Journal of Mathematical Biology manuscript No.
(will be inserted by the editor)

Yann Ponty

Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy

The boustrophedon method

the date of receipt and acceptance should be inserted later

Abstract We adapt here a surprising technique, the boustrophedon method, to speed up the sampling of RNA secondary structures from the Boltzmann low-energy ensemble. This technique is simple and its implementation straight-forward, as it only requires a permutation in the order of some operations already performed in the *stochastic traceback* stage of these algorithms. It nevertheless greatly improves their worst-case complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n\log(n))$, for n the size of the original sequence. Moreover the average-case complexity of the generation is shown to be improved from $\mathcal{O}(n\sqrt{n})$ to $\mathcal{O}(n\log(n))$ in an Boltzmann-weighted homopolymer model based on the Nussinov-Jacobson free-energy model. These results are extended to the more realistic Turner free-energy model through experiments performed on both structured (Drosophilia melanogaster mRNA 5S) and hybrid (Staphylococcus aureus RNAIII) RNA sequences, using a boustrophedon modified version of the popular software `UnaFold`. This improvement allows for the sampling of greater and more significant sets of structures in a given time.

Keywords Statistical sampling, Boltzmann free-energy ensemble, RNA structure, MFE folding

1 Introduction

1.1 Motivation

To decypher the mechanisms underlying the three dimensional folding of biopolymers is one of the great challenges of the post-genomic era. Indeed, understanding the sequence/structure relationship for these entities is a first

Biology Department, Higgins Hall 577, Boston College
140 Commonwealth Avenue
Chestnut Hill, MA 02467
E-mail: Yann.Ponty@lri.fr

step toward an automated residue-level interpretation of known interactions, which would in turn allow for an algorithmic prediction of such interactions. Such a knowledge could also lead to the computer-assisted design of more specific drugs for known diseases. In the specific context of RNA, the minimum free energy (MFE) paradigm, which states that a single stranded polymer will adopt the conformation of lowest free-energy, is at the core of successful algorithmic approaches for the prediction of RNA secondary structures [18,26,43]. These approaches are based on the nearest-neighbor model for RNA [19], which approximates the free-energy of a conformation as a sum of individual contributions, associated with each elementary loops. From such a mathematically simple model, a dynamic programming scheme allows for the exploration of the conformational landscape of a sequence in $\Theta(n^3)$ time, eventually retrieving the MFE secondary structure. Taking advantage of experimentally resolved values for the loops energies, this method has proven accurate in predicting the actual base-pairings. Recent studies claim an average 73% accuracy for these methods when applied to sequences of size up to 700 nucleotides [27].

1.2 Minimal free energy folding: Paradigm shift

However certain structures, like for instance the Natronobacterium pharaonis tRNA for alanine, remain badly mispredicted, although most always found among the suboptimals [40]. This phenomenon can be explained by one or more of the following reasons:

- The energy parameters used in the various implementations may still be lacking some accuracy. In particular, non-canonical interactions [20] are so far not taken into account by the actual model. These interactions, although poorly contributing energetically, are known to play a stabilizing role [21] neither fully understood nor modeled yet.
- For computational reasons[1], the conformational landscape explored by state-of-the-art algorithms are restricted to secondary structures only. This may favor a conformation whose best planar restriction (secondary structure) is promising over another conformation with lower free-energy, but whose best planar restriction has higher free-energy than that of the first.
- *In vivo*, a structured RNA is seldom alone and often complexed, potentially both to RNAs and proteins as can be seen in the ribosome. Furthermore, evidences have shown that some RNAs can adopt two or more alternative fully-functional conformations, as in the case of riboswitches [2, 33].

All these putative explanations plead for a focus on sub-optimal structures, so we actually experience a shift from the MFE paradigm toward the study of ensemble features (see Ding's review [7]). In this novel view of the

[1] The problem of RNA folding with general pseudo-knots has been shown to be NP-complete even in a toy version of the nearest-neighbor model [24]. However, this constraint might be practically lifted in the near future, as novel parametric approaches [42] are being developed to work around this issue.

RNA folding problem, each structure S compatible with a given sequence ω is associated with its so-called Boltzmann probability, defined such that

$$\mathbb{P}(S|\omega) = \frac{e^{\frac{-E(S)}{RT}}}{Z_\omega}$$

where $E(S)$ is the free-energy of S, R is the universal gas constant, T is the temperature in Kelvin and $Z_\omega = \sum_s e^{\frac{-E(s)}{RT}}$ is the value of the partition function for this sequence. A random set of structures compatible with the input RNA sequence, the *Boltzmann ensemble of low-energy*, and weighted with Boltzmann probabilities can then be sampled. By contrast with the suboptimals approach, for which the algorithm takes time in $\Theta(kn^3)$ for k the number suboptimals, the sampling approach relies on a stochastic traceback scheme, which can be performed in $\mathcal{O}(n^2)$ after a $\Theta(n^3)$ precomputation. Various features of this ensemble have been shown to be good separators for known classes of structural RNA [10]. Some of these features also yield promising results for the detection of riboswitches. While this new approach has already given some promising results, there is still a potential for both algorithmic and statistical improvements.

1.3 Boltzmann ensembles: Why size matters

In particular, we claim that larger Boltzmann ensembles than those of size 1000 actually used [9,10,15,36] could lead to an increase on the sensibility of the current methods. Specifically, it is unclear how to determine a generally valid, statistically sufficient size for the ensembles. In a special case involving RNA Shape [38], an ensemble of size 1000 is justified by assuming a Poisson distribution on the number of occurrence of each shape in a sampled ensemble. Namely, it ensures a deviation less than 10% with 95% confidence. Other works on this subject [9,10,15] also used this number of 1000 in radically different contexts, without neither formulating an hypothesis on the distribution for their parameters nor satisfying the conditions of Voß *et al* parameters [38]. As the numbers of RNA structures [39] and RNA shapes [23] are known to grow exponentially along with the sequence size, it would be very surprising that sampled sets of fixed cardinality actually covers the alternative conformations for larger sequences.

For instance, let us focus on the Natronobacterium pharaonis tRNA for alanine (GenBank: AB003409.1/96–167), an RNA that once got some attention from Steffen et al [36]. Its structure is known to be either mispredicted by MFE techniques or very far from the consensus cloverleaf shape generally admitted for tRNA. Running the `RNASubopt` software [18] on this RNA yields 153 suboptimals structure before the first cloverleaf structure is found. Although not necessarily the native one, its free-energy distance to the MFE structure is already of 4.2kcal/mol. Running `RNAShapes` in *statistical analysis* mode [38] on the same tRNA yields a cumulated probability of 0.0012534 for the native cloverleaf shape `[[][]]`. We conclude that a Boltzmann low-energy ensemble of size 1000 for this tRNA will **neither contain the native structure nor any other cloverleaf-like structure**

$\frac{(1-0.0012534)^{1000}}{100} = 28\%$ **of the time**. This calls for an increase of the number of samples in the sampled ensemble, as the frequency of the previous failure scenario drops to 1% for a Boltzmann ensemble of size 3672.

However, the average-case and worst-case complexities of the sampling are respectively in $\Theta(n\sqrt{n})$ and $\Theta(n^2)$, as will be shown later in this document. So the number of samples that can be afforded without significantly slowing down the whole $\Theta(n^3)$ algorithm is in $\mathcal{O}(n\sqrt{n})$. Being able to generate $\Theta(n^2)$ samples while using the same computational ressources could then provide a significant improvement of these methods.

