advanced computational methods capable of integrating structural information from diverse types of data in making holistic interpretations on the function and dynamics of *in vivo* RNA. As our ability to characterize SP data matures, it will become more important to associate structures to functional processes such as RBP binding, polyadenylation, and splicing. As such, integrative methods that aim to link structural tendencies to function are poised to blossom in the coming decades. Utilizing *patteRNA* to profile stem-loops that bind with RBPs was one such integrative analysis highlighted here. Future methods will likely be capable of directly integrating SP data with functional data (i.e., RBP signals, chromosome interactions, ligand interactions, etc.) to automatically disentangle structural trends associated with a specific function; that said, such high-dimensional analyses might be restricted to super-computing pipelines for the foreseeable future. Nevertheless, such analyses will undoubtedly uncover the mechanistic sources of countless processes across RNA biology, and it is clear more work is needed to devise the next generation of structuromic tools.

Bibliography

[1] Abdelsayed, M. M., Ho, B. T., Vu, M. M. K., Polanco, J., Spitale, R. C., and Lupták, A. Multiplex Aptamer Discovery through Apta-Seq and Its Application to ATP Aptamers Derived from Human-Genomic SELEX. *ACS Chemical Biology 12*, 8 (2017), 2149–2156.

[2] Andronescu, M., Bereg, V., Hoos, H. H., and Condon, A. RNA STRAND: The RNA secondary structure and statistical analysis database. *BMC Bioinformatics 9*, 1 (2008), 340.

[3] Aviran, S., Lucks, J. B., and Pachter, L. RNA structure characterization from chemical mapping experiments. *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (2011), 1743–1750.

[4] Aviran, S., Trapnell, C., Lucks, J. B., Mortimer, S. A., Luo, S., Schroth, G. P., Doudna, J. A., Arkin, A. P., and Pachter, L. Modeling and automation of sequencing-based characterization of RNA structure. *Proceedings of the National Academy of Sciences of the United States of America 108*, 27 (2011), 11069–11074.

[5] Aviv, T., Lin, Z., Ben-Ari, G., Smibert, C. A., and Sicheri, F. Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p. *Nature Structural and Molecular Biology 13*, 2 (2006), 168–176.

[6] Bai, Y., Tambe, A., Zhou, K., and Doudna, J. A. RNA-guided assembly of Rev-RRE nuclear export complexes. *eLife 3* (2014), e03656.

[7] Battle, D. J., and Doudna, J. A. The stem-loop binding protein forms a highly stable and specific complex with the 3' stem-loop of histone mRNAs. *Rna 7*, 1 (2001), 123–132.

[8] Beaudoin, J. D., Novoa, E. M., Vejnar, C. E., Yartseva, V., Takacs, C. M., Kellis, M., and Giraldez, A. J. Analyses of mRNA structure dynamics identify embryonic gene regulatory programs. *Nature Structural and Molecular Biology 25*, 8 (2018), 677–686.

[9] Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R., and Stadler, P. F. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics 9*, 1 (2008), 474.

[10] Bilodeau, P. S., Domsic, J. K., Mayeda, A., Krainer, A. R., and Stoltzfus, C. M. RNA Splicing at Human Immunodeficiency Virus Type 1 3' Splice Site A2 Is Regulated by Binding of hnRNP A/B Proteins to an Exonic Splicing Silencer Element. *Journal of Virology 75*, 18 (2001), 8487–8497.

[11] Breaker, R. R. Riboswitches and the RNA World. *Cold Spring Harbor Perspectives in Biology 4*, 2 (2012), a003566.

[12] Busan, S., and Weeks, K. M. Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA 24*, 2 (2017), 143–148.

[13] Busan, S., Weidmann, C. A., Sengupta, A., and Weeks, K. M. Guidelines for SHAPE Reagent Choice and Detection Strategy for RNA Structure Probing Studies. *Biochemistry 58*, 23 (2019), 2655–2664.

[14] Cambray, G., Guimaraes, J. C., and Arkin, A. P. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in escherichia coli. *Nature Biotechnology 36*, 10 (2018), 1005.

[15] Cao, J., and Xue, Y. Characteristic chemical probing patterns of loop motifs improve prediction accuracy of RNA secondary structures. *Nucleic Acids Research* (2021), gkab250–.

[16] Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B., Strein, C., Davey, N., Humphreys, D., Preiss, T., Steinmetz, L., Krijgsveld, J., and Hentze, M. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell 149*, 6 (2012), 1393–1406.

[17] Cerase, A., Pintacuda, G., Tattermusch, A., and Avner, P. Xist localization and function: new insights from multiple levels. *Genome Biology 16*, 1 (2015), 166.

[18] Charpentier, B., Stutz, F., and Rosbash, M. A dynamic in vivo view of the HIV-Rev-RRE interaction. *Journal of Molecular Biology 266*, 5 (1997), 950–962.

[19] Cheng, C. Y., Kladwang, W., Yesselman, J., and Das, R. RNA structure inference through chemical mapping after accidental or intentional mutations. *bioRxiv 114*, 37 (2017), 9876–9881.

[20] Choudhary, K., Deng, F., and Aviran, S. Comparative and integrative analysis of RNA structural profiling data: current practices and emerging questions. *Quantitative Biology 5*, 1 (2017), 3–24.

[21] Choudhary, K., Lai, Y.-H., Tran, E. J., and Aviran, S. dStruct: identifying differentially reactive regions from RNA structurome profiling data. *Genome Biology 20*, 1 (12 2019), 40.

[22] Choudhary, K., Shih, N. P., Deng, F., Ledda, M., Li, B., and Aviran, S. Metrics for rapid quality control in RNA structure probing experiments. *Bioinformatics 32*, 23 (2016), 3575–3583.

[23] Cordero, P., and Das, R. Rich RNA Structure Landscapes Revealed by Mutate-and-Map Analysis. *PLOS Computational Biology 11*, 11 (2015), e1004473.

[24] Corley, M., Burns, M. C., and Yeo, G. W. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Molecular Cell 78*, 1 (2020), 9–29.

[25] Corley, M., Flynn, R. A., Lee, B., Blue, S. M., Chang, H. Y., and Yeo, G. W. Footprinting SHAPE-eCLIP Reveals Transcriptome-wide Hydrogen Bonds at RNA-Protein Interfaces. *Molecular Cell 80*, 5 (2020), 903–914.

