An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure

Jérôme Waldispühl, Yann Ponty

▶ **To cite this version:**

Jérôme Waldispühl, Yann Ponty. An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure. RECOMB - 15th Annual International Conference on Research in Computational Molecular Biology - 2011, Mar 2011, Vancouver, Canada. pp.501-515, 10.1007/978-3-642-20036-6_45 . hal-00546847

HAL Id: hal-00546847
https://hal.science/hal-00546847v1

Submitted on 14 Dec 2010

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.

An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure

Jérôme Waldispühl[1,*] and Yann Ponty[2,⋆]

[1] School of Computer Science & McGill Center for Bioinformatics, McGill University, Montreal, Canada,
[2] Laboratoire d'Informatique, École Polytechnique, Palaiseau, France.

Abstract. The analysis of the relationship between sequences and structures (i.e. how mutations affect structures and reciprocally how structures influence mutations) is essential to decipher the principles driving molecular evolution, to infer the origins of genetic diseases or to develop bioengineering applications such as the design of artificial molecules. Because their structures can be predicted from the sequence data only, RNA molecules provide a good framework to study this sequence-structure relationship. We recently introduced a suite of algorithms called `RNAmutants` which allows, for the first time, a complete exploration of RNA sequence-structure maps in polynomial time and space. Formally, `RNAmutants` takes an input sequence (or seed) to compute the Boltzmann weighted ensembles of mutants with exactly k mutations, and sample mutations from these ensembles. However, this approach suffers from major limitations. Indeed, since the Boltzmann probabilities of the mutations depend of the free energy of the structures, `RNAmutants` has difficulties to sample mutant sequences with low `G+C`-contents. In this paper we introduce a novel unbiased adaptive sampling algorithm that enables `RNAmutants` to sample regions of the mutational landscape poorly covered by classical algorithms. We applied these methods to sample mutations with low `G+C`-contents. These adaptive sampling techniques can be easily adapted to explore other regions of the sequence and structural landscapes which are difficult to sample. Importantly, these algorithms come at a minimal computational cost. We demonstrate the insights offered by these techniques on studies of complete RNA sequence structures maps of sizes up to 40 nucleotides. Our results indicate that the `G+C`-content has a strong influence on the size and shape of the evolutionary accessible sequence and structural spaces. In particular, we show that low `G+C`-contents favor the apparition of internal loops and thus possibly the synthesis of tertiary structure motifs. On the other hand, high `G+C`-contents significantly reduce the size of the evolutionary accessible mutational landscapes.

⋆ Corresponding authors: `jeromew@cs.mcgill.ca` and `yann.ponty@lix.polytechnique.fr`.

1 Introduction

Our understanding of the mechanisms regulating cell activity has considerably improved over the last two decades. Ribonucleic acids (RNAs) have emerged as one of the most important biomolecules, playing key roles in various aspects of the gene transcription and regulation processes. For instance, ribozymes are involved in the cleavage of messenger RNAs (mRNAs), and riboswitches undergo structural changes to regulate gene expression.

To achieve their functions, RNAs use sophisticated structures which are mainly determined by their sequence. Any modification of the sequence may result in a change in its structure and a loss (or an improvement) in function. The development of tools to estimate the effect of mutations on structures, or conversely the influence of structure conservation on the mutational process, is essential for understanding the mechanisms of molecular evolution [1], the origin of genetic diseases [2] or to develop bioengineering applications such as the design of RNA molecules (a.k.a. inverse folding) [3].

To understand the role of specific nucleotides, mutagenesis experiments proceed to point-wise mutations in order to observe putative changes in the expression profile of the experiments (i.e. the experimental observation) revealing a modification of the functionality of the molecule. Such experiments are critical to identify mutations modifying the function and structure of RNAs. However, all experiments are time-consuming and have a substantial cost, and it follows that exhaustive experimental studies are impossible.

While it is not realistic to conduct large scale experimental studies on the complete RNA mutational landscape, this limitation could be circumvented in computational studies. Indeed, The structure of RNAs can be predicted from sequence data only [4,5]. More importantly, the secondary structure can be predicted with dynamic programming techniques in polynomial time and space [4,6] using a nearest neighbor energy model [7]. These algorithms are implemented in various programs [8,9,10] and enable to predict the secondary structures of thousands of sequences in a short time. Therefore it has becomes possible to compute the complete mutational landscape small RNA sequences [11] and to simulate the evolution of the structure of populations of RNAs [12,13].

Several groups intended to explore the mutational landscape of RNAs and to quantify the dependences between sequences and structures. The most representative work in this area has been achieved by P. Schuster and co-worker on the sequence-structure maps and neutral networks [14,15]. So far, all these studies were limited by brute force approaches requiring to compute individually the structure of a number of mutants growing exponentially with the length of the sequence (e.g. there is 4^n sequence of length n), thus making exhaustive exploration of the mutational landscape intractable on large sequences (≥ 20 nucleotides).

To address this issue, we have developed the program `RNAmutants` which, from an input sequence, computes the structural ensembles of all sequences with k mutations in *polynomial time and space* [16]. To achieve this algorithmic advance, we expanded the seminal dynamic programming rules introduced 30 years ago by Zuker and Stiegler [4]. The dramatic improvement of the algorithmic complexity (from an exponential to a polynomial running time) enabled us to investigate problems that could not have been addressed with previous techniques. For instance, we provided evidences that the complete sequence of the 3'UTR of the GB RNA virus C has been optimized to preserve its secondary structure from the deleterious effect of mutations [16]. `RNAmutants` has been developed upon a formal grammar-based model [17,18] which, in particular, can be used to compute k-mutants (i.e. sequences with exactly k mutations) with the lowest free energy structure.

Formally, `RNAmutants` takes an input sequence and computes the minimum free energy (MFE) structure and the Boltzmann partition function of all k-mutants sequences in the k-Hamming neigh-

borhood (i.e. ensemble of sequences with exacly k mutations) of the input sequence. In addition, it samples k-mutants together with a secondary structure. This naturally extends the seminal Zuker and Stiegler's [4], McCaskill's [6] and Ding and Lawrence's algorithms [19] which do not consider sequence variations.

In this model, the probabilities of the sequences in the k-mutants ensembles are determined by their ensemble free energies. Thus by mutants with the lowest folding energy. It follows that these ensembles are dominated by sequences with high G+C-contents and that RNAmutants has a bias towards A/U→C/G mutations. This bias is a serious drawback for a complete and rigorous analysis of RNA sequence-structure maps or the prediction of deleterious mutations (i.e. mutations altering the native structure). For instance, sampling sequences with a large number of mutations will inevitably produce sequences with high G+C-content folding into long single stem structures, while in reality a broader range of structures are observed. The nucleotide distribution can also be used to indirectly control the folding and functional properties on RNAs. Recently, Chan *et al.* showed correlation between the G+C-content and RNAi efficiency.[20].

In this paper, we develop an unbiased adaptive sampling algorithm enabling to control of the nucleotide composition of the sequences sampled from each k-neighborhood by RNAmutants. These techniques alleviates RNAmutants from its previous limitations and enable us to study mutational processes at a finer resolution level. Importantly, this algorithmic advance is achieved at a minimal computational cost and can be generalized to sample any regions of the mutational and structural landscapes which are difficult to reach with classical algorithms.

This article is organized as follows. In section 2, we formally define the problem addressed, explain why a brute force approach fails, and show how a multivariate Boltzmann model can be integrated into our RNAmutants algorithms to control the nucleotide composition of sampled sequences. Then, in section 3 we illustrate the efficiency of our technique by providing an analysis of complete sequences-structure maps of RNAs of sizes up to 40 nucleotides (we remind that previous exhaustive studies were limited to sizes of 20). Our computational experiments reveal interesting properties of RNA sequence-structure maps that can be parameterized by the G+C-content. In particular, we find that low G+C-contents favor the apparition of bulges and internal loops, thus the possible insertion of non-canonical interactions and tertiary structure motifs (Section. 3). We also show that the diversity of mutants improving the stability of the fold is effectively optimal for medium G+C-contents (around 50%) and that high G+C-contents reduce the size of the evolutionary reachable mutational landscape (Section. 3). These finding suggest that the G+C-content is essential to balance the competition between the evolutionary accessibility (i.e. the sequence diversity) and the structural stability.

