Predicting RNA Secondary Structures from Sequence and Probing Data

Ronny Lorenz[a], Michael T. Wolfinger[a,b,c], Andrea Tanzer[a,*], Ivo L. Hofacker[a,d]

[a] *Department of Theoretical Chemistry, University of Vienna, Währingerstrasse 17, 1090 Vienna, Austria*
[b] *Department of Biochemistry and Molecular Cell Biology, Max F. Perutz Laboratories, University of Vienna, Dr. Bohr-Gasse 9, 1030 Vienna, Austria*
[c] *Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Dr. Bohr-Gasse 9, A-1030 Vienna, Austria*
[d] *Research Group BCB, Faculty of Computer Science, University of Vienna, Währingerstr. 29, 1090 Vienna, Austria*

Abstract

RNA secondary structures have proven essential for understanding the regulatory functions performed by RNA such as microRNAs, bacterial small RNAs, or riboswitches. This success is in part due to the availability of efficient computational methods for predicting RNA secondary structures. Recent advances focus on dealing with the inherent uncertainty of prediction by considering the ensemble of possible structures rather than the single most stable one. Moreover, the advent of high-throughput structural probing has spurred the development of computational methods that incorporate such experimental data as auxiliary information.

Keywords: RNA secondary structure prediction, structure probing

1. Structured RNAs and RNA elements

1.1. General Concepts

In every domain of life, cellular process interpreting genetic information require RNAs. For instance, RNAs prime DNA replication, induce gene silencing and activation via DNA (de)methylation, promote cross-talk of active gene loci, serve as templates for protein synthesis, translate DNA code into peptides, enzymatically catalyze formation of peptide bonds via peptidyl transferase activity, inactivate transposable elements, and mediate target specificity in post-trascriptional gene silencing.

RNAs fold back onto themselves by forming intra-molecular base pairs. The resulting structures are composed of two fundamental building blocks: paired regions (mostly type A helices), and unpaired loops. Interaction partners such as proteins, small ligands or other RNAs recognize specific structural motifs and can trigger refolding, cleavage, and chemical modifications upon binding.

Often, the target regions themselves must be unstructured, i.e. in loop regions, or at most engaged in weak structures in order to allow for interactions. Structures and unstructuredness therefore often depend on each another [74].

1.2. MicroRNAs - Small Regulatory RNAs

The modes of function described so far are not mutually exclusive. In fact, most RNAs or RNA based systems combine multiple functions, as illustrated by microRNAs (miRNAs), which represent a class of small ncRNAs found in plants and metazoans (reviewed in [33]). They can induce post-transcriptional gene silencing (PTGS) via an RNP named RISC (RNA induced silencing complex), which in turn targets 3' UTRs of mRNAs and leads to mRNA degradation or translational repression [5]. RISC, which contains a well defined stable set of proteins, is able to target thousands of different mRNAs and binding motifs, because sequence specificity is mediated through the individual miRNAs loaded into the complex - in evolutionary terms a simple solution to increase the target range and add nodes to regulatory networks.

The target sites on mRNAs must be single stranded, i.e. mostly unstructured. Binding of the miRNA to the target can be considered as an inter-molecular folding process. The resulting structures together with specific RISC components (Ago proteins) finally determine the fate of the mRNA.

While miRNAs are unstructured when loaded into RISC, during biogenesis they reside in heavily structured precursors (pre-miRNA). These stem-loop structures consist of helices and interspersed loops (bulges and small interior loops, for definition see section 2) and are cleaved from even longer primary transcripts. It remains unknown how cleavage sites are determined, but data suggest that structural and/or resulting sterical features play a crucial role.

Micro RNA biogenesis and microRNA-mediated PTGS

[*] Corresponding author
Email addresses: ronny@tbi.univie.ac.at (Ronny Lorenz), michael.wolfinger@univie.ac.at (Michael T. Wolfinger), at@tbi.univie.ac.at (Andrea Tanzer), ivo@tbi.univie.ac.at (Ivo L. Hofacker)
URL: http://www.tbi.univie.ac.at (Andrea Tanzer)

via RISC are among the best studied RNA pathways. However, several known unknowns remain to be addressed, most of which might be solved by in-depth structural analysis, *in-silico*, *in-vitro* as well as *in-vivo*.

1.3. Riboswitches - Regulatory RNA Elements

In contrast to regulatory RNAs, structure elements are locally stable structures representing functional domains within longer RNA molecules, for instance the aminoacyltransferase activity of ribosomal rRNAs. They are also found in UTRs of protein coding mRNAs, where they serve as protein binding sites or riboswitches.

Riboswitches are locally stable structures, that bind small metabolites in a concentration dependent manner and therefore serve as environmental sensors. They are bistable, i.e. they can fold into two more or less equally stable conformations. Binding of a small ligand induces the switch to the second most favorable structure, which in turn changes the state of switch and thus respective mRNA.

In bacteria riboswitches control transcription and some also translation by sensing metabolites or substrates of the respective gene products and thus allow for autoregulatory feed-back loops [64]. Initiation of transcription requires prior activation and assembly of transcription factors at promoters. This protein based regulation sets the cell to a specific state, but operates on a much larger time scale than translation and transcription themselves. Riboswitches, however, instantaneously refold upon ligand binding and are able to immediately act on transcription and thereby stalling translation. In turn, the original state is quickly reestablished once ligand concentrations drop. In the next sections we introduce algorithms for the most common tasks related to RNA secondary structure, including predicting the structure of a single RNA and its equilibrium properties, predicting the consensus structure for a set of related RNAs, as well as interactions between RNAs, and explore how RNAs refold over time. Finally, we review methods that combine data from probing experiments with *in silico* prediction to achieve higher quality predictions.

2. Secondary Structure Prediction

The most common approach to treat RNA structures algorithmically, is to reduce them to the set of base pairs, the so-called *secondary structure*, thereby abstracting from the actual spatial arrangement of nucleotides. Moreover, we require that each nucleotide i interacts with at most one other nucleotide j to form a base pair (i, j). For the sake of reducing computational complexity, most algorithms neglect so-called pseudo-knots, which are structure motifs with at least two base pairs (i, j), and (p, q), with $i < p < j < q$. Although pseudo-knots can be important structural elements in various functional RNAs [95], considering all possible pseudo-knots in structure prediction has been shown to be NP-hard [63], and is therefore computationally infeasible. Beginning with the work of Tabaska et al. [99], and Rivas and Eddy [86], several algorithms have been developed that reduce the computational complexity by limiting the predictions to certain pseudo-knots classes. However, even today pseudo-knot aware secondary structure prediction suffers from our poor knowledge of free energies for these special kinds of structure motifs. Recent approaches towards tertiary structure prediction that involve so-called *extended secondary structures* [54, 78, 130], and the incorporation of higher-order structure motifs, such as G-quadruplexes, into secondary structure prediction algorithms [57], will also be neglected in this brief overview focusing on pseudo-knot free secondary structure prediction approaches.

Each base pair (i, j) in a secondary structure closes a loop L, thereby directly enclosing unpaired nucleotides u and, possibly, further base pairs (p, q). Here, directly means that there is no other base pair (k, l) with $i < k < l < j$ such that $k < u < l$, or $k < p < q < l$. With these requirements, the number of directly enclosed unpaired nucleotides constitute the *length*, or *size* of L, while the number of directly enclosed base pairs and the enclosing pair determine its *degree*. Below, we refer to loops of degree 1 as hairpins, loops of degree 2 as interior loops, and loops with degree > 2 as multibranch loops.

Computational prediction of RNA secondary structures has been actively researched for more than four decades, and is mainly driven by physics based models [107, 106, 113, 76, 132]. The major assumption behind all of these physics based approaches is that a good estimate of the overall stability of an RNA secondary structure s can be obtained from the additive contributions of its individual loops

$$E(s) \approx \sum_{L \in s} E_L. \tag{1}$$

In this model, the energy contribution of a base pair in a helix depends on the identity of the two adjacent pairs, giving rise to the name *Nearest Neighbor Energy Model*. Great effort has been made to experimentally determine free energy parameters from melting experiments for different types of loops with a large variety of sequence compositions [106, 28, 67, 109]. Many small interior loops with a length of up to four or five, for instance, are exhaustively tabulated in the energy parameter sets of modern prediction programs, as are a handful of extraordinarily stable hairpins, such as tetra-loops. Contributions of larger loops, and those where no explicit experimental data is available are extrapolated. For multi loops, especially few melting experiments are available. For reasons of computational efficiency they are modeled by a simple linear combination of loop length and degree, although some attempt has been made to develop more sophisticated multi-loop energies that, e.g., take into account loop asymmetry [69]. Nevertheless, this can be regarded the weakest part of the nearest neighbor model.

In recent years several methods emerged, which replace physics-based models through trained parameters [22, 3, 4, 125]. Instead of relying on experimental measurements, these methods require large sets of RNAs with known structure as training data, making them susceptible to overfitting [87].

2.1. Free Energy Minimization

The number of possible secondary structures a particular RNA can adopt grows exponentially with its sequence length and it is thus generally unfeasible to enumerate all of them in order to assign stability scores and select best candidates. In this line, limiting the number of candidate structures to some criteria of optimality is an asset. A commonly used criterion is minimal free energy (MFE). After all, structures of minimum free energy are most stable among the entire ensemble of structure candidates and, according to thermodynamics, the most probable in thermodynamic equilibrium.