[26] Crick, F. H. On protein synthesis. *Symposia of the Society for Experimental Biology 12* (1958), 138–163.

[27] Dallaire, P., Tan, H., Szulwach, K., Ma, C., Jin, P., and Major, F. Structural dynamics control the MicroRNA maturation pathway. *Nucleic Acids Research 44*, 20 (2016), 9956–9964.

[28] Dayton, E. T., Powell, D. M., and Dayton, A. I. Functional analysis of CAR, the target sequence for the Rev protein of HIV-1. *Science 246*, 4937 (1989), 1625–1629.

[29] Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences of the United States of America 106*, 1 (1 2009), 97–102.

[30] Deng, F., Ledda, M., Vaziri, S., and Aviran, S. Data-directed RNA secondary structure prediction using probabilistic modeling. *RNA 22*, 8 (2016), 1109–1119.

[31] Dethoff, E. A., Petzold, K., Chugh, J., Casiano-Negroni, A., and Al-Hashimi, H. M. Visualizing transient low-populated structures of RNA. *Nature 491*, 7426 (2012), 724–728.

[32] DiMattia, M. A., Watts, N. R., Stahl, S. J., Rader, C., Wingfield, P. T., Stuart, D. I., Steven, A. C., and Grimes, J. M. Implications of the HIV-1 Rev dimer structure at 3.2 Å resolution for multimeric binding to the Rev response element. *Proceedings of the National Academy of Sciences 107*, 13 (2010), 5810–5814.

[33] Ding, Y., Chan, C. Y., and Lawrence, C. E. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA 11*, 8 (2005), 1157–1166.

[34] Ding, Y., and Lawrence, C. E. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research 31*, 24 (2003), 7280–7301.

[35] Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., and Assmann, S. M. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature 505*, 7485 (2014), 696–700.

[36] Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N. J., Nostrand, E. L. V., Pratt, G. A., Yeo, G. W., Graveley, B. R., and Burge, C. B. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell 70*, 5 (2018), 854–867.

[37] Dowell, R. D., and Eddy, S. R. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics 7*, 1 (2006), 400.

[38] Eddy, S. R. Non–coding RNA genes and the modern RNA world. *Nature Reviews Genetics 2*, 12 (2001), 919–929.

[39] Eddy, S. R. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annual Review of Biophysics 43*, 1 (2014), 433–456.

[40] Eddy, S. R., and Durbin, R. RNA sequence analysis using covariance models. *Nucleic Acids Research 22*, 11 (1994), 2079–2088.

[41] Esteller, M. Non-coding RNAs in human disease. *Nature Reviews Genetics 12*, 12 (2011), 861–874.

[42] Fang, X., Wang, J., O'Carroll, I., Mitchell, M., Zuo, X., Wang, Y., Yu, P., Liu, Y., Rausch, J., Dyba, M., Kjems, J., Schwieters, C., Seifert, S., Winans, R., Watts, N., Stahl, S., Wingfield, P., Byrd, R., Le Grice, S., Rein, A., and Wang, Y.-X. An Unusual Topological Structure of the HIV-1 Rev Response Element. *Cell 155*, 3 (2013), 594–605.

[43] Feng, C., Chan, D., and Spitale, R. C. Assaying RNA Structure Inside Living Cells with SHAPE—mRNA Processing: Methods and Protocols. Springer New York, New York, NY, 2017, pp. 247–256.

[44] Fica, S. M., and Nagai, K. Cryo-electron microscopy snapshots of the spliceosome: Structural insights into a dynamic ribonucleoprotein machine. *Nature Structural and Molecular Biology 24*, 10 (2017), 791–799.

[45] Forster, A. C., and Symons, R. H. Self-cleavage of virusoid RNA is performed by the proposed 55-nucleotide active site. *Cell 50*, 1 (1987), 9–16.

[46] Fürtig, B., Richter, C., Wöhnert, J., and Schwalbe, H. NMR Spectroscopy of RNA. *ChemBioChem 4*, 10 (10 2003), 936–962.

[47] Gamarnik, A. V., and Andino, R. Switch from translation to RNA replication in a positive-stranded RNA virus. *Genes & Development 12*, 15 (1998), 2293–2304.

[48] Ganser, L. R., Kelly, M. L., Herschlag, D., and Al-Hashimi, H. M. The roles of structural dynamics in the cellular functions of RNAs. *Nature Reviews Molecular Cell Biology 20*, 8 (2019), 474–489.

[49] Gardner, P. P., and Giegerich, R. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics 5*, 1 (2004), 140.

[50] Gawroński, P., Pałac, A., and Scharff, L. B. Secondary structure of chloroplast mRNAs in vivo and in vitro. *Plants 9*, 3 (2020), 323.

[51] Gebauer, F., Schwarzl, T., Valcárcel, J., and Hentze, M. W. RNA-binding proteins in human genetic disease. *Nature Reviews Genetics 22*, 3 (2021), 185–198.

[52] Gerstberger, S., Hafner, M., and Tuschl, T. A census of human RNA-binding proteins. *Nature Reviews Genetics 15*, 12 (2014), 829–845.

[53] Ghut, J., Aw, A., Lim, S. W., Wang, J. X., Lambert, F. R. P., Tan, W. T., Shen, Y., Zhang, Y., Kaewsapsak, P., Li, C., Ng, S. B., Vardy, L. A., Tan, M. H., Nagarajan, N., and Wan, Y. With Nanopore Long Reads. *Nature Biotechnology* (2020), 1–11.

[54] Grechishnikova, D., and Poptsova, M. Conserved 3' UTR stem-loop structure in L1 and Alu transposons in human genome: possible role in retrotransposition. *BMC Genomics 17*, 1 (2016), 992.

[55] Groot, N. S. d., Armaos, A., Graña-Montes, R., Alriquet, M., Calloni, G., Vabulas, R. M., Tartaglia, G. G., Sanchez de Groot, N., Armaos, A., Graña-Montes, R., Alriquet, M., Calloni, G., Vabulas, R. M., and Tartaglia, G. G. RNA structure drives interaction with proteins. *Nature Communications 10*, 1 (2019), 3246.

