JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 25, Number 11, 2018
© Mary Ann Liebert, Inc.
Pp. 1179–1192
DOI: 10.1089/cmb.2018.0075

An Efficient Dual Sampling Algorithm with Hamming Distance Filtration

CHRISTOPHER BARRETT,[1,2] QIJUN HE,[1]
FENIX W. HUANG,[1] and CHRISTIAN M. REIDYS[1,3,4]

ABSTRACT

Recently, a framework considering ribonucleic acid (RNA) sequences and their RNA secondary structures as pairs has led to new information theoretic perspectives on how the semantics encoded in RNA sequences can be inferred. In this context, the pairing arises naturally from the energy model of RNA secondary structures. Fixing the sequence in the pairing produces the RNA energy landscape, whose partition function was discovered by McCaskill. Dually, fixing the structure induces the energy landscape of sequences. The latter has been considered for designing more efficient inverse folding algorithms. In this work, we present the dual partition function filtered by Hamming distance, together with a Boltzmann sampler using novel dynamic programming routines for the loop-based energy model. The time complexity of the algorithm is $O(h^2n)$, where h, n are Hamming distance and sequence length, respectively, reducing the time complexity of samplers, reported in the literature by $O(n^2)$. We then present two applications, the first in the context of the evolution of natural sequence–structure pairs of microRNAs and the second in constructing neutral paths. The former studies the inverse folding rate (IFR) of sequence–structure pairs, filtered by Hamming distance, observing that such pairs evolve toward higher levels of robustness, that is, increasing IFR. The latter is an algorithm that constructs neutral paths: given two sequences in a neutral network, we employ the sampler to construct short paths connecting them, consisting of sequences all contained in the neutral network.

Keywords: Boltzmann sampler, inverse folding, neutral path, partition function, sequence–structure pairs.

1. INTRODUCTION

Ribonucleic acid (RNA) is a polymeric molecule essential in various biological roles. RNA consists of a single strand of nucleotides (**A, C, G, U**) that can fold and bond to itself through base pairings. At first, RNA was regarded as a simple messenger—the conveyor of genetic information from its repository in DNA to the ribosomes. Over the last several decades, however, researchers have discovered an increasing number

[1] Biocomplexity Institute of Virginia Tech, Blacksburg, Virginia.
Departments of [2] Computer Science and [3] Mathematics, Virginia Tech, Blacksburg, Virginia.
[4] Thermo Fisher Scientific Fellow in Advanced Systems for Information Biology, Thermo Fisher Scientific, Waltham, Massachusetts.

of important roles carried out by RNA. RNAs have been found to have catalytic activities, to participate in processing of messenger RNAs, to help maintain the telomeres of eukaryotic chromosomes, and to influence gene expression in multiple ways (Breaker and Joyce, 1994; Breaker, 1996; Serganov and Patel, 2007; Darnell, 2011). The specific shape into which RNAs fold plays a major role in their function, which makes RNA folding of prime interest to scientists. An understanding of RNA's three-dimensional structure will allow a greater understanding of RNA function. However, obtaining these three-dimensional structures through crystallization is often costly and time consuming. Accordingly, we consider coarse-grained RNA structures, the most prominent of which being RNA secondary structures. The latter are contact structures with noncrossing arcs when presented as a diagram (Fig. 1).

The key feature of RNA secondary structures is that they can be inductively constructed[1] (Stein and Everett, 1978). Waterman and other researchers (Kleitman, 1970; Nussinov et al., 1978; Stein and Everett, 1978; Waterman, 1978) studied the combinatorics and folding of RNA secondary structures. The noncrossing arcs of RNA secondary structures allow for a recursive build as follows. Let $S_2(n)$ denote the number of RNA secondary structures over n nucleotides, then we have (Waterman, 1978): $S_2(n)=S_2(n-1)+\sum_{j=0}^{n-3} S_2(n-2-j)S_2(j)$, where $S_2(n)=1$ for $0 \leq n \leq 2$. The recursion forms the basis of more than three decades of research resulting in what is referred to as the dynamic programming (DP) paradigm. The DP paradigm allows one to compute minimum free energy (MFE) structure in $O(n^3)$ time and $O(n^2)$ space. Implementations of these DP folding algorithms are mfold and *ViennaRNA* (Zuker and Stiegler, 1981; Hofacker et al., 1994), employing the energy values derived in Mathews et al. (1999) and Turner and Mathews (2010). The so-called inverse folding, that is, identifying sequences that realize a given structure as an MFE structure, has been studied in Hofacker et al. (1994) and Busch and Backofen (2006).