2 Methods

Notations, definitions and existing works Throughout this document, we will abstract an RNA molecule ω as a sequence of bases chosen from $\mathbb{B} := \{A, C, G, U\}$. The length of the RNA sequence will be denoted by $n = |\omega|$. Following standard notations, we will denote by ω_i the base at position i. A secondary structure s for an RNA ω is defined as a set of base pairs of the form $(i, j) \in [1, n]^2$ with $i < j$, such that any two base pairs $\{(i, j), (k, l)\} \subset s$ do not share an extremity ($\{i, j\} \cap \{k, l\} = \varnothing$), and are either non-overlapping ($[i, j] \cap [k, l] = \varnothing$) or inclusive ($[i, j] \subset [k, l]$ or $[k, l] \subset [i, j]$). Moreover in order to avoir steric clashes, a minimal number of bases θ is usually required between the two extremities of a base pair (i, j) $(i + \theta < j)$. Finally let us denote by $\mathcal{S}_{\omega,\theta}$ the set of all secondary structures compatible with a given RNA ω under the θ constraint.

Free-energy model. For the sake of clarity, we will illustrate our claims and algorithms on a generalization of the energy model proposed by Nussinov and Jacobson [21], assigning additive free-energy contributions to each base-pair. This model may appear overly simplistic in comparison with the Turner model [4], but it is sufficient to capture the key algorithmic elements while remaining easier to grasp. It should however be noted that the implementations used for our experiments make use of the full Turner model, as was described in the initial presentation of `RNAMutants` [16].

In this section, each base-pair $(a, b) \in s$ within a sequence ω is associated with a free-energy contribution $\Delta_{\omega_a,\omega_b}$ and unpaired bases are not taken into account by the model. Consequently the overall free-energy $E(\omega, s)$ of a structure s over a sequence ω is given by $E(\omega, s) = \sum_{(i,j)\in s} \Delta_{\omega_i,\omega_j}$. Note that this energy model captures the incompatibility of a base-pair $(x, y) \in \mathbb{B}^2$ upon setting $\Delta_{x,y} = +\infty$.

Thermodynamics. Following McCaskill [6], one can define a Boltzmann distribution and assign to each structure s a Boltzmann factor $\mathcal{B}_\omega(s) := e^{\frac{-E(\omega,s)}{RT}}$ where is T the temperature and R the universal gaz constant. This induces a Boltzmann probability distribution on the set $\mathcal{S}_{\omega,\theta}$ of structures compatible with ω such that

$$P(s \mid \omega) = \frac{\mathcal{B}_\omega(s)}{\mathcal{Z}_\omega} \tag{1}$$

where $\mathcal{Z}_\omega$ is the partition function, defined as $\mathcal{Z}_\omega = \sum_{s\in\mathcal{S}_{\omega,\theta}} \mathcal{B}_\omega(s)$.

Restricting our attention to an interval $[i, j]$ of ω, we can easily observe that within a secondary structure on $[i, j]$, the first position i is either unpaired and is followed by a secondary structure on $[i+1, j]$, or paired to some position $l \in [i+\theta+1, j]$, in which case the non-crossing condition forces the existence of two independent structures on intervals $[i+1, l-1]$ and $[l+1, j]$. Furthermore this case decomposition is complete as shown by Waterman [22]. The partition function is then locally defined recursively by

$$\mathcal{Z}_{[i,j]} = \mathcal{Z}_{[i+1,j]} + \sum_{l=i+\theta+1}^{j} e^{-\frac{\Delta_{\omega_i,\omega_l}}{RT}} \mathcal{Z}_{[i+1,l-1]} \cdot \mathcal{Z}_{[l+1,j]}. \tag{2}$$

and by $\mathcal{Z}_{[i,i-1]} = 1$. The partition function $\mathcal{Z}_\omega := \mathcal{Z}_{[1,n]}$ can therefore be computed in $\Theta(n^3)/\Theta(n^2)$ time and space. Direct applications of this algorithm described include the derivation of base-pairing probabilities [6] and statistical sampling [19].

RNAMutants. For the sake of completeness, let us remind that the `RNAMutants` algorithm [16] starts from an initial sequence ω and traverses the space of all sequences parameterized by their Hamming distance to ω (equivalent to the minimal number of mutations required). A parameterized analogue of the partition function is then obtained by summing over sequences/structures couples that are compatible with a given interval (i, j) and a prescribed number of mutations k.

Let us remind that the Hamming distance $\sigma : \mathbb{B}^n \times \mathbb{B}^n \to \mathbb{N}$ between two sequences of equal length is defined by $\sigma_{\varepsilon,\varepsilon} = 0$ and by $\sigma_{x.X',y.Y'} = (1 - \mathbb{1}_{x,y}) + \sigma_{X',Y'}$. Then the partition function over k mutants is recursively defined by

$$\mathcal{Z}_{\left[\begin{smallmatrix} i,j \\ k \end{smallmatrix}\right]} = \sum_{b\in\mathbb{B}} \mathcal{Z}_{\left[\begin{smallmatrix} i+1,j \\ k-\sigma_{\omega_i,b} \end{smallmatrix}\right]} + \sum_{b,b'\in\mathbb{B}^2} \sum_{l=i+\theta+1}^{j} \sum_{k'=0}^{k-\sigma_{\omega_i\omega_l,bb'}} e^{-\frac{\Delta_{b,b'}}{RT}} \cdot \mathcal{Z}_{\left[\begin{smallmatrix} i+1,l-1 \\ k' \end{smallmatrix}\right]} \cdot \mathcal{Z}_{\left[\begin{smallmatrix} l+1,j \\ k-k'-\sigma_{\omega_i\omega_l,bb'} \end{smallmatrix}\right]} \tag{3}$$

with limit conditions $\mathcal{Z}_{\left[\begin{smallmatrix} i,i-1 \\ 0 \end{smallmatrix}\right]} = 1$ and $\mathcal{Z}_{\left[\begin{smallmatrix} i,i-1 \\ k \end{smallmatrix}\right]} = 0, \forall k > 0$. A direct computation of the recursion yields a $\Theta(n^3 \cdot k^2)/\Theta(n^2 \cdot k)$ time/space algorithm with k the maximal number of mutations.