1.4 Plan of the paper

At first we transpose the statistical sampling algorithm to the Nussinov-Jacobson model for free-energy. This simple model allows for a clear definition of the cost associated with a sampling scenario. Such a simplification first allows us to isolate and explain the $\Theta(n^2)$ worst-case complexity already pointed out by previous works [10,25]. Then generating function techniques are used to derive a $\Theta(n\sqrt{n})$ complexity for the average-case complexity.

Then we present the boustrophedon approach in the context of the random generation of decomposable objects. We then propose a transposition to the statistical sampling of RNA structures in the Boltzmann ensemble of low-energy. We show that its implementation reduces the worst-case complexity from $\Theta(n^2)$ to $n\log(n)$.

Lastly, we present a *proof of concept* for the boustrophedon optimization. We applied the boustrophedon modification to the `Unafold` software, a state-of-the-art implementation of statistical sampling based on Turner model, and compared the complexities of the original and modified version. Results are consistant with the complexities derived in the Nussinov-Jacobson model, although the average-case complexity of the non-boustrophedon approach even seems to scale like $\mathcal{O}(n^2)$.

2 Sequential sampling from the Boltzmann ensemble: Complexities

2.1 Statistical sampling in the Nussinov-Jacobson model

We quickly introduce the principles of sampling from the Boltzmann ensemble of low energy associated with a given sequence ω under the Nussinov-Jacobson model [32]. We claim that this simplified model is sufficiently expressive to derive algorithmic properties of state-of-the-art implementations [8,26] while avoiding some details that increase the mathematical complexity of such an analysis.

In the original version of the **Nussinov-Jacobson model** [32], the free-energy $E(S)$ of a secondary structure S is defined such that $E(S) = -|bp(S)|$, where $bp(S)$ is the set of base-pairs in S.

The **partition function** $Z_{[i,j]}$ for an interval $[i,j] \subseteq [1,n]$ is such that

$$Z_{[i,j]} = \sum_{S \in \mathcal{S}_{i,j}} e^{\frac{-E(S)}{RT}} \tag{2.1}$$

where $S_{i,j}$ is the set of secondary structures compatible with the interval $[i,j]$, R is the universal gas constant and T is the temperature.

A secondary structure, defined as a set of base-pairing positions, is said to be **compatible** with a sequence when two bases involved in a base pair are **complementary** (`G-C`, `A-U` or `G-U`), thus forming *canonical* interactions, a minimal distance θ is ensured between two base-pairing bases. Finally, it is required that the base-pairs do not form pseudoknots, i.e. are either non-overlapping or in inclusion relationship pairwise.

The classical RNA folding algorithm in the Nussinov-Jacobson model relies on a clever exploration of the conformational space. It states that, for a subsequence $[i,j]$ of the original one, the first[2] base i can either be unpaired, or paired to a base k at distance greater than θ.

Then, a transposition of the previous decomposition allows for the recursive computation of the partition function during a **precomputation stage**. Namely , it can be shown [28] that $Z_{[i,j]}$ obeys the following recurrence

$$Z_{[i,j]} = \begin{cases} Z_{[i+1,j]} + \displaystyle\sum_{k=i+1+\theta}^{j} e^{\frac{1}{RT}} \lambda(\omega_i, \omega_k) Z_{[i+1,k-1]} Z_{[k+1,j]} & \text{If } j-i > \theta \\ 1 & \text{Otherwise} \end{cases} \tag{2.2}$$

where θ is the minimal number of unpaired bases in terminal loops, and $\lambda(b,b')$ is the function that returns 1 if b and b' can form a canonical base-pair, and 0 otherwise. To our opinion, the beauty of such an approach lies in the fact that, thanks to the independence of conformation in intervals $[i+1,k-1]$ and $[k+1,j]$, and to the additivity of the energy functions, an exponential number of contributions (i.e. Boltzmann probabilities for compatible secondary structures) can be summed in a polynomial $\Theta(n^3)$ time.

Once these values are available, it is then possible to perform **stochastic traceback**, by *inverting* the recurrence in order to sample structures according to their Boltzmann probability. Namely, it will consist in, starting from the interval $[1,n]$, choosing one of eligible decompositions according to suitable probabilities, i.e. w.p. proportional to the contributions of the different decompositions in the sum of the equation 2.2. After such a step, it is known whether or not the first base i in the general case $[i,j]$ is paired or not. If the base is paired, it has been determined to which valid base k it is paired. At that specific point, an order has to be chosen in order to investigate potential values for k. We will call **sequential strategy** the exploration of candidate values for k in the sequential order

$$(i+\theta+1) \to (i+\theta+2) \to (i+\theta+3) \to \ldots \to (j-2) \to (j-1) \to (j).$$

[2] Some versions of this historical algorithm alternatively focus on the last base. The explored landscape is then exactly equivalent, and the algorithm based on this alternative decomposition yields identical results.

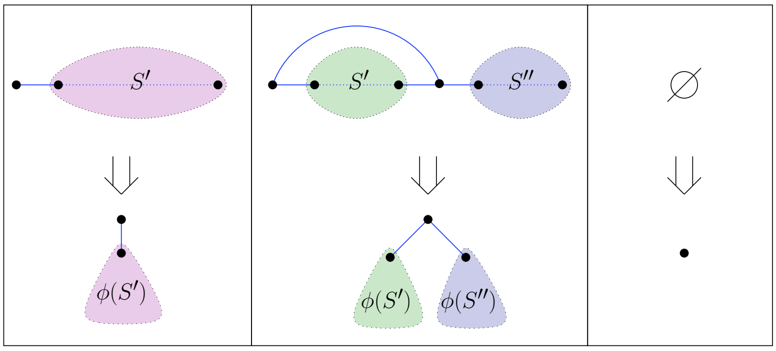


Fig. 1 Transform between RNA secondary structures and unary/binary trees, interpreted as stochastic traceback scenarios.

Until further notice, we will assume the sampling algorithms to implement such an order.

Independently from the strategy used to choose k, the process is then iterated on the remaining interval $[i+1, j]$ in the unpaired case, or on intervals $[i+1, k-1]$ and $[k+1, n]$, until intervals of size less than θ are encountered. In this case an empty structure is issued as these bases cannot by definition form base pairs at a distance less than θ. The emission probability of a structure S compatible with the original RNA sequence ω can then be shown to be equal to its Boltzmann probability

$$\mathbb{P}(S|\omega) = \frac{e^{\frac{-E(S)}{RT}}}{Z_{[1,n]}}.$$

2.2 Tree representation of a random sampling scenario

An insightful way to analyze the complexity of the traceback stage consists in drawing an unary/binary tree $\phi(S)$ associated with each sampled structure S. It can be built in the following way, also summarized in Figure 1, from a secondary structure S:

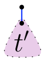
- Empty: If S is the empty structure, then $\phi(S)$ is a simple leaf.
- Unpaired case: If the first base is unpaired, $S = \bullet S'$, then the tree $\phi(S)$ is an unary node, whose unique child is the unary/binary tree $\phi(S')$ associated with S'.
- Paired case: If the first base is paired with another base, $S = \texttt{[}\,S'\,\texttt{]}\,S''$, return the tree starting with a binary node, whose left (resp. right) child is the unary/binary tree associated with S' (resp. S'').

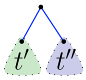
This representation was previously introduced in [30] in order to perform a very clever analysis of the *order* of RNA secondary structures, a parameter introduced by Waterman [39].

We will assume that, while sampling from an interval $[i, j]$, the unpaired case is investigated first, followed by the potential values for k in increasing order, starting from $i + \theta + 1$ all the way to j, according to the **sequential strategy** introduced hereabove. We claim that the complexity, expressed in term of number of comparisons, of a stochastic traceback scenario resulting in a given structure S can be computed directly from its tree-representation $\phi(S)$.