The first efficient dynamic programming (DP) algorithm to compute the minimum free energy (MFE) of an RNA, and a corresponding secondary structure was published in 1981 by Zuker and Stiegler [132], about a decade after the first attempts to predict secondary structures using experimentally determined loop energy contributions. For sequences of length n, the Zuker algorithm has an asymptotic time and memory complexity of $\mathcal{O}(n^3)$, and $\mathcal{O}(n^2)$, respectively.

Prediction Accuracy. Structure predictions are generally far from perfect. For moderately long RNAs, one can expect some 70% of predicted base pairs to be correct [67], a number that can fall as low as 40% for longer RNAs [23]. The reasons for limited accuracy are multifold, including effects such as simplifications in the energy model, inaccuracies of parameters, ignoring the effect of binding to ions (such as Mg^{2+}), proteins and other ligands, as well non-equilibrium states of the RNA. Fundamentally, the exponential growth of the number of possible structures means that even very small errors in the model can have strong effects. Many variants to the basic folding algorithms have therefore been developed chiefly to deal with limited accuracy.

Suboptimal Structures. A straightforward way to deal with uncertainty in structure prediction is to generate a set of of plausible structures instead of a single optimal one. The first approach to suboptimal structure was `mfold` [131], which produces structures that are optimal given that one base pair is enforced. This results in a small set of (hopefully) representative structures. A different, more exhaustive but computationally more expensive method was introduced in `RNAsubopt` [123], which enumerates all secondary structures within an energy band $[\mathrm{MFE}, \mathrm{MFE}+\delta]$.

2.2. The Thermodynamic Ensemble of Structures

The probability of a secondary structure s in equilibrium follows the laws of thermodynamics, specifically the Boltzmann distribution:

$$p(s) \propto e^{-E(s)/RT} \tag{2}$$

where $E(s)$ is the free energy of the structure, R the gas constant and T the thermodynamic temperature of the system. Given that the right-hand side of (2) can be easily computed for a particular structure s, it is straightforward to obtain the partition function Z by summing over all possible structures:

$$Z = \sum_s e^{-E(s)/RT} \tag{3}$$

The latter can then be used as the normalization factor for obtaining the equilibrium probability of a secondary structure s

$$p(s) = \frac{e^{-E(s)/RT}}{Z} \tag{4}$$

Equation 3 is impractical, since it requires summing over all possible structures. In 1990, McCaskill [70] realized that the problem can be solved by a variant of the DP recursions for MFE prediction. The essential point lies in using a unique decomposition of the secondary structure space, ensuring that no structure is counted twice. This paved the way to apply a broad variety of statistical methods from thermodynamics to RNA secondary structures, such as the computation of base pair probabilities [70] and statistical sampling of secondary structures according to their equilibrium probabilities [18]. However, the first practical implementation of the partition function and base pair probability computations that could be applied to RNAs of reasonable size became available in 1994 with the `RNAfold` program [41].

Accessibility. An example for an important statistical property that can be derived from the partition function is the *accessibility* of a region along the RNA, such as a binding motifs. In bacteria, for instance, translation is initiated upon binding of the ribosome to the Shine-Dalgarno sequence. There are several RNAs that control gene expression by differentially sequestering this motif through strong secondary structure formation, thus making it inaccessible for the ribosome. Furthermore, trans-acting RNAs such as miRNAs, sRNA, or siRNAs, but also proteins, and other ligands may bind specifically to single stranded regions to control the RNAs function As such, computing the accessibility of binding motifs is crucial for detection of potential RNA-RNA interaction targets and RNA-ligand binding.

Accessibility can be quantified as the probability that a region $i \dots j$ on an RNA is single stranded, or equivalently the free energy needed to force the region to be single stranded. First attempts to compute accessibilities where based on sampling [18], which however introduces sampling errors for longer regions. Mückstein et al. [75] introduced an exact (but still inefficient) computation of accessibilities via the partition function, while Bernhart et al. [9] showed that the accessibilities of *all* intervals of an RNA can be computed in $\mathcal{O}(n^3)$ time, i.e. the same complexity as simple MFE folding. This enables exact computation of accessibilities and opening free energies in short time for RNAs of reasonable length.

2.3. Reliability, Optimality, and Prediction Performance

An important approach to deal with uncertainty in prediction, is to provide reliability information that informs the user how trustworthy a prediction (or part of a prediction) is. Several such reliability measures can be conveniently derived from the partition function and base pair probabilities.

Ensemble Diversity. A simple yet powerful measure for the diversity of secondary structure ensembles is the average distance $\langle d \rangle$ of two structures drawn from the Boltzmann ensemble. The simplest distance measure is the *base pair distance* which counts the number of pairs present in one, but not both structures. Using the base pair distance, the average $\langle d \rangle$ can be expressed in terms of base pair probabilities p_{ij}.

$$\langle d \rangle = \sum_{s,t} p(s)p(t)d(s,t) = \sum_{i,j} p_{ij}(1 - p_{ij}) \qquad (5)$$

Likewise, the expected distance $\langle d(s) \rangle$ of a particular structure s to the entire ensemble can be computed

$$\langle d(s) \rangle = \sum_{i,j \in s} (1 - p_{ij}) + \sum_{i,j \notin s} p_{ij} \qquad (6)$$

Both measures provide reasonable information to which extent the ensemble is dominated by single structures, or whether there exist alternative low free energy structures. A scaling factor of $\frac{1}{n}$ can be used for both measures in order to compare RNAs of different lengths.

Positional Entropy. Reliability can also be measured locally for each nucleotide. The positional entropy $S(i)$ is a measure that captures whether a particular nucleotide i is found mainly in the same paired or unpaired configuration.

$$S(i) = -\sum_k p_{ik} \log_2 p_{ik} - p_i^u \log_2 p_i^u \qquad (7)$$

where $p_i^u = 1 - \sum)kp_{ik}$ is the probability that nucleotide i is unpaired. The positional entropy is 0 for a nucleotide that is always unpaired or always paired with the same partner. Thus, positions with low entropy are predicted with high confidence.

Ensemble Centroids. Even when only a single optimal structure is desired, the MFE is not the only choice available. In probabilistic terms, the MFE simply represents the most likely structure in the ensemble. However, other optimality criteria exist and could yield structure more representative of the ensemble. One idea for such a representative is the *centroid* structure s_c. Formally, the centroid is the structure that minimizes the weighted average distance to all other structures:

$$s_c = \underset{s}{\operatorname{argmin}} \langle d(s) \rangle = \sum_{t \in \Omega} p(t)d(s,t) \qquad (8)$$

The construction of s_c becomes trivial when the distance between structures is measured in terms of the number of base pairs both structures differ in. In this case, s_c simply consists of all base pairs with $p_{ij} > 0.5$. Note, that for very diverse ensembles, it may well be that no pair has probability > 0.5 and thus the centroid structure contains no base pairs. For such diverse ensembles, a possibility is to first subdivide Ω into clusters and compute a centroid for each cluster separately [16]. The latter approach, however, relies again on sampling.

Maximizing the Expected Accuracy. Another type of optimal representative is the so called maximum expected accuracy (MEA) structure. Suppose, we define the accuracy of a structure as the number of correct base pairs. The *expected* accuracy is then $\text{EA}(s) = \sum_{(i,j) \in s} p_{ij}$, and the structure maximizing the expected accuracy is

$$s_{MEA} = \underset{s}{\operatorname{argmax}}\ \text{EA}(s) \qquad (9)$$

In order to avoid overpredicting base pairs, a more general from for the expected accuracy is commonly used:

$$\text{EA}(s) = \sum_{(i,j) \in s} 2\gamma p_{ij} + \sum_{i \nexists (i,j) \in s} q_i. \qquad (10)$$

Here, $q_i = 1 - \sum_j p_{ij}$ is the probability that nucleotide i is unpaired, and γ is a weighting factor that balances between paired and unpaired positions. A simplified version of the MFE prediction DP algorithm [22] can be applied to efficiently solve 9.

As an example to emphasize the variety in selection of secondary structure representatives for different prediction methods, we depict their results for the 57nt long spliced leader RNA of *Leptomonas collosoma* [53] in Figure 1.