Function GenMuts$(i, j, k, \mathbf{w}, \omega)$: Returns a sequence/structure couple over interval (i, j) at distance k of ω, drawn with respect to a $\mathbf{w}$-weighted Boltzmann probability.

if $i > j$ **then return** ε (Empty sequence); // Terminal case

rand $\leftarrow$ Random($\mathcal{Z}_{\left[\begin{smallmatrix} i,j \\ k \end{smallmatrix}\right]}$);

for $b \in \mathbb{B}$ **do** // Unpaired case

  rand $\leftarrow$ rand $- \mathbf{w}^{|b|_{GC}} \cdot \mathcal{Z}_{\left[\begin{smallmatrix} i+1,j \\ k-\sigma_{\omega_i,b} \end{smallmatrix}\right]}$;

  if rand < 0 **then return** $\left[\begin{smallmatrix} \bullet \\ b \end{smallmatrix}\right] \cdot$ GenMuts$(i+1, j, k-\sigma_{\omega_i,b}, \mathbf{w}, \omega)$;

for $b, b' \in \mathbb{B}^2$ **do** // Paired case

  for $l' \leftarrow i + \theta + 1$ **to** j **do**

    delta $\rightarrow l' - (i + \theta + 1)$; // Boustrophedon search

    if delta *is even* **then** $l \leftarrow i + \theta + 1 + \left\lfloor \frac{l'}{2} \right\rfloor$ **else** $l \leftarrow j - \left\lfloor \frac{l'-1}{2} \right\rfloor$;

    for $k' \leftarrow 0$ **to** $k - \sigma_{\omega_i\omega_l,bb'}$ **do**

      rand $\leftarrow$ rand $- \mathbf{w}^{|bb'|_{GC}} \cdot e^{-\frac{\Delta_{b,b'}}{RT}} \cdot \mathcal{Z}_{\left[\begin{smallmatrix} i+1,l-1 \\ k' \end{smallmatrix}\right]} \cdot \mathcal{Z}_{\left[\begin{smallmatrix} l+1,j \\ k-k'-\sigma_{\omega_i\omega_l,bb'} \end{smallmatrix}\right]}$;

      if rand < 0 **then**

        return $\left[\begin{smallmatrix} (\\ b \end{smallmatrix}\right] \cdot$ GenMuts$(i+1, l-1, k', \mathbf{w}, \omega) \cdot \left[\begin{smallmatrix}) \\ b' \end{smallmatrix}\right] \cdot$ GenMuts$(l+1, j, k-k'-\sigma_{\omega_i\omega_l,bb'}, \mathbf{w}, \omega)$;

Improved statistical sampling. Statistical sampling was introduced by Ding and Lawrence [19] and implemented within the `SFold` software. By contrast with previous algorithms which considered only the minimal free energy structure [4] or a deterministic subset of its suboptimals [23], this algorithm performs a stochastic backtrack and generates any suboptimal structure s for a sequence ω with respect to its Boltzmann probability (see Equation 1). Following a general weighted sampling scheme [24], the algorithm starts from an interval $[1, n]$, and chooses at each step one of the possible cases (First base being either unpaired or paired to some l) with probability proportional to the contribution of the case to the local partition function.

A direct adaptation of this principle based on Equation 3 gives Function `GenMuts` (setting $\mathbf{w} := 1$). By contrast with its original implementation [16], this sampling procedure uses a Boustrophedon search [25], decreasing the worst-case complexity of the stochastic backtrack from $\Theta(n^2 k)$ [16] to $\Theta(nk \log n)$. Therefore the generation of m structure/sequence couples at Hamming distance k of ω can be performed in $\Theta(n^3 \cdot k^2 + m \cdot nk \log n)$ worst-case complexity.

Reaching regions of predefined G+C-content. Now let us address the problem of sampling sequence/structure couples (ω', s') having predefined G+C-content $GC(s) = \frac{\#G(\omega)+\#C(\omega)}{|\omega|}$. As is is illustrated by Figure 1, the main difficulty is that the interplay between the Boltzmann distribution and the combinatorial explosion of sequences induces a drift of the expected G+C-content. Furthermore the G+C-content distribution is concentrated around its mean. Thus, a suitable sequence/structure will seldom be obtained by chance if the expected G+C-content does not match the targeted one. Our sampling procedure must also remain unbiased within areas of targeted G+C-content, i.e. generate each sequence/structure (ω', s') such that $\sigma_{\omega',\omega} = k$ and $GC(\omega') = gc^*$ with probability

$$p(\omega', s' \mid k, gc^*) = \frac{\mathcal{B}_{\omega'}(s')}{\sum_{\substack{(\omega'',s'') \text{ s.t.} \\ GC(\omega'')=gc^* \\ \text{and } \sigma_{\omega'',\omega}=k}} \mathcal{B}_{\omega''}(s'')}. \tag{4}$$

Algorithm 1: Rejection algorithm

Input : RNA ω, targeted G+C-content gc^*, number of samples m, number of mutations k and weight $\mathbf{w}$.
Output: Set of m sequence/structure samples

```
FillMatrices(ω,k,w);
samples ← ∅;
while |samples| < m do
    candidate ← GenMuts(1, n, ω, k, w);
    if GC(candidate) = gc* then samples ← samples ∪ {candidate};
return samples;
```

Direct rejection yields exponential-time sampling. A natural idea for achieving an unbiased sampling consists in sampling from the complete set of structure/sequence and reject sequences of unsuitable G+C-content. Since an unsuitable couple can be generated repeatedly, the worst-case complexity (infinite) of such an algorithm is perhaps not very informative. Therefore we propose an average-case analysis, using methods developed in the *analysis of algorithms* community to determine the asymptotical limit of the G+C-content distribution.

Theorem 1. *Assuming an homopolymer model (any base pair can form), a Nussinov-style energy function and an unconstrained number of mutations, the distribution of the number of* G + C *is asymptotically normal of mean $\mu \cdot n$ and standard deviation $\sigma\sqrt{n}$, for μ and σ positive real constants. The probability of sampling a sequence/structure of* G+C*-content gc^* is then*

$$p(gc^* \mid n) \sim \frac{1}{\sigma\sqrt{2\pi n}} \cdot e^{-\frac{n(gc^*-\mu)^2}{2\sigma^2}}. \tag{5}$$

The trials of Algorithm 1 are mutually independent, therefore we know that its expected number of calls to GenMuts is the inverse of the probability assigned to a G+C-content of gc^*. Unless $\mu = gc^*$ the average-case complexity is then dominated asymptotically by a term which is exponential in n and Algorithm 1 has exponential complexity for some (most) targeted G+C-contents.

A weighted sampling approach. We adapt a general approach recently proposed by Bodini *et al* [26], which uses weights to efficiently bias a random generation process towards areas of interest, while respecting a (renormalized) prior distribution. Namely let $\mathbf{w} \in \mathbb{R}^+$ be a weight associated with each occurrence of G or C, we define the $\mathbf{w}$-weighted partition function as

$$\mathcal{Z}_{[k]}^{[\mathbf{w}]} := \mathcal{Z}_{\begin{bmatrix}1,n\\k\end{bmatrix}}^{[\mathbf{w}]} = \sum_{\substack{\omega' \text{ s.t.}\\ \sigma_{\omega,\omega'}=k}} \sum_{s' \in \mathcal{S}_{\omega',\theta}} \mathcal{B}_{\omega'}(s') \cdot \mathbf{w}^{|\omega'|_{GC}}. \tag{6}$$

which can be computed by the following recurrence

$$\mathcal{Z}_{\begin{bmatrix}i,j\\k\end{bmatrix}}^{[\mathbf{w}]} = \sum_{b \in \mathbb{B}} \mathbf{w}^{|b|_{GC}} \mathcal{Z}_{\begin{bmatrix}i+1,j\\k-\sigma_{\omega_i,b}\end{bmatrix}}^{[\mathbf{w}]} + \sum_{b,b' \in \mathbb{B}^2} \sum_{l=i+\theta+1}^{j} \sum_{k'=0}^{k-\sigma_{\omega_i\omega_l,bb'}} \mathbf{w}^{|bb'|_{GC}} e^{-\frac{\Delta_{b,b'}}{RT}} \mathcal{Z}_{\begin{bmatrix}i+1,l-1\\k'\end{bmatrix}}^{[\mathbf{w}]} \mathcal{Z}_{\begin{bmatrix}l+1,j\\k-k'-\sigma_{\omega_i\omega_l,bb'}\end{bmatrix}}^{[\mathbf{w}]} \tag{7}$$

where $|x|_{GC} := n \cdot GC(x)$ denotes the number of occurrences of G or C within x. Upon multiplying by a weight $\mathbf{w}$ whenever a Guanine or Cytosine is generated, a $\mathbf{w}$-weighted probability distribution is induced on the sequence/structure and any sequence/structure (ω', s') such that $GC(\omega') = gc^*$ and $\sigma_{\omega,\omega'} = k$ has probability

$$p(\omega', s' \mid \mathbf{w}, k) = \frac{\mathbf{w}^{|x|_{GC}} \cdot \mathcal{B}_{\omega'}(s)}{\mathcal{Z}_{[k]}^{[\mathbf{w}]}}. \tag{8}$$