Theorem 21 *Let $S \in \mathcal{S}$ be a secondary structure whose tree representation is $\phi(S) \in \mathcal{T}$. Let $c : \mathcal{T} \to \mathbb{N}$ be a cost-function that associates to each tree $t \in \mathcal{T}$ a cost $c(t)$ such that*

$$c(t) = \begin{cases} 0 & \text{If } t = \bullet \\ 1 + c(t') & \text{If } t = \\ 1 + |t'| - \theta + c(t') + c(t'') & \text{If } t = \end{cases}$$





where $|t'|$ is the number of edges in t'. Let $K(S)$ the complexity of the generation of S, expressed in term of comparisons. Then

$$K(S) = c(\phi(S)).$$

Proof First, it be can easily shown that the image $\phi(S)$ of an RNA structure S on an interval $[i, j]$ is such that $|\phi(S)| = |S| = j - i + 1$ where $|\phi(S)|$ stands for the number of edges in $\phi(S)$.

We point out that the *terminal* case $t = \bullet$ corresponds to the generation of the structure having size 0, whose cost is null by definition. Then let us assume, as an inductive hypothesis, that for any RNA structure S over an interval $[i, j]$ such that $|S| \equiv j - i + 1 \leq n$, the equality $K(S) = c(\phi(S))$ holds. Let us then consider an RNA structure S' over $[i', j']$ such that $|S'| = n + 1$. One of the two cases above arises, depending on the pairing status of the first base i':

- *Unpaired:* As described previously, this case is investigated first, so it requires only one comparison in addition to the generation of the structure S'' from $[i'+1, j']$ and thus $K(S') = 1 + K(S'')$. As the image of S' through ϕ is an unary node whose only child is $\phi(S'')$, then $c(\phi(S')) = 1 + c(\phi(S''))$. Because the size of $[i' + 1, j']$ equals to n, we can apply the induction hypothesis to prove that $K(S'') = c(\phi(S''))$, thus
$$K(S') = 1 + K(S'') = 1 + c(\phi(S'')) = c(\phi(S')).$$
- *Paired to a base k':* This case has been investigated and chosen after $k' - i' - \theta$ comparisons. It remains to add the cost of the recursive generation of the structures S'' and S''', respectively on intervals $[i' + 1, k' - 1]$ and $[k' + 1, j']$, thus $K(S') = k' - i' - \theta + K(S'') + K(S''')$. As $|S''| = |\phi(S'')| = k' - i' - 1$, $K(S'') = c(\phi(S''))$ and $K(S''') = c(\phi(S'''))$ (induction hypothesis), then
$$\begin{aligned} K(S') &= k' - i' - \theta + K(S'') + K(S''') \\ &= 1 + |\phi(S'')| - \theta + c(\phi(S'')) + c(\phi(S''')) \\ &= c(\phi(S')). \end{aligned}$$

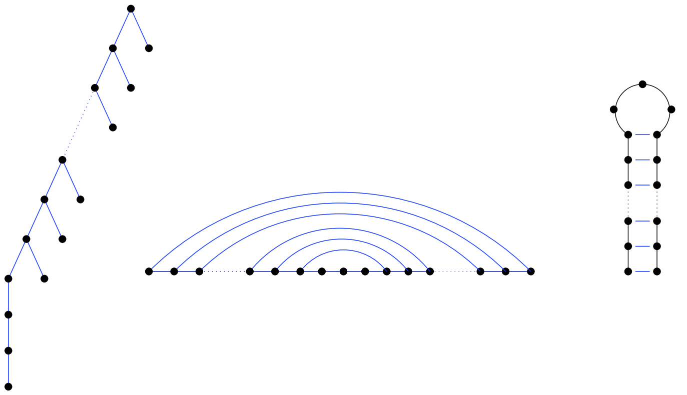


Fig. 2 A *left pseudo-linear* tree, that achieves a $\Theta(n^2)$ complexity in the sequential strategy, along with his corresponding secondary structure.

From the recurrence of Theorem 21, the complexity of the generation for a given structure in the sequential strategy is strongly related to the sum of size of each left subtree found in its tree-representation . Namely, let $\{t_1, \ldots, t_m\}$ be the set of left subtrees in the tree-representation $\phi(S)$ of a structure S, then

$$K(S) = c(\phi(S)) = in(S) - m\theta + \sum_{i=1}^{m} |t_i| \tag{2.3}$$

where $in(S)$ is the number of internal nodes of S.

2.3 Worst-case analysis

Theorem 22 *The worst-case scenario for a statistical sampling using a* sequential strategy *based on a sequence of size n has a complexity $\Theta(n^2)$.*

Proof A run that achieves such a complexity consists in a systematic pairing of the first element of the range (i) with the last one (j), after $(j - i - \theta)$ failed comparisons. The algorithm then samples for intervals $[1, n]$, $[2, n-1]$, $[3, n-2]$, ... Prior to the recursive calls, sequences of $(n-1-\theta)$, $(n-3-\theta)$, $(n-5-\theta)$, ... comparisons are made, followed by θ comparisons for the last unpaired bases, so that the overall complexity for the structure is

$$\begin{aligned} K(S) &= \theta + \sum_{i=1}^{\lfloor \frac{n-\theta}{2} \rfloor} (1 + n - 2i - \theta) \\ &= \theta + n \left\lfloor \frac{n-\theta}{2} \right\rfloor - \left\lfloor \frac{n-\theta}{2} \right\rfloor^2 - \theta \left\lfloor \frac{n-\theta}{2} \right\rfloor \in \Theta(n^2). \end{aligned}$$

An alternative way to see this scenario is to consider the left pseudo-linear tree depicted by Figure 2. The number of internal nodes is $\lfloor \frac{n-\theta}{2} \rfloor + \theta$, the number of left subtrees is $\lfloor\frac{n-\theta}{2}\rfloor$ and the numbers of edges in the left subtrees are respectively $n-2$, $n-4$, ... so that the application of formula 2.3 yields the previous result.

2.4 Average case-analysis

To analyse the average-case complexity of an algorithm requires an assumption over the distribution of the objects at the heart of the method. Here we will assume an **Boltzmann-weighted homopolymer model** based on the Nussinov-Jacobson free-energy model. Namely, every secondary structure having same size as the input sequence can be generated with the Boltzmann probability computed from its free-energy in the Nussinov model. Experiments presented in section 4 suggest similar, yet harder to get analytically, results for the Turner model.

Theorem 23 *The average-case complexity of the statistical sampling using a* sequential strategy *from a sequence of size n is in $\mu \cdot n\sqrt{n}(1+\mathcal{O}(1/\sqrt{n}))$, for $\mu \in \mathbb{R}$ an explicit constant of n.*

Proof Flajolet *et al* [14], analyzed the complexity of the sequential strategy for the random generation of combinatorial structures. They concluded on an $\mathcal{O}(n\sqrt{n})$ complexity for combinatorial classes analogous to trees with limited type of nodes, identified by their degree. However, their analysis applied to the uniform case, and could potentially be altered by the Boltzmann distribution. Thus we present below a full proof for this complexity.

The proof uses generating functions (g.f.) techniques coupled with analysis of singularity. Applications of this general framework to evaluate asymptotic behaviors in the context of computational biology can be found in [30] and [23], the latter containing an introduction on these techniques. We will address the average number of comparisons, which can be formerly defined as the expectancy $\mathbb{E}(K \mid n)$ of the random variable K that holds the number of comparisons dedicated to the sampling of a random structure of size n in the Boltzmann distribution of probability. This expectancy is the relevant parameter for a study of the sampling complexity, as it is the only one which doesn't grow linearly along with n. Formally, this expectancy is such that

$$\mathbb{E}(K \mid n) = \sum_{\substack{\omega\in\mathcal{S}\\|\omega|=n}} \frac{e^{\frac{E(\omega)}{RT}}}{Z_n} c(S)$$

where $Z_n = \sum_{\substack{\omega\in\mathcal{S}\\|\omega|=n}} e^{\frac{-E(\omega)}{RT}}$ is the partition function for the sequence.