Global and Local Secondary Structures. Discovery of novel functional RNAs, and putative targets for RNA-RNA interactions based on motif accessibilities require fast and efficient algorithms for genome-wide applications. For such purposes, variations of the MFE and partition function algorithm can be applied, that limit the maximum base pair span along the backbone of the RNA to a certain number L. Consequently, the asymptotic time and memory complexity for MFE prediction and partition function

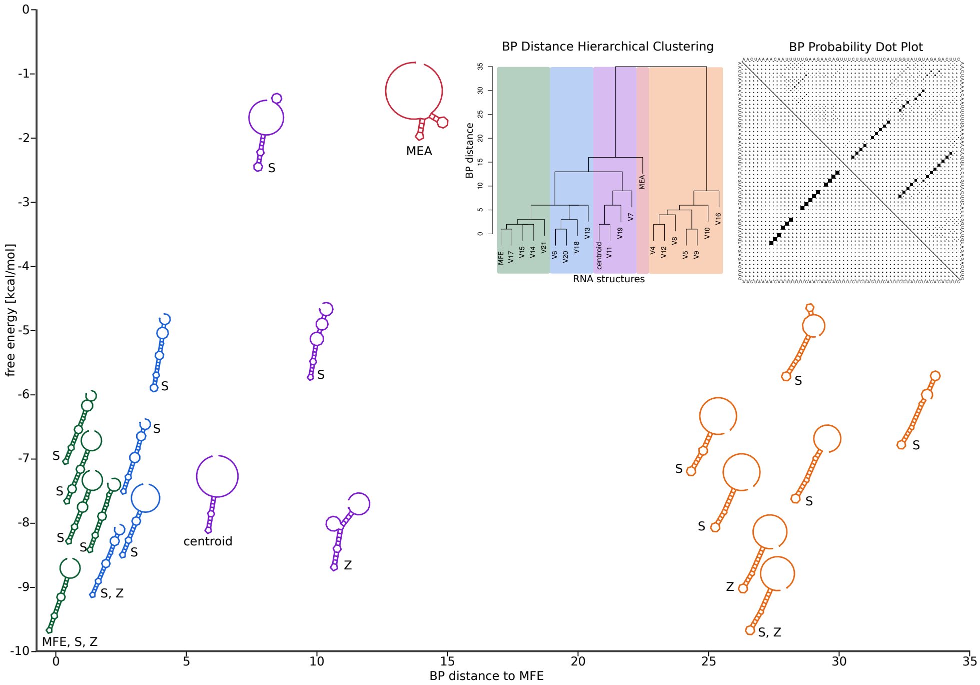


Figure 1: Secondary structure predictions for the spliced leader RNA from *Leptomonas collosoma* [53]. Displayed are secondary structures predicted by various methods, such as MFE, ensemble centroid, MEA structure, as well as suboptimal structures obtained from stochastic backtracking (marked by S), and the 5 best suboptimals sensu Zuker (marked by Z), all implemented in the programs `RNAfold`, and `RNAsubopt` of the `ViennaRNA Package` [41, 58]. To account for their respective pairwise base pair distance, a hierarchical cluster tree is shown. Furthermore, equilibrium base pair probabilities are shown in the upper triangle of the respective dot-plot. The lower triangle displays the base pairs of the MFE structure.

computation becomes $\mathcal{O}(n \cdot L^2)$ [42, 7]. As such, this approach is applicable to genome-wide surveys not only for bacteria, but even for chromosome lengths found in the human genome.

2.4. Consensus Structures

An entirely different approach towards better structure predictions, especially of MFE structures, is to consider phylogenetic information whenever possible. In the most common approach one starts from an alignment of several homologous sequences and asks for a consensus structure, i.e. the optimal structure that can be adopted by all sequences. This approach is followed for example by the `Pfold` [49, 97] and `RNAalifold` [39, 8] programs. A slightly different strategy is adopted by `Turbofold` [35], which predicts a structure for each individual input sequence guided by the pair probabilities of all other sequences in the set. The predictions can be refined iteratively by repeating this cycle several times.

The above methods rely on sequence alignments, whose quality in turn limits the accuracy of consensus structure prediction. As proposed already in 1985 by Sankoff [89], the most principled approach would be to determine the optimal alignment and structure simultaneously. Unfortunately, the Sankoff algorithm is computationally very expensive with a time complexity of $\mathcal{O}(n^6)$ already for two sequences. Nevertheless, a number of practically useful implementations exist today, all of which use heuristics to restrict the search space and thus reduce time and/or space complexity. One of the earliest approaches, `Dynalign` [68], limits the difference in length of aligned subsequences to a maximum value M resulting in a $\mathcal{O}(M^3N^3)$ algorithm. The `pmcomp` and `pmmulti` tools [38] restrict structure search space by ignoring low probability base pairs resulting in a run time of $\mathcal{O}(n^4)$, an idea that was extended in `LocARNA` [119] to also reduce the memory consumption from quartic to quadratic. Finally, `RAF` [21] restricts both structure and alignment search space to achieve quadratic run time.

2.5. RNA-RNA interactions

A natural extension of folding algorithms is to consider the interaction between two or more RNAs. This is of particular interest, since most regulatory RNAs work through interaction with an RNA target. The simplest (and fastest) methods, such as `RNAhybrid` [84], `RNAduplex`, or `RISearch` [116] only consider the hybridization energy between two RNAs. This, however, neglects that intermolecular structure always competes with intra-molecular structure formation. A straightforward way to view RNA-RNA hybridization is to assume a two step process, first intra-molecular structure has to opened in order generate single stranded regions that can then hybridize. The total free energy change is thus given by the sum of a (positive) opening energy and a (negative) hybridization energy. Opening energies can be computed in accessibility predictions, as described above. The approach is used e.g.

in RNAup [75] and IntARNA [12], a fast approximate version is available in RNAplex [101, 100].

The above methods limit the search to a single interacting region. In contrast, RNAcofold [41, 10], proceeds by artificially linking the two interacting RNAs, and running a standard folding algorithm, modified to correctly treat the loop containing the linker element. In this case, the search space is limited to structures that are pseudoknot free after linking, thus excluding the important case of kissing hairpins.

Some restriction of the search space is indeed necessary, since the general RNA-RNA interaction problem is NP hard [1]. A number of works have tried to allow a broader set of interaction structures [80, 1, 44, 13]. Because of the high computational cost of $\mathcal{O}(n^3m^3)$ these methods are less frequently used. A generalization to arbitrary numbers of interacting nucleic acid sequences was proposed by Dirks et al. [20], and is implemented as part of the NUPACK suite [124].

2.6. Kinetic folding of RNA

While thermodynamic modeling allows for detailed investigation of equilibrium properties of RNA, many biological processes are governed by non-equilibrium processes, e.g. long-lived folding intermediates resulting from stable helices. Unfortunately, the number of methods that explicitly model folding dynamics is still limited, moreover these methods are generally much more computationally demanding than the DP algorithms for equilibrium folding. In the following we summarize some of the available approaches. For a more comprehensive review see, e.g., [26].

Biopolymer folding can be viewed as walk on an energy landscape. Formally, such a landscape consists of a finite state space of structures X, an function $E(x)$ that assigns an energy to each state $x \in X$, and a move set that describes which states are connected by elementary transitions. Each possible move $x \to y$ is associated with a rate k_{yx}. For RNA, the simplest move set consists of opening or closing of a single base pair.

Folding dynamics can then be modeled by a continuous-time Markov process based on a master equation which describes the change in state probabilities $P_t(x)$ to see state x at time t

$$\frac{dP_t(x)}{dt} = \sum_{y \neq x} [P_t(y)k_{xy} - P_t(x)k_{yx}] \qquad (11)$$

Solving eq. 11 directly is impractical for anything except toy examples, since the dimension of the rate matrix is equal to the number of possible structures. One possibility to address this issues is to perform stochastic simulations of RNA folding using a Monte Carlo method. This approach is taken e.g. by Kinfold [25] and KineFold [46]. While the outcome of this method can be regarded as a gold standard, computing and analyzing a large number of trajectories can be time consuming and tedious. An alternative approach is available through direct investigation of the energy landscape in terms of local minima, energy barriers and transition rates, as done by the barriers program [27]. The local minima can be used as the basis for a coarse graining, reducing the number of states to a few hundred or thousand, thus allowing for direct numerical integration of eq. 11 [120].

Since barriers relies on an exhaustive enumeration of low energy structures, it is in practice applicable only to RNAs of less than 100nt. Recently, a number of heuristic approaches have been reported that attempt to raise this limit based on flooding techniques [121, 65] or sampling of local minima [103, 104, 50].

Another important, yet often neglected, aspect is the fact that RNA structure is formed already during its synthesis, i.e. it folds back on itself co-transcriptionally. Co-transcriptional folding is fairly easy to implement in simulation approaches [32, 25, 46]. In the landscape view, co-transcriptional folding induces a landscape that varies over time. A framework to deal with such scenarios has been presented in [40]. Finally, methods such as Kinwalker [30] attempt to construct a single, most likely, folding trajectory for the growing RNA chain. While this introduces fairly drastic approximations it can be applied to RNAs up to ≈ 1500nt length.

3. Guiding RNA Secondary Structure Prediction with Experimental Data

3.1. Experimental approaches

Experimental technologies to elucidate RNA structure by means of chemical and enzymatic probing were established long before the first computational approaches toward RNA structure prediction have become available [96]. Ribonucleases (RNAse) are highly specific at recognizing single-stranded (ss) or double stranded (ds) RNA regions and modify them by adding functional groups or by cleaving them at their recognition sites. Treated RNAs are then analyzed on sequencing gels in order to characterize sites of modification or cleavage.

While the first chemical probing workflows based on 1-cyclohexyl-3'-(2-morpholinoethyl) carbodiimide (CMCT) [73, 79] and lead(II) probing [31, 56] have been available for decades, more recent approaches including protocols based on hydroxyl radicals [108], inline probing [93, 83], kethoxal[11], dimethyl sulfate (DMS) [127, 115, 14] and selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) [72, 118] have become available.

Chemical probes bind single-stranded nucleotides and hence allow for fine-grained experimental elucidation of RNA secondary and tertiary structure. Several chemical probing chemistries, each targeting distinct regions of nucleotides in a specific manner have been described [114]. SHAPE reagents, for example, query the backbone by acylating the ribose 2'-hydroxyl group of flexible nucleotides, thereby forming 2'-O-ester adducts which cause subsequent reverse transcription to terminate at the site of modification. Single stranded or conformationally unconstrained RNA regions show high 2'-hydroxyl reactivity, thus designating them as a primary targets to measure the dynamics of local RNA structure.