[56] Guenther, U.-P., Weinberg, D. E., Zubradt, M. M., Tedeschi, F. A., Stawicki, B. N., Zagore, L. L., Brar, G. A., Licatalosi, D. D., Bartel, D. P., Weissman, J. S., and Jankowsky, E. The helicase Ded1p controls use of near-cognate translation initiation codons in 5' UTRs. *Nature 559*, 7712 (2018), 130–134.

[57] Guo, J. U., and Bartel, D. P. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science 353*, 6306 (2016), aaf5371.

[58] Gutell, R. R., Lee, J. C., and Cannone, J. J. The accuracy of ribosomal RNA comparative structure models. *Current Opinion in Structural Biology 12*, 3 (2002), 301–310.

[59] Hajdin, C. E., Bellaousov, S., Huggins, W., Leonard, C. W., Mathews, D. H., and Weeks, K. M. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proceedings of the National Academy of Sciences of the United States of America 110*, 14 (4 2013), 5498–5503.

[60] Hallberg, Z. F., Su, Y., Kitto, R. Z., and Hammond, M. C. Engineering and In Vivo Applications of Riboswitches. *Annual Review of Biochemistry 86*, 1 (2015), 1–25.

[61] Halvorsen, M., Martin, J. S., Broadaway, S., and Laederach, A. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genetics 6*, 8 (2010), e1001074.

[62] Helm, M., and Motorin, Y. Detecting RNA modifications in the epitranscriptome: predict and validate. *Nature Reviews Genetics 18*, 5 (2017), 275–291.

[63] Higgs, P. G., and Lehman, N. The RNA World: molecular cooperation at the origins of life. *Nature Reviews Genetics 16*, 1 (2015), 7–17.

[64] Hiller, M., Zhang, Z., Backofen, R., and Stamm, S. Pre-mRNA secondary structures influence exon recognition. *PLoS Genetics 3*, 11 (2007), 2147–2155.

[65] Hjelm, B. E., Rollins, B., Morgan, L., Sequeira, A., Mamdani, F., Pereira, F., Damas, J., Webb, M. G., Weber, M. D., Schatzberg, A. F., Barchas, J. D., Lee, F. S., Akil, H., Watson, S. J., Myers, R. M., Chao, E. C., Kimonis, V., Thompson, P. M., Bunney, W. E., and Vawter, M. P. Splice-Break: Exploiting an RNA-seq splice junction algorithm to discover mitochondrial DNA deletion breakpoints and analyses of psychiatric disorders. *Nucleic Acids Research 47*, 10 (2019), 26.

[66] Hofacker, I. L., Bernhart, S. H. F., and Stadler, P. F. Alignment of RNA base pairing probability matrices. *Bioinformatics 20*, 14 (9 2004), 2222–2227.

[67] Hofacker, I. L., Priwitzer, B., and Stadler, P. F. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics 20*, 2 (2004), 186–190.

[68] Holbrook, S. R., and Kim, S.-H. RNA crystallography. *Biopolymers 44*, 1 (1997), 3–21.

[69] Homan, P. J., Favorov, O. V., Lavender, C. A., Kursun, O., Ge, X., Busan, S., Dokholyan, N. V., and Weeks, K. M. Single-molecule correlated chemical probing of RNA. *Proceedings of the National Academy of Sciences 111*, 38 (2014), 13858–13863.

[70] Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics 26*, 5 (2010), 680–682.

[71] Huston, N. C., Wan, H., Strine, M. S., de Cesaris Araujo Tavares, R., Wilen, C. B., and Pyle, A. M. Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *Molecular Cell* (2021), 2020.07.10.197079.

[72] Ignatova, Z., and Narberhaus, F. Systematic probing of the bacterial RNA structurome to reveal new functions. *Current Opinion in Microbiology 36* (2017), 14–19.

[73] Incarnato, D., Morandi, E., Anselmi, F., Simon, L. M., Basile, G., and Oliviero, S. In vivo probing of nascent RNA structures reveals principles of cotranscriptional folding. *Nucleic Acids Research 45*, 16 (2017), gkx617–.

[74] Iserman, C., Roden, C., Boerneke, M., Sealfon, R., McLaughlin, G., Jungreis, I., Park, C., Boppana, A., Fritch, E., Hou, Y. J., Theesfeld, C., Troyanskaya, O. G., Baric, R. S., Sheahan, T. P., Weeks, K., and Gladfelter, A. S. Specific viral RNA drives the SARS CoV-2 nucleocapsid to phase separate. *bioRxiv* (1 2020), 2020.06.11.147199.

[75] Jayaraman, B., Crosby, D. C., Homer, C., Ribeiro, I., Mavor, D., and Frankel, A. D. RNA-directed remodeling of the HIV-1 protein Rev orchestrates assembly of the Rev–Rev response element complex. *eLife 3* (2014), e04120.

[76] Jayaraman, B., Mavor, D., Gross, J. D., and Frankel, A. D. Thermodynamics of Rev–RNA Interactions in HIV-1 Rev–RRE Assembly. *Biochemistry 54*, 42 (2015), 6545–6554.

[77] Jolma, A., Zhang, J., Mondragón, E., Morgunova, E., Kivioja, T., Laverty, K. U., Yin, Y., Zhu, F., Bourenkov, G., Morris, Q., Hughes, T. R., Maher, L. J., and Taipale, J. Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. *Genome Research 30*, 7 (2020), 962–973.

[78] Jones, S. Protein-RNA interactions: a structural analysis. *Nucleic Acids Research 29*, 4 (2 2001), 943–954.

[79] Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., Rivas, E., Eddy, S. R., Finn, R., Bateman, A., and Petrov, A. I. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research 49*, D1 (2020), D192–D200.

[80] Karn, J., and Stoltzfus, C. M. Transcriptional and Posttranscriptional Regulation of HIV-1 Gene Expression. *Cold Spring Harbor Perspectives in Medicine 2*, 2 (2012), a006916.

[81] Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. Genome-wide measurement of RNA secondary structure in yeast. *Nature 467*, 7311 (2010), 103–107.

[82] Kingma, D. P., and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* (2014).

[83] Kjems, J., Brown, M., Chang, D. D., and Sharp, P. A. Structural analysis of the interaction between the human immunodeficiency virus Rev protein and the Rev response element. *Proceedings of the National Academy of Sciences 88*, 3 (1991), 683–687.

[84] Knapp, G. Enzymatic approaches to probing of RNA secondary and tertiary structure. In *Methods in Enzymology*, vol. 180. Academic Press, 1989, pp. 192–212.