MFE folding naturally induces a genotype–phenotype (sequence to structure) map in which the preimage of a structure is called the *neutral network*. Neutral networks are closely related to the *neutral theory* of Motoo Kimura (Kimura, 1968), which stipulates that evolution is driven by mutations that do not change the phenotype. The properties of neutral networks as subsets of sequences in sequence space allow one to study how genotypes evolve. Neutral networks have been studied theoretically through random graph theory (Reidys, 1997), in the context of the molecular quasi-species (Reidys et al., 1997) and by exhaustive enumeration (Grüner et al., 1996; Göbel, 2000). A neutral network represents the set of all inverse folding solutions of a fixed structure. Graph properties, including size, density, and connectivity are of crucial functionality in molecular evolution. Clearly, a vast, extended neutral network is more accessible than a small, localized one and on a connected and dense neutral network, neutral evolution can easily be facilitated through point and pair mutations. On such a network, a population of RNA sequences can explore sequence space through gradual genotypic changes while maintaining its phenotype.

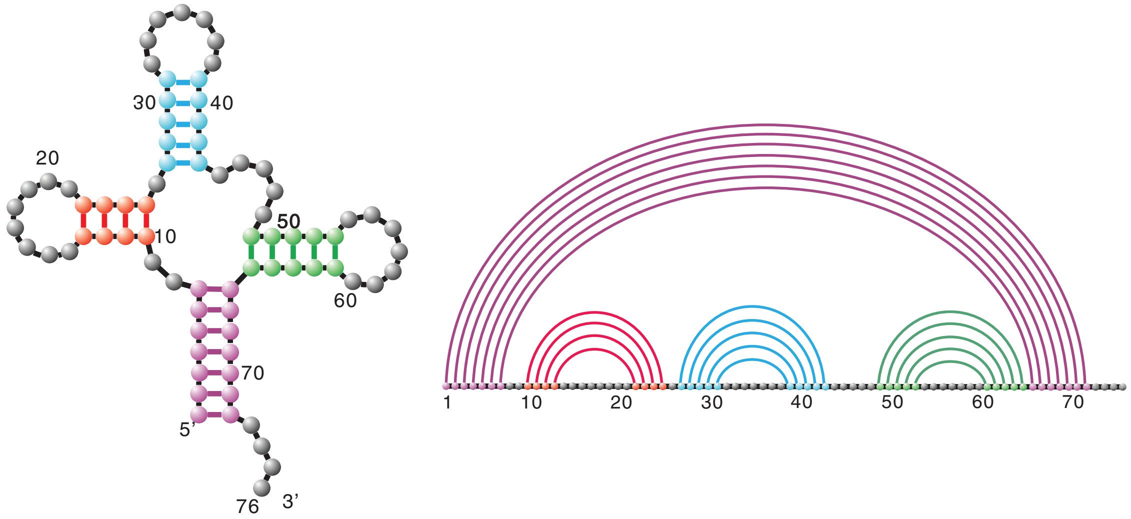


FIG. 1. The secondary structure and its diagram representation of a transfer RNA (tRNA).

[1] Considered as fatgraphs of genus zero they are the Poincaré dual of planar trees.

However, there is more to sequences and structures than MFE folding: certain RNA sequences exhibit multiple, distinctively different, stable configurations (Baumstark et al., 1997; Schultesand Bartel, 2000), as for example, riboswitches (Mandal and Breaker, 2004; Serganov and Patel, 2007). Recently evolutionary trajectories, so called *drift walks* that are obtained by either neutral evolution or switching between a multiplicity of MFE structures present at a fixed sequence have been considered (Rezazadegan et al., 2018). Such sequences indicate that it may not be sufficient to consider merely the MFE structure, but rather to broaden the scope to the entire RNA energy landscape. Energy landscapes of sequences, that is, the spectrum of free energies of the associated secondary structures of a fixed sequence have been studied in physics, chemistry, and biochemistry, and play a key role in understanding the dynamics of both RNA and protein folding (Martinez, 1984; Dill and Chan, 1997; Onuchic et al., 1997; Wolfinger et al., 2004).