3.2. High-throughput RNA structure probing

With the advent of novel genome-wide sequencing technologies and the availability of whole transcriptome data for various model and non-model organisms came demand for reliable, large-scale RNA structure prediction methods. It only took the scientific community a couple of years to come up with first *in vitro* approaches for high-throughput transcriptome-wide RNA probing, where next-generation sequencing (NGS) technologies are employed for readout instead of gel or capillary electrophoresis. Parallel analysis of RNA structure (PARS) [47], parallel analysis of RNA structures with temperature elevation (PARTE) [111], fragmentation sequencing (Frag-seq) [110] and ss/dsRNA-seq [129] form a class of experimental approaches that combine RNAse treatment with NGS. The chemical inference of RNA followed by massive sequencing (CIRS-seq) [45], multiplexed accessibility probing-sequencing (MAP-seq) [90] and chemical modification-sequencing (ChemMod-seq) [37] methods employ CMCT and DMS probing, whereas hydroxyl radicals are used within the hydroxyl radical footprinting-sequencing (HRF-seq) method [48] in the context of RNA tertiary structure analysis. Combination of SHAPE chemical probing with NGS (SHAPE-seq) [62, 6, 60] provides highly reproducible reactivity data over a wide rage of RNA structural contexts without apparent biases.

Large-scale *de novo* identification of RNA functional motifs has recently become accessible through the SHAPE-MaP approach [91, 92], where chemically modified sites are quantified in a single direct step by modification in the RNA backbone. The method makes use of the fact that noncomplementary nucleotides are included into the newly synthesized cDNA during reverse transcription, thus documenting qualitative and quantitative information of SHAPE adducts in a SHAPE-MaP. Similarly, the RNA interacting groups mutational profiling method (RING-MaP)[43] employs DMS treatment followed by special buffer conditions that allow read-through at positions of DMS modification in combination with incorporation of non-complementary nucleotides.

3.3. RNA structure probing in vivo

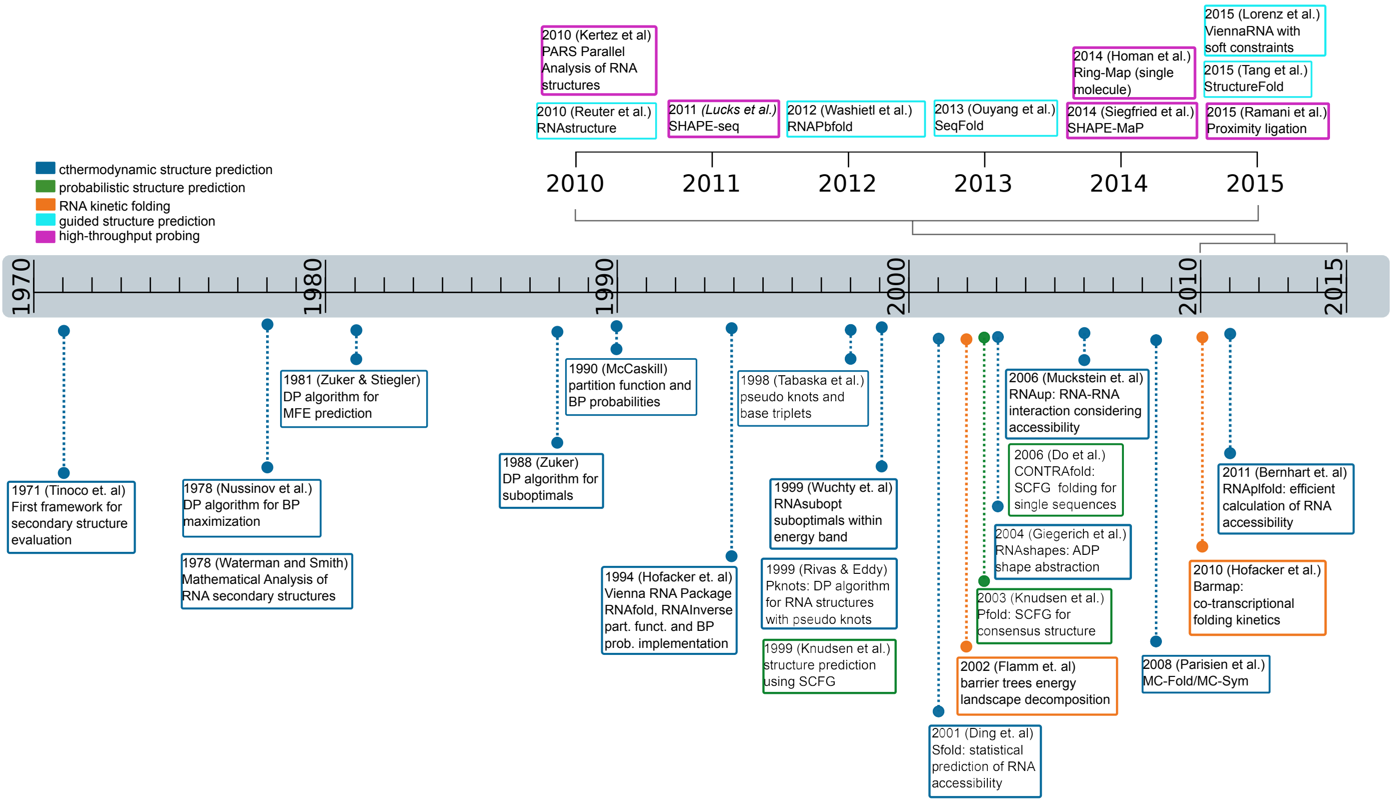
While *in vitro* RNA probing approaches have improved our understanding of the complex inference among RNA structure and function, *in vivo* probing allows to interrogate RNA structure in a native environment under the influence of various cellular processes such as transcription, splicing, binding of small molecules and proteins [117]. Although classical DMS probing *in vivo* has been available for several years [115, 55, 128], high-throughput variants have been reported recently, including Structure-seq [19, 17], DMS-seq [88] and Mod-seq [102]. A detailed comparison of these methods, along with computational procedures for data analysis is available in [52]. *In vivo* SHAPE probing has been reported for abundant [94, 71] and low-abundant [51] transcripts.

All methods mentioned so far allow researchers to determine to what extent specific nucleotides are paired, however they do not reveal pairing partners. To address this problem, a novel method for resolving RNA structure by proximity ligation has recently been described [82]. Here, pairs of interacting RNA regions are ligated after initial RNAse digestion, thus forming chimeric molecules of RNA sequences that were initially forming secondary structure. Subsequent high-throughput sequencing and quantification of the relative abundance of specific intramolecular ligation junctions provides a decent picture of short- and long-range interactions of RNA secondary structure.

3.4. Combining Experimental Data with RNA Secondary Structure Prediction

As many RNA structure probing methods became a mainstream technology, the demand for efficient and precise methods to combine them with computational methods in RNA structure determination is evident. Chemical probing, such as SHAPE, or DMS, usually yields per-nucleotide reactivities that, to some extent, reflect the structural context of a nucleotide. These reactivities are then used to either guide *in silico* RNA structure prediction methods directly, or determine which representatives fit the experiments best. Today, several approaches to incorporate chemical probing data into thermodynamics-based computational tools have been suggested [29]. Available programs that allow for probing data guided structure prediction include `Fold` of the `RNAstructure` package [85], the `MC-Fold / MC-Sym` [78] pipeline, `RNAsc` [126], `RNApbfold` [112], `SeqFold` [77], and `StructureFold` [105]. A historic overview of RNA structure prediction methods with, and without the possibility to incorporate probing data is shown in Figure 2. Below, we will review current concepts of probing data guided structure prediction.

RNA folding with hard and soft constraints. Historically, the first attempts to guide RNA structure prediction based on prior knowledge, such as experimental probing data or covariation within homologous sequences, were based on so-called *hard constraints.* These constraints restrict the folding space on the level of the generating function, e.g. through exclusion or enforcement of specific base pairs [132, 41, 66]. However, experimental data usually comes with some amount of uncertainty, that easily translate into errors in such binary restraints. Unfortunately, for hard constraints, even small errors in the input easily lead to entirely wrong predictions. To overcome issues with ambiguous data, more elaborate approaches use *soft constraints* that instead target the energy evaluation of



cthermodynamic structure prediction
probabilistic structure prediction
RNA kinetic folding
guided structure prediction
high-throughput probing
2010 (Kertez et al) PARS Parallel Analysis of RNA structures
2010 (Reuter et al.) RNAstructure
2011 (Lucks et al.) SHAPE-seq
2012 (Washietl et al.) RNAPbfold
2013 (Ouyang et al.) SeqFold
2014 (Homan et al.) Ring-Map (single molecule)
2014 (Siegfried et al.) SHAPE-MaP
2015 (Lorenz et al.) ViennaRNA with soft constraints
2015 (Tang et al.) StructureFold
2015 (Ramani et al.) Proximity ligation
2010
2011
2012
2013
2014
2015
1970
1980
1990
2000
2010
2015
1971 (Tinoco et. al) First framework for secondary structure evaluation
1978 (Nussinov et al.) DP algorithm for BP maximization
1978 (Waterman and Smith) Mathematical Analysis of RNA secondary structures
1981 (Zuker & Stiegler) DP algorithm for MFE prediction
1988 (Zuker) DP algorithm for suboptimals
1990 (McCaskill) partition function and BP probabilities
1994 (Hofacker et. al) Vienna RNA Package RNAfold, RNAInverse part. funct. and BP prob. implementation
1998 (Tabaska et al.) pseudo knots and base triplets
1999 (Wuchty et. al) RNAsubopt suboptimals within energy band
1999 (Rivas & Eddy) Pknots: DP algorithm for RNA structures with pseudo knots
1999 (Knudsen et al.) structure prediction using SCFG
2001 (Ding et. al) Sfold: statistical prediction of RNA accessibility
2002 (Flamm et. al) barrier trees energy landscape decomposition
2003 (Knudsen et al.) Pfold: SCFG for consensus structure
2004 (Giegerich et al.) RNAshapes: ADP shape abstraction
2006 (Do et al.) CONTRAfold: SCFG folding for single sequences
2006 (Muckstein et. al) RNAup: RNA-RNA interaction considering accessibility
2008 (Parisien et al.) MC-Fold/MC-Sym
2010 (Hofacker et al.) Barmap: co-transcriptional folding kinetics
2011 (Bernhart et. al) RNAplfold: efficient calculation of RNA accessibility