[85] Knudsen, B., and Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic acids research 31*, 13 (7 2003), 3423–3428.

[86] Kramer, M. C., and Gregory, B. D. Does RNA secondary structure drive translation or vice versa? *Nature Structural and Molecular Biology 25*, 8 (2018), 641–643.

[87] Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell 31*, 1 (1982), 147–157.

[88] Kubota, M., Tran, C., and Spitale, R. C. Progress and challenges for chemical probing of RNA structure inside living cells. *Nature Chemical Biology 11*, 12 (2015), 933–941.

[89] Kuhn, J., Tengler, U., and Binder, S. Transcript Lifetime Is Balanced between Stabilizing Stem-Loop Structures and Degradation-Promoting Polyadenylation in Plant Mitochondria. *Molecular and Cellular Biology 21*, 3 (2001), 731–742.

[90] Kullback, S., and Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics 22*, 1 (1951), 79–86.

[91] Kutchko, K. M., and Laederach, A. Transcending the prediction paradigm: novel applications of SHAPE to RNA function and evolution. *Wiley Interdisciplinary Reviews: RNA 8*, 1 (2017), e1374.

[92] Kutchko, K. M., Madden, E. A., Morrison, C., Plante, K. S., Sanders, W., Vincent, H. A., Cruz Cisneros, M. C., Long, K. M., Moorman, N. J., Heise, M. T., and Laederach, A. Structural divergence creates new functional features in alphavirus genomes. *Nucleic Acids Research 46*, 7 (2018), gky012–.

[93] Kwok, C. K. Dawn of the in vivo RNA structurome and interactome. *Biochemical Society Transactions 44*, 5 (2016), 1395–1410.

[94] Kwok, C. K., Ding, Y., Tang, Y., Assmann, S. M., and Bevilacqua, P. C. Determination of in vivo RNA structure in low-abundance transcripts. *Nature Communications 4*, 1 (2013), 2971.

[95] Kwok, C. K., Marsico, G., Sahakyan, A. B., Chambers, V. S., and Balasubramanian, S. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nature Methods 13*, 10 (2016), 841–844.

[96] Lai, Y.-H., Choudhary, K., Cloutier, S. C., Xing, Z., Aviran, S., and Tran, E. J. Genome-Wide Discovery of DEAD-Box RNA Helicase Targets Reveals RNA Structural Remodeling in Transcription Termination. *Genetics 212*, 1 (2019), genetics.302058.2019.

[97] Lan, T., Allan, M., Malsick, L., Khandwala, S., Nyeo, S., Bathe, M., Griffiths, A., and Rouskin, S. Structure of the full SARS-CoV-2 RNA genome in infected cells. *bioRxiv* (2020), 2020.06.29.178343.

[98] Larsen, N., and Zwieb, C. SRP-RNA sequence alignment and secondary structure. *Nucleic Acids Research 19*, 2 (1991), 209–215.

[99] Latham, M. P., Brown, D. J., McCallum, S. A., and Pardi, A. NMR methods for studying the structure and dynamics of RNA. *Chembiochem : a European journal of chemical biology 6*, 9 (9 2005), 1492–1505.

[100] Lavender, C. A., Lorenz, R., Zhang, G., Tamayo, R., Hofacker, I. L., and Weeks, K. M. Model-Free RNA Sequence and Structure Alignment Informed by SHAPE Probing Reveals a Conserved Alternate Secondary Structure for 16S rRNA. *PLOS Computational Biology 11*, 5 (5 2015), e1004126.

[101] Ledda, M., and Aviran, S. PATTERNA: Transcriptome-wide search for functional RNA elements via structural data signatures. *Genome Biology 19*, 1 (2018), 28.

[102] Lee, Y. J., Wang, Q., and Rio, D. C. Coordinate regulation of alternative pre-mRNA splicing events by the human RNA chaperone proteins hnRNPA1 and DDX5. *Genes & Development 32*, 15-16 (2018), 1060–1074.

[103] Legiewicz, M., Badorrek, C. S., Turner, K. B., Fabris, D., Hamm, T. E., Rekosh, D., Hammarskjöld, M.-L., and Grice, S. F. J. L. Resistance to RevM10 inhibition reflects a conformational switch in the HIV-1 Rev response element. *Proceedings of the National Academy of Sciences 105*, 38 (2008), 14365–14370.

[104] Lewis, C. J. T., Pan, T., and Kalsotra, A. RNA modifications and structures cooperate to guide RNA–protein interactions. *Nature Reviews Molecular Cell Biology 18*, 3 (2017), 202–210.

[105] Li, B., Tambe, A., Aviran, S., and Pachter, L. PROBer Provides a General Toolkit for Analyzing Sequencing-Based Toeprinting Assays. *Cell Systems 4*, 5 (2017), 568–574.

[106] Li, H., and Aviran, S. Statistical modeling of RNA structure profiling experiments enables parsimonious reconstruction of structure landscapes. *Nature Communications 9*, 1 (2018), 606.

[107] Long, D., Lee, R., Williams, P., Chan, C. Y., Ambros, V., and Ding, Y. Potent effect of target structure on microRNA function. *Nature Structural & Molecular Biology 14*, 4 (2007), 287–294.

[108] Lorenz, R., Bernhart, S. H., Siederdissen, C. H. z., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. ViennaRNA Package 2.0. *Algorithms for Molecular Biology 6*, 1 (2011), 26.

[109] Lorenz, R., Luntzer, D., Hofacker, I. L., Stadler, P. F., and Wolfinger, M. T. SHAPE directed RNA folding. *Bioinformatics 32*, 1 (2016), 145–147.

[110] Lu, W., Tang, Y., Wu, H., Huang, H., Fu, Q., Qiu, J., and Li, H. Predicting RNA secondary structure via adaptive deep recurrent neural networks with energy-based filter. *BMC Bioinformatics 20*, Suppl 25 (2019), 684.

[111] Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., and Arkin, A. P. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences of the United States of America 108*, 27 (2011), 11063–11068.

[112] Mailler, E., Paillart, J. C., Marquet, R., Smyth, R. P., and Vivet-Boudou, V. The evolution of RNA structural probing methods: From gels to next-generation sequencing. *Wiley Interdisciplinary Reviews: RNA 10*, 2 (2019), e1518.