McCaskill (1990) observed that the tropicalization of the DP routine that computes the MFE structure produced the partition function of structures for a given sequence. This allows one to study statistical features, for instance, base pairing probabilities of RNA energy landscapes by means of Boltzmann sampling (Tacker et al., 1996; Ding and Lawrence, 2003), enhancing structure prediction (Ding and Lawrence, 2003; Bernhart et al., 2006; Rogers and Heitsch, 2014). Aside from global features, local features are being studied as well. For instance, local minima of the energy landscape, that is, ''energy traps'' are crucial to the understanding of folding dynamics as they represent the metastable configurations (Tinoco and Bustamante, 1999; Chen and Dill, 2000). Statistical features of constrained energy landscapes, corresponding to conditional distributions can also be Boltzmann sampled (Hofacker et al., 1994; Freyhult et al., 2007; Lorenz et al., 2009).

Accordingly, the partition function is tantamount to computing the probability space of structures with which a fixed sequence is compatible. This gives rise to the pairing (Barrett et al., 2017):

$$\eta : \mathcal{N}^n \times \mathcal{S}_n \to \mathbb{R}, \tag{1}$$

that maps a fixed sequence–structure pair into its free energy. Here, $\mathcal{N}^n$ and $\mathcal{S}_n$ denote the space of sequences, σ, and the space of secondary structures, S, respectively. The pairing illuminates the symmetry between sequences and structures, suggesting we consider the ''dual'' of RNA energy landscape, that is, the spectrum of free energies of sequences for a fixed structure. This dual has been employed for designing more efficient inverse folding algorithms: Busch and Backofen (2006) discovers that using the MFE sequence of a fixed structure, as starting point for the inverse folding, significantly accelerates the algorithm. In other words, the global minimum of the RNA dual energy landscape is typically very close in sequence space to the corresponding neutral network. This line of work motivated the use of the dual RNA energy landscape[2] in inverse folding algorithms (Levin et al., 2012; Garcia-Martin et al., 2016). Recently, Barrett et al. (2017) proposed a framework considering RNA sequences and their RNA secondary structures simultaneously, as pairs. The RNA dual energy landscape in this context gives rise to an information theoretic framework for RNA sequences.

In practice, the exhaustive exploration of the dual RNA energy landscape is not feasible, whence specific localizations, for instance the point mutant neighborhood of a natural RNA sequence (Borenstein and Ruppin, 2006; Rodrigo and Fares, 2012), have been studied.

To conduct a systematic and biologically meaningful study of the dual RNA energy landscape, we present in this article an efficient Boltzmann sampling algorithm with a Hamming distance filtration. This filtration facilitates the analysis of Hamming classes of sequences in the dual RNA energy landscape that would otherwise be impossible to access (Fig. 4). Instead of being restricted to neighborhoods of point mutants (Borenstein and Ruppin, 2006; Rodrigo and Fares, 2012), we have now access to arbitrary Hamming classes. Such a dual sampler has, to our knowledge, first been derived in Levin et al. (2012). In fact, the sampler arises as the restriction of Waldispühl et al. (2008), where the structure partition function of sets of sequences with fixed Hamming distance is computed. As a result, its recursions over subintervals that form the conceptual backbone, lead to a time complexity of $O(h^2n^3)$, where h and n denote Hamming distance and sequence length, respectively. In contrast, the Boltzmann sampler presented in this study is based on the loop decomposition of the fixed structure and has a time complexity of $O(h^2n)$.

The article is organized as follows. In Section 2, we discuss prerequisites for the derivation of the algorithm. In Section 3, we present the sampling algorithm and analyze its time and space complexity. In

[2] By sampling,

Section 4, we apply the dual sampler to study the inverse folding rates (IFRs) as a function of Hamming distance. In Section 5, we employ the dual sampler for constructing paths in neutral networks.

2. PRELIMINARIES

Recall the graph presentation of RNA secondary structures. RNA secondary structures can be represented as diagrams, where vertices are drawn in a horizontal line and arcs in the upper half-plane. In a diagram, vertices are representing nucleotides and arcs are representing base pairs (Fig. 1). Vertices are labeled $[n] = \{1, 2, \ldots, n\}$ from left to right, indicating the orientation of the backbone from the $5'$-end to $3'$-end. A base pair, denoted (i, j), is an arc connecting vertices i and j. Two arcs (i, j) and (r, s) are called *crossing* if for $i < r$, the inequality $i < r < j < s$ also holds. An RNA secondary structure contains exclusively noncrossing arcs and thus induces the partial order: $(r, s) \prec (i, j)$ if and only if $i < r < s < j$.

