

Non-Redundant Sampling and Statistical Estimators for RNA Structural Properties at the Thermodynamic Equilibrium

Christelle Rovetta, Juraj Michálik, Ronny Lorenz, Andrea Tanzer, Yann Ponty

▶ **To cite this version:**

Christelle Rovetta, Juraj Michálik, Ronny Lorenz, Andrea Tanzer, Yann Ponty. Non-Redundant Sampling and Statistical Estimators for RNA Structural Properties at the Thermodynamic Equilibrium. 2019. hal-02288811

HAL Id: hal-02288811

https://inria.hal.science/hal-02288811

Preprint submitted on 16 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Non-Redundant Sampling and Statistical Estimators for RNA Structural Properties at the Thermodynamic Equilibrium

Christelle Rovetta*, Juraj Michalik*, Ronny Lorenz, Andrea Tanzer, and Yann Ponty,

Abstract—The computation of statistical properties of RNA structure at the thermodynamic equilibrium, or Boltzmann ensemble of low free-energy, represents an essential step to understand and harness the selective pressure weighing on RNA evolution. However, classic methods for sampling representative conformations are frequently crippled by large levels of redundancy, which are uninformative and detrimental to downstream analyses.
In this work, we adapt and implement, within the Vienna RNA package, an efficient non-redundant backtracking procedure to produce collections of unique secondary structures generated within a well-defined distribution. This procedure is coupled with a novel statistical estimator, which we prove is unbiased, consistent and has lower variance (better convergence) than the classic estimator.
We demonstrate the efficiency of our coupled non-redundant sampler/estimator by revisiting several applications of sampling in RNA bioinformatics, and demonstrate its practical superiority over previous estimators. We conclude by discussing the choice of the number of samples required to produce reliable estimates.

Index Terms—RNA secondary structure, Boltzmann equilibrium, Non-redundant sampling, Statistical estimator

1 Introduction

Structural properties of RNAs are crucial to build a mechanical understanding of their function. Aside from their role in mediating genetic information from genome to the protein levels, RNAs are associated with multiple enzymatic and regulatory functions, leading recent versions of the RFAM database to enumerate more than 3,000 functional families [1]. This collection will likely expand in the upcoming years, following the discovery of hundreds of thousands of long non-coding RNAs [2]. In many of those families, an evolutionary pressure can be observed towards the adoption of one or several important folds, leading to an instrumental part being played by a consensus secondary structure in the definition of functional families.

RNA functional architectures are adopted as the final outcome of a folding process governed by thermodynamics. Multiple copies of an RNA alternate between their stable structures, inducing an equilibrium which favors a subset of stables conformations, the Boltzmann ensemble of low-energy including the minimum free-energy structure. The concept of Boltzmann ensemble is found at the core of recent computational methods, allowing to embrace the full conformational diversity. Following the seminal work of McCaskill [3], essential thermodynamics quantities, namely the partition function and base-pairs probabilities, can be computed in polynomial time using dynamic programming (DP). Systematic modifications of the McCaskill DP scheme have been proposed over the past decades to compute other properties, including the expected 5′–3′ distance [4], base-pair distance [5], [6] or mutation-classified [7] partition functions, moments of the free-energy distribution [8] and of general additive features [9]. However, more complex quantities, such as the graph distance distribution, may require algorithms whose complexity, albeit polynomial, become prohibitively large [10]. Moreover, the study of new quantities requires the development of new *ad hoc* algorithmic DP schemes.

As an alternative, statistical approaches are increasingly used to estimate features of the Boltzmann ensemble. First introduced by Ding and Lawrence [11], a stochastic backtrack procedure samples RNA secondary structures from the exact Boltzmann distribution by recursively performing local random choices, using precomputed probabilities. Boltzmann sampling procedures are now implemented in most libraries for RNA secondary structure analysis, including the `ViennaRNA` package [12], `RNAStructure` [13] and `Unafold` [14]. Such approaches are also used to sample from reduced subsets of secondary structures sharing a certain property, such as locally optimal structures [15], [16], or within partitioned sets [17]. Sampling methods possess a wide range of applications, including RNA kinetics studies [18], evolutionary neutrality [19], structure modeling from experimental probing data [20], gradient-based optimization strategies [21], and RNA design [22]. By only requiring a capacity to compute the quantity of interest, such methods represent a flexible, if approximate, alternative to exact DP-based computations. Sampling may even represent the only available option for the production of dominant conformers in coarse-grained abstractions [23].

Statistical estimates for the Boltzmann ensemble are

* *Both authors contributed equally*

- *Y. Ponty, J. Michalik and C. Rovetta were with the Department of Computer Science (LIX), Ecole Polytechnique, 91 120 Palaiseau, France. E-mail: yann.ponty@lix.polytechnique.fr*
- *R. Lorenz and A. Tanzer were with the Theoretical Biochemistry Institute, University of Vienna, Austria.*

Manuscript received XXXX; revised YYYY.

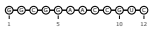
Id	Structure	Bolz. Prob.	$\mathbb{P}$(Abs.)	$\mathbb{E}$ (#Occ)	#Occ.
S_1		0.808	$1.12\ 10^{-36}$	40.44	35
S_2		0.0981	$5.72\ 10^{-3}$	4.91	8
S_3		0.0371	0.151	1.85	3
S_4		0.0228	0.316	1.14	1
S_5		0.0193	0.376	0.969	2
S_6		$7.32\ 10^{-3}$	0.693	0.366	1
⋮		⋮	⋮	⋮	⋮
S_{84}		$9.01\ 10^{-10}$	1.000	0.0	0

Fig. 1. **Redundancy within a Boltzmann sample consisting of 50 structures for the RNA GGCGGAACCGUC.** Out of the 84 available conformations for this RNA, only 6 distinct structures are represented in the redundant sample (expected value = 4.76).

classically computed by first generating a fixed number of structures, followed by an evaluation of features of interest. Redundancy within the sample, *i.e.* the presence of multiple copies of the same conformation, is typically used to estimate emission probabilities. However, in Boltzmann-Gibbs distributions, the exact emission probability of a given structure only depends on its free-energy, and the partition function, both of which are readily available after each generation. Redundancy is therefore uninformative, since estimated probabilities are detrimental to the accuracy of derived. Moreover, a high level of redundancy within sampled sets in Boltzmann-like distributions is theoretically expected [24], [25], leading one to anticipate a substantial impact of redundancy on performances.

In previous work [26], [16], we introduced general principles for the non-redundant generation of statistically-sound sampled sets. However, while the overall sample distribution remained well-defined, these work left open the computation of statistical estimates from non-redundant samples. Computing statistics on such a non-redundant set is not a trivial problem since, by breaking the assumption of independence between consecutive samples, non-redundancy sampling forbids usage of classic naive estimators.

In this work, we address the computation of statistical estimates from samples generated using non-redundant sampling. We introduce a new statistical estimator, which we prove is unbiased, asymptotically converges and has lower empirical variance than the naive estimator for the same sample set. We adapt non-redundant principles to the statistical sampling introduced by Ding and Lawrence [11], which considers the realistic Turner energy model, and implement our algorithm in the popular `ViennaRNA` package [12]. Extensive empirical analysis demonstrates that our non-redundant sampler and estimator enables more precise estimates for RNA structural features. Typical estimates include base-pair probabilities matrices (dot-plots), shape probabilities [23] and graph distance distribution [10], for which exact dynamic-programming alternatives are prohibitively costly. We finally discuss the number of samples required to achieve a given precision for the estimates.

2 Material and Methods

2.1 Notations and model

An RNA is abstracted as a sequence w of nucleotides of size $|w| = n$. Consider as **valid** any base pair in $\{\{\mathsf{A},\mathsf{U}\},\{\mathsf{C},\mathsf{G}\},\{\mathsf{G},\mathsf{U}\}\}$. Denote by $\mathcal{P}$ the set of all **valid base pairs** in w *ie.* all the pairs (i,j), such that $(w[i],w[j])$ form a valid base pair.

A **secondary structure** s is defined as a subset of $\mathcal{P}$ satisfying the following constraints: i) $\forall (i,j),(k,l) \in s, i \geq k$ one either has $i < k < l < j$ or $i < j < k < l$ (absence of pseudoknots) and ii) any base can participate at most in one base pair within the same secondary structure. For instance, the secondary structure S_1 in Figure 1, can be represented as $\{(1,12),(2,11),(3,10),(4,9)\}$. In the following, we use Ω to denote the set of all valid secondary structures for w.

An **energy model** associates a free-energy $E(s)$ to any structure $s \in \Omega$ for an RNA w. At the **thermodynamic equilibrium**, the conformations in Ω are expected to follow a Boltzmann distribution, where any given structure has probability proportional to its **Boltzmann factor** $\mathcal{B}(s)$

$$\mathcal{B}(s) = e^{-\beta E(s)}$$

with $\beta := \frac{1}{RT}$, R the gaz constant, and T the temperature in coherent units. The **partition function** $\mathcal{Z}$ is then defined as

$$\mathcal{Z} = \sum_{s \in \Omega} \mathcal{B}(s)$$

The **Boltzmann probability** $\mathbb{P}(s)$ of observing a given conformation s is then simply

$$\mathbb{P}(s) := \frac{\mathcal{B}(s)}{\mathcal{Z}}.$$

2.2 The challenge of redundancy

For a given RNA sequence, the free energies of compatible secondary structures greatly vary. Thus, under the Boltzmann distribution, the probabilities of secondary structures cover a wide range of values. In particular, it is not uncommon for a small subset of structures to accumulate a substantial proportion of the probability mass, and therefore have overwhelming accumulated probability.