[113] Manfredonia, I., and Incarnato, D. Structure and regulation of coronavirus genomes: state-of-the-art and novel insights from SARS-CoV-2 studies. *Biochemical Society Transactions 0*, November (12 2020), 1–12.

[114] Manfredonia, I., Nithin, C., Ponce-Salvatierra, A., Ghosh, P., Wirecki, T. K., Marinus, T., Ogando, N. S., Snijder, E. J., van Hemert, M. J., Bujnicki, J. M., and Incarnato, D. Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Research 48*, 22 (2020), 12436–12452.

[115] Mann, D. A., Mikaélian, I., Zemmel, R. W., Green, S. M., Lowe, A. D., Kimura, T., Singh, M., Jonathan, P., Butler, G., Gait, M. J., and Karn, J. A Molecular Rheostat Co-operative Rev Binding to Stem I of the Rev-response Element Modulates Human Immunodeficiency Virus Type-1 Late Gene Expression. *Journal of Molecular Biology 241*, 2 (1994), 193–207.

[116] Marangio, P., Law, K. Y. T., Sanguinetti, G., and Granneman, S. Differential BUM-HMM: a robust statistical modelling approach for detecting RNA flexibility changes in high-throughput structure probing data. *bioRxiv* (2020), 2020.07.30.229179.

[117] Marinus, T., Fessler, A. B., Ogle, C. A., and Incarnato, D. A novel SHAPE reagent enables the analysis of RNA structure in living cells with unprecedented accuracy. *Nucleic Acids Research* (2021).

[118] Markham, N. R., and Zuker, M. UNAFold—Bioinformatics: Structure, Function and Applications. Humana Press, Totowa, NJ, 2008, pp. 3–31.

[119] Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology 288*, 5 (1999), 911–940.

[120] Maticzka, D., Lange, S. J., Costa, F., and Backofen, R. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biology 15*, 1 (2014), R17.

[121] Mauger, D. M., Golden, M., Yamane, D., Williford, S., Lemon, S. M., Martin, D. P., and Weeks, K. M. Functionally conserved architecture of

hepatitis C virus RNA genomes. *Proceedings of the National Academy of Sciences of the United States of America 112*, 12 (2015), 3692–3697.

[122] Mauger, D. M., Joseph Cabral, B., Presnyak, V., Su, S. V., Reid, D. W., Goodman, B., Link, K., Khatwani, N., Reynders, J., Moore, M. J., and McFadyen, I. J. mRNA structure regulates protein expression through changes in functional half-life. *Proceedings of the National Academy of Sciences of the United States of America 116*, 48 (2019), 24075–24083.

[123] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics 9*, 5 (2008), 356–369.

[124] McCown, P. J., Corbino, K. A., Stav, S., Sherlock, M. E., and Breaker, R. R. Riboswitch diversity and distribution. *RNA 23*, 7 (2017), 995–1011.

[125] Melnick, M., Gonzales, P., Cabral, J., Allen, M. A., Dowell, R. D., and Link, C. D. Heat shock in C. elegans induces downstream of gene transcription and accumulation of double-stranded RNA. *PLoS ONE 14*, 4 (2019), e0206715.

[126] Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M. RNA Structure Analysis at Single Nucleotide Resolution by Selective 2‘-Hydroxyl Acylation and Primer Extension (SHAPE). *Journal of the American Chemical Society 127*, 12 (3 2005), 4223–4231.

[127] Miao, Z., and Westhof, E. RNA Structure: Advances and Assessment of 3D Structure Prediction. *Annual Review of Biophysics 46*, 1 (2017), 483–503.

[128] Montavon, G., Samek, W., and Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing 73* (2018), 1–15.

[129] Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., and Worm, B. How Many Species Are There on Earth and in the Ocean? *PLoS Biology 9*, 8 (2011), e1001127.

[130] Mortimer, S. A., Kidwell, M. A., and Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics 15*, 7 (2014), 469–479.

[131] Mustoe, A. M., Brooks, C. L., and Al-Hashimi, H. M. Hierarchy of RNA functional dynamics. *Annual Review of Biochemistry 83*, 1 (2014), 441–466.

[132] Mustoe, A. M., Busan, S., Rice, G. M., Hajdin, C. E., Peterson, B. K., Ruda, V. M., Kubica, N., Nutiu, R., Baryza, J. L., and Weeks, K. M. Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. *Cell 173*, 1 (2018), 181–195.

[133] Narberhaus, F., Waldminghaus, T., and Chowdhury, S. RNA thermometers. *FEMS Microbiology Reviews 30*, 1 (1 2006), 3–16.

[134] Nawrocki, E. P., and Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics 29*, 22 (2013), 2933–2935.

[135] Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics 25*, 10 (2009), 1335–1337.

[136] Nostrand, E. L. V., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J.-Y., Cody, N. A. L., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C., Bouvrette, L. P. B., Bergalet, J., Duff, M. O., Garcia, K. E., Gelboin-Burkhart, C., Hochman, M., Lambert, N. J., Li, H., McGurk, M. P., Nguyen, T. B., Palden, T., Rabano, I., Sathe, S., Stanton, R., Su, A., Wang, R., Yee, B. A., Zhou, B., Louie, A. L., Aigner, S., Fu, X.-D., Lécuyer, E., Burge, C. B., Graveley, B. R., and Yeo, G. W. A large-scale binding and functional map of human RNA-binding proteins. *Nature 583*, 7818 (2020), 711–719.

[137] Nussinov, R., and Jacobson, A. B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences of the United States of America 77*, 11 (1980), 6309–6313.

[138] Ochsenreiter, R., Hofacker, I. L., and Wolfinger, M. T. Functional RNA Structures in the 3'UTR of Tick-Borne, Insect-Specific and No-Known-Vector Flaviviruses. *bioRxiv 11*, 3 (2019), 298.

[139] Oliphant, T. E. Python for Scientific Computing. *Computing in Science & Engineering 9*, 3 (2007), 10–20.

[140] Pace, N. R., Thomas, B. C., and Woese, C. R. Probing RNA structure, function, and history by comparative analysis. *Cold Spring Harbor Monograph Series 37* (1999), 113–142.

[141] Palazzo, A. F., and Lee, E. S. Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics 6* (2015), 2.