The energy of a sequence–structure pair $\eta(\sigma, S)$ can be computed as the sum of the energy contributions of individual base pairs (Nussinov et al., 1978). A more elaborate model (Mathews et al., 1999; Turner and Mathews, 2010) evaluates the total free energy to be the sum of the energies of loops. A loop L in a secondary structure is a sequence of intervals $([a_i, b_i])_i$, $1 \leq i \leq k$, where (a_1, b_k), (b_i, a_{i+1}) are base pairs, for all $1 \leq i \leq k-1$. Since crossing arcs are not allowed, nucleotides in the interval $([a_i+1, b_i-1])_i$ are unpaired. In particular, for $k=1$, L is called a hairpin loop, for $k=2$, either an interior loop, bulge loop, or helix, depending on how many unpaired vertices are contained in the respective intervals, and for $k > 2$, a multiloop (Fig. 2). Note that the arc (a_1, b_k) is the maximal arc of the loop, that is, $(b_i, a_{i+1}) \prec (a_1, b_k)$ for all $1 \leq i \leq k-1$, whence L can be represented by (a_1, b_k). The intersection of two distinct loops is either empty or consists of exactly one base pair. Each base pair is contained in exactly two loops and is maximal in exactly one of these two. There is a particular loop, the *exterior* loop, consisting of all maximal arcs in a secondary structure. As a matter of convention, we shall assume that any diagram is "closed" by the arc, $(0, n+1)$, referred to as its rainbow and by convention, there are the two "formal" nucleotides N_0, N_{n+1} associated with positions 0 and $n+1$, respectively.

In Turner's energy model, the energy of a loop, $\eta(\sigma, L)$, is determined by its loop type (hairpin, helix, bulge loop, interior loop, exterior loop, or multiloop), the specific nucleotide composition of its base pairs, as well as a certain number of unpaired bases contained in it. Those unpaired bases are typically adjacent to a base pair. Accordingly, the energy of a sequence–structure pair equals the sum of the energies of all the associated loops, that is,

$$\eta(\sigma, S) = \sum_{L \in S} \eta(\sigma, L). \tag{2}$$

A secondary structure can be decomposed by successively removing arcs from the outside to the inside (top to bottom) (Fig. 2). Since any base pair is maximal in exactly one loop, removing a base pair is tantamount to removing its associated loop.

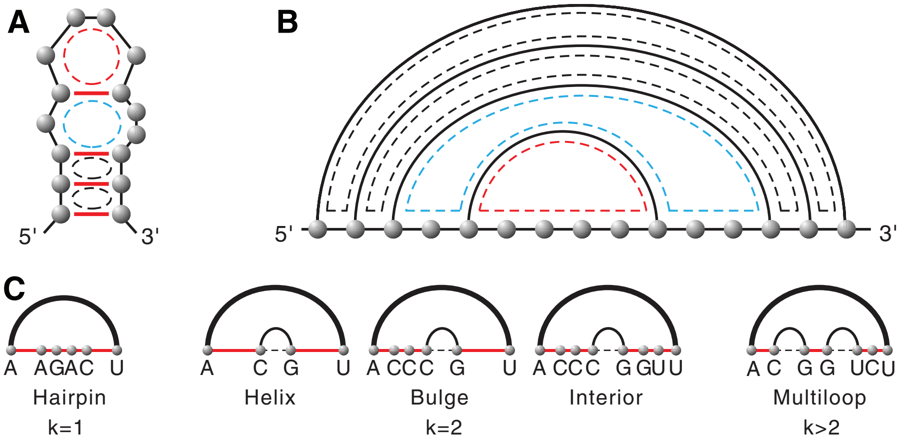


FIG. 2. Loops in an RNA secondary structure and their classification. Loops in the planar graph presentation **(A)**. Loops in the arc diagram **(B)**. The classification of loop types **(C)**. Loops are represented as dashed circles. Unpaired intervals are represented as red lines on the backbone.

By viewing a secondary structure S as a diagram we observe that any interval $[i, j]$ induces a substructure containing all arcs that have both endpoints contained in $[i, j]$ and we denote such substructures by $X^S_{i,j}$. If the interval $[i, j]$ contains no arcs, we simply refer to the substructure $X^S_{i,j}$ again as an interval. Given S, the concatenation of the two substructures $X^S_{i,j} \cup X^S_{j+1,k}$ is the substructure $X^S_{i,k}$. In the following, we shall simply write $X_{i,j}$ instead of $X^S_{i,j}$.