At first, let us consider

$$P_f(z) = \sum_{\omega\in\mathcal{S}} e^{\frac{bp(\omega)}{RT}} z^{|\omega|}$$

the **partition function generating function** of RNA secondary structures, where $p_n := [z^n]P_f(z) = Z_n$ as $E(S) = -bp(S)$ in the Nussinov-Jacobson model, and $P_f^{\geq\theta}(z)$ the restriction of $P_f(z)$ to terms having degree greater or equal to θ. A way to enumerate RNA secondary structures is to use a *context-free grammar*, as pointed out by Viennot and Vauchaussade de Chaumont [37]. Here, we present a slightly modified (yet equivalent for $\theta = 1$) version of the grammar, suitable for capturing the minimal number θ of unpaired bases in a terminal loop:

$$\begin{aligned} A &\rightarrow \texttt{(}\, B \,\texttt{)}\, A \mid \bullet A \mid \varepsilon \\ B &\rightarrow \texttt{(}\, B \,\texttt{)}\, A \mid \bullet B \mid \bullet^{\theta} \end{aligned}$$

It is part of the combinatorial folklore [12] that an unambiguous context-free grammar can be interpreted as a system of functional equations involving the length generating function associated with its non-terminals (See also [23]). Furthermore, thanks to the nice regular structure (algebraicity) of the objects and probabilities at stake here, it is also possible to embed the contribution of a base-pair to the Boltzmann probability during the classical transposition of the grammar into a system of functional equations

$$\begin{cases} A(z) = z^2 e^{\frac{1}{RT}} B(z)A(z) + zA(z) + 1 \\ B(z) = z^2 e^{\frac{1}{RT}} B(z)A(z) + zA(z) + z^{\theta}. \end{cases}$$

Solving the system yields the following *positive* solution for $P_f(z) := A(z)$:

$$\begin{aligned} P_f(z) &= \frac{1 - 2z + \Gamma z^2 + z^2 - \Gamma z^{\theta+2} - \sqrt{\Omega(z)}}{2\Gamma z^2(1-z)} \\ \Omega(z) &:= 1-4z-6z^2-2\Gamma z^2-4z^3-4\Gamma z^3+(1-\Gamma)^2 z^4-2\Gamma z^{\theta+2}+4\Gamma z^{\theta+3}-2\Gamma(1+\Gamma)z^{\theta+4}+\Gamma^2 z^{2\theta+4} \\ \Gamma &:= e^{\frac{1}{RT}}. \end{aligned}$$

This instantly yields an expression for the g.f. of $P_f^{\geq\theta}(z)$, since the only structures of size lower than θ are empty structures of size in $[0, \theta - 1]$, for which energy is null. Thus

$$P_f^{\geq\theta}(z) = P_f(z) - \sum_{i=0}^{\theta-1} z^i = P_f(z) - \frac{1 - z^{\theta}}{1 - z}.$$

Then, let us consider the generating function $C(z) = \sum_{\omega \in \mathcal{S}} e^{\frac{bp(\omega)}{RT}} c(\omega) z^{|\omega|}$ whose n-th coefficient $c_n := [z^n]C(z)$ is the unnormalized average cost dedicated to generating an RNA secondary structure of size n. $C^{\geq\theta}(z)$ is the natural restriction to terms of $C(z)$ having degree greater than θ. By transposing the recurrence 2.3 on the secondary structures, we find that

$$\begin{aligned} c(\varepsilon) &= 0 \\ c(\bullet\, \omega) &= 1 + c(\omega) \\ c(\texttt{(}\, \omega \,\texttt{)}\, \omega') &= 1 + |\omega| - \theta + c(\omega) + c(\omega'). \end{aligned}$$

We then expand the former definition of $C(z)$ by distinguishing the contributions of the three cases above and obtain the following expression for $C(z)$

$$\begin{aligned} C(z) = &\sum_{\omega=\varepsilon} e^{\frac{0}{RT}} c(\varepsilon) z^0 + \sum_{\substack{\omega=\bullet\,\omega' \\ \omega'\in\mathcal{S}}} e^{\frac{bp(\omega')}{RT}} (1+c(\omega')) z^{|\omega'|+1} \\ &+ \sum_{\substack{\omega=\,(\,\omega'\,)\,\omega'' \\ \omega'\in\mathcal{S}^{\geq\theta} \\ \omega''\in\mathcal{S}}} e^{\frac{bp(\omega')+bp(\omega'')+1}{RT}} (1+|\omega'|-\theta+c(\omega')+c(\omega'')) z^{|\omega'|+|\omega''|+2}. \end{aligned}$$

This expression, upon developing on the additive contributions of $c(\omega)$, yields

$$\begin{aligned} C(z) = &\left(\sum_{\omega'\in\mathcal{S}} e^{\frac{bp(\omega')}{RT}} z^{|\omega'|}\right) z + \left(\sum_{\omega'\in\mathcal{S}} e^{\frac{bp(\omega')}{RT}} c(\omega') z^{|\omega'|}\right) z \\ &+ \left(\sum_{\omega'\in\mathcal{S}^{\geq\theta}} e^{\frac{bp(\omega')}{RT}} z^{|\omega'|}\right) \left(\sum_{\omega''\in\mathcal{S}} e^{\frac{bp(\omega'')}{RT}} z^{|\omega''|}\right) e^{\frac{1}{RT}} z^2 \\ &+ \left(\sum_{\omega'\in\mathcal{S}^{\geq\theta}} e^{\frac{bp(\omega')}{RT}} |\omega'| z^{|\omega'|}\right) \left(\sum_{\omega''\in\mathcal{S}} e^{\frac{bp(\omega'')}{RT}} z^{|\omega''|}\right) e^{\frac{1}{RT}} z^2 \\ &- \theta \left(\sum_{\omega'\in\mathcal{S}^{\geq\theta}} e^{\frac{bp(\omega')}{RT}} z^{|\omega'|}\right) \left(\sum_{\omega''\in\mathcal{S}} e^{\frac{bp(\omega'')}{RT}} z^{|\omega''|}\right) e^{\frac{1}{RT}} z^2 \\ &+ \left(\sum_{\omega'\in\mathcal{S}^{\geq\theta}} e^{\frac{bp(\omega')}{RT}} c(\omega') z^{|\omega'|}\right) \left(\sum_{\omega''\in\mathcal{S}} e^{\frac{bp(\omega'')}{RT}} z^{|\omega''|}\right) e^{\frac{1}{RT}} z^2 \\ &+ \left(\sum_{\omega'\in\mathcal{S}^{\geq\theta}} e^{\frac{bp(\omega')}{RT}} z^{|\omega'|}\right) \left(\sum_{\omega''\in\mathcal{S}} e^{\frac{bp(\omega'')}{RT}} c(\omega'') z^{|\omega''|}\right) e^{\frac{1}{RT}} z^2. \end{aligned}$$