Figure 2: History of RNA structure prediction methods. Upper panel: guided structure prediction using high-throughput probing data. Lower panel: major algorithms and implementations of the past decades. For details about individual approaches see text.

loop motifs through additional pseudo free energy terms [67, 39, 8, 36]. The transformation of chemical footprinting data into soft constraints for secondary structure prediction has only been developed recently, mainly driven by the advances in (high-throughput) SHAPE experiments.

Directly derived Pseudo Free Energies. In 2009, Deigan et al. [15] were the first to pick up the observation that SHAPE reactivities are roughly inversely proportional to the probability that a nucleotide forms a canonical base pair [72]. Therefore, their linear ansatz directly converts the SHAPE reactivity $r_S(i)$ of each nucleotide i into a pseudo free energy term

$$\Delta\ G_S(i) = m \log(r_S(i) + 1) + b. \qquad (12)$$

In the prediction algorithm, these contributions are then applied to each of the four nucleotides involved in a base pair stack. The idea is to penalize the formation of stacked pairs whenever high reactivity values from the experiment suggest that these nucleotides should be unpaired. The intercept b, and the slope m are then carefully adjusted in such a way, that low reactivity valued positions receive little penalty, while those with high reactivity values are penalized a lot. Latest parameters for this pseudo free energy term are $m = 1.8$, and $b = -0.6$ [34], but several parametrizations that sometimes differ substantially can be found in recent literature, e.g. in [81, 61]. Because consecutive base pair stacks in a helix are evaluated for two adjacent pairs at a time, pairing nucleotides inside a helix are penalized twice, compared to those at the ends of a helix. Energy evaluations of any of the remaining loop motifs remain unchanged, even if the motif is in disagreement with experimental data.

Probing Data and Pairing Probabilities. While the above ad-hoc conversion of SHAPE reactivities into pseudo free energies has no direct physical justification, later approaches to incorporate SHAPE reactivities first convert them into likelihoods to be paired, or unpaired. Subsequently, the corresponding pseudo energies are computed from these probabilities. Thus, the actual probing data is detached from the pseudo energy conversions, and different methods of probability estimation from probing data may be applied. Consequently, this ansatz is applicable to other probing methods, such as DMS, or PARS, as well.

However, the conversion of probing data into probabilities to be paired, or unpaired is not trivial since SHAPE reactivities, for instance, do not distinguish paired from unpaired positions unambiguously. In fact, the distributions of reactivities for unpaired and paired positions have a rather large overlap [98]. To account for this ambiguity, Eddy [24] suggested to use conditional probabilities $P(r_S(i)|\pi_i)$ to observe a reactivity $r_S(i)$ given nucleotide i is in a particular context π_i. These posterior probabilities can be obtained from prior models for the reactivity distributions of the respective contexts π, that have been fitted to a training set of structure / reactivity data pairs. In turn, this conditional probability may then be readily converted into a pseudo energy

$$\Delta G_{\mathcal{S}}(\pi_i, i) = RT \log P(r_{\mathcal{S}}(i)|\pi_i) \qquad (13)$$

and applied to each derivation where i is added to a growing substructure. Note, that $\Delta G_{\mathcal{S}}(i)$ yields a penalty that grows with smaller posterior probabilities of the probing reactivities, whereas it diminishes for posteriors close to 1.

On the other hand, according to Bayesian statistics, the likelihood of nucleotide i being in a particular structural context π_i, given the experimentally determined reactivity value $r_{\mathcal{S}}(i)$ is

$$p(\pi_i|r_{\mathcal{S}}(i)) = \frac{P(r_{\mathcal{S}}(i)|\pi_i) \cdot p(\pi_i)}{p(r_{\mathcal{S}}(i))}. \qquad (14)$$

Unfortunately, this conversion requires additional parameters that need to be fitted from training data. However, the probability $p(\pi)$ to observe a nucleotide in context π, and the probability $p(r_{\mathcal{S}}(i))$ to observe a reactivity value $r_{\mathcal{S}}(i)$ can be determined from the training set for the prior distributions.

Among the first using a probabilistic, yet still ad-hoc, strategy is the method suggested by Zarringhalam et al. [126]. Here, the authors use a non-linear piece-wise mapping technique to convert $r_{\mathcal{S}}(i)$ into probabilities to be unpaired q_i. They proceed to predict a structure s with minimal distance to the probing data, where the distance for each nucleotide i is defined as $|\pi_i - q_i|$, with $\pi_i = 0$ if i is paired, and $\pi_i = 1$ if it is unpaired in s. This directly leads to a pseudo energy term of

$$\Delta\ G_{\mathcal{S}}(\pi_i, i) = \beta\ |\pi_i - q_i|. \qquad (15)$$

where β serves as a scaling factor to adjust the magnitude of penalty for disagreement between prediction and probing data. Given that SHAPE experiments are not free of errors, one must not put too much weight on the experimental data, since errors in probing data directly lead to errors in secondary structure prediction. Moreover, it is arguable, whether their distance measure is well chosen. After all, empirical data on SHAPE reactivities shows, that both, paired and unpaired nucleotides, are more likely to have low reactivities [98]. Thus one can not directly infer a small likelihood to be unpaired from low reactivity.

A more recent implementation that takes up the idea of Eddy [24] to incorporate SHAPE, DMS, and PARS data was introduced with the RME program [122]. Following a Bayesian approach to determine posterior pairing probabilities $p(\pi_i|r_{\mathcal{S}}(i))$ the authors first fit prior distributions for the respective probing methods from known data. From that, they compute conditional probabilities $\hat{p}(i) = p(\pi_i = 0|r_{\mathcal{S}}(i))$ to observe a nucleotide i being paired. Only then, the resulting probabilities are converted into pseudo energies

$$\Delta G_{\mathcal{S}}(i) = -RT \cdot m \cdot \log \frac{\hat{p}(i)}{1 - \hat{p}(i)} \qquad (16)$$

to guide a partition function computation, where m serves as a scaling factor. Although their soft constraint makes use of the likelihood $\hat{p}(i)$ of a nucleotide to be paired, they apply the pseudo energy term only to nucleotides involved in base pair stacks, analogously to the method of Deigan et al. [15]. Consequently, nucleotides at the end of helices receive only half of the pseudo energy correction compared to those within a helix. In a post-processing step, the base pair probabilities predicted under this model are corrected by their deviation from the probing data, and finally used for the construction of a MEA structure.

An entirely different approach was proposed by Washietl et al. [112]. Instead of converting the probing data into a pseudo energy term, the authors draw an optimization problem that aims to find a perturbation vector $\vec{\epsilon}$ that (i) minimizes the changes to the nearest neighbor free energy model required, while (ii) at the same time maximizing the agreement between predicted probabilities and observed data. For that purpose, they convert shape reactivities into probabilities to be unpaired q_i using a thresholding approach. An appropriate perturbation vector thus satisfies

$$F(\vec{\epsilon}) = \sum_{\mu} \frac{\epsilon_\mu^2}{\tau^2} + \sum_{i=1}^{n} \frac{(p_i(\vec{\epsilon}) - q_i)^2}{\sigma^2} \rightarrow \min,$$

where ϵ_μ is the perturbation energy for structural element μ, $p_i(\vec{\epsilon})$ is the predicted probability to be unpaired given $\vec{\epsilon}$, and the variances τ_μ^2 and σ_i^2 serve as weighting factors to account for the trade-off between the relative uncertainties inherent in the experimental measurements and the energy model. Ideally, $\vec{\epsilon}$ shows close-to-zero values for positions with good agreement between model and experiment. Thus, it may directly reveal sequence positions that require adjustment, indicating potential post-transcriptional modification site, or intermolecular interactions that are not explicitly handled by the nearest neighbor model. In contrast to the other methods discussed above, the impact of $\vec{\epsilon}$ on the diversity of the structure ensemble can be usually considered small, making it applicable to RNAs with several distinct low energy structures. Though, this method has a relatively high asymptotic time complexity of $\mathcal{O}(n^4)$, uses an ad-hoc probing data conversion, and strongly depends on the optimization technique to alleviate exploration of the rugged landscape of solutions. Our implementation of this method in the ViennaRNA Package alleviates these drawbacks by estimating $p_i(\vec{\epsilon})$ from structure samples, and allows for a variety of optimization techniques and conversions from probing data into probabilities (see supplementary material, Section 5 of [59]).