In particular, let (i, j) be a base pair, L be the loop that is represented by (i, j), and $S_{i,j}$ be the substructure for which (i, j) is the unique maximal arc. Suppose (p_r, q_r), $1 \leq r \leq k$ are base pairs in L, different from (i, j). Removing the arc (i, j) produces a sequence of substructures $S_{p_1,q_1}, \ldots S_{p_k,q_k}$ as well as a sequence of intervals $[i+1, p_1-1], [q_1+1, p_2-1], \ldots, [q_k+1, j-1]$ (Fig. 3A).

Let $q_0 = i$, concatenating the interval $[q_{r-1}+1, p_r-1]$ with S_{p_r,q_r} produces a substructure, denoted by $M^r_{q_{r-1}+1,q_r}$, $1 \leq r \leq k$. Let $R^1_{q_0+1,q_k}$ be the substructure obtained by concatenating all $M^r_{q_{r-1}+1,q_r}$ for $1 \leq r \leq k$, that is, $\bigcup_r M^r_{q_{r-1}+1,q_r}$ (Fig. 3A). By construction, removing (i, j) from $S_{i,j}$ generates $R^1_{q_0+1,q_k} \cup [q_k+1, j-1]$ (Fig. 3B).

Note that $R^1_{q_0+1,q_k}$ can be obtained by recursively concatenating $M^r_{q_{r-1}+1,q_r}$, $1 \leq r \leq k$. We use the superscript w to represent the intermediates (recursively concatenating from right to left):

$$R^w_{q_{w-1}+1,q_k} = \bigcup_{w \leq r \leq k} M^r_{q_{r-1}+1,q_r}.$$

Clearly, we have the following bipartition (Fig. 3B):

$$R^w_{q_{w-1}+1,q_k} = M^w_{q_{w-1}+1,q_w} \bigcup R^{w+1}_{q_w+1,q_k}.$$

Rules shown in Figure 3B facilitate the decomposition of a secondary structure, S, recursively into arcs and unpaired intervals. These decompositions lead to efficient DP routines.

3. DUAL SAMPLING WITH HAMMING DISTANCE FILTRATION

In Busch and Backofen (2006), an MFE sequence for a given structure is derived by means of DP. The algorithm facilitates the arc decomposition of a secondary structure (Waterman, 1978) by computing an MFE sequence recursively. In analogy to the partition function of structures for a given sequence, the dual partition function, that is, the partition function of sequences for a given structure, has been computed in Garcia-Martin et al. (2016) and Barrett et al. (2017), where, in addition, Boltzmann samplers were derived (Garcia-Martin et al., 2016; Barrett et al., 2017).

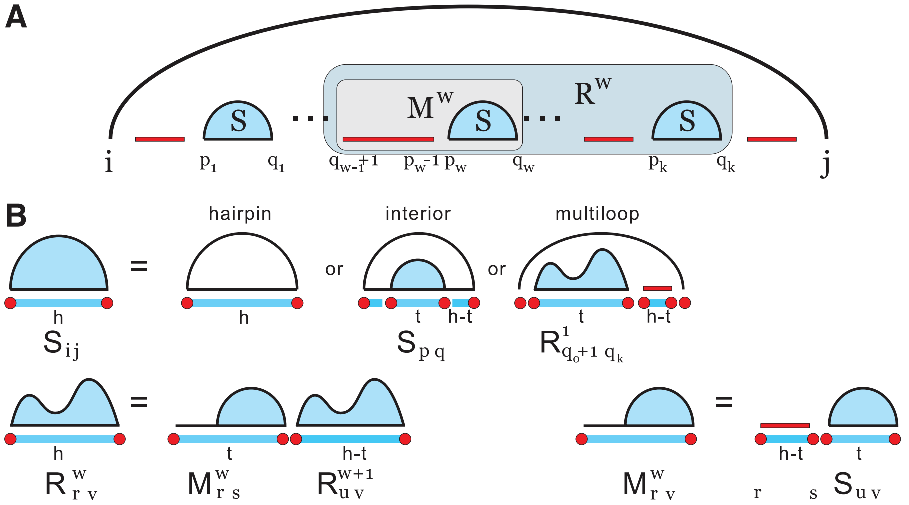


FIG. 3. **(A)** The substructures $S_{a,b}$, $M^w_{a,b}$, and $R^w_{a,b}$ and their relationship. **(B)** Decomposition rules of an RNA secondary structure.