We point out that $\sum_{\omega\in\mathcal{S}^{\geq\theta}} e^{\frac{bp(\omega)}{RT}} |\omega| z^{|\omega|} = z\frac{\partial P_f^{\geq\theta}(z)}{\partial z}$ and identify the individual formal series with known generating functions, which yields

$$\begin{aligned} C(z) = &\, z\left(P_f(z)+C(z)\right) + z^2 e^{\frac{1}{RT}} (1-\theta) P_f^{\geq\theta}(z) P_f(z) + z^3 e^{\frac{1}{RT}} \frac{\partial P_f^{\geq\theta}(z)}{\partial z} P_f(z) \\ &+ z^2 e^{\frac{1}{RT}} C^{\geq\theta}(z) P_f(z) + z^2 e^{\frac{1}{RT}} P_f^{\geq\theta}(z) C(z). \end{aligned}$$

As for $P_f(z)$, it is noticeable that the only objects of size less than θ contributing to $C(z)$ are structures without base pairs, having energy 0 and respective values for the cost function 1, 2, …, $\theta - 1$, thus

$$C^{\geq\theta}(z) = C(z) - \sum_{i=0}^{\theta-1} i z^i = C(z) - \frac{z^\theta(\theta z - z - \theta) + z}{(1-z)^2}.$$

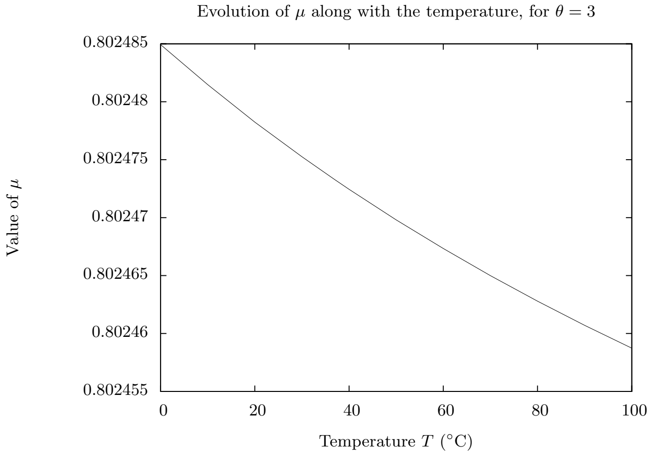


Fig. 3 Influence of temperature over the multiplicative constant μ involved in the $\mu \cdot n\sqrt{n}$ average-case complexity of the sequential strategy.

It remains then to solve a system that is linear in $C(z)$. This can easily be done using any symbolic mathematical software. Let $\rho \in \mathbb{C}$ be the smallest-modulus value[3] of z such that $\Omega(\rho) = 0$, then the solution can be transformed into the following form

$$C(z) = f(z) + \frac{g(z)}{h(z)\Omega(z)}$$

for f,g and h explicit functions that are analytic on the disc of modulus ρ centered in 0. From such an expansion, singularity analysis [13] can be performed, and we find the following expansions for $p_n := [z^n]P_f(z)$ and $c_n := [z^n]C(z)$:

$$p_n \sim \frac{\kappa}{\rho^n n\sqrt{n}}(1 + \mathcal{O}(1/n)) \qquad c_n \sim \frac{\kappa'}{\rho^n}(1 + \mathcal{O}(1/\sqrt{n}))$$

for some explicit constants $\kappa \in \mathbb{R}$ and $\kappa' \in \mathbb{R}$. From the definition of p_n and c_n, it is then obvious that

$$\mathbb{E}(K \mid n) = \frac{c_n}{p_n} \sim \mu \cdot n\sqrt{n}(1 + \mathcal{O}(1/\sqrt{n}))$$

with $\mu = \frac{\kappa'}{\kappa}$.

From the equations derived for the generating functions $C(z)$ and $P_f(z)$, it is possible to compute automatically the generating functions and their asymptotic expansions, using the `GFun` package [35], for any given value of T and θ. Thus we plot in Figure 3 the influence of the temperature over the constant μ involved in the asymptotics of the sampling average-case complexity.

[3] It is a known fact that $\rho \in \mathbb{R}^+$, thanks to a Pringsheim theorem.

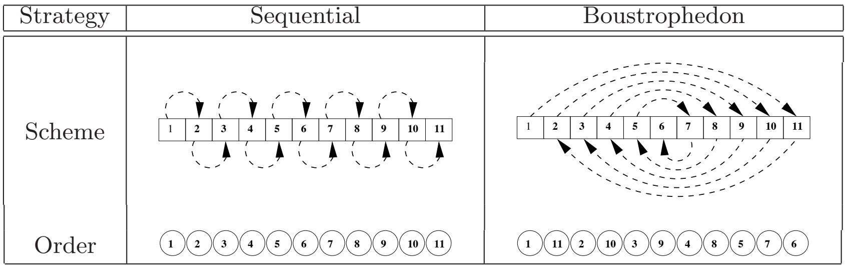


Fig. 4 Sequential and boustrophedon strategies for the range [1..11]

Surprisingly, μ is almost a constant of T which means that an alteration of the distribution of structures, such as obtained by modifying the temperature, will have an almost negligible impact on the complexity. In other words, *the average-case complexity of the sequential strategy doesn't depend on the temperature.*

3 A Boustrophedon optimization

3.1 Context

In ancient Greek, Boustrophedon means *turning like oxen in ploughing* and has been formerly used to describe an ancient style of writing where lines read from left to right and from right to left are alternated. This adjective then occurs in Mathematics as a description of an operation on integer sequences [29] that allows for instance for an easy computation of a Taylor expansion for $tan(x)$ [1]. It is later found in works related to the random generation of combinatorial structures [14], where it denotes a way of investigating potential decompositions of an object into two parts prior to calling inductively.

We illustrate this process on the recursive approach to the uniform random generation from a context-free language. Suppose we want to draw, uniformly and at random, a word of size n from the language generated by a non-terminal A having rule of the **product** form

$$A \rightarrow B\ C$$

Assume that we know how to draw uniformly words of any size up to n from B and C. Then we only need to find appropriate (i.e. with probabilities consistant with the uniformity) sizes i and $n-i$ for the words issued from B and C. Once such sizes are chosen, it suffices to draw a word of size i from B, a word of size $n-i$ from C and to concatenate them into a word from A of length n. We refer to [14] for the choice on the appropriate probabilities, which are related to the cardinalities of languages generated from B and C. The order of investigation, illustrated by Figure 4, of the candidate values for i is also shown to have an important impact on the complexity of the generation. Namely, it has been shown [14] that investigating values for i in

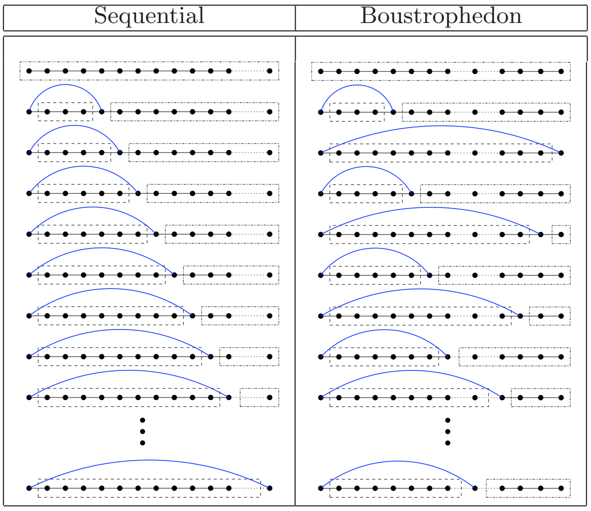


Fig. 5 Comparison of the sequential and boustrophedon orders for conformation sampling.

a *sequential* $(1, 2, 3, \ldots, n)$ order yields a $\Theta(n^2)$ worst-case complexity and a $\mathcal{O}(n\sqrt{n})$ average complexity, whereas preferring a *boustrophedon* $(1, n, 2, n-1, \ldots, \lceil n/2 \rceil)$ order yields $\mathcal{O}(n \log(n))$ average and worst-case complexities.