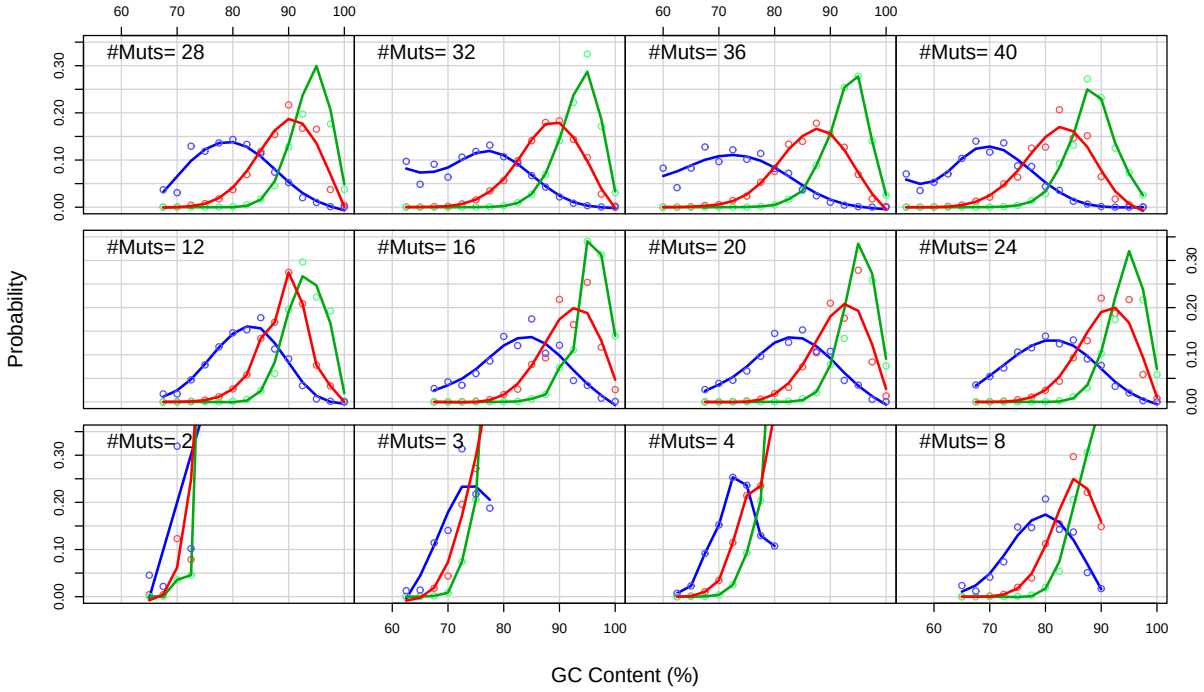

Fig. 1. Evolution of G+C-content along with the number of mutation within domain IIA of the internal ribosome entry site (Classical swine fever virus, PDB: 2HUA_A, seq: GGCCUCCAGCGACGGCCUUCGGGACUAGCAAACGGAGGCC). Red: Original Boltzmann distribution [16]; Green: Weighted sampling using $\mathbf{w} = 2$; Blue: Weighted sampling using $\mathbf{w} = 1/2$; In the unweighted model, a significant drift of the G+C-content ($70\% \rightarrow \sim 85\%$) is observed. This drift can be confined within a weighted model $\mathbf{w} = 1/2$ (Blue) or be accentuated $\mathbf{w} = 2$ (Green).

Function `GenMuts` implements a sampling procedure for the $\mathbf{w}$-weighted distribution. Processing its output with Algorithm 1 discards any structure/sequence whose G+C-content differs from gc^*, and the probability of sampling a structure/sequence (ω', s') of G+C-content gc^* is then

$$p'(\omega', s' \mid gc^*, \mathbf{w}, k) = \frac{\mathbf{w}^{|\omega'|_{GC}} \mathcal{B}_{\omega'}(s')}{\sum\limits_{\substack{(\omega'', s'') \\ \text{s.t. } GC(\omega'')=gc^* \\ \text{and } \sigma_{\omega'',\omega}=k}} \mathbf{w}^{|\omega''|_{GC}} \mathcal{B}_{\omega''}(s'')} = p(\omega', s' \mid gc^*, k)$$

since $|\omega'|_{GC} = |\omega''|_{GC} = n \cdot gc^*$. Our weighted sampling/rejection pipeline is consequently unbiased within the sequence/structures subset having the targeted G+C-content.

Let us now discussing the algorithmic gain achieved by this approach. Here, we assume that we have a weight $\mathbf{w}$ such that $gc^* = \mu_{\mathbf{w}}$. First, let us point out that the proof of Theorem 1 does not rely on any specificity of the energy model/weighted scheme, but rather on intrinsic properties (strong connectedness and aperiodicity) of the context-free grammar used to model the structure/sequence space. It follows that Theorem 1 holds even in the presence of weights, with an additional dependency in $\mathbf{w}$ for $\mu_{\mathbf{w}}$ the expected G+C-content and $\sigma_{\mathbf{w}}$ its standard deviation. It also follows that the exponential part of the complexity cancels out, and the expected number of calls to `GenMuts` drops to $\Theta(\sqrt{n})$ per sample. Consequently, the generation of m structure/sequence couples

Algorithm 2: Bisection algorithm

Input : RNA ω, targeted G+C-content gc^*, number of samples m and number of mutations k
Output: Set of m sequence/structure samples having G+C-content gc^* at Hamming distance k of ω

$(\mu_L, \mu_R) \leftarrow (0,0); (\mathbf{w}_L, \mathbf{w}_R) \leftarrow (0,0); \mathbf{w} \leftarrow 1;$
while |samples| $< m$ **do**
  `FillMatrices(`ω,k,$\mathbf{w}$`);`
  candidates $\leftarrow \varnothing$;
  for $x \leftarrow 1$ **to** M **do** `// Generate` $M := K \times m$ `candidates in the weighted distribution`
    candidates $\leftarrow$ candidates $\cup$ `GenMuts(`$1, n, k, \mathbf{w}, \omega$`);`
  for cand $\in$ candidates **do if** $GC(\text{cand}) = gc^*$ **then** samples $\leftarrow$ samples $\cup$ cand; `// Filter on G+C-Content`
  $\mu \leftarrow$ `EstimateMeanGC(candidates);`
  if $\mu < gc^*$ and $\mu_R \leq \mu$ **then** $(\mu_R, \mathbf{w}_R, \mathbf{w}) \leftarrow (\mu, \mathbf{w}, 2 \cdot \mathbf{w})$; `// Update weights`
  else
    if $\mu_L < \mu < gc^*$ **then** $(\mu_L, \mathbf{w}_L) \leftarrow (\mu, \mathbf{w})$;
    if $gc^* < \mu < \mu_R$ **then** $(\mu_R, \mathbf{w}_R) \leftarrow (\mu, \mathbf{w})$;
    $\mathbf{w} \leftarrow (\mathbf{w}_L + \mathbf{w}_R)/2$;
return samples;

at Hamming distance k of ω and G+C-content gc^* can be performed in $\Theta(n^3 \cdot k^2 + m \cdot n\sqrt{n} \cdot k \log n)$ average-case complexity.