In this section, we introduce an algorithm refining the Boltzmann sampler in Barrett et al. (2017) that constructs RNA sequences from the Boltzmann ensemble of a structure S, subject to a Hamming distance constraint.[3] The straightforward approach would be to run a rejection sampler based on the sampler introduced in Garcia-Martin et al. (2016) and Barrett et al. (2017). However, as we shall prove later in this section, this would result in a rather inefficient algorithm. Instead, we follow a different approach, introducing a new parameter h associated to a subsequence, representing the Hamming distance.

Now we begin the derivation of the dual partition function with Hamming distance filtration.

Definition 1. *Given a structure S and a reference sequence $\bar{\sigma}$, the dual partition function for S with Hamming distance filtration is given by*

$$Q_h^{S,\bar{\sigma}} = \sum_{\sigma,\, d(\sigma,\bar{\sigma})=h} e^{\frac{-\eta(\sigma,S)}{RT}},$$

where $\eta(\sigma, S)$ is the energy of S on σ, $d(\sigma, \bar{\sigma})$ denotes the Hamming distance between σ and $\bar{\sigma}$, R is the universal gas constant, and T is the temperature.

Remark 1. *Note that $Q_h^{S,\bar{\sigma}}$ allows us to further stratify the dual partition function Q^S in Garcia-Martin et al. (2016) and Barrett et al. (2017) as follows:*

$$Q^S = \sum_{0\le h\le n} Q_h^{S,\bar{\sigma}}.$$

In the following, we omit the explicit reference to $\bar{\sigma}$ and simply write Q_h^S. We shall compute Q_h^S following the secondary structure decomposition recursively. Suppose (i, j) is a base pair with its induced substructure $S_{i,j}$. Since the specific nucleotide composition of (i, j) may be involved in energy calculation of more than one loop, we introduce the partition functions of substructures $X_{a,b}$, $Q_h^{X_{a,b}}(N_a, N_b)$, where $X_{a,b} = S_{a,b}$, $R^w_{a,b}$, or $M^w_{a,b}$, whose left and right endpoints, $\sigma_a = N_a$ and $\sigma_b = N_b$, are determined and contributes h to Hamming distance. We consider the set of subsequences

$$\{\sigma_{a,b} \in \mathcal{N}^{b-a+1} | d(\sigma_{a,b}, \bar{\sigma}_{a,b}) = h,\ \sigma_a = N_a,\ \sigma_b = N_b\},$$

which we refer to as $\mathcal{S}_h^{a,b}(N_a, N_b)$. Summing over all $\sigma_{a,b} \in \mathcal{S}_h^{a,b}(N_a, N_b)$ we derive

$$Q_h^{X_{a,b}}(N_a, N_b) = \sum_{\sigma_{a,b}\in\mathcal{S}_h^{a,b}(N_a,N_b)} e^{\frac{-\eta(\sigma_{a,b},X_{a,b})}{RT}}, \tag{3}$$

where $N_a, N_b \in \mathcal{N}$, $\mathcal{N} = \{\mathbf{A}, \mathbf{U}, \mathbf{C}, \mathbf{G}\}$.

We next derive the recursion for $Q_h^{S_{i,j}}(N_i, N_j)$, computed from bottom to top (Fig. 3B).

Case 1: (i, j) is $\prec$-minimal, that is, $S_{i,j}$ is a hairpin loop ($k = 0$). By Equation (3), summing over all subsequence $\sigma_{i,j} \in \mathcal{S}_h^{i,j}(N_i, N_j)$ we derive

$$Q_h^{S_{i,j}}(N_i, N_j) = \sum_{\sigma_{i,j}\in\mathcal{S}_h^{i,j}(N_i,N_j)} e^{\frac{-\eta(\sigma_{i,j},S_{i,j})}{RT}}.$$