3.2 Boustrophedon strategy for the statistical sampling

The boustrophedon order, at the core of the fruitful optimization described in section 3.1 can be used in the context of sampling from the Boltzmann ensemble of low-energy. Instead of investigating potential partners k for the first base i of a subsequence $[i, j]$ in a sequential manner, we will use a **boustrophedon strategy**. Namely, we will investigate potential values for k from $[i + \theta + 1, j]$ alternatively from both ends, progressing toward the middle of the subsequence of interest (See Figure 5):

$$(i{+}\theta{+}1) \to (j) \to (i{+}\theta{+}2) \to (j{-}1) \to \ldots \to \left(i + \theta + 1 + \left\lfloor \frac{j - i + \theta + 1}{2} \right\rfloor\right)$$

The implementation of this optimization is very easy, as a boustrophedon order can be easily simulated from a sequential one, yielding only a minor modification in an already existing implementation.

3.3 Worst-case complexity analysis

Surprisingly, applying such a strategy while investigating potential candidates at a given stage yields a significant improvement of the sampling algorithm.

Theorem 31 *The worst-case complexity, in term of number of comparisons, of the statistical sampling implementing a boustrophedon strategy is in $\mathcal{O}(n\log(n))$.*

Proof Flajolet *et al* [14] analyzed the worst-case complexity of the boustrophedon approach, in the more general case of context-free grammar. They showed that the cost function $f(n)$ for the worst-case generation of a structure of size n is such that

$$f(n) = \max_{k\in[1,n-1]} (f(k) + f(n-k) + 2\min(k, n-k)). \tag{3.1}$$

Intuitively, the equation states that a value k is investigated after $2\min(k, n-k)$ comparisons, with min capturing the fact that the final k can either be reached first from the left or from the right in the Boustrophedon walk. Recursive calls then independently cost $f(k)$ and $f(n-k)$. It can be easily shown that this result holds for the analysis of the sampling, where n stands for the overall size of the subsequence.

This class of equations has been solved by Knuth [16], and is known to have a $\mathcal{O}(n\log(n))$ solution. Let us summarize the argument:

- The maximum for this equation (the worst-case complexity) is reached at $k = \lfloor \frac{n}{2} \rfloor$.
- Equation 3.1 can then be analyzed using generating functions techniques and singularity analysis, or by drawing a parallel with the sum of *ones* in the binary representation of all numbers from n to 0, which is in $\mathcal{O}(n\log(n))$ for obvious reasons.

In our precise context, the potential structures associated with a given sequence are clearly a subset of the trees considered in [14] equipped with probabilities. As the probability distribution does not affect the generation cost of a given structure[4], neither does it affect the generation cost for the worst-case scenario.

We try to convey an intuitive explanation for this result by stating that decomposing the range into two *uneven* parts now takes a limited amount of comparisons, whereas it could take the whole size of the range $[i+\theta+1, j]$ in the former, sequential version. Furthermore, paying many ($\mathcal{O}(n)$) comparisons at a certain step leads to an almost even decomposition of the range for the next recursive calls. If we keep on dividing almost evenly, the tree drawn from the recursive calls will then have heights $\Theta(\log(n))$. As it is obvious that the total number of comparisons over all nodes at a given height is $\mathcal{O}(n)$, then the overall complexity is $\mathcal{O}(n\log(n))$.

[4] A probability distribution of the instances may affect the average-case complexity of an algorithm, but won't affect the worst. As the average-case complexity cannot exceed the worst-case complexity, it is reasonable to expect a $\mathcal{O}(n\log(n))$ average-case complexity.

4 Experimental validation

4.1 Turner model

The Turner model [41] is at the core of current state-of-the-art free-energy approaches for the prediction of RNA folding. Although exploring roughly the same conformation landscape as the Nussinov-Jacobson model, it uses different elementary contributions to the free-energy, which allow for a more accurate approximate of the free-energy. Namely, instead of mimicking the base-pairing process, as does the Nussinov-Jacobson model, it uses the so-called *loop decomposition*, focusing on the inner faces of the map naturally associated with an RNA secondary structure. This fundamental difference might in theory result in different behaviors for the complexities of the statistical sampling, both in the sequential and boustrophedon approaches.

Therefore, we tested the practical efficiency of the boustrophedon optimization on the `UnaFold` software [26], that includes sampling from the Boltzmann ensemble of low-energy. This software has been chosen as it is so far the only available implementation accounting correctly for dangles in the computation of the partition function (see N. Markham's thesis [25]). This condition is essential in order to perform an unbiased sampling.

4.2 Method

First, we retrieved the source-code of `UnaFold v3.3` [26] from the DINAMelt website at RPI. We duplicated the code related to stochastic traceback and modified the copy to implement the boustrophedon strategy described above and compiled both versions. Only two lines from the original code were modified, and two lines added. Then we retrieved from RFAM [17] the sequences for a Drosophila melanogaster 5S rRNA (GenBank: X06937/117–251) and a Staphylococcus aureus RNAIII (GenBank: AJ223774.1/71–632). While the former has well-known structure and function, as being part of the small sub-unit of the ribosome, the latter has a more unusual, dual regulatory/messenger function. Indeed it is known to be at the same time the effector of the agr system (accessory gene regulator) and the mRNA of the 26 amino acid delta-haemolysin. Therefore, deep structural differences are expected, that might challenge the robustness of our theoretical results.

In order to get an estimate of the complexities for both approaches, we needed several sets of sequences for increasing non-bounded sizes. Previous works [5] have focused on the dinucleotide frequency to distinguish between structural and coding RNAs in the Turner model. It can also be shown that shuffling models preserving dinucleotide frequencies converge quickly to Markov models of order 1. As we were interested in the behavior of the complexity for higher instance sizes, we used `GenRGenS` [34] to create Markov models $\mathcal{M}_{5S}$ and $\mathcal{M}_{\text{RNAIII}}$ of order 1 for both RNAs described above. We drew 100 sequences from size 10 to size 3000 (using step 10) in both models $\mathcal{M}_{5S}$ and $\mathcal{M}_{\text{RNAIII}}$.

For each of these 60000 sequences we ran both the original and modified versions of `UnaFold`, and counted the overall number of comparisons per-

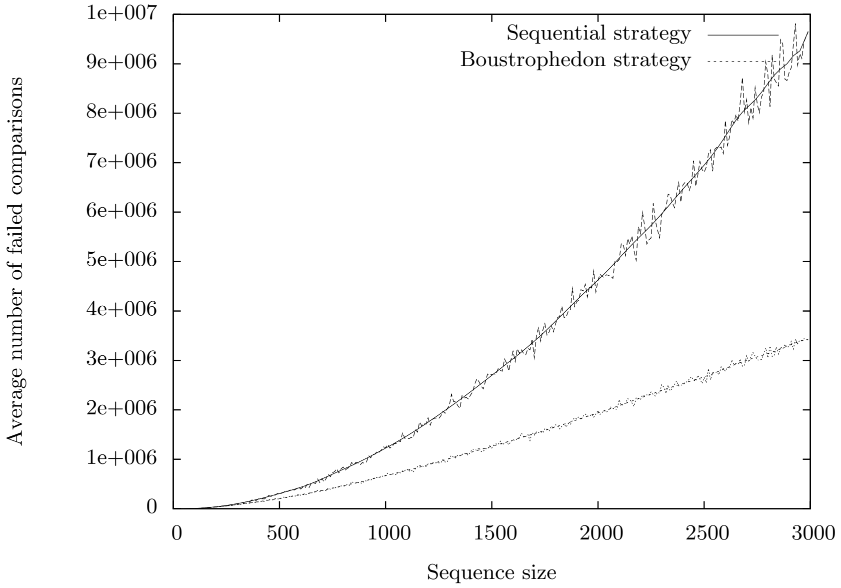


Fig. 6 Experimentally determined complexities for both approaches on sequences based on a Drosophila melanogaster 5S rRNA.

formed prior to two inductive calls on the same sequence. This quantity is the relevant one in the complexity analysis, as the other contributions will sum up to a linear term in the size of the original sequence, and thus will become negligible for high sizes. We then averaged these numbers for each size, and obtained the values plotted in Figures 6 and 7.