Figure 1 illustrates the practical impact of redundancy. While sampling structures, the first collision (*i.e.* first duplicate in a dataset) is observed, on average, only after a theoretical 3.5 generations [24]. The number of distinct structures grows at a painstakingly slow pace, with an expected 4.75 distinct structures within a sample of 50 structures, only increasing to 9 structures for 10^3 structures, and 15 structures for 10^6 structures. Generating all the

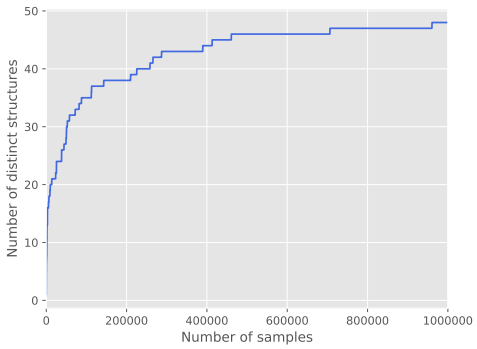


Fig. 2. **Distinct structures within a redundant sample.** Generation of 10^6 Boltzmann-distributed structures for the RNA of Figure 1. The number of distinct structures quickly reaches a plateau after 170 000 generations, and only 48 out of the possible 84 possible structures are represented in the final output.

structure (84 secondary structures) is expected to require between $1.09 \cdot 10^9$ and $6.8 \cdot 10^{10}$ structures, using standard estimates [24].

To mitigate the wastefulness of redundancy, one could perform a **non-redundant sampling**, to record generated structures and avoid them during future generations. Denote by Θ the set of structures previously generated, then the probability for a non-redundant sampling algorithm to generate a structure $t \in \Omega \setminus \Theta$ is given by:

$$\mathbb{P}_\Theta(t) = \begin{cases} \frac{\mathbb{P}(t)}{1-\mathrm{Cov}(\Theta)} & \text{if } s \notin \Theta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\mathrm{Cov}(\Theta)$ is the **coverage** of Θ, *i.e.* the probability accumulated by a collection Θ of distinct structures:

$$\mathrm{Cov}(\Theta) = \sum_{s \in \Theta} \mathbb{P}(s)$$

In this setting, the probability to generate a non-redundant collection of m structures $\mathbf{t} := (t_1, t_2, \ldots, t_m)$ (in this order) is :

$$\mathbb{P}(\mathbf{t}) = \mathbb{P}(t_1) \frac{\mathbb{P}(t_2)}{1-\mathbb{P}(t_1)} \cdots \frac{\mathbb{P}(t_m)}{1-\mathbb{P}(t_1)\cdots-\mathbb{P}(t_{m-1})}. \quad (2)$$

2.3 Statistical sampling

Secondary structures compatible with w can be randomly generated in a Boltzmann distribution, using a (redundant) **stochastic backtrack** introduced by Ding and Lawrence [11]. For the sake of simplicity, we present the general principle reduced to its key elements, using a base-pair based, additive energy model, where any base pair (i,j) has associated contribution $E_{i,j}$. The underlying principle easily generalizes to the loop-based Turner model [27], which is supported by our implementation.

2.3.1 *Redundant stochastic backtrack*

The set of all secondary structures Ω can be recursively generated. Consider $w_{i,j}$ the subsequence of w on the interval $[i,j] \subseteq [1,n]$ ($w_{1,n} = w$) and $\Omega_{i,j}$ the set of all secondary structures generated from $w_{i,j}$. Then we have

$$\Omega_{i,j} = \Omega_{i+1,j} \bigcup_{\substack{(i,k)\in\mathcal{P} \\ \text{s.t. } i<k\leq j}} \{(i,k)\} \times \Omega_{i+1,k-1} \times \Omega_{k+1,j}. \quad (3)$$

The first term represents structures leaving i unpaired, and the second one covers all possible partners for i. The set of all secondary structures is then given by $\Omega := \Omega_{1,n}$.

The construction described in Equation 3 is then used to build a **tree of valid structures**, where each node is indexed by a couple (I,P) where I is a set of **non-overlapping** subintervals of $[1,n]$, and $P \subseteq \mathcal{P}$ a subset of pairwise non-crossing **pairs**. We distinguish two types of nodes: leaves and internal nodes. A leaf is a node $(\varnothing, P)$, where P represents a single secondary structure. An internal node represents a set of secondary structures obtained by choosing a set of suitable base pairs for each of the intervals in I. The **root** of the tree is then the node $v_\mathrm{r} := ([1,n], \varnothing)$ from which all secondary structures defined in Equation (3) are accessible.

The tree is constructed from the root to the leaves using Equation (3). The **children** of an internal node $v := ([i,j] \cup I', P)$ are produced using the following formulas:

$$\mathcal{C}(v)) := \{([i+1,j] \cup I', P)\}$$
$$\bigcup_{\substack{i<k\leq j \\ \text{s. t. } (i,k)\in\mathcal{P}}} \{([i+1,k-1] \cup [k+1,j] \cup I', P \cup \{(i,k)\})\}.$$

We now define the **partition function** of a node. It is computed recursively from the leaves to the root as follow:

$$\mathcal{B}((I,P)) = \begin{cases} \mathcal{B}(P) & \text{if } I = \varnothing \text{ (leaf)}) \\ \sum_{v' \in \mathcal{C}((I,P))} \mathcal{B}(v') & \text{otherwise.} \end{cases}$$

To randomly generate a secondary structure, we perform a random walk within the tree. Starting at the root v_r, at each step a child v' of a node v is chosen with probability

$$p_{v \to v'} := \frac{\mathcal{B}(v')}{\mathcal{B}(v)} \quad (4)$$

and we iterate this process until until a leaf is reached. The probability to generate a given structure s – to reach a leaf $(\varnothing, s)$ – is then:

$$\mathbb{P}(s) = \frac{\mathcal{B}(v_1)}{\mathcal{B}(v_\mathrm{r})} \times \frac{\mathcal{B}(v_2)}{\mathcal{B}(v_1)} \times \cdots \times \frac{\mathcal{B}(v_p)}{\mathcal{B}(v_{p-1})} \times \frac{\mathcal{B}(s)}{\mathcal{B}(v_p)} = \frac{\mathcal{B}(s)}{\mathcal{Z}}$$

where $v_\mathrm{r} =: v_0, v_1, \ldots, v_p$ is the sequence of nodes encountered from the root to the leaf. In other words, using probabilities in Equation (4) ensures that the emission probability coincides with the Boltzmann distribution.

Of course, since the number of secondary structures grows exponentially with the length of the sequence [28], we cannot memorize explicitly the entirety of such a tree. However, the local aspect of our energy model leads to simpler expressions for the partition function associated with a node (I,P):

$$\mathcal{B}((I,P)) = \mathcal{B}(P) \prod_{[i,j]\in I} \mathcal{B}_{i,j} \quad (5)$$

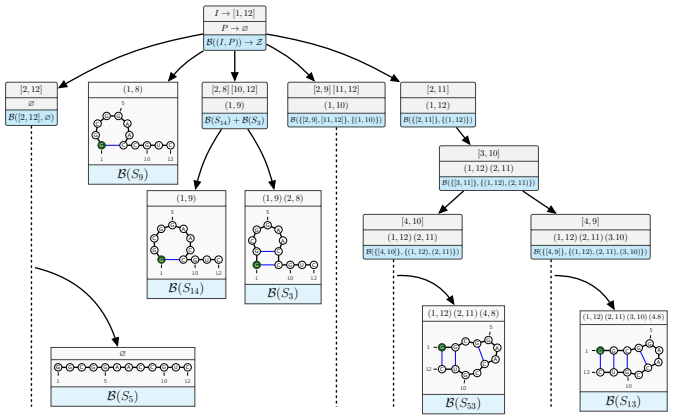


Fig. 3. **Tree of valid secondary structures** (truncated) for the RNA sequence of Figure 1. Classic stochastic backtrack can be seen as the process of generating a random path from the root to a leaf, choosing at each step one of the alternatives proportionally to its partition function. For instance, S_3 results from two local choices, and has probability $\frac{\mathcal{B}(S_{14})+\mathcal{B}(S_3)}{\mathcal{Z}} \frac{\mathcal{B}(S_3)}{\mathcal{B}(S_{14})+\mathcal{B}(S_3)} = \frac{\mathcal{B}(S_3)}{\mathcal{Z}}$, which is indeed the targeted Boltzmann probability.

where $\mathcal{B}_{i,j}$ corresponds to the partition function of all structures confined to an interval $[i,j]$, computable in polynomial time using dynamic programming through:

$$\mathcal{B}_{i,j} = \mathcal{B}_{i+1,j} + \sum_{i<k\leq j} \mathcal{B}((i,k)) \times \mathcal{B}_{i+1,k-1} \times \mathcal{B}_{k+1,j}.$$

Interestingly, plugging the expression of Equation (5) in the transition probability greatly simplifies the expression of the transition probability. For instance, in a node $v := (I,P)$, if a base pair (i,k) is chosen for the first interval $[i,j]$ $I := [i,j] \cup I'$ of I, then we reach a new node $v' := ([i+1,k-1] \cup [k+1,j] \cup I', P \cup \{(i,k)\})$, and we get

$$\begin{aligned} p_{v\to v'} :=& \frac{\mathcal{B}(v')}{\mathcal{B}(v)} \\ =& \frac{\mathcal{B}((i,k))\,\mathcal{B}(P)\,\mathcal{B}_{i+1,k-1}\,\mathcal{B}_{k+1,j}\,\prod_{[a',b']\in I'}\mathcal{B}_{a',b'}}{\mathcal{B}(P)\,\mathcal{B}_{i,j}\prod_{[a,b]\in I'}\mathcal{B}_{a,b}} \\ =& \frac{\mathcal{B}((i,k))\,\mathcal{B}_{i+1,k-1}\mathcal{B}_{k+1,j}}{\mathcal{B}_{i,j}} \end{aligned}$$

Remark that the set of previously assigned base pairs no longer influences the probability of choosing a given alternative. This crucial observation is at the core of the efficient stochastic backtrack procedure of Ding and Lawrence [11]. Unfortunately, this simplification no longer holds when a set Θ of structures needs to be avoided. Indeed, the above probabilities then need to account for structures in Θ in a way that depend on previous choices, calling for an alternative strategy.