Case 2: (i, j) is nonminimal and $k = 1$, that is, L is an interior loop (helix, bulge loop). Removing (i, j) produces a single $S_{p,q}$ as well as two intervals $[i+1, p-1]$ and $[q+1, j-1]$, either of which being possibly empty. Suppose $d(\sigma_{i,j}, \bar{\sigma}_{i,j}) = h$ and $d(\sigma_{p,q}, \bar{\sigma}_{p,q}) = t$, where $0 \le t \le h$. Then the distance contribution from the intervals $[i, p-1]$ and $[q+1, j]$, t_1 and t_2, satisfies $t_1 + t_2 = h - t$. Then $Q_h^{S_{i,j}}(N_i, N_j)$ equals

$$\sum_{t,t_1,t_2}\sum_{N_p,N_q}\sum_{\sigma_{i,p}}\sum_{\sigma_{q,j}} e^{\frac{-\eta(\sigma_{i,j},L)}{RT}} Q_t^{S_{p,q}}(N_p, N_q),$$

where $t + t_1 + t_2 = h$, $N_p, N_q \in \mathcal{N}$, $\sigma_{i,p} \in \mathcal{S}^{i,p}_{t_1+\delta_p}(N_i, N_p)$, and $\sigma_{q,j} \in \mathcal{S}^{q,j}_{t_2+\delta_q}(N_q, N_j)$.

Here $\delta_x = 1$ if $N_x = \bar{\sigma}_x$, and $\delta_x = 0$, otherwise, for $x = p, q$.

[3] To a given reference sequence $\bar{\sigma}$, say.

Case 3: (i, j) is nonminimal and $k \geq 2$, that is, L is a multiloop. In this case (in difference to the interior loops analyzed above) the Turner energy model allows us to further decompose the energy of $\eta(\sigma, L)$ into independent components, which in turn allows us to compute $Q_h^{S_{i,j}}(N_i, N_j)$ through recursive bipartitioning. Removing (i, j) produces $R^1_{q_0+1,\, q_k}$ as well as $[q_k+1, j-1]$ (Fig. 3B). The energy $\eta(\sigma_{i,j}, S_{i,j})$ is then given by

$$\eta(\sigma, R^1_{q_0+1,\, q_k}) + \alpha_{\mathrm{mul}} + \eta_{\mathrm{mul}}((i, j)) + \eta_{\mathrm{mul}}([q_k+1, j-1]),$$

where α_{mul} is the energy contribution of forming a multiloop, $\eta_{\mathrm{mul}}((i, j))$ is the energy contribution of base pair (i, j) in a multiloop, and $\eta_{\mathrm{mul}}([q_k+1, j-1])$ is the energy contribution from the unpaired base interval in a multiloop. The sum of the latter three components is denoted by η^0.

Suppose $d(\sigma_{q_0+1,\, q_k}, \bar{\sigma}_{q_0+1,\, q_k}) = t$ and $d(\sigma_{i,j}, \bar{\sigma}_{i,j}) = h$. Then the distance contribution from the unpaired interval $[q_k+1, j-1]$ is $h - t - \delta_i - \delta_j$. Then $Q_h^{S_{i,j}}(N_i, N_j)$ equals

$$\sum_t \sum_{N_{q_0+1},\, N_{q_k}} \sum_{\sigma_{q_k, j}} e^{\frac{-\eta^0}{RT}} Q_t^{R^1_{q_0+1,\, q_k}}(N_{q_0+1}, N_{q_k}),$$

where $0 \leq t \leq h$, $N_{q_0+1}, N_{q_k} \in \mathcal{N}$ and $\sigma_{q_k, j} \in \mathcal{S}^{q_k, j}_{h-t-\delta_i+\delta_{q_k}}(N_{q_k}, N_j)$.

This brings us to substructures $R^w_{q_{w-1}+1,\, q_k}$, $1 \leq w \leq k$, which decompose into (are concatenations of) $M^w_{q_{w-1}+1,\, q_w}$ and $R^{w+1}_{q_w+1,\, q_k}$. For notational convenience we set $r = q_{w-1}+1$, $s = q_w$, $u = q_w + 1$, and $v = q_k$. Suppose $d(\sigma_{r,s}, \bar{\sigma}_{r,s}) = t$ and $d(\sigma_{r,v}, \bar{\sigma}_{r,v}) = h$, then $d(\sigma_{u,v}, \bar{\sigma}_{u,v}) = h - t$. We obtain for $Q_h^{R^w_{r,v}}(N_r, N_v)$ the expression

$$\sum_t \sum_{N_s,\, N_u} Q_t^{M^w_{r,s}}(N_r, N_s) Q_{h-t}^{R^{w+1}_{u,v}}(N_u, N_v),$$

where $0 \leq t \leq h$ and $N_s, N_u \in \mathcal{N}$.