[142] Pedregosa, F., Grisel, O., Weiss, R., Passos, A., Brucher, M., Varoquax, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., and Brucher, M. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[143] Perrigue, P. M., Erdmann, V. A., and Barciszewski, J. Alexander Rich: In Memoriam. *Trends in Biochemical Sciences 40*, 11 (2015), 623–624.

[144] Phillips, C., Kyriakopoulou, C. B., and Virtanen, A. Identification of a stem-loop structure important for polyadenylation at the murine IgM secretory poly(A) site. *Nucleic Acids Research 27*, 2 (1999), 429–438.

[145] Pirakitikulr, N., Kohlway, A., Lindenbach, B. D., and Pyle, A. M. The Coding Region of the HCV Genome Contains a Network of Regulatory RNA Structures. *Molecular Cell 62*, 1 (2016), 111–120.

[146] Pollard, V. W., and Malim, M. H. THE HIV-1 REV PROTEIN. *Annual Review of Microbiology 52*, 1 (1998), 491–532.

[147] Ponce-Salvatierra, A., Astha, Merdas, K., Nithin, C., Ghosh, P., Mukherjee, S., and Bujnicki, J. M. Computational modeling of RNA 3D structure based on experimental data. *Bioscience Reports 39*, 2 (2 2019), BSR20180430.

[148] Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 2 (1989), 257–286.

[149] Radecki, P., Ledda, M., and Aviran, S. Automated recognition of RNA structure motifs by their SHAPE data signatures. *Genes 9*, 6 (2018), 300.

[150] Radecki, P., Ledda, M., and Aviran, S. Dataset: Automated recognition of RNA structure motifs by their SHAPE data signatures. *10.5281/zenodo.1256867* (5 2018).

[151] Radecki, P., Uppuluri, R., and Aviran, S. Rapid structure-function insights via hairpin-centric analysis of big RNA structure probing datasets. *bioRxiv* (1 2021), 2021.04.27.441661.

[152] Radecki, P., Uppuluri, R., Deshpande, K., and Aviran, S. Accurate Detection of RNA Stem-Loops in Structurome Data Reveals Widespread Association with Protein Binding Sites. *bioRxiv* (1 2021), 2021.04.28.441809.

[153] Radecki, P., Uppuluri, R., Deshpande, K., and Aviran, S. Dataset: Accurate Detection of RNA Stem-Loops in Structurome Data Reveals Widespread Association with Protein Binding Sites. *Zenodo* (4 2021).

[154] Ramanouskaya, T. V., and Grinev, V. V. The determinants of alternative RNA splicing in human cells. *Molecular Genetics and Genomics 292*, 6 (2017), 1175–1195.

[155] Rausch, J. W., and Grice, S. F. J. L. HIV Rev Assembly on the Rev Response Element (RRE): A Structural Perspective. *Viruses 7*, 6 (2015), 3053–3075.

[156] Reuter, J. S., and Mathews, D. H. RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinformatics 11*, 1 (2010), 129.

[157] Rich, A. On the problems of evolution and biochemical information transfer. *Horizons in biochemistry* (1962), 103–126.

[158] Righetti, F., Nuss, A. M., Twittenhoff, C., Beele, S., Urban, K., Will, S., Bernhart, S. H., Stadler, P. F., Dersch, P., and Narberhaus, F. Temperature-responsive in vitro RNA structurome of Yersinia pseudotuberculosis. *Proceedings of the National Academy of Sciences 113*, 26 (6 2016), 7237 LP – 7242.

[159] Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature 505*, 7485 (2014), 701–705.

[160] Ruggiero, E., and Richter, S. N. G-quadruplexes and G-quadruplex ligands: targets and tools in antiviral therapy. *Nucleic Acids Research 46*, 7 (2018), gky187–.

[161] Saha, K., England, W., Fernandez, M. M., Biswas, T., Spitale, R. C., and Ghosh, G. Structural disruption of exonic stem-loops immediately upstream of the intron regulates mammalian splicing. *Nucleic acids research 48*, 11 (2020), 6294–6309.

[162] Sanders, W., Fritch, E. J., Madden, E. A., Graham, R. L., Vincent, H. A., Heise, M. T., Baric, R. S., and Moorman, N. J. Comparative analysis of coronavirus genomic RNA structure reveals conservation in SARS-like coronaviruses. *bioRxiv* (2020), 2020.06.15.153197.

[163] Sankoff, D. Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM Journal on Applied Mathematics 45*, 5 (10 1985), 810–825.

[164] Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics 6*, 2 (1978).

[165] Selega, A., Sirocchi, C., Iosub, I., Granneman, S., and Sanguinetti, G. Robust statistical modeling improves sensitivity of high-throughput RNA structure probing experiments. *Nature Methods 14*, 1 (2017), 83–89.

[166] Serganov, A., and Patel, D. J. Ribozymes, riboswitches and beyond: Regulation of gene expression without proteins. *Nature Reviews Genetics 8*, 10 (2007), 776–790.

[167] Sexton, A. N., Wang, P. Y., Rutenberg-Schoenberg, M., and Simon, M. D. Interpreting Reverse Transcriptase Termination and Mutation Events for Greater Insight into the Chemical Probing of RNA. *Biochemistry 56*, 35 (2017), 4713–4721.

[168] Sharp, P. A. The Centrality of RNA. *Cell 136*, 4 (2009), 577–580.

[169] Sherpa, C., Rausch, J. W., Le Grice, S. F. J., Hammarskjold, M.-L., and Rekosh, D. The HIV-1 Rev response element (RRE) adopts alternative conformations that promote different rates of virus replication. *Nucleic Acids Research 43*, 9 (2015), 4676–4686.

[170] Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E., and Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods 11*, 9 (2014), 959–965.

[171] Simon, L. M., Morandi, E., Luganini, A., Gribaudo, G., Martinez-Sobrido, L., Turner, D. H., Oliviero, S., and Incarnato, D. In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. *Nucleic Acids Research 47*, 13 (2019), 7003–7017.

[172] Singh, J., Hanson, J., Paliwal, K., and Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications 10*, 1 (2019), 5407.

[173] Sloma, M. F., and Mathews, D. H. Chapter Four: Improving RNA Secondary Structure Prediction with Structure Mapping Data. In *Computational Methods for Understanding Riboswitches*, S.-J. Chen and D. H. B. T. M. i. E. Burke-Aguero, Eds., vol. 553. Academic Press, 2015, pp. 91–114.