Adaptive weighted sampling. To conclude, we need to find a weight $\mathbf{w}^*$ such that $gc^* = \mu_{\mathbf{w}^*}$. We claim that $\mathbf{w}^*$ can be computed by Algorithm 2 using a bisection method, based on the observation that $\mu_{\mathbf{w}}$ is a strictly increasing function of $\mathbf{w}$. Indeed, let $u_{gc^*,k}$ be the cumulated Boltzmann factors over all structures/sequences at distance k of ω, having G+C-content gc^*, then the probability of generating a sequence with G+C-content gc is exactly $p''_{\mathbf{w},gc,k} := u_{gc,k} \cdot \mathbf{w}^{n \cdot gc} / \mathcal{Z}^{[\mathbf{w}]}_{[k]}$. It follows that

$$\mu_{\mathbf{w},k} = \sum_{x=0}^{n} \frac{x}{n} \cdot p''_{\mathbf{w},gc,k} = \sum_{x=0}^{n} \frac{x}{n} \cdot \frac{u_{x/n,k} \cdot \mathbf{w}^x}{\mathcal{Z}^{[\mathbf{w}]}_{[k]}} \quad \Rightarrow \quad \frac{\partial \mu_{\mathbf{w},k}}{\partial \mathbf{w}} = \sum_{x=0}^{n} \frac{x^2 \cdot u_{x/n,k} \cdot \mathbf{w}^{x-1}}{n \cdot \mathcal{Z}^{[\mathbf{w}]}_{[k]}} > 0, \forall k > 0. \quad (9)$$

Implementation remarks. Our implementation of Algorithm 2 uses sampled sets to estimate expected G+C-contents. Since the G+C-content asymptotically follows a normal law of standard deviation in $\sigma\sqrt{n}$, a sampled set of size $M := K \times m \in \Omega(4n\sigma^2/\varepsilon^2)$, for some $K > 1$, will guarantee a 95% probability of falling within a confidence interval of $[(1-\varepsilon)\mathbf{w}^*, (1+\varepsilon)\mathbf{w}^*], \forall \varepsilon > 0$. The generation of such a growing number of samples will however remain negligible compared to the computation of the partition function. The expected value of $\mu_{\mathbf{w}}$ can also be computed exactly in $\Theta(n^3 \cdot k^2)$ using dynamic programming, following ideas pioneered in Miklos *et al* [27].

The value of $\mathbf{w}^*$ can also be exactly computed. Indeed the partition function can be expressed as $\mathcal{Z}^{\mathbf{w}}_{[k]} = \sum_{x=0}^{n} u_{x/n,k} \cdot \mathbf{w}^x$, i.e. a polynomial of degree n in $\mathbf{w}$. Therefore it suffices to evaluate $\mathcal{Z}^{\mathbf{w}}_{[k]}$ at n different values of $\mathbf{w}$ to determine the coefficients $u_{x/n,k}$ using Gaussian elimination. From there, we can use numerical recipes (e.g. Grobner bases [28]) to find the unique root $\mathbf{w}^*$ of the polynom:

$$gc^* = \sum_{x=0}^{n} \frac{x \cdot u_{x/n,k} \cdot \mathbf{w}^{*x}}{n \cdot \mathcal{Z}^{\mathbf{w}^*}_{[k]}} \Leftrightarrow 0 = \sum_{x=0}^{n} (x - n \cdot gc^*) \cdot u_{x/n,k} \cdot \mathbf{w}^{*x}.$$

Since the weighted partition functions $\mathcal{Z}^{\mathbf{w}}_{[k]}$ are computed prior to sampling, we can adopt an hybrid approach. We initially apply the bisection and we switch to an exact computation after n computations of $\mathcal{Z}^{\mathbf{w}}_{[k]}$.

Fig. 2. X-axis: Number of mutations in mutants. Y-axis: Number of stacks in mutant secondary structures. Blue: 10% GC, Green: 30%, Yellow: 50%, Orange: 70%, Red: 90%.

Finally let us remark that the relative probabilities of structure/sequences within the set of suitable G+C-contents are not affected by the introduction of weights. Therefore samples obtained during any iteration of the bisection method can be accumulated into a sample set, and returned when the targeted number of samples m is reached. This sampling strategy provably yields an unbiased sampled set.

3 Results

Now we illustrate how `RNAmutants` can be used to explore RNA sequence-structure maps and analyze an evolutionary scenario based on the improvement of the structure stability. This study can be motivated by a recent work of Cowperthwaite *et al.* [12] showing that energetically stable single stem structures correlate with the abundances of RNA sequences in the Rfam database [29].

Benchmark methodology. In these experiments, we analyzed sequences of size 20, 30 and 40 nucleotides. We also defined five G+C-content regimes at 10%, 30%, 50%, 70% and 90% (±10%). For each G+C-content we generated 20 seeds of length 20 and 30, and 10 seeds of length 40. Thus yielding a total of 250 seeds.

For each seed we ran `RNAmutants` and sampled at least 200 secondary structures in each k-neighborhood[3]. Each run explores the complete mutational landscape (i.e. 4^n sequences where n is the length of the sequence) and currently takes less than a minute for a size of 20 nucleotides, about 45 minutes for a 30 nucleotides, and about 5 hours for 40 nucleotides. In each experiment, we report the evolution of four parameters for each value of k (i.e. number of mutations). Namely, the number of stacks in the secondary structures sampled with the mutants (See Fig. 2), the number of bulges and internal loops (See Fig. 3), and the entropy of the sampled sequences (See Fig. 4).

Low G+C-contents favor structural diversity. In these experiments, we seek to characterize how sequences may constrain the variety of structures. RNA secondary structures can be characterized by their number of hairpins, stacks, bulges, internal loops and multi-loops. Here, because our sequences are relatively small, the large majority of the structures have a single stem shape. Thus, they have a single hairpin and no multi-loop, and we choose to report only the number of stacks and loops (bulges and internal loops). In Fig. 2 and 3 we report these statistics in each k-neighborhood of the seed.

Since the number of stacks correlates with single stems structures and thus more stable structures, one could expect that the number of stacks will naturally increase with the number of mutations. This intuition explains the results of simulations performed on sequences of length 20 (See Fig. **??**). However, surprisingly, this property does not hold for longer sequences with low G+C-contents. In Fig. **??** and **??**, we observe that the number of stacks increases first, and then drops for large numbers of mutations (approximately $k \geq n/3$). Symmetrically, the number of bulges and internal loops initially drops and then increases.

[3] Our implementation allows to input a minimal number of sequences to sample at a targeted G+C-content in each k neighborhood. Since we keep all samples generated at each round, this lower bound typically produces about 1000 samples per value of k.

Fig. 3. X-axis: Number of mutations in mutants. Y-axis: Number of bulges and internal loops in mutant secondary structures. Blue: 10% GC, Green: 30%, Yellow: 50%, Orange: 70%, Red: 90%.

Fig. 4. X-axis: Number of mutations in mutants. Y-axis: Entropy of sampled mutant sequences. Blue: 10% GC, Green: 30%, Yellow: 50%, Orange: 70%, Red: 90%. Dotted line represents the maximal entropy value that can be obtained for GC contents of 30% and 70%. And the dashed line represents the maximal entropy value for GC contents of 10% and 90%.

These experiments enable us estimate the strength of an evolutionary pressure which stems from an improvement of the stability of the folds. Our data indicate that for short period of evolution this "structural" pressure is always dominant. But after longer periods of evolution, low G+C-contents enable more diversity in the structural ensembles. In other words, if we make the assumption that bulges and internal loops represent more sophisticated structures that could be associated to functional shapes. Then, under this scenario, we showed that the structures are first stabilized (i.e. backbone is created) and subsequently refined for functions.

Our results suggest a couple of hypothesis. First the size is an important factor of the structural diversity, and the analysis of sequence-structure maps of sequences of length larger than 20 may result in very different conclusions than those drawn for small sequences [15,12]. Next, sequences with a low G+C-content (below 40%) may allow a broader "choice" of structures. Low G+C-contents seem to favor the apparition of bulges and internal loops, making the apparition of non-canonical interactions and RNA 3D motifs [30,31,32] easier. Such tertiary structure motifs are frequently associated with specific RNA functions, and we conjecture that low G+C-contents favor their synthesis.