2.3.2 *Sequential non-redundant (NR) sampling*

For non-redundant sampling, we need to generate within the Boltzmann distribution, normalized to avoid a predefined set Θ of forbidden structures (see Equation (1)). A first remark is that Equation (4) induces a Boltzmann distribution as long as the individual partition functions are correctly computed. A similar strategy can then be used, using partition functions that reflect the unavailability of structures in Θ, to generate a structure within the distribution of Equation (1). However, the efficient update of those partition functions remains a complex task, and requires an explicit construction of a structure *on the fly* during the sampling, followed by a back propagation of the Boltzmann weights.

Namely, we introduced in [16] a dedicated data structure $\widetilde{\mathcal{B}}$ to gather and access the contributions of forbidden structures during the stochastic backtrack. Built similarly as the graph of secondary structures, it is restricted to structures in Θ, and thus grants access to the partition function $\widetilde{\mathcal{B}}(v)$ of structures from Θ that could (but should not) be generated from a given node v. The values in $\widetilde{\mathcal{B}}$ are (implicitly) initialized to 0 for all putative nodes, and are updated after each generation of a structure s, incrementing by $\mathcal{B}(s)$ the value $\widetilde{\mathcal{B}}(v)$ of any node v encountered during the generation of s. Figure 4 superimposes the contributions of the data structure onto the tree of valid secondary structures.

Duration the generation, the probability of selecting a possible child $v' \in \mathcal{C}(v)$ is modified to

$$\tilde{p}_{v\to v'} := \frac{\mathcal{B}(v') - \widetilde{\mathcal{B}}(v')}{\mathcal{B}(v) - \widetilde{\mathcal{B}}(v)}.$$

The probability of generating a structure s becomes

$$\mathbb{P}(s) = \frac{\mathcal{B}(v_1) - \widetilde{\mathcal{B}}(v_1)}{\mathcal{Z} - \mathcal{B}(\Theta)} \frac{\mathcal{B}(v_2) - \widetilde{\mathcal{B}}(v_2)}{\mathcal{B}(v_1) - \widetilde{\mathcal{B}}(v_1)} \cdots \frac{\mathcal{B}(s)}{\mathcal{B}(v_{p-1}) - \widetilde{\mathcal{B}}(v_{p-1})}.$$

Adjacent numerator and denominator terms cancel in a pairwise manner, and we are left with

$$\mathbb{P}(s) = \frac{\mathcal{B}(s)}{\mathcal{Z} - \mathcal{B}(\Theta)} = \frac{\mathbb{P}(s)}{1 - \sum_{s'\in\Theta}\mathbb{P}(s')}$$

in which one recognizes the targeted distribution described in Equation (1).

The generation procedure, described in Algorithm 1, can be optimized to offer non-redundant implementations for the Turner energy model (see Supp. mat. for more details) having the same asymptotic complexity as the classic – redundant – backtrack [11].

Algorithm 1: Non-redundant stochastic backtrack.

Data: Forbidden structures $\Theta \subset \Omega$, gathered into data structure $\widetilde{\mathcal{B}}$
Result: $s \in \Omega - \Theta$, randomly generated with probability proportional to $\mathcal{B}(s)$

$v \leftarrow v_{\mathrm{r}}$ (root $([1,n],\varnothing)$ of T) ;
$L \leftarrow [v]$ (L is a list) ;
while *v is not a leaf* **do**
  Choose $v' \in \mathcal{C}(v)$ with probability $\frac{\mathcal{B}(v')-\widetilde{\mathcal{B}}(v')}{\mathcal{B}(v)-\widetilde{\mathcal{B}}(v)}$;
  $L \leftarrow L + v'$;
  $v \leftarrow v'$;
$(\varnothing, s) \leftarrow v$ (v is a leaf, *i.e.* structure s) ;
for $v' \in L$ **do**
  $\widetilde{\mathcal{B}}(v') \leftarrow \widetilde{\mathcal{B}}(v') + \mathcal{B}(s)$;
return s

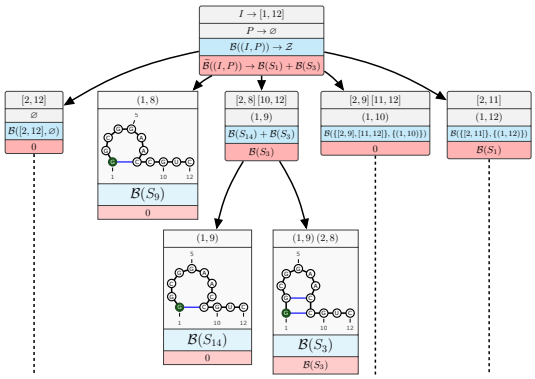


Fig. 4. **Boltzmann factor update** after generating structures S_1 and S_3) using Algorithm 1. Blue boxes indicate initial partition functions, while red boxes represent the values stored in our dedicated data structure $\widetilde{\mathcal{B}}$.

2.4 Estimating equilibrium properties

Stochastic backtrack algorithms are frequently used to estimate statistical properties of RNAs at the thermodynamic equilibrium. While such quantities can sometimes be computed exactly through dynamic programming schemes [8], [9], others seem to induce impractical complexities [10], or are even believed to be associate with NP-hard problems [23]. In any case, sampling-based estimates only require a capacity to evaluate the feature on any given structure, and thus provide a very flexible solution to perform a statistical analysis of RNA thermodynamics.

Formally, an **RNA feature** can be any function $F : \Omega \to \mathbb{R}$ that measures some characteristic of an RNA. Features of interest include structural descriptors (presence/absence of base pairs, #helices...), thermodynamic stability (free-energy)... The goal of our analysis is to compute the **expected value** $\mathbb{E}(F(S))$ of F, where S is a Boltzmann-distributed random structure, defined as

$$\mathbb{E}(F(S)) = \sum_{s \in \Omega} \mathbb{P}(s) \times F(s). \tag{6}$$

This formulation is very general, and captures any computable property of RNA structures. For instance, to estimate the probability of occurrence of a motif M within a Boltzmann-distributed random structure, one simply takes a boolean feature function F_M, taking value $F_M(s) = 1$ if the motif M occurs in s, and 0 otherwise. The expectation of F_M then simplifies into the accumulated probability of all structures having M as a motif. More complex statistical quantities, such as the standard deviation of a feature, the Pearson correlation of features, or even higher-order statistics, can be simply obtained by estimating, and combining, powers of the feature(s) of interest.

2.4.1 Empirical mean: The classic redundant (R) estimator

Generally, consider $\mathbf{s} := (s_1, s_2, \ldots, s_m)$ a vector of independent, uniformly-distributed, structures obtained by a redundant (R) sampling. The **empirical mean**, further referred to as the **R estimator**, is given by

$$\widehat{F}(\mathbf{s}) = \frac{1}{m} \sum_{i=1}^{m} F(s_i)$$

For instance one could estimate, from the redundant sample of Figure 1, the probability that, at the thermodynamic equilibrium, the first nucleotide forms some base pair. A Boolean feature F_1 would take value 1 when the first nucleotide is paired (*e.g.* structure S_1), and 0 otherwise (*e.g.* structure S_2). The estimator would then simplify into the empirical proportion of structures pairing their first position, leading to an estimated $\widehat{F_1} = 0.78$ probability. This estimate differs substantially from the real value of 0.87. The origin of this problem is that, in the context of RNA, few structures concentrate a large part of the probability mass, leading the classic sampling to converge quite slowly.

2.4.2 Estimating from a non-redundant (NR) sample

Intuitively, the empirical mean (R estimator) can be seen as solving two tasks simultaneously:

- It estimates the probability of structures;
- It computes a weighted average of the feature values.

However, within Boltzmann distributions, the exact probability of a structure is entirely known, and readily available (or typically computable in $\Theta(n)$ time) as soon as the structure is produced. Redundancy is then, in theory, uninformative and could be safely avoided.

Unfortunately, one cannot simply use the R estimator to process a non-redundant sampled set of structures. Indeed, the R estimator implicitly assigns the same weight to all structures found in the sample, and redundancy is then crucial to account for the Boltzmann probabilities of structures. Other natural ideas, such as weighting each structure with its Boltzmann factor (possibly followed by a renormalization step), also demonstrably falter, leading to biased estimates beyond trivial cases.

We now introduce our novel **non-redundant (NR) estimator** $\widetilde{F}(\mathbf{t})$ for the expected value of a feature F from a non-redundant sample generated according to the distribution 2. Given a non-redundant sequence of sampled structures $\mathbf{t} := (t_1, t_2, \ldots, t_m)$, it is defined as:

$$\widetilde{F}(\mathbf{t}) = \frac{1}{m} \sum_{i=1}^{m} F(t_i) \left(1 - \bar{F}^{(0)}_{\Theta_{i-1}} + (m-i) \times \mathbb{P}(t_i)\right) \tag{7}$$

where $\Theta_i := (t_1, \ldots, t_i)$ and $\bar{F}^{(x)}_{\Theta} := \sum_{t \in \Theta} \mathbb{P}(t) F(s)^x$.