[174] Smola, M. J., Christy, T. W., Inoue, K., Nicholson, C. O., Friedersdorf, M., Keene, J. D., Lee, D. M., Calabrese, J. M., and Weeks, K. M. 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 113*, 37 (2016), 10322–10327.

[175] Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A., and Weeks, K. M. 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 (2015), 1643–1669.

[176] Spasic, A., Assmann, S. M., Bevilacqua, P. C., and Mathews, D. H. Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Research 46*, 1 (2017), 314–323.

[177] Spitale, R. C., Crisalli, P., Flynn, R. A., Torre, E. A., Kool, E. T., and Chang, H. Y. RNA SHAPE analysis in living cells. *Nature Chemical Biology 9*, 1 (2013), 18–20.

[178] Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H. Y., Batista, P. J., Torre, E. A., Kool, E. T.,

and Chang, H. Y. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature 519*, 7544 (3 2015), 486–490.

[179] Šponer, J., Bussi, G., Krepl, M., Banáš, P., Bottaro, S., Cunha, R. A., Gil-Ley, A., Pinamonti, G., Poblete, S., Jurečka, P., Walter, N. G., and Otyepka, M. RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. *Chemical Reviews 118*, 8 (2018), 4177–4338.

[180] Sükösd, Z., Swenson, M. S., Kjems, J., and Heitsch, C. E. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic Acids Research 41*, 5 (2013), 2807–2816.

[181] Sun, L., Xu, K., Huang, W., Yang, Y. T., Li, P., Tang, L., Xiong, T., and Zhang, Q. C. Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures. *Cell Research* (2021), 1–22.

[182] Sysoev, V. O., Fischer, B., Frese, C. K., Gupta, I., Krijgsveld, J., Hentze, M. W., Castello, A., and Ephrussi, A. Global changes of the RNA-bound proteome during the maternal-to-zygotic transition in Drosophila. *Nature Communications 7*, 1 (2016), 12128.

[183] Tang, Y., Bouvier, E., Kwok, C. K., Ding, Y., Nekrutenko, A., Bevilacqua, P. C., and Assmann, S. M. StructureFold: genome-wide RNA secondary structure mapping and reconstruction in vivo. *Bioinformatics 31*, 16 (2015), 2668–2675.

[184] Tijerina, P., Mohr, S., and Russell, R. DMS footprinting of structured RNAs and RNA-protein complexes. *Nature Protocols 2*, 10 (2007), 2608–2623.

[185] Tomezsko, P., Swaminathan, H., and Rouskin, S. DMS-MaPseq for Genome-Wide or Targeted RNA Structure Probing In Vitro and In Vivo. *Methods in Molecular Biology 2254*, 1 (2021), 219–238.

[186] Turner, D. H., and Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research 38*, suppl_1 (1 2010), D280–D282.

[187] Twittenhoff, C., Brandenburg, V. B., Righetti, F., Nuss, A. M., Mosig, A., Dersch, P., and Narberhaus, F. Lead-seq: Transcriptome-wide

structure probing in vivo using lead(II) ions. *Nucleic Acids Research 48*, 12 (7 2020), E71–E71.

[188] Underwood, J. G., Uzilov, A. V., Katzman, S., Onodera, C. S., Mainzer, J. E., Mathews, D. H., Lowe, T. M., Salama, S. R., and Haussler, D. FragSeq: Transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature Methods 7*, 12 (2010), 995–1001.

[189] Vasilyev, N., Polonskaia, A., Darnell, J. C., Darnell, R. B., Patel, D. J., and Serganov, A. Crystal structure reveals specific recognition of a G-quadruplex RNA by a β-turn in the RGG motif of FMRP. *Proceedings of the National Academy of Sciences 112*, 39 (2015), E5391–E5400.

[190] Velagapudi, S. P., Cameron, M. D., Haga, C. L., Rosenberg, L. H., Lafitte, M., Duckett, D. R., Phinney, D. G., and Disney, M. D. Design of a small molecule against an oncogenic noncoding RNA. *Proceedings of the National Academy of Sciences 113*, 21 (2016), 5898–5903.

[191] Velagapudi, S. P., Gallo, S. M., and Disney, M. D. Sequence-based design of bioactive small molecules that target precursor microRNAs. *Nature Chemical Biology 10*, 4 (2014), 291–297.

[192] Wacker, A., Weigand, J. E., Akabayov, S. R., Altincekic, N., Bains, J. K., Banijamali, E., Binas, O., Castillo-Martinez, J., Cetiner, E., Ceylan, B., Chiu, L. Y., Davila-Calderon, J., Dhamotharan, K., Duchardt-Ferner, E., Ferner, J., Frydman, L., Fürtig, B., Gallego, J., Grün, J. T., Hacker, C., Haddad, C., Hähnke, M., Hengesbach, M., Hiller, F., Hohmann, K. F., Hymon, D., de Jesus, V., Jonker, H., Keller, H., Knezic, B., Landgraf, T., Löhr, F., Luo, L., Mertinkus, K. R., Muhs, C., Novakovic, M., Oxenfarth, A., Palomino-Schätzlein, M., Petzold, K., Peter, S. A., Pyper, D. J., Qureshi, N. S., Riad, M., Richter, C., Saxena, K., Schamber, T., Scherf, T., Schlagnitweit, J., Schlundt, A., Schnieders, R., Schwalbe, H., Simba-Lahuasi, A., Sreeramulu, S., Stirnal, E., Sudakov, A., Tants, J. N., Tolbert, B. S., Vögele, J., Weiss, L., Wirmer-Bartoschek, J., Wirtz Martin, M. A., Wöhnert, J., and Zetzsche, H. Secondary structure determination of con-

served SARS-CoV-2 RNA elements by NMR spectroscopy. *Nucleic acids research 48*, 22 (2020), 12415–12435.

[193] Waldron, J. A., Tack, D. C., Ritchey, L. E., Gillen, S. L., Wilczynska, A., Turro, E., Bevilacqua, P. C., Assmann, S. M., Bushell, M., and Quesne, J. L. mRNA structural elements immediately upstream of the start codon dictate dependence upon eIF4A helicase activity. *Genome Biology 20*, 1 (2019), 300.