High G+C-contents reduce the sequence diversity Our next analysis aims to reveal how the structural stability (i.e. the folding energy) may influence the diversity of sequences and then the mutational space explored across evolution. We need for that to compute the entropy of the sequences in each k-neighborhood. First, we align all k-mutants and compute the Shannon entropy at position i: $\sigma(i) = \sum_{x=\{A,C,G,U\}} -f_i(x) \cdot \log_4(f_i(x))$, where $f_i(x)$ is the frequency of the nucleotide x in the i-th column of the alignment. Then, we average these measures and compute the average entropy per position $1/N \cdot \sum_{i=i}^{N} \sigma(i)$, where N is the length of the alignment (i.e. also the length the sequences and the target structure since no gaps are allowed). Our results are shown in Fig. 4.

Before discussing these results, we note that the G+C-content bias the entropy values. Indeed, when the distribution of nucleotides is no longer uniform among all nucleotides (i.e. when the G+C-content is shifted away from 50%), the maximal entropy value decreases. We report the theoretical limits reachable for G+C-contents of 30% and 70% (roughly equal to 0.94 and indicated with a dotted line in Fig. 4), and 10% and 90% (approximately 0.74 and indicated with a dashed line in Fig. 4). Obviously, the upper bound for a GC content of 50% is 1.

Once again, as expected the maximum entropy is reached for sequences with a G+C-content of 50%. Medium GC contents offer a larger sequence accessibility. More interestingly, the maximal entropy value reached in these experiments seems to vary between extreme G+C-contents regimes. We observe that sequences at 10% of GC achieve the optimal entropy value, but that sequences at 90% GC significantly fail to explore the complete mutational landscape. This remark suggests that high G+C-contents reduce the evolutionary accessibility and the variety of sequences designed under this scenario. Finally, unlike our previous experiments (cf. section 3), we notice that the size of the sequence has no influence on these results.

4 Conclusion

In this paper, we showed how adaptive sampling techniques can be used to explore regions poorly covered by classical sampling algorithms. We applied this methodology to `RNAmutants`, and showed how regions of the mutational landscape with low `G+C`-contents could be efficiently sampled and analyzed.

Importantly, the techniques developed in this work can be generalized to many other sequential and structural additive properties, such as the number of mutations, number of base pairs or the free energy. The versatility of these techniques suggests a broad range of novel applications as well as algorithm improvements.

This methodology is particularly well-suited to the exploration of large sequence-structure maps. We expect that their application in various ways will reveal novel properties of the RNA evolutionary landscapes [14,15,1,13]. More practically, as recently reported by Barash and Churkin, our algorithms are also particularly well suited to predict deleterious mutations in structural RNAs [33]. We expect that our adaptive sampling algorithm will help improve our current prediction accuracy.

All these algorithms have been implemented in a new version of our `RNAmutants` software suite available at `http://csb.cs.mcgill.ca/RNAmutants`. This new distribution includes various new features such as an RNA duplex model for simple hybridizations and weighted substitution events.

References

1. Cowperthwaite, M., Meyers, L.: How mutational networks shape evolution: Lessons from RNA models. Annual Review of Ecology, Evolution, and Systematics (2008) 203–230
2. Halvorsen, M., Martin, J.S., Broadaway, S., Laederach, A.: Disease-associated mutations that alter the RNA structural ensemble. PLoS Genet **6**(8) (2010)
3. Westhof, E.: Toward atomic accuracy in RNA design. Nat Methods **7**(4) (Apr 2010) 272–3
4. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res **9**(1) (Jan 1981) 133–48
5. Parisien, M., Major, F.: The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature **452**(7183) (Mar 2008) 51–5
6. McCaskill, J.S.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers **29**(6-7) (1990) 1105–19
7. Turner, D.H., Mathews, D.H.: NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res **38**(Database issue) (Jan 2010) D280–2
8. Mathews, D.H.: RNA secondary structure analysis using RNAstructure. Curr Protoc Bioinformatics **Chapter 12** (Mar 2006) Unit 12.6
9. Hofacker, I.L.: RNA secondary structure analysis using the vienna RNA package. Curr Protoc Bioinformatics **Chapter 12** (Jun 2009) Unit12.2
10. Markham, N.R., Zuker, M.: UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol **453** (2008) 3–31
11. Grüner, W. Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I., Stadler, P., Schuster, P.: Analysis of RNA sequence structure maps by exhaustive enumeration i. neutral networks. Monatshefte f. Chemie **127**(4) (1995) 355–374
12. Cowperthwaite, M.C., Economo, E.P., Harcombe, W.R., Miller, E.L., Meyers, L.A.: The ascent of the abundant: how mutational networks constrain evolution. PLoS Comput Biol **4**(7) (2008) e1000110
13. Stich, M., Lázaro, E., Manrubia, S.C.: Phenotypic effect of mutations in evolving populations of RNA molecules. BMC Evol Biol **10** (2010) 46
14. Schuster, P., Fontana, W., Stadler, P.F., Hofacker, I.L.: From sequences to shapes and back: a case study in RNA secondary structures. Proc Biol Sci **255**(1344) (Mar 1994) 279–84
15. Reidys, C., Stadler, P.F., Schuster, P.: Generic properties of combinatory maps: neutral networks of RNA secondary structures. Bull Math Biol **59**(2) (Mar 1997) 339–97
16. Waldispühl, J., Devadas, S., Berger, B., Clote, P.: Efficient algorithms for probing the RNA mutation landscape. PLoS Comput Biol **4**(8) (2008) e1000124
17. Waldispühl, J., Behzadi, B., Steyaert, J.M.: An approximate matching algorithm for finding (sub-)optimal sequences in S-attributed grammars. Bioinformatics **18 Suppl 2** (2002) S250–9
18. Clote, P., Waldispühl, J., Behzadi, B., Steyaert, J.M.: Energy landscape of k-point mutants of an RNA molecule. Bioinformatics **21**(22) (Nov 2005) 4140–7
19. Ding, Y., Lawrence, C.E.: A bayesian statistical algorithm for RNA secondary structure prediction. Comput Chem **23**(3-4) (Jun 1999) 387–400
20. Chan, C.Y., Carmack, C.S., Long, D.D., Maliyekkel, A., Shao, Y., Roninson, I.B., Ding, Y.: A structural interpretation of the effect of gc-content on efficiency of RNA interference. BMC Bioinformatics **10 Suppl 1** (2009) S33
21. Nussinov, R., Jacobson, A.: Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci U S A **77** (1980) 6903–13
22. Waterman, M.S.: Secondary structure of single stranded nucleic acids. Advances in Mathematics Supplementary Studies **1**(1) (1978) 167–212
23. Wuchty, S., Fontana, W., Hofacker, I., Schuster, P.: Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers **49** (1999) 145–164
24. Denise, A., Ponty, Y., Termier, M.: Controlled non uniform random generation of decomposable structures. Theoretical Computer Science **411**(40-42) (September 2010) 3527–3552
25. Ponty, Y.: Efficient sampling of RNAsecondary structures from the boltzmann ensemble of low-energy: The boustrophedon method. Journal of Mathematical Biology **56**(1-2) (Jan 2008) 107–127
26. Bodini, O., Ponty, Y.: Multi-dimensional boltzmann sampling of languages. DMTCS Proceedings **0**(01) (2010)
27. Mikls, I., Meyer, I.M., Nagy, B.: Moments of the boltzmann distribution for rna secondary structures. Bull Math Biol **67**(5) (Sep 2005) 1031–1047
28. Faugere, J.C.: A new efficient algorithm for computing Gröbner bases (f4). Journal of Pure and Applied Algebra **139**(1–3) (June 1999) 61–88

29. Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R., Bateman, A.: Rfam: updates to the rna families database. Nucleic Acids Res **37**(Database issue) (Jan 2009) D136–40
30. Djelloul, M., Denise, A.: Automated motif extraction and classification in RNA tertiary structures. RNA **14**(12) (Dec 2008) 2489–97
31. Lemieux, S., Major, F.: RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. Nucleic Acids Res **30**(19) (Oct 2002) 4250–63
32. Leontis, N.B., Lescoute, A., Westhof, E.: The building blocks and motifs of RNA architecture. Curr Opin Struct Biol **16**(3) (Jun 2006) 279–87
33. Barash, D., Churkin, A.: Mutational analysis in RNAs: comparing programs for RNA deleterious mutation prediction. Briefings in Bioinformatics (2010)
34. Lorenz, W., Ponty, Y., Clote, P.: Asymptotics of RNA shapes. Journal of Computational Biology **15**(1) (Jan–Feb 2008) 31–63
35. Drmota, M.: Systems of functional equations. Random Struct. Alg. **10** (1997) 103–124

A Appendix

A.1 Proof of convergence of G+C-content toward a normal law

Theorem 2. *Assuming an homopolymer model (any base pair can form), a Nussinov-style energy function and an unconstrained number of mutations, the distribution of the* G+C*-content is asymptotically normal, and the probability of sampling at any* G+C*-content* $gc^* \in [0,1]$ *is*

$$p(gc^* \mid n) \sim \frac{1}{\sigma\sqrt{2\pi n}} \cdot e^{-\frac{n(gc^*-\mu)^2}{2\sigma^2}} \tag{10}$$

for μ *and* σ *real numbers independent of* n*.*

Proof. Let n be the length of our input sequence, gc the targeted G+C-content, and let $\Delta_{a,b}$ be the free-energy contribution of a base pair (a,b). Remark that the entire set of sequence/secondary structure couples can be generated by the following context-free grammar:

$$S \to \begin{bmatrix}\bullet\\ \texttt{A}\end{bmatrix} S \mid \begin{bmatrix}\bullet\\ \texttt{C}\end{bmatrix} S \mid \begin{bmatrix}\bullet\\ \texttt{G}\end{bmatrix} S \mid \begin{bmatrix}\bullet\\ \texttt{U}\end{bmatrix} S \mid \begin{bmatrix}(\\ \texttt{A}\end{bmatrix} T \begin{bmatrix})\\ \texttt{U}\end{bmatrix} S \mid \begin{bmatrix}(\\ \texttt{C}\end{bmatrix} T \begin{bmatrix})\\ \texttt{G}\end{bmatrix} S \mid \begin{bmatrix}(\\ \texttt{G}\end{bmatrix} T \begin{bmatrix})\\ \texttt{C}\end{bmatrix} S \mid \begin{bmatrix}(\\ \texttt{G}\end{bmatrix} T \begin{bmatrix})\\ \texttt{U}\end{bmatrix} S \mid \begin{bmatrix}(\\ \texttt{U}\end{bmatrix} T \begin{bmatrix})\\ \texttt{A}\end{bmatrix} S \mid \begin{bmatrix}(\\ \texttt{U}\end{bmatrix} T \begin{bmatrix})\\ \texttt{G}\end{bmatrix} S \mid \varepsilon$$

$$T \to \begin{bmatrix}\bullet\\ \texttt{A}\end{bmatrix} S \mid \begin{bmatrix}\bullet\\ \texttt{C}\end{bmatrix} S \mid \begin{bmatrix}\bullet\\ \texttt{G}\end{bmatrix} S \mid \begin{bmatrix}\bullet\\ \texttt{U}\end{bmatrix} S \mid \begin{bmatrix}(\\ \texttt{A}\end{bmatrix} T \begin{bmatrix})\\ \texttt{U}\end{bmatrix} S \mid \begin{bmatrix}(\\ \texttt{C}\end{bmatrix} T \begin{bmatrix})\\ \texttt{G}\end{bmatrix} S \mid \begin{bmatrix}(\\ \texttt{G}\end{bmatrix} T \begin{bmatrix})\\ \texttt{C}\end{bmatrix} S \mid \begin{bmatrix}(\\ \texttt{G}\end{bmatrix} T \begin{bmatrix})\\ \texttt{U}\end{bmatrix} S \mid \begin{bmatrix}(\\ \texttt{U}\end{bmatrix} T \begin{bmatrix})\\ \texttt{A}\end{bmatrix} S \mid \begin{bmatrix}(\\ \texttt{U}\end{bmatrix} T \begin{bmatrix})\\ \texttt{G}\end{bmatrix} S \tag{11}$$

Under the hypotheses of Theorem 1, the process of producing a sample using stochastic backtrack within `RNAMutants` is provably equivalent to drawing a word of length n from the grammar with respect to a suitable probability distribution. Namely, a Boltzmann distribution can be exactly reproduced by adjoining weights to the productions of the grammar. The reader is referred to previous works by one of the authors who performed a similar analysis of a statistical sampling algorithm [25], and described a general weighted framework for context-free grammars [24].

Consider the bivariate generating functions $S(z,u) = \sum_{n\geq 0}\sum_{k\geq 0} s_{n,k} z^n u^k$ (resp. $T(z,u) = \sum_{n\geq \theta}\sum_{k\geq 0} s_{n,k} z^n u^k$) which counts the cumulated weight $s_{n,k}$ of all words of length n having k occurrences of G+C generated from the non-terminal S (resp. T). Since the grammar in Equation 11 is unambiguous, a system of equations involving the generating functions $S(z,u)$ and $T(z,u)$ can be established through a direct translation of the grammar productions [34,24]. One then obtains

$$\begin{cases} S(z,u) = z(2+2u)S(z,u) \\ \qquad\qquad +z^2\left(2e^{\frac{-\Delta_{A,U}}{RT}} + 2e^{\frac{-\Delta_{G,U}}{RT}}u + 2e^{\frac{-\Delta_{G,C}}{RT}}u^2\right) S(z,u)T(z,u) + 1 \\ T(z,u) = z(2+2u)S(z,u) \\ \qquad\qquad +z^2\left(2e^{\frac{-\Delta_{A,U}}{RT}} + 2e^{\frac{-\Delta_{G,U}}{RT}}u + 2e^{\frac{-\Delta_{G,C}}{RT}}u^2\right) S(z,u)T(z,u). \end{cases} \tag{12}$$

Remark that the system is strongly connected (Each non-terminal makes use of the other) and aperiodic (Words of any parity can be generated). Applying a striking result by Drmota [35], we directly conclude that the distribution of the number of occurrence of G+C is asymptotically normal of mean μn and standard deviation $\sigma\sqrt{n}$, for some constants $\mu, \sigma > 0$. Remarking that

the probability of a given G+C-content gc^* is also the probability of observing $n \cdot gc^*$ occurrence of $\mathtt{G}+\mathtt{C}$ and consequently follows

$$p(gc^* \mid n) \sim \frac{1}{\sigma\sqrt{2\pi n}} e^{-\frac{n(gc^*-\mu)^2}{2\sigma^2}}. \tag{13}$$