Limitations and Future Perspectives. In theory, one would expect that perfect one-dimensional probing data, i.e. data that binarily distinguishes between paired and unpaired, yields almost perfect structure predictions. To test this hypothesis, we collected a dataset of about 1,900 pseudoknot free sequence/structure pairs from the RNAstrand database [2]. For each reference structure, we constructed

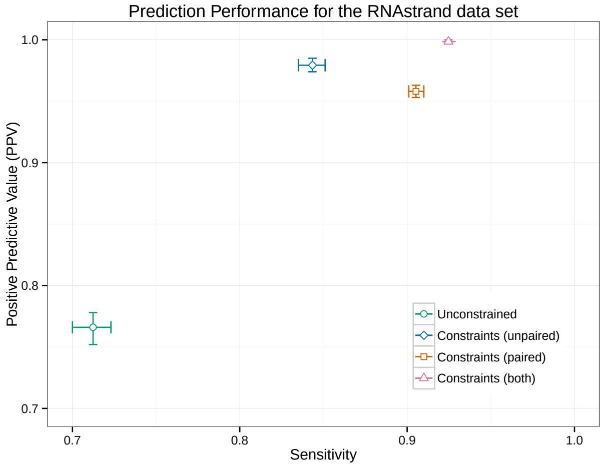


Figure 3: Average prediction performance of `RNAfold` using perfect hard constraints. The benchmark set for the prediction consists of about $1,900$ pseudo-knot free sequence/structure pairs taken from RNA Strand database [2]. The 95% confidence intervals for PPV, and sensitivity were estimated using bootstrapping with 1000 iterations. As visible in the plot, perfect (one-dimensional) hard constraints extracted from reference structures substantially increases the prediction performance, even if only unpaired positions in the reference are constrained during the prediction. Surprisingly, constraining just paired positions in the reference yields slightly lower PPV, while the sensitivity increases as expected. Application of constraints to both, paired and unpaired positions yields almost perfect predictions.

three sets of (one-dimensional) hard structure constraints to (i) prohibit unpaired positions in the reference from being paired, (ii) enforce paired positions in the reference to be paired, and (iii) a combination of both, respectively. We then applied the resulting constraints to MFE structure predictions using `RNAfold`, and assessed the prediction performance by means of Positive Predictive Value (PPV), and Sensitivity. For the preparation of the hard constraints, we removed all non-canonical base pairs, and hairpin loops with a size $u < 3$, since `RNAfold`, as most other secondary structure prediction programs, can not predict such motifs. However, the corresponding base pairs remained in the reference structures for the assessment of prediction performance.

As visible in the benchmark results shown in Figure 3, perfect hard constraints are capable to yield almost perfect prediction performance. Though, sensitivity does not exceed 0.925 due to unusually small hairpin loops, and a variety of non-canonical base pairs in the reference data set. It should be noted, however, that in contrast to soft constraints, i.e. pseudo energy contributions as used in latest probing data guided structure predictions, hard constraints are not robust. Even the slightest error in hard constraints might yield an entirely wrong prediction, while this effect is much less pronounced when using soft constraints. This property has to be kept in mind, whenever constraints are used for guided secondary structure prediction, since the concept of secondary structures does not account for tertiary effects such as non-canonical base pairs, extremely short hairpin loops, or very long interior loops, that are implicitly included in experimental probing data.

Still, all of the above methods assume that the probing data can essentially distinguish between paired, and unpaired nucleotides. However, SHAPE reactivities, for instance, have been shown to display distinct distributions for at least three different states, namely paired inside a stack, paired at the end of helices, and unpaired [98]. Unfortunately, none of the existing approaches takes these findings into account. Therefore, it remains unclear, whether more elaborate methods that distinguish more than two pairing states help to increase prediction performances. We have recently implemented a generic approach for guided structure prediction by means of hard and/or soft constraints into several programs of the `ViennaRNA Package` [59]. This allows, for instance, an easy application of SHAPE data using the methods of Deigan et al. [15], Zarringhalam et al. [126], and Washietl et al. [112] with the programs `RNAfold`, `RNAsubopt`, and `RNAalifold`.

4. Discussion

Computational methods for RNA structure prediction have evolved rapidly over the past decades, primarily due to fundamental improvements of the underlying algorithms. At the same time, advances in structure probing technologies allowed for high-throughput screening of the RNA 'structure-ome' both *in vivo* and *in vitro*. In recent years, these two approaches have been combined to further increase the accuracy of both 2D and 3D structure predictions. In this paper, we reviewed the concepts of computational RNA structure prediction and discuss current challenges focusing on integration of experimentally derived footprinting information.

However, integration of chemical probing data does not necessarily yield better predictions. In fact, we observed in a recent benchmark that incorporating SHAPE data into MFE prediction does in some cases lead to decreased accuracy of the resulting secondary structures, as shown for group II intron and the Lysine riboswitch in Lorenz et al. [59], supplementary material.

Like other experimental approaches, chemical probing is inherently noisy and reproducibility still remains an issue. Furthermore, the experimental condition such as pH, ionic strength or consentrations of co-factors might differ from those under which the reference structure was derived. As a consequence, the experiment might probe a structure different from the reference. Also, the concept of reference structures silently assumes that a given sequence folds into exactly one structure, even though alternative low free energy states may exist. It is currently unclear how best to deal with cases where the RNA forms an ensemble of diverse structures. Quite possibly, probing data will be less useful in such cases: Even an equilibrium of just two structures could in the worst case result in pairing probabilities of exactly 50% for *every* nucleotide, thus yielding a comletely uninformative probing signal. One the up side, it is likely that current methods do not yet make best possible use of probing data, since they assume a binary distinction between paired and unpaired positions. Clearly, probing reactivity will depend on more structural details and should therefore give information on more classes of structural context. The observed distribution of reactivities in SHAPE experiments suggests that at least a ternary distinction between unpaired, helix-end, and stacked nucleotides might be advantageous [98].

Acknowledgements

This work was partly funded by the Austrian science fund FWF project F43 "RNA regulation of the transcriptome", and the Austrian / French project "RNALands", FWF-I-1804-N28 and ANR-14-CE34-0011. Dedicated to Prof. Bernhard Keppler on the occasion of his 60th birthday.

[1] Alkan, C., Karakoc, E., Nadeau, J. H., Sahinalp, S. C., Zhang, K., 2006. RNA-RNA interaction prediction and antisense RNA target search. Journal of Computational Biology 13 (2), 267–282.

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

[3] Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H., Murphy, K. P., Jul 2007. Efficient parameter estimation for RNA secondary structure prediction. Bioinformatics 23 (13), i19–28.

[4] Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H., Murphy, K. P., Dec 2010. Computational approaches for RNA energy parameter estimation. RNA 16 (12), 2304–18.

[5] Antic, S., Wolfinger, M. T., Skucha, A., Hosiner, S., Dorner, S., 2015. General and miRNA-mediated mRNA degradation occurs on ribosome complexes in Drosophila cells. Mol. Cell Biol., MCB–01346.

[6] Aviran, S., Trapnell, C., Lucks, J., Mortimer, S., Luo, S., Schroth, G., Doudna, J., Arkin, A., Pachter, L., 2011. Modeling and automation of sequencing-based characterization of RNA structure. PNAS 108 (27), 11069–11074.

[7] Bernhart, S. H., Hofacker, I. L., Stadler, P. F., 2006. Local RNA base pairing probabilities in large sequences. Bioinformatics 22 (5), 614–615.

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

[9] Bernhart, S. H., Mückstein, U., Hofacker, I. L., 2011. RNA accessibility in cubic time. Algorithms for Molecular Biology 6 (1).

[10] Bernhart, S. H., Tafer, H., Mückstein, U., Flamm, C., Stadler, P. F., Hofacker, I. L., 2006. Partition function and base pairing probabilities of RNA heterodimers. Algorithms for Molecular Biology 1 (1), 3.

[11] Brow, D. A., Noller, H. F., 1983. Protection of ribosomal RNA from kethoxal in polyribosomes: Implication of specific sites in ribosome function. J Mol Biol 163 (1), 27–46.

[12] Busch, A., Richter, A. S., Backofen, R., 2008. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. Bioinformatics 24 (24), 2849–2856.

[13] Chitsaz, H., Salari, R., Sahinalp, S. C., Backofen, R., 2009. A partition function algorithm for interacting nucleic acid strands. Bioinformatics 25 (12), i365–i373.

[14] Cordero, P., Kladwang, W., Vanlang, C., Das, R., 2012. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. Biochemistry 51 (36), 7037–7039.

[15] Deigan, K. E., Li, T. W., Mathews, D. H., Weeks, K. M., 2009. Accurate SHAPE-directed RNA structure determination. PNAS 106, 97–102.

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

[17] Ding, Y., Kwok, C. K., Tang, Y., Bevilacqua, P. C., Assmann, S. M., 2015. Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. Nat Protoc 10, 1050–1066.

[18] Ding, Y., Lawrence, C. E., 2001. Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. Nucleic Acids Res 29 (5), 1034–1046.

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

[20] Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E., Pierce, N. A., 2007. Thermodynamic analysis of interacting nucleic acid strands. SIAM review 49 (1), 65–88.