[194] Wan, Y., Qu, K., Zhang, Q. C., Flynn, R. A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R. C., Snyder, M. P., Segal, E., and Chang, H. Y. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature 505*, 7485 (2014), 706–709.

[195] Wang, P. Y., Sexton, A. N., Culligan, W. J., and Simon, M. D. Carbodiimide reagents for the chemical probing of RNA structure in cells. *RNA 25*, 1 (1 2019), 135–146.

[196] Warner, K., Homan, P., Weeks, K., Smith, A., Abell, C., and Ferré-D'Amaré, A. Validating Fragment-Based Drug Discovery for Biological RNAs: Lead Fragments Bind and Remodel the TPP Riboswitch Specifically. *Chemistry & Biology 21*, 5 (2014), 591–595.

[197] Watters, K. E., Fellmann, C., Bai, H. B., Ren, S. M., and Doudna, J. A. Systematic discovery of natural CRISPR-Cas12a inhibitors. *Science 362*, 6411 (2018), eaau5138.

[198] Watters, K. E., Yu, A. M., Strobel, E. J., Settle, A. H., and Lucks, J. B. Characterizing RNA structures in vitro and in vivo with selective 2-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Methods 103* (2016), 34–48.

[199] Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Jr, J. W. B., Swanstrom, R., Burch, C. L., and Weeks, K. M. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature 460*, 7256 (2009), 711–716.

[200] Weeks, K. M. Advances in RNA structure analysis by chemical probing. *Current Opinion in Structural Biology 20*, 3 (2010), 295–304.

[201] Wei, C., Xiao, R., Chen, L., Cui, H., Zhou, Y., Xue, Y., Hu, J., Zhou, B., Tsutsui, T., Qiu, J., Li, H., Tang, L., and Fu, X.-D. RBFox2 Binds Nascent RNA to Globally Regulate Polycomb Complex 2 Targeting in Mammalian Genomes. *Molecular Cell 62*, 6 (2016), 875–889.

[202] Weinberg, Z., Lünse, C. E., Corbino, K. A., Ames, T. D., Nelson, J. W., Roth, A., Perkins, K. R., Sherlock, M. E., and Breaker, R. R. Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic regions. *Nucleic Acids Research 45*, 18 (2017), gkx699.

[203] Weinberg, Z., Nelson, J. W., Lünse, C. E., Sherlock, M. E., and Breaker, R. R. Bioinformatic analysis of riboswitch structures uncovers variant classes with altered ligand specificity. *Proceedings of the National Academy of Sciences 114*, 11 (2017), E2077–E2085.

[204] Weng, X., Gong, J., Chen, Y., Wu, T., Wang, F., Yang, S., Yuan, Y., Luo, G., Chen, K., Hu, L., Ma, H., Wang, P., Zhang, Q. C., Zhou, X., and He, C. Keth-seq for transcriptome-wide RNA structure mapping. *Nature Chemical Biology 16*, 5 (2020), 489–492.

[205] Wilkinson, K. A., Merino, E. J., and Weeks, K. M. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols 1*, 3 (8 2006), 1610–1616.

[206] Wuchty, S., Fontana, W., Hofacker, I. L., and Schuster, P. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers 49*, 2 (2 1999), 145–165.

[207] Xue, Z., Hennelly, S., Doyle, B., Gulati, A., Novikova, I., Sanbonmatsu, K., and Boyer, L. A G-Rich Motif in the lncRNA Braveheart Interacts with a Zinc-Finger Transcription Factor to Specify the Cardiovascular Lineage. *Molecular Cell 64*, 1 (2016), 37–50.

[208] Yang, D., Liu, P., Wudeck, E. V., Giedroc, D. P., and Leibowitz, J. L. Shape analysis of the rna secondary structure of the mouse hepatitis virus 5' untranslated region and n-terminal nsp1 coding sequences. *Virology 475* (2015), 15–27.

[209] Yang, M., Woolfenden, H. C., Zhang, Y., Fang, X., Liu, Q., Vigh, M. L., Cheema, J., Yang, X., Norris, M., Yu, S., Carbonell, A.,

Brodersen, P., Wang, J., and Ding, Y. Intact RNA structurome reveals mRNA structure-mediated regulation of miRNA cleavage in vivo. *Nucleic Acids Research 48*, 15 (2020), 8767–8781.

[210] Zhang, H., Zhang, C., Li, Z., Li, C., Wei, X., Zhang, B., and Liu, Y. A new method of RNA secondary structure prediction based on convolutional neural network and dynamic programming. *Frontiers in Genetics 10*, MAY (2019), 467.

[211] Zhang, K., Li, S., Kappel, K., Pintilie, G., Su, Z., Mou, T.-C., Schmid, M. F., Das, R., and Chiu, W. Cryo-EM structure of a 40 kDa SAM-IV riboswitch RNA at 3.7 Å resolution. *Nature Communications 10*, 1 (12 2019), 5511.

[212] Zhao, B., Guffy, S. L., Williams, B., and Zhang, Q. An excited state underlies gene regulation of a transcriptional riboswitch. *Nature Chemical Biology 13*, 9 (2017), 968–974.

[213] Zinshteyn, B., Chan, D., England, W., Feng, C., Green, R., and Spitale, R. C. Assaying RNA structure with LASER-Seq. *Nucleic Acids Research 47*, 1 (2019), 43–55.

[214] Ziv, O., Gabryelska, M. M., Lun, A. T., Gebert, L. F., Sheu-Gruttadauria, J., Meredith, L. W., Liu, Z. Y., Kwok, C. K., Qin, C. F., MacRae, I. J., Goodfellow, I., Marioni, J. C., Kudla, G., and Miska, E. A. COMRADES determines in vivo RNA structures and interactions. *Nature Methods 15*, 10 (2018), 785–788.

[215] Ziv, O., Price, J., Shalamova, L., Kamenova, T., Goodfellow, I., Weber, F., and Miska, E. A. The Short- and Long-Range RNA-RNA Interactome of SARS-CoV-2. *Molecular Cell 80*, 6 (2020), 1067–1077.

[216] Zubradt, M., Gupta, P., Persad, S., Lambowitz, A. M., Weissman, J. S., and Rouskin, S. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nature Methods 14*, 1 (2017), 75–82.

[217] Zuker, M. On finding all suboptimal foldings of an RNA molecule. *Science 244*, 4900 (1989), 48–52.