A.2 Multivariate normal joint distributions for G+C-content and mutations

It is worth noticing that Drmota's theorem [35] also covers the multidimensional case, where similar conditions are shown to yield multivariate normal distributions. For instance, if we assume the initial sequence to be $\omega := \mathtt{A}^n$, then any production of $\{\mathtt{C}, \mathtt{G}, \mathtt{U}\}$ constitutes a mutation, and the grammar shown in Equation 11 translates into the following system involving the trivariate generating functions

$$\begin{cases} S(z,u,v) = z(1+v+2uv)S(z,u,v) \\ \qquad\qquad +z^2\left(2ve^{\frac{-\Delta_{A,U}}{RT}} + 2uv^2e^{\frac{-\Delta_{G,U}}{RT}} + 2u^2v^2e^{\frac{-\Delta_{G,C}}{RT}}\right)S(z,u,v)T(z,u,v) + 1 \\ T(z,u,v) = z(1+v+2uv)S(z,u,v) \\ \qquad\qquad +z^2\left(2ve^{\frac{-\Delta_{A,U}}{RT}} + 2uv^2e^{\frac{-\Delta_{G,U}}{RT}} + 2u^2v^2e^{\frac{-\Delta_{G,C}}{RT}}\right)S(z,u,v)T(z,u,v). \end{cases} \tag{14}$$

where $S(z,u,v) = \sum_{n\geq 0}\sum_{g\geq 0}\sum_{k\geq 0} s_{n,g,k} z^n u^g v^k$ and $T(z,u,v) = \sum_{n\geq \theta}\sum_{g\geq 0}\sum_{k\geq 0} s_{n,g,k} z^n u^g v^k$ with $s_{n,g,k}$ the total weight of sequence/structures couples of length n with g occurrences of $\mathtt{G}+\mathtt{C}$ and k mutations. The above system is again aperiodic and strongly-connected, therefore Drmota's theorem [35] applies and the couple of random variables (G_n, K_n), denoting respectively the expected G+C-content and expected proportion of mutations, follow a bivariate normal distribution of mean vector $\boldsymbol{\mu}_n := \mathbb{E}\left((G_n, K_n) \mid n\right)$ and covariance matrix $\boldsymbol{\Sigma}_n$ such that

$$\boldsymbol{\mu}_n = n \cdot (\mu_G, \mu_K) \quad \text{and} \quad \boldsymbol{\Sigma}_n = n \cdot \begin{pmatrix} \sigma_{GG} & \sigma_{GK} \\ \sigma_{KG} & \sigma_{KK} \end{pmatrix}$$

where μ_X and σ_{XY} are positive real values independent on n.

A.3 Dynamic programming alternative to the adaptive sampling algorithm

An exact, targeted G+C-content can also be enforced explicitly at the level of the dynamic programming equation. Namely one only needs to introduce the number of occurrences of $\mathtt{G}+\mathtt{C}$ as an additional parameter gc^*. The partition function $\mathcal{Z}_{[i,j,k,gc^*]}$ restricted to the interval (i,j), allowing for exactly k mutations, and considering only mutated sequences with gc^* G+C-content follows

$$\begin{aligned} \mathcal{Z}_{[i,j,k,gc]} &= \sum_{b\in\mathbb{B}} \mathcal{Z}_{\begin{bmatrix} i+1,j \\ k-\sigma_{\omega_i,b} \\ gc-|b|_G \end{bmatrix}} \\ &\quad + \sum_{b,b'\in\mathbb{B}^2} \sum_{l=i+\theta+1}^{j} \sum_{k'=0}^{k-\sigma_{\omega_i\cdot\omega_l,b.b'}} \sum_{gc'=0}^{gc-|b.b'|_{G.G}} e^{-\frac{\Delta_{b,b'}}{RT}} \mathcal{Z}_{\begin{bmatrix} i+1,l-1 \\ k' \\ gc' \end{bmatrix}} \mathcal{Z}_{\begin{bmatrix} l+1,j \\ k-k'-\sigma_{\omega_i\cdot\omega_l,b.b'} \\ gc-gc'-|b.b'|_{G.G} \end{bmatrix}}. \end{aligned}$$

Statistical sampling can then be performed through a slight refinement of Function `GenMuts`.

The increase in complexity due to this explicit control of the G+C-content is in $\Theta(n^2)$ in time and $\Theta(n)$ in memory, bringing the overall time complexity of the precomputation to $\Theta(n^7)$ (assuming $k = \Theta(n)$) for the computation of the partition function. Although this approach is impractical for large values of n, it is exact and can easily be transformed (through a change of algebra dear to R. Giegerich) into an algorithm for computing the minimal free energy structure/sequence for any G+C-content.

A.4 Exact computation of expected G+C-content

Adapting an idea of Miklos*et al* [27], one can extract exactly the expectation of the G+C-content. Indeed let us observe that the expectation $\mu_{\mathbf{w},k}$ of G+C-content for k mutations obeys

$$\mu_{\mathbf{w},k} = \sum_{x=0}^{n} \frac{x}{n} \cdot p''_{\mathbf{w},gc,k} = \sum_{x=0}^{n} \frac{x}{n} \cdot \frac{u_{x/n,k} \cdot \mathbf{w}^x}{\mathcal{Z}^{[\mathbf{w}]}_{[k]}} = \frac{\sum_{x=0}^{n} x \cdot u_{x/n,k} \cdot \mathbf{w}^x}{n \cdot \mathcal{Z}^{[\mathbf{w}]}_{[k]}} := \frac{\mathcal{Z}^{\bullet[\mathbf{w}]}_{[k]}}{n \cdot \mathcal{Z}^{[\mathbf{w}]}_{[k]}}. \tag{15}$$

Using a formal derivative construct (To be presented in a future paper), we readily obtain from Equation 7 the following recurrence for $\mathcal{Z}^{\bullet[\mathbf{w}]}_{[k]} := \mathcal{Z}^{\bullet[\mathbf{w}]}_{\begin{bmatrix}1,n\\k\end{bmatrix}}$ through

$$\begin{aligned}
\mathcal{Z}^{\bullet[\mathbf{w}]}_{\begin{bmatrix}i,j\\k\end{bmatrix}} = \sum_{b\in\mathbb{B}} \mathbf{w}^{|b|_{GC}} & \left(|b|_{GC} \cdot \mathcal{Z}^{[\mathbf{w}]}_{\begin{bmatrix}i+1,j\\k-\sigma_{\omega_i,b}\end{bmatrix}} + \mathcal{Z}^{\bullet[\mathbf{w}]}_{\begin{bmatrix}i+1,j\\k-\sigma_{\omega_i,b}\end{bmatrix}} \right) \\
+ \sum_{b,b'\in\mathbb{B}^2} \sum_{l=i+\theta+1}^{j} \sum_{k'=0}^{k-\sigma_{\omega_i\omega_l,bb'}} & \mathbf{w}^{|bb'|_{GC}} \cdot e^{-\frac{\Delta_{b,b'}}{RT}} \left(|bb'|_{GC} \cdot \mathcal{Z}^{[\mathbf{w}]}_{\begin{bmatrix}i+1,l-1\\k'\end{bmatrix}} \cdot \mathcal{Z}^{[\mathbf{w}]}_{\begin{bmatrix}l+1,j\\k-k'-\sigma_{\omega_i\omega_l,bb'}\end{bmatrix}} \right. \\
& \left. + \mathcal{Z}^{\bullet[\mathbf{w}]}_{\begin{bmatrix}i+1,l-1\\k'\end{bmatrix}} \cdot \mathcal{Z}^{[\mathbf{w}]}_{\begin{bmatrix}l+1,j\\k-k'-\sigma_{\omega_i\omega_l,bb'}\end{bmatrix}} + \mathcal{Z}^{[\mathbf{w}]}_{\begin{bmatrix}i+1,l-1\\k'\end{bmatrix}} \cdot \mathcal{Z}^{\bullet[\mathbf{w}]}_{\begin{bmatrix}l+1,j\\k-k'-\sigma_{\omega_i\omega_l,bb'}\end{bmatrix}} \right)
\end{aligned}$$

where $\mathcal{Z}^{\bullet[\mathbf{w}]}_{\begin{bmatrix}i,i-1\\k\end{bmatrix}} = 0$ and $\mathcal{Z}^{[\mathbf{w}]}_{\begin{bmatrix}i,j\\k\end{bmatrix}}$ is computed as stated in Equation 7. It follows that $\mathcal{Z}^{\bullet[\mathbf{w}]}_{[k]}$, and then $\mu_{\mathbf{w},k}$, can be computed in time $\Theta(n^3k^2)$.