Intuitively, the $(1 - \bar{F}^{(0)}_{\Theta})$ term can be interpreted as correcting for the fact that, at the i-th iteration of the non-redundant sampling algorithm, an overall probability mass of $\bar{F}^{(0)}_{\Theta}$ has already been generated, increasing the probability of generating t_i by a factor $1/(1 - \bar{F}^{(0)}_{\Theta})$. Similarly, the term $(m-i) \times \mathbb{P}(t)$ can be interpreted as the expected number of futures occurrences for t_i using classic sampling, the absence of which within the non-redundant sample must be counterbalanced.

Note that, while the accumulated probability $\bar{F}^{(0)}_{\Theta}$ of previously-generated structures remains negligible, as happens for longer RNAs, the term $(m-i) \times \mathbb{P}(t_i)$ stays close to zero. The NR estimator is then equivalent to the empirical mean, consistent with the fact that redundant sampling then typically yields a non-redundant sequence of structures.

Computing the NR estimator typically induces **negligible time and space consumptions** compared to the sampling itself. Indeed, the sum in Equation (7) can be computed in time $\Theta(m \times C_F(n))$, where $C_F(n)$ is the time required for evaluating the feature F (usually $C_F(n) \in \mathcal{O}(n)$). Indeed, any probability $\mathbb{P}(t_i)$ can be computed in constant time, and the sums involved in the computation of $\bar{F}^{(0)}_{\Theta}$ can be incrementally updated in constant time anytime a new value becomes available, and so can $\widetilde{F}(\mathbf{t})$. This complexity compares favorably against the $\Theta(n^3+m\times n^2)$ time required by the (non-redundant) sampling itself.

2.4.3 *Correctness and efficiency of the NR estimator*

Firstly, the NR estimator is **unbiased**, meaning that the expected value of the estimator over a sequence of random structures matches the true expectation of the feature. Indeed, we establish in Supp. mat. that

$$\mathbb{E}\left(\widetilde{F}(\mathbf{t})\right) = \mathbb{E}\left(F(S)\right)$$

where S is a random Boltzmann-distributed structure, and $\mathbf{t}$ a non-redundant sequence of random structures consisting of at least one structure. This follows from the fact that the estimator can be rewritten as

$$\widetilde{F}(\mathbf{t}) = \frac{1}{m}\left(\sum_{i=1}^{m} F(v_i) \times (1 - \bar{F}^{(0)}_{\Theta_{i-1}}) + \bar{F}^{(1)}_{\Theta_{i-1}}\right), \quad (8)$$

and one easily shows that, for any set Θ of avoided structures, one has

$$\mathbb{E}\left(F(T) \times (1 - \bar{F}^{(0)}_{\Theta}) + \bar{F}^{(1)}_{\Theta} \mid \Theta\right) = \mathbb{E}\left(F(S)\right).$$

The NR estimator can then be reformulated as the average of a sequence of random variables, each having $\mathbb{E}(F(S))$ as its expected value, and the absence of bias immediately follows.

The absence of bias also implies that the NR estimator can also be used to **estimate the variance** of a feature $F(T)$. To that end, simply compute the estimators $\widetilde{F}(\mathbf{t})$ and $\widetilde{F^2}(\mathbf{t})$ to estimate the expectations of $F(S)$ and $F(S)^2$ respectively, and use the formula

$$\mathbb{V}\left(F(S)\right) = \mathbb{E}\left(F(S)^2\right) - \mathbb{E}\left(F(S)\right)^2$$

to recover an estimate for the variance.

Secondly, the NR estimator is **statistically consistent**, *i.e.* as the number of sample grows, the estimated value gets increasingly and arbitrarily close to the real expected value of the feature. Since this property only formally holds for infinite sources, we consider a generalized version of the NR sampling process, which repeatedly returns a fake structure $\perp$ with probability 0 and value $F(\perp) = 0$ once the full collection of structure has been generated. Remark that, for any i-th sample such that $i \geq |\Omega|$ samples, the contribution to the sum in the estimator greatly simplifies, since $1 - \bar{F}^{(0)}_{\Theta_{i-1}} = 0$ and $\bar{F}^{(1)}_{\Theta_{i-1}} = \mathbb{E}(F(S))$. It follows that, denoting by A_Ω the accumulated value of the sum over the $|\Omega|$ first samples, one has

$$\begin{aligned}\widetilde{F}(\mathbf{t}) &= \frac{1}{m}\left(A_\Omega + \sum_{i=|\Omega|+1}^{m} \mathbb{E}\left(F(S)\right)\right) \\ &= \frac{A_\Omega}{m} + \frac{(m - |\Omega|)}{m} \cdot \mathbb{E}\left(F(S)\right).\end{aligned}$$

It immediately follows that

$$\lim_{m\to\infty} \widetilde{F}(\mathbf{t}) = \mathbb{E}\left(F(S)\right),$$

implying the consistency of the estimator.

Finally, the NR estimator provably has **lower variance** than the empirical mean computed from a redundant sample, using the same number of structures. Formally, one has

$$\mathbb{V}(\widetilde{F}(\mathbf{t})) \leq \mathbb{V}(\widehat{F}(\mathbf{s})), \forall |\mathbf{t}| = |\mathbf{s}| \geq 1,$$

the inequality being strict as soon as $|\mathbf{t}| = |\mathbf{s}| > 1$, the property implies a lower dispersion for the values obtained using the NR estimator. A formal proof of this property is slightly involved, and can be found in Supp. mat. This theoretical superiority has concrete practical consequences, as can be observed by Figure 5.

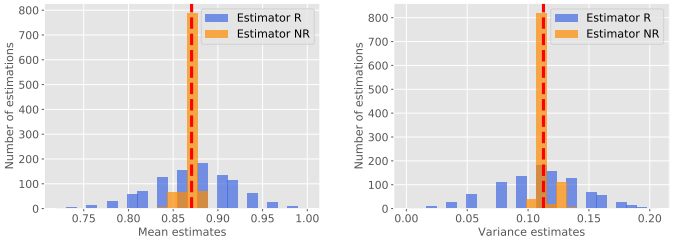


Fig. 5. **Compared accuracies of R (empirical mean) and NR estimators.** For the running example of Fig. 1, the expectation (left) and variance (right) of the Boolean feature indicating the base pairing status of the first nucleotide, were computed using both estimators for 1 000 representative samples of 50 secondary structures each. Dashed lines indicate the exact value of both statistics.

3 Results

Our non-redundant backtracking procedure was implemented within the Vienna RNA package [12], and is available within the `RNAsubopt`, `RNApvmin` and `RNAalifold` utilities using the `-N` modifier. The Vienna RNA package can be compiled from freely-accessible sources, or downloaded as a bundle of binaries for virtually any architectures at:

https://www.tbi.univie.ac.at/RNA/

The paper experiments has been implemented in `python`. Implementations for NR generation (Simulation using `ViennaRNA`, NR estimator, ...) and a tutorial are freely-accessible sources at:

https://gitlab.com/christelle.rovetta/rnanr-stats

3.1 Inferring dot-plots

As a first illustration of the potential of our NR methodology, we consider the fast computation of estimates for the

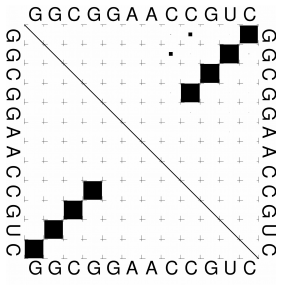


Fig. 6. **Example of a dot plot.** Dot plot is associated to the RNA of Fig. 1. The surface covered by black squares in the upper-right triangular matrix indicates the probability of a base pair to occur at the thermodynamic equilibrium. The lower left triangular matrix indicates the base pairs of the most stable/probable structure S_1.

base pair probability matrices, also know as **dot plots**. Such matrices are at the core of reference computational methods in RNA bioinformatics, including structural alignment [29] and design [30], and contain the probabilities $p_{i,j}$ of forming a base pair between positions i and j at the thermodynamic equilibrium, such that $p_{i,j} = \sum_{\substack{s\in\Omega,\\(i,j)\in s}} \mathbb{P}(s)$.

Exact probabilities can be computed using a variant of the inside/outside algorithm, practically doubling the time of the costly compution of the partition function [3]. As an alternative, we consider the estimation of such probabilities, introducing Boolean features $D_{i,j}$ that indicates presence/absence of a base pair (i,j):

$$D_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in s \\ 0 & \text{otherwise.} \end{cases}$$

The expectation of any such feature is given by

$$\mathbb{E}(D_{i,j}) = \sum_{s\in\Omega} D_{i,j} \times \mathbb{P}(s) = \sum_{\substack{s\in\Omega,\\(i,j)\in s}} 1 \times \mathbb{P}(s) = p_{i,j}.$$

An estimator for the expectation of the proposed feature is therefore also an estimator for the base-pair probabilities.

We consider the dot plots estimates $\{\widehat{D}_{i,j}(\mathbf{s})\}_{i,j}$ and $\{\widetilde{D}_{i,j}(\mathbf{t})\}_{i,j}$, obtained using the empirical mean and our NR estimator respectively. We assess the accuracy of both estimators by comparing their inferred probabilities to the exact base pair probabilities, computed using the `ViennaRNA` implementation of the McCaskill algorithm [3]. The **overall error** e_{R} and e_{NR}, respectively induced by the empirical mean and the NR estimator, are defined as

$$e_{\mathrm{R}} = \frac{\sqrt{\sum_{i,j}(\widehat{D}_{i,j} - p_{i,j})^2}}{n(n-1)} \text{ and } e_{\mathrm{NR}} = \frac{\sqrt{\sum_{i,j}(\widetilde{D}_{i,j} - p_{i,j})^2}}{n(n-1)}.$$

The renormalization by $n(n-1)$ is used to mitigate the influence of the sequence length, and thus assess the average error per base pair probability.