4.3 Results

A first glimpse at the results clearly indicates that the boustrophedon version of the software outperforms the original implementation. For instance, for sequences of size 1000 generated from the model $\mathcal{M}_{\text{5S}}$, the boustrophedon sampling already saves **half of the time** needed by the sequential approach.

From regression calculuses, it turns out that the estimated growth of the sequential strategy on these examples scales like respectively $1.1n^2$ and $0.53n^2$ for models $\mathcal{M}_{\text{5S}}$ and $\mathcal{M}_{\text{RNAIII}}$. The average-case complexity of the sequential approach for the Turner-based sampling model even seems to be quadratic on these experiments. However this interpretation might just be a representation artefact and needs to be confirmed by further experiments, it is certain that a complexity greater than $\Theta(n\sqrt(n))$ is observed for the sequential approach.

Furthermore, we observe on these graphics a significant difference of behaviors between the two random sequences models $\mathcal{M}_{\text{5S}}$ and $\mathcal{M}_{\text{RNAIII}}$. Namely, the performance gap between the Boustrophedon and sequential

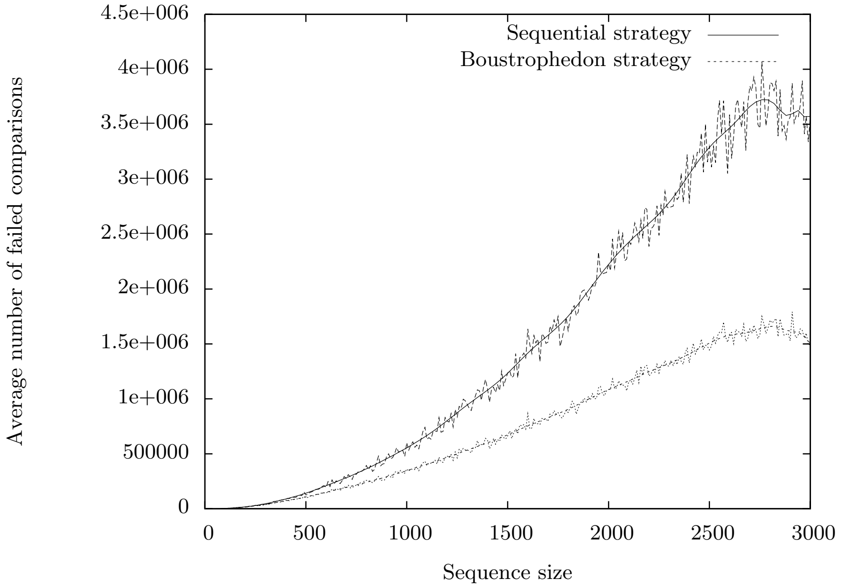


Fig. 7 Experimentally determined complexities for both approaches on sequences based on a Staphylococcus aureus RNAIII.

approaches seems to widen quicker in the $\mathcal{M}_{\text{5S}}$ model than in the $\mathcal{M}_{\text{RNAIII}}$ model. This observed difference of behavior might be attributed to the different GC contents for this two models. Namely, the mononucleotide frequencies for $\mathcal{M}_{\text{5S}}$ are {A:21%, U:22%, G:28%, C:29%} whereas those of $\mathcal{M}_{\text{RNAIII}}$ are {A:35%, U:39%, G:13%, C:13%}. This G-C poverty of the latter might be responsible for a low average number of base pairings in the sampled structures, thus pulling down the constants involved in the overall complexity[5].

5 Conclusion and perspectives

We presented here an adaptation of an optimizing technique, the boustrophedon approach, that improved the stochastic traceback stage of the sampling from the Boltzmann ensemble of low-energy. We drew a parallel with a certain class of unary/binary trees to analyze the complexity of a simplified version of the sampling and found again the $\Theta(n^2)$ complexity pointed out by the authors of this algorithm [10]. Using generating function techniques, we proved that the average-case complexity of our sampling was in $\Theta(n\sqrt{n})$. To our best knowledge, this is the first study of the average-case complexity for this classe of algorithms. We then showed that implementing the boustrophedon strategy dramatically decreases the worst-case complexity of the

[5] The base-pairing process and the subsequent recursive calls to the sampling functions are responsible for the non-linear behavior of the complexities.

classical algorithms from $\Theta(n^2)$ to $\mathcal{O}(n\log(n))$, and the average-case complexity from $\Theta(n\sqrt{n})$ to $\mathcal{O}(n\log(n))$. As this implementation yields a quasi-linear complexity for each *stochastic traceback* performed, it is now possible to sample Boltzmann ensembles of low-energy of $\Theta(n^2)$ elements without any significant overhead to the overall $\Theta(n^3)$ complexity of the *precomputation stage*. Implementation of this technique will increase, by allowing sampling of bigger sets, the coverage (sensitivity) of existing approaches based on these ensembles. It will also allow for more significant results for large sequences when the computational ressources required to run an $\Theta(n^3)$ algorithm will become available.

More generally, if Δ is the free-energy distance between the two *shreps*[6] for the best-scoring shapes, then the number of structures that need to be sampled before getting at least the best-scoring couple of shapes is believed to grow polynomially on n the sequenced length and exponentially on Δ. A natural perspective of these works would then consist in an analysis of the evolution of Δ for sequences of increasing lengths. If it remains roughly constant, a sampled set that captures at least two alternative conformations with high probability will have to have size at least polynomial on the sequence length. This pleads for a more general reflection on the relationship between the sequence length and the size of a *statistically significant* ensemble.

Understanding the influence of the nucleotide frequency over the behavior of the complexity observed in the experiments remains a challenge. This might be performed using the idea of stickiness [22,31] which allows to turn the nucleotide frequencies into base-pairing probabilities, later injected inside functional equations for the generating series. However, it is unclear how this distribution might interact with the currently embedded Boltzmann probability.

Moreover, the Boltzmann ensemble is a multiset, which reveals useful to evaluate the probability of a given shape, but useless in the context of clustering, or more generally when a coverage of the low-energy conformational landscape of a given size is required. It might then be useful to prevent already sampled structures to be drawn again. However it is still unclear how to perform such an random generation without introducing a global bias in the generation.

Another promising perspective of this work would consist in generalizing the boustrophedon approach to higher-dimensional search space for indices. Namely, parametric approaches are actually developed by Clote *et al* [3,4, 6,11] that use an additional parameter x, which can be understood as a way to classify structures in the conformational landscape. The complexities of these approaches are then increased and equations for the parameterized partition function now involve double-nested sums over k the base-pairing point and x the parameter, both exploring values of the order of n. As the stochastic traceback can be interpreted as an inversion of the sums involved in the dynamic programming equation, followed by recursive calls, a natural sequential implementation of the sampling in that case would have a $\Theta(n^3)$ complexity for the worst-case scenario. A natural way to transpose the boustrophedon philosophy to these cases would be to *radiate from the corners.*

[6] The *shrep* of a shape π is the lowest free-energy structure having shape π [36].