[21] Do, C. B., Foo, C.-S., Batzoglou, S., 2008. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. Bioinformatics 24 (13), i68–76.

[22] Do, C. B., Woods, D. A., Batzoglou, S., 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics 22 (14), e90–e98.

[23] Doshi, K. J., Cannone, J. J., Cobaugh, C. W., Gutell, R. R., 2004. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinformatics 5 (1), 105.

[24] Eddy, S. R., 2014. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. Annu Rev Biophys 43, 433–456.

[25] Flamm, C., Fontana, W., Hofacker, I. L., Schuster, P., 2000. RNA folding at elementary step resolution. RNA 6 (03), 325–338.

[26] Flamm, C., Hofacker, I. L., 2008. Beyond energy minimization: Approaches to the kinetic folding of RNA. Monatsh. f. Chemie 139 (4), 447–457.

[27] Flamm, C., Hofacker, I. L., Stadler, P. F., Wolfinger, M. T., 2002. Barrier trees of degenerate landscapes. Z Phys Chem 216, 155–173.

[28] Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T., Turner, D. H., 1986. Improved free-energy parameters for predictions of RNA duplex stability. PNAS 83 (24), 9373–9377.

[29] Ge, P., Zhang, S., 2015. Computational analysis of RNA structures with chemical probing data. Methods 79, 60–66.

[30] Geis, M., Flamm, C., Wolfinger, M. T., Tanzer, A., Hofacker, I. L., Middendorf, M., Mandl, C., Stadler, P. F., Thurner, C., 2008. Folding kinetics of large RNAs. J Mol Biol 379 (1), 160–173.

[31] Gornicki, P., Baudin, F., Romby, P., Wiewiorowski, M., Kryzosiak, W., Ebel, J. P., Ehresmann, C., Ehresmann, B., 1989. Use of lead(II) to probe the structure of large RNAs. conformation of the 3' terminal domain of e. coli 16s rRNA and its involvement in building the tRNA binding sites. J Biomol Struct Dyn 6 (5), 971–984.

[32] Gultyaev, A. P., Van Batenburg, F., Pleij, C. W., 1995. The computer simulation of RNA folding pathways using a genetic algorithm. J Mol Biol 250 (1), 37–51.

[33] Ha, M., Kim, V. N., Aug 2014. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol 15 (8), 509–24.

[34] Hajdin, C. E., Bellaousov, S., Huggins, W., Leonard, C. W., Mathews, D. H., Weeks, K. M., 2013. Accurate SHAPE-directed RNA secondary structure modelling, including pseudoknots. PNAS 110 (14).

[35] Harmanci, A. O., Sharma, G., Mathews, D. H., 2011. Turbofold: iterative probabilistic estimation of secondary structures for multiple RNA sequences. BMC Bioinformatics 12, 108.

[36] Harmanci, A. O., Sharma, G., Mathews, D. H., 2011. TurboFold: Iterative probabilistic estimation of secondary structures for multiple RNA sequences. BMC Bioinformatics 12, 108.

[37] Hector, R. D., Burlacu, E., Aitken, S., Le Bihan, T., Tuijtel, M., Zaplatina, A., Cook, A. G., Granneman, S., 2014. Snapshots of pre-rRNA structural flexibility reveal eukaryotic 40S assembly dynamics at nucleotide resolution. Nucleic Acids Res, gku815.

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

[39] Hofacker, I. L., Fekete, M., Stadler, P. F., 2002. Secondary structure prediction for aligned RNA sequences. J Mol Biol 319, 1059–1066.

[40] Hofacker, I. L., Flamm, C., Heine, C., Wolfinger, M. T., Scheuermann, G., Stadler, P. F., Jul 2010. BarMap: RNA folding on dynamic energy landscapes. RNA 16 (7), 1308–1316.

[41] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., Schuster, P., 1994. Fast folding and comparison of RNA secondary structures. Chemical Monthly 125, 167–188.

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

[43] Homan, P. J., Favorov, O. V., Lavender, C. A., Kursun, O., Ge, X., Busan, S., Dokholyan, N. V., Weeks, K. M., 2014. Single-molecule correlated chemical probing of RNA. PNAS, 201407306.

[44] Huang, F. W., Qin, J., Reidys, C. M., Stadler, P. F., 2009. Partition function and base pairing probabilities for RNA–RNA interaction prediction. Bioinformatics 25 (20), 2646–2654.

[45] Incarnato, D., Neri, F., Anselmi, F., Oliviero, S., 2014. Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. Genome Biol 15, 491.

[46] Isambert, H., Siggia, E. D., 2000. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. PNAS 97 (12), 6515–6520.

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

[48] Kielpinski, L., Vinther, J., 2014. Massive parallel-sequencing-based hydroxyl radical probing of RNA accessibility. Nucleic Acids Res 42 (8), e70.

[49] Knudsen, B., Hein, J., 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics 15 (6), 446–454.

[50] Kucharík, M., Hofacker, I. L., Stadler, P. F., Qin, J., 2014. Basin hopping graph: a computational framework to characterize RNA folding landscapes. Bioinformatics 30 (14), 2009–2017.

[51] Kwok, C. K., Ding, Y., Tang, Y., Assmann, S. M., Bevilacqua, P. C., 2013. Determination of in vivo RNA structure in low-abundance transcripts. Nat Commun 4.

[52] Kwok, C. K., Tang, Y., Assmann, S. M., Bevilacqua, P. C., 2015. The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. Trends Biochem Sci 40 (4), 221 – 232.

[53] LeCuyer, K. A., Crothers, D. M., 1993. The leptomonas collosoma spliced leader RNA can switch between two alternate structural forms. Biochemistry 32 (20), 5301–5311.

[54] Leontis, N. B., Westhof, E., 2001. Geometric nomenclature and classification of RNA base pairs. RNA 7 (04), 499–512.

[55] Liebeg, A., Waldsich, C., 2009. Probing RNA structure within living cells. Method Enzymol 468, 219–238.

[56] Lindell, M., Romby, P., Wagner, E. G. H., 2002. Lead(II) as a probe for investigating RNA structure in vivo. RNA 8 (04), 534–541.

[57] Lorenz, R., Bernhart, S., Qin, J., Höner zu Siederdissen, C., Tanzer, A., Amman, F., Hofacker, I., Stadler, P., 2013. 2d meets 4g: G-quadruplexes in rna secondary structure prediction. Computational Biology and Bioinformatics, IEEE/ACM Transactions on PP (99), 1.

[58] Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., Hofacker, I. L., 2011. ViennaRNA package 2.0. Algorithms Mol Biol 6 (1).

[59] Lorenz, R., Luntzer, D., Hofacker, I. L., Stadler, P. F., Wolfinger, M. T., 2015. SHAPE directed RNA folding. Bioinformatics, btv523.

[60] Loughrey, D., Watters, K. E., Settle, A. H., Lucks, J. B., 2014. SHAPE-seq 2.0: Systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. Nucleic Acids Res 42 (21), e165.

[61] Low, J. T., Garcia-Miranda, P., Mouzakis, K. D., Gorelick, R. J., Butcher, S. E., Weeks, K. M., 2014. Structure and dynamics of the HIV-1 frameshift element RNA. Biochemistry 53 (26), 4282–4291.

[62] Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., Arkin, A. P., 2011. Multiplexed RNA structure characterization with selective 2-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-seq). PNAS 108 (27), 11063–11068.

[63] Lyngsø, R. B., Pedersen, C. N., 2000. RNA pseudoknot prediction in energy-based models. J Comput Biol 7 (3-4), 409–427.

[64] Mandal, M., Breaker, R. R., Jun 2004. Gene regulation by riboswitches. Nat Rev Mol Cell Biol 5 (6), 451–63.

[65] Mann, M., Kucharík, M., Flamm, C., Wolfinger, M. T., 2014. Memory efficient RNA energy landscape exploration. Bioinformatics 30 (18), 2584–2591.

[66] Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. PNAS 101, 7287–7292.

[67] Mathews, D. H., Sabina, J., Zuker, M., Turner, D. H., 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288, 911–940.

[68] Mathews, D. H., Turner, D. H., 2002. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol 317 (2), 191–203.

[69] Mathews, D. H., Turner, D. H., 2002. Experimentally derived nearest-neighbor parameters for the stability of RNA three-and four-way multibranch loops. Biochemistry 41 (3), 869–880.

[70] McCaskill, J. S., 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29 (6-7), 1105–1119.

[71] McGinnis, J. L., Weeks, K. M., 2014. Ribosome RNA assembly intermediates visualized in living cells. Biochemistry 53 (19), 3237–3247.

[72] Merino, E. J., Wilkinson, K. A., Coughian, J. L., Weeks, K. M., 2005. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). JACS 127, 4223–4231.

[73] Metz, D. H., Brown, G. L., 1969. The investigation of nucleic acid secondary structure by means of chemical modification with a carbodiimide reagent. i. the reaction between n-cyclohexyl-n'-beta-(4-methylmorpholinium)ethylcarbodiimide and model nucleotides. Biochemistry 8, 2312–2328.

[74] Mortimer, S. A., Kidwell, M. A., Doudna, J. A., 2014. Insights into RNA structure and function from genome-wide studies. Nat Rev Gen 15 (7), 469–479.

[75] Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F., Hofacker, I. L., 2006. Thermodynamics of RNA–RNA binding. Bioinformatics 22 (10), 1177–1182.