Next, we turn to the analysis of both estimators in two settings: First, we compare the accuracy of both estimates, computed from two sets of structures having equal cardinality; Then, to account for the fact that non-redundant sampling typically induces a 10 to 20% computational overhead [31], we allocate the same time to both generators, and compute estimates for a given elapsed time.

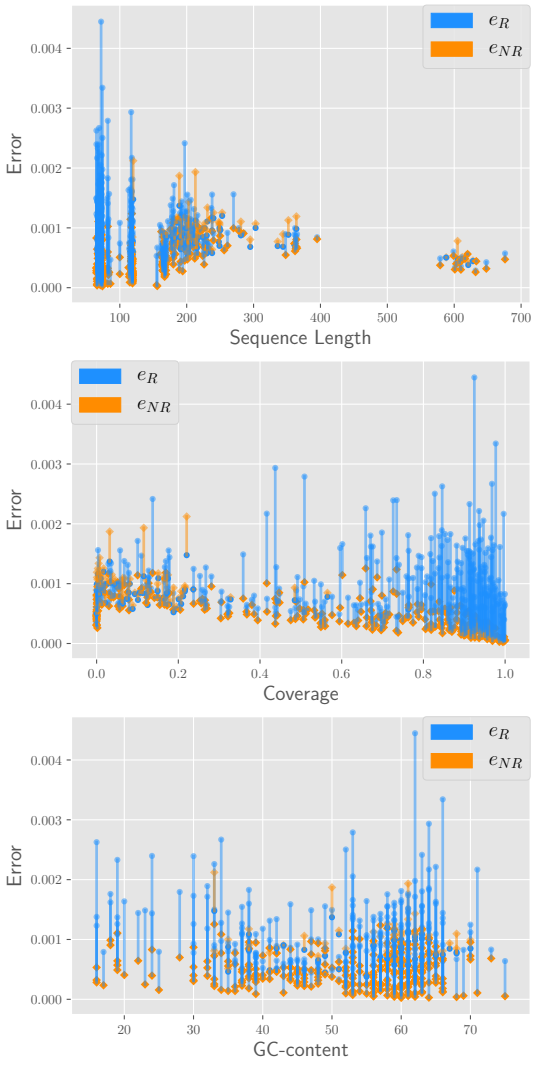


Fig. 7. **Effect of sequence length, coverage and GC% on the accuracy of redundant and non-redundant dot-plot estimators.** Comparison of errors e_{R} (redundant estimator $\widehat{D}$, in blue) and e_{NR} (non-redundant estimator $\widetilde{D}$, in orange), for $m = 1000$. To emphasize the difference between the two estimators for a single sequence, a line is drawn between the difference between e_{R} and e_{NR} is shown in color of more important error. The sequences are ordered by their length (top), coverage (middle) and %GC (bottom)

3.1.1 *For a given sample size*

Here we consider the errors observed by analyzing a sample of fixed cardinality $m = 1000$, generated using R and NR sampling. We gathered a data set of 365 sequences, extracted from the *seed* alignments of selected `RFAM` [1] families, chosen to cover a wide range of lengths and GC-contents:

- 63 sequences from RF00001 – lengths 91 to 135 nts;
- 100 sequences from RF00005 – lengths 62 to 93 nts;
- 60 sequences from RF00061 – lengths 177 to 365 nts;
- 72 sequences from RF00174 – lengths 168 to 248 nts;
- 20 sequences from RF01071 – lengths 395 to 676 nts;
- 50 sequences from RF01731 – lengths 66 to 173 nts.

We evaluated both estimators with respect to the above metrics, sampling $m = 1000$ secondary structures for each of the 365 sequences, and computing $\widehat{D}$ and $\widetilde{D}$. We also ex-

ecuted the partition function version of RNAfold to compute the exact value of D, from which we derived the values of e_{R} and e_{NR}.

The results, shown in Figure 7, reveal that, for samples of equal cardinality, the error difference $e_{\mathrm{R}} - e_{\mathrm{NR}}$ is rarely negative. This means that the non-redundant estimate $\widetilde{D}$ typically produces better approximations of the reference dot-plot than the redundant estimator $\widehat{D}$, achieving lower error values for 83.3% of the sequences, as shown by the histogram in Figure 8. Specifically, these estimations are better for higher values of coverage, and shorter sequences, the NR estimator being especially powerful on sequences that are shorter than 200 nt nucleotides. Interestingly, we did not observe any obvious relationship between the GC-content, and the efficiency of both estimators.

This demonstrates that non-redundant sampling, when combined with a non-redundant estimator, can provide more accurate estimates than using the empirical mean based on a classic redundant sample.

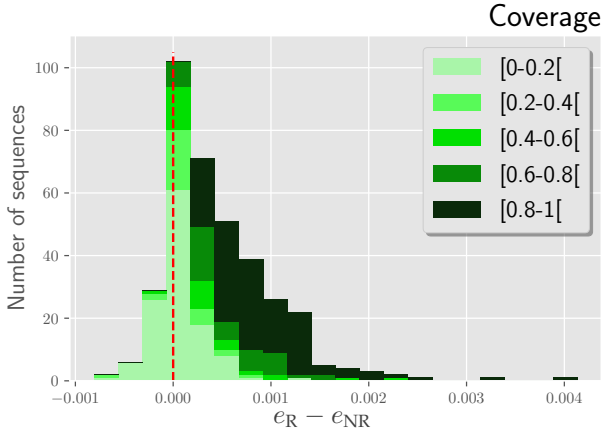


Fig. 8. **Histogram of the $e_{\mathrm{R}} - e_{\mathrm{NR}}$ difference**. Samples of a fixed cardinality $m = 1000$ are provided to both estimators for all sequences in our RFAM-based dataset.

3.1.2 *For a given execution time*

In practice, producing a NR sample requires slightly more time than the production of a redundant sample of the same cardinality. This computational overhead, which only represents a fraction of the original running time of the sampling phase, is due to the additional operations involved in maintaining the data structure and accessing its values during the sampling. For this reason, it seems more fair an assessment to compare the performances on both estimators when allocating the same amount of time to both sampling procedures.

In order to assess the evolution of error as a function of the elapsed time, we performed a detailed analysis of a subset of sequences, selected to cover a wide range of sequence length. The reduced data set includes sequences:

- **a** X06837.1/1-119 (100nt – RF00001);
- **b** M30199.1/68-167 (119nt – RF00001);
- **c** BAAU01027214.1/624-783 (160nt – RF01731).
- **d** CP000283.1/2593935-2594143 (209nt–RF00174);

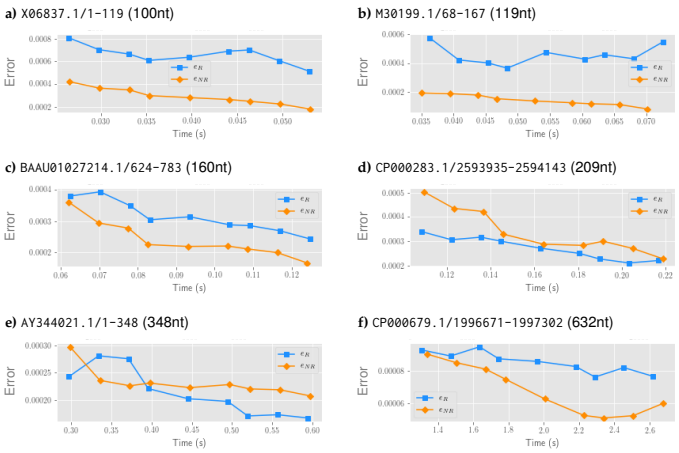


Fig. 9. **Prediction error of base-pair estimators, plotting against sampling time.** For 6 sequence spanning a large range of lengths, we report the errors of the redundant ($\widehat{F}$ – blue) and non-redundant ($\widetilde{F}$ – orange) estimators.

- **e** AY344021.1/1-348 (348nt – RF00061);
- **f** CP000679.1/1996671-1997302 (632nt – RF01071).

We first sampled a nominal number of unique structures using NR sampling, storing the final elapsed time $t^\star$. We then computed the NR estimator for each subset of structure generated after time t representing fractions of $t^\star$, with $t/t^\star \in [0, 1/8, 1/4, 1/3, 1/2, 2/3, 3/4, 7/8, 1]$. Finally, we computed the error for a redundant set of structures generated using the same time.

As can be seen in Figure 9, the NR estimator clearly outperforms its competitor in 4 out of the 6 cases, and essentially matches its competitor for one the remaining two. Surprisingly, the NR estimator yields a smaller error e_{NR} for CP000679.1/1996671-1997302, the longest among the sequences presented here. This behavior may be attributed to the natural stochasticity of statistical estimators, noting that the NR estimator gently degrades into the empirical mean when the sampled structure only represent a negligible proportion of the Boltzmann probability distribution. By contrast, the dominance of the NR estimator over the empirical mean is much clearer, and robust, within smaller sequences, where redundancy has a much greater chance to manifest itself.

3.2 Sampling distinct shapes and estimating shape probabilities

RNA Shapes are abstractions of secondary structures, initially introduced as tool for the comparative folding of RNA, developed within the Giegerich group over the course of a series of works spanning a nearly decade [23], [32]. RNA shapes represent coarse-grain representations of the classic secondary structures. However, by focusing on the high-level organization of RNA architectures, RNA Shapes are less sensitive to insertions and deletions of individual nucleotides than classic secondary structures, and can thus be used to extract recurrent conformations across homologous RNAs without having to align RNAs.