Namely, if $k \in [i, j]$ and $x \in [i', j']$ in the dynamic programming equation involving the double sum, then we start by exploring *corners* candidates $\{(i, i'), (i, j'), (j, i'), (j, j')\}$ for (k, x). If no suitable candidate is found, we investigate all candidates at distance 1,2,3,... of these corners, until an appropriate decomposition is found. Such a strategy could be easily generalized to a higher-dimension decomposition space, which appears in the context of random generation of words of fixed compositional frequency under constraint of complying with context-free grammar.

Acknowledgements The author would like to thank A. Lorenz for reviewing this document and P. Clote for his advise and guidance.

References

1. André, D.: Développements de $sec(x)$ et de $tan(x)$. C. R. Acad. Sci. Paris **88**, 965–967 (1879)
2. Barrick, J., Corbino, K., Winkler, W., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I., Wickiser, J., Breaker, R.: New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proc. Natl. Acad. Sci. USA **101(17)**, 6421–6426 (2004)
3. Clote, P.: An efficient algorithm to compute the landscape of locally optimal rna secondary structures with respect to the Nussinov-Jacobson energy model. Journal of Computational Biology **12**(1), 83–101 (2005)
4. Clote, P.: RNALOSS: A web server for RNA locally optimal secondary structures. Nucleic Acids Res. **33(Web Server issue)**, W600–604 (2005)
5. Clote, P., Ferre, F., Kranakis, E., Krizanc, D.: Structural rna has lower folding energy than random rna of the same dinucleotide frequency. RNA **11**(5), 578–591 (2005)
6. Clote, P., Waldispühl, J., Behzadi, B., Steyaert, J.M.: Energy landscape of k-point mutants of an rna molecule. Bioinformatics **21**(22), 4140–4147 (2005)
7. Ding, Y.: Statistical and bayesian approaches to rna secondary structure prediction. RNA **12**(3), 323–331 (2006)
8. Ding, Y., Chan, C., Lawrence, C.: SFold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res. - Web Server Issue **32**, 135–141 (2004)
9. Ding, Y., Chan, C.Y., Lawrence, C.E.: Rna secondary structure prediction by centroids in a boltzmann weighted ensemble. RNA **11**, 1157–1166 (2005)
10. Ding, Y., Lawrence, E.: A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Research **31**(24), 7280–7301 (2003)
11. E.Freyhult, V.Moulton, P.Clote: Rnabor: A web server for rna structural neighbors. Nucleic Acids Research (2007). To appear ...
12. Flajolet, P.: Singular combinatorics. In: Proceedings of the International Congress of Mathematicians, vol. 3, pp. 561–571 (2002)
13. Flajolet, P., Odlyzko, A.: Singularity analysis of generating functions. SIAM J. Discrete Math. **3**(2), 216–240 (1990)
14. Flajolet, P., Zimmermann, P., Van Cutsem, B.: Calculus for the random generation of labelled combinatorial structures. Theoretical Computer Science **132**, 1–35 (1994). A preliminary version is available in INRIA Research Report RR-1830
15. Gan, N.K.H.H., Schlick, T.: A computational proposal for designing structured rna pools for in vitro selection of rnas. RNA **13**, 478–492 (2007)
16. Greene, D.H., Knuth, D.E.: Mathematics for the Analysis of Algorithms. Birkhauser Boston (1981)
17. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R.: Rfam: an RNA family database. Nucleic Acids Research **31**(1), 439–441 (2003)

18. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., Schuster, P.: Fast folding and comparison of RNA secondary structures. Monatsch. Chem. **125**, 167–188 (1994)
19. I. Tinoco, J., Borer, P., Dengler, B., Levin, M., Uhlenbeck, O., Crothers, D., J.Bralla: Improved estimation of secondary structure in ribonucleic acids. Nat New Biol. **246**(150), 40–41 (1973)
20. Leontis, N., Westhof, E.: Geometric nomenclature and classification of rna base pairs. RNA **7**, 499–512 (2001)
21. Lescoute, A., Westhof, E.: Topology of three-way junctions in folded rnas. RNA **12**(1), 83–93 (2006)
22. Lesk, A.M.: A combinatorial study of the effects of admitting non-watson-crick base pairings and of base compositions on the helix-forming potential of polynucleotides of random sequences. J. Theor. Biol. **44**, 7–17 (1974)
23. Lorenz, W., Ponty, Y., Clote, P.: Asymptotics of RNA shapes. Journal of Computational Biology (2007). To appear
24. Lyngs, R.B., Pedersen, C.N.S.: Rna pseudoknot prediction in energy-based models. Journal of Computational Biology **7**(3-4), 409–427 (2000)
25. Markham, N.R.: Algorithms and software for nucleic acid sequences. Ph.D. thesis, Rensselaer Polytechnic Institute (2006)
26. Markham, N.R., Zuker, M.: Dinamelt web server for nucleic acid melting prediction. Nucleic Acids Res. **33**, W577–W581 (2005)
27. Mathews, D.: Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA **10**, 1178–1190 (2004)
28. McCaskill, J.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers **29**, 1105–1119 (1990)
29. Millar, J., Sloane, N., Young, N.: A new operation on sequences: The boustrophedon transform. J. Combin. Th. Ser. **A 76**, 44–54 (1996)
30. Nebel, M.: Combinatorial properties of RNA secondary structures. Journal of Computational Biology **3**(9), 541–574 (2003)
31. Nebel, M.E.: Investigation of the bernoulli model for rna secondary structures. Bulletin of Mathematical Biology **66**(5), 925–964 (2004)
32. Nussinov, R., Jacobson, A.: Fast algorithm for predicting the secondary structure of single-stranded rna. Proc Natl Acad Sci U S A **77**, 6903–13 (1980)
33. Penchovsky, R., Breaker, R.: Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. Nature Biotechnology **23**(11), 1424–1431 (2005)
34. Ponty, Y., Termier, M., Denise, A.: GenRGenS: Software for generating random genomic sequences and structures. Bioinformatics **22**(12), 1534–1535 (2006)
35. Salvy, B., Zimmerman, P.: Gfun: a maple package for the manipulation of generating and holonomic functions in one variable. ACM Transactions on Mathematical Softwares **20**(2), 163–177 (1994). DOI http://doi.acm.org/10.1145/178365.178368
36. Steffen, P., B.Voss, Rehmsmeier, M., Reeder, J., Giegerich, R.: RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics **22**(4), 500–503 (2006)
37. Vauchaussade de Chaumont, M., Viennot, X.: Enumeration of RNA's secondary structures by complexity. In: V. Capasso, E. Grosso, S. Paven-Fontana (eds.) Mathematics in Medecine and Biology, *Lecture Notes in Biomathematics*, vol. 57, pp. 360–365 (1985)
38. Voss, B., Giegerich, R., Rehmsmeier, M.: Complete probabilistic analysis of RNA shapes. BMC Biol. **4**(5) (2006)
39. Waterman, M.S.: Secondary structure of single stranded nucleic acids. Advances in Mathematics Supplementary Studies **1**(1), 167–212 (1978)
40. Wuchty, S., Fontana, W., Hofacker, I.L., Schuster, P.: Complete suboptimal folding of rna and the stability of secondary structures. Biopolymers **49**, 145–165 (1999)
41. Xia, T., J. SantaLucia, J., Burkard, M., Kierzek, R., Schroeder, S., Jiao, X., Cox, C., Turner, D.: Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry **37**, 14,719–35 (1999)

42. Zhao, J., Malmberg, R., Cai, L.: Rapid ab initio rna folding including pseudoknots via graph tree decomposition. In: Proceedings of the 6th Workshop on Algorithms in Bioinformatics (WABI 2006), vol. 4175, pp. 262–273 (2006)
43. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. **9**, 133–148 (1981)