[76] Nussinov, R., Pieczenik, G., Griggs, J. R., Kleitman, D. J., 1978. Algorithms for loop matchings. SIAM Journal on Applied Mathematics 35 (1), 68–82.

[77] Ouyang, Z., Snyder, M. P., Chang, H. Y., 2013. SeqFold: Genome-scale reconstruction of RNA secondary structure inte-

grating high-throughput sequencing data. Genome Res 23 (2), 377–387.

[78] Parisien, M., Major, F., 2008. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 452 (7183), 51–55.

[79] Peattie, D., Gilbert, W., 1980. Chemical probes for higher-order structure in RNA. PNAS 77 (8), 4679–4682.

[80] Pervouchine, D. D., 2004. IRIS: intermolecular RNA interaction search. Genome Informatics Series 15 (2), 92.

[81] Qi, L., Lucks, J. B., Liu, C. C., Mutalik, V. K., Arkin, A. P., 2012. Engineering naturally occurring trans-acting non-coding RNAs to sense molecular signals. Nucleic Acids Res 40 (12), 5775–5786.

[82] Ramani, V., Qiu, R., Shendure, J., 2015. High-throughput determination of RNA structure by proximity ligation. Nat Biotech 33, 980–984.

[83] Regulski, E. E., Breaker, R. R., 2008. In-line probing analysis of riboswitches. In: Post-transcriptional Gene Regulation. Springer, pp. 53–67.

[84] Rehmsmeier, M., Steffen, P., Höchsmann, M., Giegerich, R., 2004. Fast and effective prediction of microRNA/target duplexes. RNA 10 (10), 1507–1517.

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

[86] Rivas, E., Eddy, S. R., 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol 285 (5), 2053–2068.

[87] Rivas, E., Lang, R., Eddy, S. R., Feb 2012. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. RNA 18 (2), 193–212.

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

[89] Sankoff, D., 1985. Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM Journal on Applied Mathematics 45 (5), 810–825.

[90] Seetin, M., Kladwang, W., Bida, J., Das, R., 2014. Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol. In: RNA Folding. pp. 95–117.

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

[92] Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A., Weeks, K. M., 2015. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. Nat Protocols 10, 1643–1669.

[93] Soukup, G. A., Breaker, R. R., Oct 1999. Relationship between internucleotide linkage geometry and the stability of RNA. RNA 5 (10), 1308–25.

[94] Spitale, R. C., Crisalli, P., Flynn, R. A., Torre, E. A., Kool, E. T., Chang, H. Y., 2013. RNA SHAPE analysis in living cells. Nat Chem Biol 9, 18–20.

[95] Staple, D. W., Butcher, S. E., 06 2005. Pseudoknots: RNA structures with diverse functions. PLoS Biol 3 (6), e213.

[96] Stern, S., Moazed, D., Noller, H., 1988. Structural analysis of RNA using chemical and enzymatic probing monitored by primer extension. Method Enzymol 164, 481–489.

[97] Sükösd, Z., Knudsen, B., Kjems, J., Pedersen, C. N., Oct 2012. PPfold 3.0: Fast RNA secondary structure prediction using phylogeny and auxiliary data. Bioinformatics 28 (20), 2691–2.

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

[99] Tabaska, J. E., Cary, R. B., Gabow, H. N., Stormo, G. D., 1998. An RNA folding method capable of identifying pseudoknots and base triples. Bioinformatics 14 (8), 691–699.

[100] Tafer, H., Amman, F., Eggenhoffer, F., Stadler, P. F., Hofacker, I. L., 2011. Fast accessibility-based prediction of RNA-RNA interactions. Bioinformatics 27, 1924–1940.

[101] Tafer, H., Hofacker, I. L., 2008. RNAplex: a fast tool for RNA–RNA interaction search. Bioinformatics 24 (22), 2657–2663.

[102] Talkish, J., May, G., Lin, Y., Woolford Jr., J., McManus, C., 2014. Mod-seq: High-throughput sequencing for chemical probing of RNA structure. RNA 20 (5), 713–720.

[103] Tang, X., Kirkpatrick, B., Thomas, S., Song, G., Amato, N. M., 2005. Using motion planning to study RNA folding kinetics. J Comput Biol 12 (6), 862–881.

[104] Tang, X., Thomas, S., Tapia, L., Giedroc, D. P., Amato, N. M., 2008. Simulating RNA folding kinetics on approximated energy landscapes. J Mol Biol 381 (4), 1055–1067.

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

[106] Tinoco, I., Borer, P. N., Dengler, B., Levin, M. D., Uhlenbeck, O. C., Crothers, D. M., Bralla, J., Nov 1973. Improved estimation of secondary structure in ribonucleic acids. Nat New Biol 246 (150), 40–41.

[107] Tinoco, I., Uhlenbeck, O. C., Levine, M. D., April 1971. Estimation of secondary structure in ribonucleic acids. Nature 230 (5293), 362–367.

[108] Tullius, T., Greenbaum, J., 2005. Mapping nucleic acid structure by hydroxyl radical cleavage. Curr Opin Chem Biol 9 (2), 127–134.

[109] Turner, D. H., Mathews, D. H., 2010. NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res 38 (suppl 1), D280–D282.

[110] Underwood, J., Uzilov, A., Katzman, S., Onodera, C., Mainzer, J., Mathews, D., Lowe, T., Salama, S., Haussler, D., 2010. Fragseq: Transcriptome-wide RNA structure probing using high-throughput sequencing. Nat Methods 7 (12), 995–1001.

[111] Wan, Y., Qu, K., Ouyang, Z., Kertesz, M., Li, J., Tibshirani, R., Makino, D. L., Nutter, R. C., Segal, E., Chang, H. Y., 2012. Genome-wide measurement of RNA folding energies. Molecular Cell 48 (2), 169–181.

[112] Washietl, S., Hofacker, I. L., Stadler, P. F., Kellis, M., 2012. RNA folding with soft constraints: reconciliation of probing data and thermodynamics secondary structure prediction. Nucleic Acids Res 40 (10), 4261–4272.

[113] Waterman, M. S., Smith, T. F., 1978. RNA secondary structure: A complete mathematical analysis. Math Biosci 42, 257–266.

[114] Weeks, K. M., 2010. Advances in RNA structure analysis by chemical probing. Curr Opin Struc Biol 20 (3), 295–304.

[115] Wells, S. E., Hughes, J. M., Igel, H. A., Ares, M. J., 2000. Use of dimethyl sulfate to probe RNA structure *in vivo*. In: RNA-Ligand Interactions Part B. Vol. 318 of Meth Enzymol. Academic Press, pp. 479 – 493.

[116] Wenzel, A., Akbasli, E., Gorodkin, J., Nov 2012. Risearch: fast rna-rna interaction search using a simplified nearest-neighbor energy model. Bioinformatics 28 (21), 2738–46.

[117] Wildauer, M., Zemora, G., Liebeg, A., Heisig, V., Waldsich, C., 2014. Chemical probing of RNA in living cells. In: RNA Folding. Vol. 1086. pp. 159–176.

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

[119] Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., Backofen, R., 2007. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol 3 (4), e65.

[120] Wolfinger, M. T., Svrcek-Seiler, W. A., Flamm, C., Hofacker, I. L., Stadler, P. F., 2004. Efficient computation of RNA folding dynamics. J. Phys. A: Math. Gen. 37 (17), 4731.

[121] Wolfinger, M. T., Will, S., Hofacker, I. L., Backofen, R.,

Stadler, P. F., 2006. Exploring the lower part of discrete polymer model energy landscapes. Europhys Lett 74 (4), 726–732.

[122] Wu, Y., Shi, B., Ding, X., Liu, T., Hu, X., Yip, K. Y., Yang, Z. R., Mathews, D. H., Lu, Z. J., 2015. Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. Nucleic Acids Res 43 (15), 7247–7259.

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

[124] Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., Pierce, M. B., Khan, A. R., Dirks, R. M., Pierce, N. A., 2011. NUPACK: analysis and design of nucleic acid systems. J Comput Chem 32 (1), 170–173.

[125] Zakov, S., Goldberg, Y., Elhadad, M., Ziv-Ukelson, M., Nov 2011. Rich parameterization improves rna structure prediction. J Comput Biol 18 (11), 1525–42.

[126] Zarringhalam, K., Meyer, M. M., Dotu, I., Chuang, J. H., Clote, P., 2012. Integrating chemical footprinting data into RNA secondary structure prediction. PLOS ONE 7 (10).

[127] Zaug, A. J., Cech, T. R., 1995. Analysis of the structure of tetrahymena nuclear RNAs in vivo: telomerase RNA, the self-splicing rRNA intron, and u2 snRNA. RNA 1 (4), 363–74.

[128] Zemora, G., Waldsich, C., 2010. RNA folding in living cells. RNA Biology 7 (6), 634–641.

[129] Zheng, Q., Ryvkin, P., Li, F., Dragomir, I., Valladares, O., Yang, J., Cao, K., Wang, L.-S., Gregory, B. D., 09 2010. Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in arabidopsis. PLoS Genetics 6 (9).

[130] zu Siederdissen, C. H., Bernhart, S. H., Stadler, P. F., Hofacker, I. L., 2011. A folding algorithm for extended RNA secondary structures. Bioinformatics 27 (13), i129–i136.

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

[132] Zuker, M., Stiegler, P., 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxilary information. Nucleic Acids Res 9 (1), 133–147.