At its coarsest level, the **RNA Shapes** associated with a structure is obtained by suppressing all consecutive unpaired positions, and contracting consecutive (*aka* stacking) pairs into a single pair of matching brackets. Shapes can be represented using a notation that is analogous to the classic dot-parenthesis notation for secondary structures, using brackets instead of parenthesis. Base pairs are encoded using matching parentheses/brackets, and unpaired positions correspond to dots.

For instance, the structure S_1 in Figure 1, which consists of the set of base pairs $\{(1,12),(2,11),(3,10)\}$ admits a representation `(((.....)))` in dot-parenthesis notation, and the shape associated with S_1 is simply denoted by

$$\text{SHAPE}(S_1) = \texttt{[]}.$$

For a more complex example, the secondary structure $S^\star$ = `(((..)))..(((...)))` admits $\text{SHAPE}(S^\star)$ = `[[][]]` as its coarsest shape representation.

Notably, a single shape typically represent a large number of, structurally similar, secondary structures. While computing the overall Boltzmann probability of a shape can be done in polynomial time [33], it requires a complex and shape-specific computation. Moreover, no deterministic efficient algorithm is currently known for computing the list of shapes having (sub-)optimal Boltzmann probability, and current methods typically resort to (redundant) statistical sampling to identify promising shape candidates [23]. This suggests two contexts in which non-redundant sampling and estimator could be beneficial: i) The computation of a list of dominant shapes; and ii) The estimation of the probability of a given shape.

3.2.1 Comprehensive lists of dominant shapes

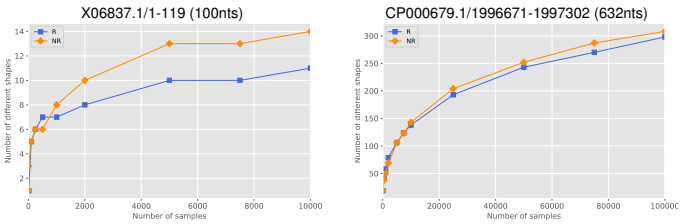


Fig. 10. **Number of different RNA shapes** populated by at least one secondary structure within samples of increasing cardinality, produced using R (blue) and NR sampling (orange).

As a straightforward application, one can use NR sampling in order to establish a more comprehensive list of shapes supported by an RNA sequence, *i.e.* shapes π such that $\text{SHAPE}(S) = \pi$ for some secondary structure S. In this context, redundancy is clearly uninformative, and using NR sampling seems natural choice.

Indeed, as can be seen in Figure 10, using NR sampling induces a sizable gain, allowing to access a more comprehensive list of shapes for smaller RNAs. This benefit decreases for longer RNAs, where redundancy seldom occurs, although some gain can still be observed. Nevertheless, even for longer RNAs, the non-linear behavior of both curves suggests a high level of redundancy at the shape level, while the superposition of curves indicates an absence of redundancy at the secondary structure level.

3.2.2 Estimating the probability of the most stable shape

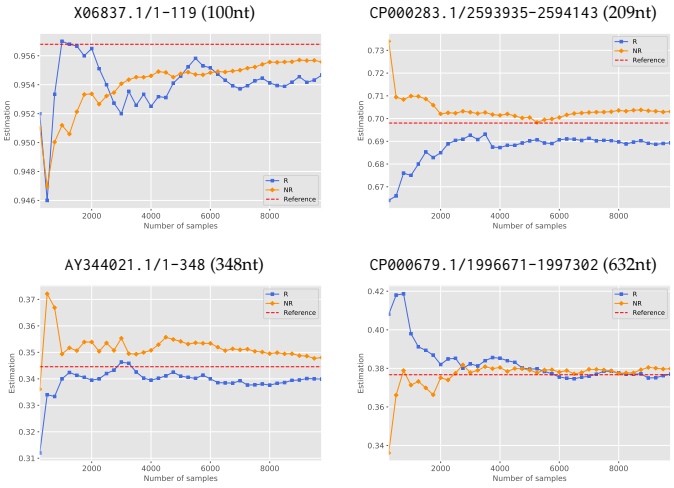


Fig. 11. **Estimation of MFE shape probability** for sample size m ranging from 250 to 10 000. Empirical mean (R) estimates $\widehat{F}$ drawn in blue, with non-redundant estimates $\widetilde{F}$ in orange. Reference value colored in red.

In this second application, we use NR sampling to estimate the Boltzmann probability associated with a given shape. While any shape can theoretically be considered, we choose to focus our attention on **the shape associated with the minimum free-energy (MFE)** secondary structure, *i.e.* the most stable structure, and also the most probable at the thermodynamic equilibrium.

To achieve this objective, we define a feature function $F_{\text{SHAPE}} : \Omega \to \mathbb{B}$ such that

$$F_{\text{SHAPE}}(S) = \begin{cases} 1 & \text{if } \in \text{SHAPE}(S) = \text{SHAPE}(\text{MFE}) \\ 0 & \text{otherwise.} \end{cases}$$

Again, it can be shown that the èxpected value of F_{SHAPE} coincides with the probability of generating a structure that admits the MFE shape as its representative.

To compare the quality of estimates, we consider sequences **a**, **d**, **e**, and **f** from our reduced dataset, and report the evolution of the R/NR estimates for the expectation of F_{SHAPE}, *a.k.a.* the MFE shape probability, for sample sizes $m \in [250, 500, \ldots, 10000]$. We used the empirical estimate computed for $\widehat{F}$ for m =1 000 000 samples as a ground truth.

The results, which can be visualized in Figure 11, reveal once again that NR estimates typically outperform the classic empirical mean. In particular, NR estimate tend to show a smoother convergence towards the reference value, as shown in Figure 12, as well as a better concentration around the ground truth for a given sample size.

3.3 Efficient approximations of graph distance

As a final illustration, we consider the **graph distance** $\text{dist}_{i,j}(S)$, *i.e.* the minimal number of pairs and backbone bonds that must be traversed to span positions i and j

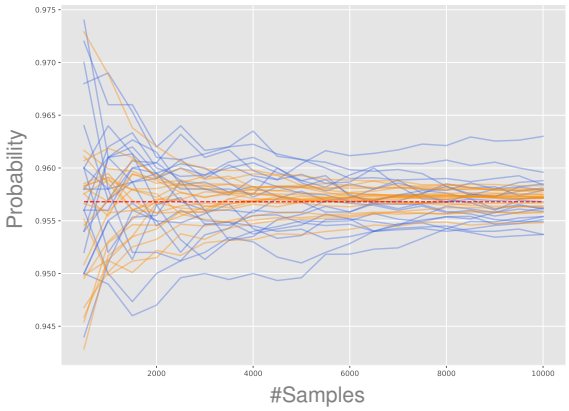


Fig. 12. **Concentrations of R (blue) and NR (orange) estimates.** For sequence `X06837.1/1-119` (100nt), 20 independent sampling of m =10 000 structures are performed using both generator, and the evolution of estimates with an increasing sample size is plotted. A red line indicates the reference probability.

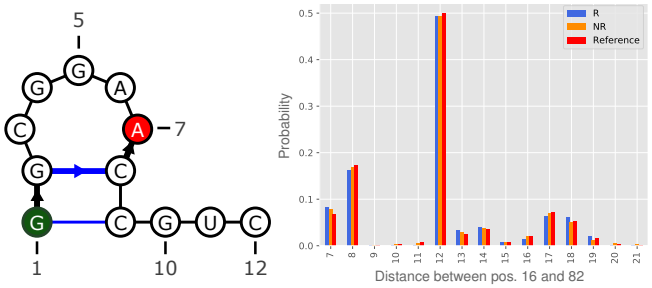


Fig. 13. **Examples of graph distance and distribution**. Left – The secondary structure S_3 induces a distance of 3 between nucleotides 1 and 7. Right – For sequence `X06837.1/1-119` in RF00001, reference (red) and estimated distributions of distances between nucleotides 16 and 82, using redundant (blue) and non-redundant (orange) sampling

within a structure S. This quantity approximates the actual geometric distance under the assumption of a unit distance between both consecutive and paired positions, and statistics of the graph distance at the thermodynamic equilibrium can be used to refine our interpretation of experimental data produced in the context of structural biology [10].

To illustrate the concept of graph distance, we consider in Figure 13 the nucleotides 1 and 7 in structure S_3. All base pairs and all transition between consecutive backbone positions are associate with a unit increment to the distance. For example, if we consider the structure $s = \{(1, 9), (2, 8)\}$, the graph distance $\text{dist}_{1,7}(s)$ is 3, because the shortest path to reach 7 from 1 is $1 \rightarrow 2 \rightarrow 8 \rightarrow 7$.

As seen in this example, the shortest path between two positions may require forward and backward (resp. inward and outward) moves within a given structure, making the computation of the distance distribution a complex task. Indeed, while an elegant and exact polynomial algorithm has been proposed by Qin *et al* [10], its complexity scales like $\mathcal{O}(n^{11})$ and, to the best of our knowledge, has never been fully implemented.

To illustrate the benefits of non-redundant sampling, we focus on sequence **a** from our reduced data (`X06837.1/1-119`), a 100nts long sequence featured in the

Experiment	#Distance classes	TVD
Reference	39	0
1 000 R samples	18	0.044
1 000 NR samples	22	0.026

Fig. 14. **Comparison of populated distance classes, and total variation distance (TVD) between reference distribution, and estimates produced using redundant and non-redundant sampling.** Non-redundant estimates are more similar to the reference distribution than the empirical mean, and populates a higher number of distance classes.

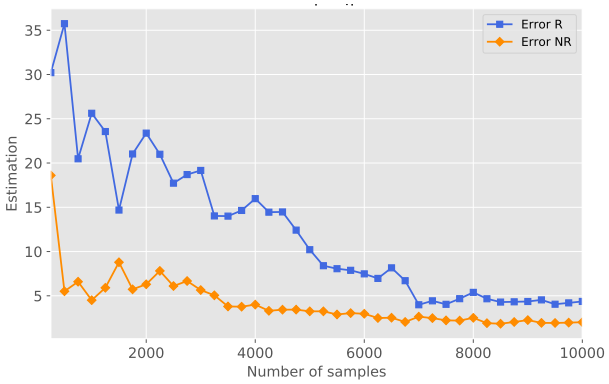


Fig. 15. **Errors for the estimator on redundant and non-redundant samples**. For sequence sequence `X06837.1/1-119`, the error of the non-redundant estimator is marked in orange, while that of redundant one is colored blue. The larger variance of the redundant estimator/samples visibly results in bigger deviations than the non-redundant strategy.

seed alignment of the RFAM family RF00001. Again, we use as a reference (ground truth) the distributions/expectations obtained using the empirical mean for a sample of 1 000 000 structures.

3.3.1 *Graph-distance distribution for a pair of nucleotides*

In a first experiment, we estimate the distribution of the graph distance between the nucleotides 16 and 82. We generate 1 000 samples using both the R and NR generators, and estimate the probability of being at any given distance, *i.e.* the joint expectations, for any distance $d \leq n$, of the feature functions

$$F_d(S) = \begin{cases} 1 & \text{if } \in \text{dist}_{16,82}(S) = d \\ 0 & \text{otherwise.} \end{cases}$$

The results are shown on Figure 13, and more details can be found in Figure 14. From these results, we see that, although both estimated distributions achieve a substantial similarity with the reference, the distribution induced by non-redundant sampling features is more precise, as assessed by the total variation distance. It also populates a higher number of graph distances, implying the presence of more diverse structures in the non-redundant sample. We conclude that non-redundant sampling, combined with our NR estimator, represents a better option to estimate the graph-distance distribution for a set of structures.

3.3.2 Expected distance for all pairs

In this final application, we want to estimate M a matrix of size $n \times n$ whose cells contain the expected graph-distances of each pair of nucleotides, defined as

$$M = (M_{i,j})_{1 \leq i \leq j \leq n} \quad \text{where} \quad M_{i,j} := \sum_{s \in \Omega} d_{i,j}(S) \times \mathbb{P}(s).$$

Formally, we simply need to introduce specific features

$$M_{i,j}(S) = d_{i,j}(S)$$

for all $1 \leq i \leq j \leq n$.

We consider sequence **a** of our reduced data set, leading to the estimation of 100×100 upper-triangular matrices. We evaluate both estimators for the introduced features, using the following error function:

$$e_{\text{R}} = \|M^\star - \widehat{M}\| \qquad \text{and} \qquad e_{\text{NR}} = \|M^\star - \widetilde{M}\|.$$

where $M^\star$ is the reference matrix of expected distances, estimated from a sample of 1 000 000 structures.

Figure 15 summarizes the evolution of e_{R} and e_{NR} for samples of increasing size ($m \leq 10000$). We can observe a much earlier convergence of the NR sampler/estimator than the empirical mean. This behavior quite clearly illustrates the relevance of the demonstrably lower variance achieved by our strategy.

4 On the Right Number of Structures

When using sampling to estimate statistical properties of RNA structures at the thermodynamic equilibrium, a recurrent – crucial – question is to choose the number of generated samples as to produce accurate estimates. Historically, and in many subsequent works, a sample size of 1 000 structures has been proposed [11], somewhat irrespectively of the context. However, such a *one size fits all* may not yield accurate, or reasonably reproducible results, motivating the probabilistic analysis below.

Before stating our recommendations, we need to remind the crucial concept of **confidence interval**. In general, a sample size needs to be chosen in order to achieve a desired level of precision. However, since the process of sampling is stochastic in nature, it is impossible to unconditionally guarantee a given precision since, out of the possible sequences of generated structures, some may typically induce arbitrarily large errors. For instance, in the running example of Figure 1, redundant sampling may generate structure S_{84} m times, leading to an (erroneous) estimated probability of 1 for base pairing the first nucleotide. However, this scenario has an abysmal probability, lower than 10^{-9m}, so one needs to adopt a **confidence intervals** perspective, considering the trade-off between the precision and how often this precision is achieved while estimating from a random sample.

In the case of the R estimator, the empirical mean is essentially a sum of independent variables, meaning that classic concentration inequalities contributed by the field of probability theory, can be used with minimal modifications.

Tolerated Error	Frequency within tolerance 90%	95%	99%
$\varepsilon = 20\%$	37	46	66
$\varepsilon = 10\%$	150	184	265
$\varepsilon = 5\%$	599	738	1 060
$\varepsilon = 2.5\%$	2 397	2 951	4 239
$\varepsilon = 1\%$	14 979	18 444	26 492
$\varepsilon = 5‰$	59 915	73 778	105 966
$\varepsilon = 1‰$	1 497 866	1 844 440	2 649 159

Fig. 16. **Recommended number of samples for estimating equilibrium probabilities (boolean features).** For instance, to ensure that the estimate falls within 1% of the true value for 95% of the runs, a large number of m =18 444 structures should be generated.

In particular, the **Hoeffding inequality** implies that, for any feature F:

$$\mathbb{P}\left(|\widehat{F}(\mathbf{S}) - \mathbb{E}(F(S))| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2m\varepsilon^2}{c}\right), \quad (9)$$

where ε is a tolerated level of error, $\mathbf{S}$ is a random sample of size m and $c := (\max_S(F(S)) - \min_S(F(S)))^2$. Note that when a feature function only takes values 0 or 1, as in many of our experiments, then one has $c = 1$. Equation (9) can be used to build a **confidence interval** at level $(1-\alpha)$, for any value $\alpha \in [0,1]$:

$$\left[\widehat{F}(\mathbf{S}) - \sqrt{\frac{c}{2m}\log\left(\frac{2}{\alpha}\right)}, \widehat{F}(\mathbf{S}) + \sqrt{\frac{c}{2m}\log\left(\frac{2}{\alpha}\right)}\right].$$

This means that, over multiple executions of the R sampling/estimation, at least a fraction $(1-\alpha)$ of the runs will feature an error smaller than $\sqrt{\frac{c}{2m}\log\left(\frac{2}{\alpha}\right)}$. This function can be inverted numerically to estimate the number m of samples that achieve an error bounded by ε at least $(1-\alpha)$ of the times.

We report in Figure 16 some typical sample sizes required to achieve a given precision with reasonable probability when estimating probabilities (*i.e.* expectations of 0/1-valued features). For instance, to reach a **90% chance** of estimating a base pair probability **within 0.5% of its true value**, a total of **59 915 structures** should be generated.

By contrast, the **1 000 structures** usually considered in the literature will guarantee a value **within 3%** of the true probability **only 2/3 of the times**, although this sample size will **almost always** (99%) return estimates **within 5%** of the correct value. Finally, the formula can be used to more complex features, taking values in a wider range. For instance, to compute the **expected distance** for sequence **a** (100 nts $\rightarrow c = 99$), a **sample of 263 structures** will produce an estimated distance **within one step from the true value** in more than **99% of executions**.

Due to its lower variance, our NR sampling/estimator achieves strictly better estimate qualities for a given sample size, but a refined analysis would be much more challenging due to the dependence of consecutive samples. Therefore, we recommend sampling the same number of structures as described above for the R sampling, but expect more accurate results.

5 Conclusion

In this work, we have described an algorithm for the non-redundant sampling of secondary structures in the Boltzmann ensemble, using algorithmic principles introduced by some of the authors [15], [16]. This algorithm was implemented in the Vienna RNA package [12] as an extension to the `RNAsubopt` tool. While the non-redundant sampling allows to produce more diverse samples, it induces dependencies between structures generated during the sampling, forbidding the use of classic estimators, such as the empirical mean. We have thus introduced a statistical estimator for non-redundant sampling which is unbiased, consistent and easy to compute. By exploiting an explicit knowledge of the emission probability of structures within the Boltzmann distribution, our new estimator produces higher-quality estimates (lower variance) than classic estimators based on redundant samples of the same cardinality. Our non-redundant sampler and estimator achieve better estimates for various quantities of interest at the thermodynamic equilibrium. We concluded our study with recommendations regarding the sample size to achieve reproducibility.

While this work describes specific applications of non-redundant sampling to RNA bioinformatics, its scope of application is much wider. Boltzmann-Gibbs distributions are quite frequent in Bioinformatics [34], [22], [35], and could be explored to study the stability of predictions in any context where unambiguous dynamic-programming schemes [9], [36] exist. Our novel estimator can then be used without any modifications, for instance to estimate the diversity of near-optimal solutions.

Acknowledgments

The authors are greatly indebted to Ivo Hofacker for pointing out a critical flaw in an earlier version of this work, motivating the search and discovery of a specific estimator from non-redundant samples. This work was part of the `RNALands` project, jointly supported by the French *Agence Nationale de la Recherche* (ANR-14-CE34-0011) and the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung* (I 1804-N28).

References

[1] I. Kalvari, E. P. Nawrocki, J. Argasinska, N. Quinones-Olvera, R. D. Finn, A. Bateman, and A. I. Petrov, "Non-coding RNA analysis using the RFAM database." *Current protocols in bioinformatics*, vol. 62, p. e51, Jun. 2018.

[2] L.-L. Zheng, J.-H. Li, J. Wu, W.-J. Sun, S. Liu, Z.-L. Wang, H. Zhou, J.-H. Yang, and L.-H. Qu, "deepBase v2. 0: identification, expression, evolution and function of small RNAs, LncRNAs and circular RNAs from deep-sequencing data," *Nucleic acids research*, vol. 44, no. D1, pp. D196–D202, 2015.

[3] J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," *Biopolymers: Original Research on Biomolecules*, vol. 29, no. 6-7, pp. 1105–1119, 1990.

[4] P. Clote, Y. Ponty, and J.-M. Steyaert, "Expected distance between terminal nucleotides of RNA secondary structures." *Journal of mathematical biology*, vol. 65, pp. 581–599, Sep. 2012.

[5] E. Freyhult, V. Moulton, and P. Clote, "RNAbor: a web server for RNA structural neighbors." *Nucleic acids research*, vol. 35, pp. W305–W309, Jul. 2007.

[6] R. Lorenz, C. Flamm, and I. L. Hofacker, "2d projections of RNA folding landscapes," in *German conference on bioinformatics 2009*. Gesellschaft für Informatik eV, 2009.

[7] J. Waldispühl, S. Devadas, B. Berger, and P. Clote, "RNAmutants: a web server to explore the mutational landscape of RNA secondary structures." *Nucleic acids research*, vol. 37, pp. W281–W286, Jul. 2009.

[8] I. Miklós, I. M. Meyer, and B. Nagy, "Moments of the boltzmann distribution for RNA secondary structures." *Bulletin of mathematical biology*, vol. 67, pp. 1031–1047, Sep. 2005.

[9] Y. Ponty and C. Saule, "A combinatorial framework for designing (pseudoknotted) RNA algorithms," in *Algorithms in Bioinformatics*, ser. LNBI, M.-F. S. T. Przytycka, Ed. Springer, Jan. 2011, no. 6833, pp. 250–269. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-23038-7_22

[10] J. Qin, M. Fricke, M. Marz, P. F. Stadler, and R. Backofen, "Graph-distance distribution of the boltzmann ensemble of RNA secondary structures." *Algorithms for molecular biology : AMB*, vol. 9, p. 19, 2014.

[11] Y. Ding and C. E. Lawrence, "A statistical sampling algorithm for RNA secondary structure prediction." *Nucleic acids research*, vol. 31, pp. 7280–7301, Dec. 2003.

[12] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "ViennaRNA package 2.0." *Algorithms for molecular biology : AMB*, vol. 6, p. 26, Nov. 2011.

[13] J. S. Reuter and D. H. Mathews, "RNAstructure: software for RNA secondary structure prediction and analysis." *BMC bioinformatics*, vol. 11, p. 129, Mar. 2010.

[14] N. R. Markham and M. Zuker, "UNAFold: software for nucleic acid folding and hybridization." *Methods in molecular biology (Clifton, N.J.)*, vol. 453, pp. 3–31, 2008.

[15] W. A. Lorenz and P. Clote, "Computing the partition function for kinetically trapped RNA secondary structures." *PloS one*, vol. 6, p. e16178, Jan. 2011.

[16] J. Michálik, H. Touzet, and Y. Ponty, "Efficient approximations of RNA kinetics landscape using non-redundant sampling." *Bioinformatics (Oxford, England)*, vol. 33, pp. i283–i292, Jul. 2017.

[17] J. Waldispühl and Y. Ponty, "An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 18, pp. 1465–1479, 2011.

[18] M. Kucharík, I. L. Hofacker, P. F. Stadler, and J. Qin, "Basin hopping graph: a computational framework to characterize RNA folding landscapes." *Bioinformatics (Oxford, England)*, vol. 30, pp. 2009–2017, Jul. 2014.

[19] S. Pei, J. S. Anthony, and M. M. Meyer, "Sampled ensemble neutrality as a feature to classify potential structured RNAs." *BMC genomics*, vol. 16, p. 35, Feb. 2015.

[20] A. Spasic, S. M. Assmann, P. C. Bevilacqua, and D. H. Mathews, "Modeling RNA secondary structure folding ensembles using SHAPE mapping data." *Nucleic acids research*, vol. 46, pp. 314–323, Jan. 2018.

[21] S. Washietl, I. L. Hofacker, P. F. Stadler, and M. Kellis, "Rna folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction." *Nucleic acids research*, vol. 40, pp. 4261–4272, May 2012.

[22] V. Reinharz, Y. Ponty, and J. Waldispühl, "A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotide distribution." *Bioinformatics (Oxford, England)*, vol. 29, pp. i308–i315, Jul. 2013.

[23] B. Voss, R. Giegerich, and M. Rehmsmeier, "Complete probabilistic analysis of RNA shapes." *BMC biology*, vol. 4, p. 5, Feb. 2006.

[24] D. Gardy and Y. Ponty, "Weighted random generation of context-free languages: Analysis of collisions in random urn occupancy models," in *GASCOM - 8th conference on random generation of combinatorial structures - 2010*. Montréal, Canada: LACIM, UQAM, Sep. 2010, p. 14pp. [Online]. Available: https://hal.inria.fr/inria-00543150

[25] J. Du Boisberranger, D. Gardy, and Y. Ponty, "The weighted words collector," in *AOFA - 23rd International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms - 2012*, ser. DMTCS Proceedings, Nicolas, Broutin, Luc, and Devroye, Eds., vol. AQ. Montreal, Canada: DMTCS, Jun. 2012, pp. 243–264. [Online]. Available: https://hal.inria.fr/hal-00666399

[26] W. A. Lorenz and Y. Ponty, "Non-redundant random generation algorithms for weighted context-free grammars," *Theoretical Computer Science*, vol. 502, pp. 177–194, 2013.

[27] D. H. Turner and D. H. Mathews, "NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure." *Nucleic acids research*, vol. 38, pp. D280–D282, Jan. 2010.
[28] M. Zuker and D. Sankoff, "RNA secondary structures and their prediction," *Bulletin of mathematical biology*, vol. 46, no. 4, pp. 591–621, 1984.
[29] S. Will, K. Reiche, I. Hofacker, P. Stadler, and R. Backofen, "Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering." *PLoS Comput. Biol.*, vol. 3, p. e65, Apr. 2007.
[30] J. N. Zadeh, B. R. Wolfe, and N. A. Pierce, "Nucleic acid sequence design via efficient ensemble defect optimization," *Journal of Computational Chemistry*, vol. 32, no. 3, pp. 439–52, 2011.
[31] J. Michalik, "Non-redundant sampling in RNA bioinformatics," Ph.D. dissertation, Université Paris-Saclay, Apr. 2019.
[32] S. Janssen and R. Giegerich, "The RNA shapes studio." *Bioinformatics (Oxford, England)*, vol. 31, pp. 423–425, Feb. 2015.
[33] ——, "Faster computation of exact RNA shape probabilities." *Bioinformatics (Oxford, England)*, vol. 26, pp. 632–639, Mar. 2010.
[34] M. Vingron and P. Argos, "Determination of reliable regions in protein sequence alignments." *Protein engineering*, vol. 3, pp. 565–569, Jul. 1990.
[35] E. Jacox, C. Chauve, G. J. Szöllősi, Y. Ponty, and C. Scornavacca, "ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony." *Bioinformatics (Oxford, England)*, vol. 32, pp. 2056–2058, Jul. 2016.
[36] C. Chauve, J. Courtiel, and Y. Ponty, "Counting, generating, analyzing and sampling tree alignments," *International Journal of Foundations of Computer Science*, vol. 29, no. 05, pp. 741–767, 2018.



Christelle Rovetta has been a postdoc at LRI (Univ. Paris-Saclay) and at LIX (Ecole Polytechnique). She received her Ph.D. at Ecole Normale Supérieure de Paris, France in June 2017, following Master of Science degree in Applied Mathematics and another M.Sc. in Computer Science. Her main research interests are Simulation and Algorithms for Markov Chains analysis, and RNA secondary structure.



Juraj Michalik Juraj Michalik is a bioinformatician, who defended a PhD in computer science from Ecole Polytechnique, France, following initial engineering studies at Institut National des Sciences Appliquées in Lyon, France. He currently holds a postdoc position at Institute of Molecular Biology, Czech Republic, where he studies the dependency of T-cell receptor sequences, their type and pathogen affinity.



Ronny Lorenz Ronny Lorenz is a bioinformatician who received his PhD in molecular biology in 2014. He currently holds a University Assistant position at the Department of Theoretical Chemistry, and is working on various aspects of RNA secondary structure prediction algorithms. Since 2010, he is the leading developer of the ViennaRNA Package. His main research interests are algorithmic aspects of RNA secondary structure prediction and the development of novel methods in RNA folding kinetics prediction.



Andrea Tanzer Andrea Tanzer is an Elise Richter fellow at the Medical University of Vienna at the Center for Anatomy and Cell Biology. She holds an MSc in Biology/Genetics from the University of Vienna and a PhD in Computer Science/Bioinformatics from the University of Leipzig. She was a member of ENCODE pilot and phase2, and Austrian PI of the RNALands consortium. Her research focuses on RNA bioinformatics and genomics, including big data analysis in transcriptomics, RNA folding kinetics, ncRNA detection/annotation, RNA structure prediction in vertebrates, plants and viruses, RNA G-quadruplex analysis and epitranscriptomics.



Yann Ponty Yann Ponty is a tenured research scientist since 2009 at the French center for scientific research (CNRS), based at Ecole Polytechnique, where he currently leads the AMIBio research group in computational biology. He received his PhD in computer science from Université Paris-Saclay, and has held postdoctoral positions at Boston College and Sorbonne University. His main research interests include Discrete Mathematics and Algorithms applied to Bioinformatics research, where he has contributed more than 50 manuscripts in journals and international conferences.