The substructures $M^w_{q_{w-1}+1,\, q_w}$ are concatenations of $[q_{w-1}+1, p_w - 1]$ and S_{p_w, q_w}, for $1 \leq w \leq k$. For notational convenience we set $r = q_{w-1}+1$, $s = p_w - 1$, $u = p_w$, and $v = q_w$.

Suppose $d(\sigma_{u,v}, \bar{\sigma}_{u,v}) = t$ and $d(\sigma_{r,v}, \bar{\sigma}_{r,v}) = h$, then the Hamming distance of $[r, s]$ to the corresponding $\overline{\sigma}$-interval is $h - t$.

Summing over $0 \leq t \leq h$, all $N_{p_w-1}, N_{p_w} \in \mathcal{N}$, all $\sigma_{q_{w-1}+1,\, p_w-1} \in \mathcal{S}^{q_{w-1}+1,\, p_w-1}_{h-t}(N_{q_{w-1}+1}, N_{p_w-1})$, we derive for $Q_h^{M^w_{r,v}}(N_r, N_v)$

$$\sum_t \sum_{N_s,\, N_u} \sum_{\sigma_{r,s}} Q_{h-t}^{S_{u,v}}(N_u, N_v) e^{\frac{-\eta^w}{RT}},$$

where $0 \leq t \leq h$, $N_s, N_u \in \mathcal{N}$, $\sigma_{r,s} \in \mathcal{S}^{r,s}_{h-t}(N_r, N_s)$ and $\eta^w = \eta_{\mathrm{mul}}((u, v)) + \eta_{\mathrm{mul}}([r, s])$. Here, $\eta_{\mathrm{mul}}((u, v))$ is the energy contribution of base pair (u, v) in a multiloop and $\eta_{\mathrm{mul}}([r, s])$ is the contribution of segment of unpaired bases in a multiloop. We present the recursions in Figure 3B.

As for analyzing the time and space complexity, the introduction of the intermediate substructures $M^w_{q_{w-1}+1,\, q_w}$ and $R^w_{q_{w-1}+1,\, q_k}$ avoids processing concatenation of substructures simultaneously, which would result in a $O(h^{k-1})$ time complexity. The family of intermediate substructures $M^w_{q_{w-1}+1,\, q_w}$ and $R^w_{q_{w-1}+1,\, q_k}$ remedies this problem by executing one concatenation at each step, effectively bipartitioning and requiring a time complexity of $O(h)$. In total, we encounter $k-1$ such bipartition, resulting in a $(k-1)O(h)$ time complexity. Since there are $O(n)$ base pairs in a structure and each entails to compute $O(h)$ partition functions, we have to consider $O(hn)$ partition functions. As a result, the time complexity of the algorithms is $O(h^2 n)$.

Following this recursion, $Q_h^{S_{i,j}}(N_i, N_j)$ can be computed from bottom to top as claimed. The recursion terminates when reaching the rainbow $(0, n+1)$. The partition function of S with Hamming distance filtration h to $\overline{\sigma}$ is given by $Q_h^S = Q_h^{S_{0,n+1}}(N_0, N_{n+1})$, where N_0 and N_{n+1} are "formal" nucleotides, discussed above.

Having computed the partition function $Q_h^{S, \overline{\sigma}}$, we implement the Boltzmann sampler of RNA sequences having a fixed Hamming distance h to $\overline{\sigma}$ from S following the classical stochastic backtracking method introduced by Ding and Lawrence (2003). This process first samples the nucleotides in the exterior loop and then subsequently samples the loops following the partial order $\prec$ from top to bottom until reaching the hairpin loops.

Since the time complexity of computing a loop energy in Turner's model is constant, the worst case time complexity of the sampling process is $O(n^2)$ (Ding and Lawrence, 2003), and applying the Boustrophedon technique for Boltzmann sampling, introduced in Ponty (2008) and Nebel et al. (2011), reduces the time complexity to $O(n \log n)$ on average. The source code of the sampler is freely available on our group website (HamSampler).

Remark 2. *To illustrate the utility of the Hamming distance filtration, we display some data obtained using the unrestricted dual sampler given in Barrett et al. (2017). For this purpose, we consider* 12 *sequence–structure pairs from the human microRNA let-7 family in miRBase (Kozomara and Griffiths-Jones, 2013). For each pair we sample* 5×10^4 *sequences using the unrestricted sequences sampler. Then we compute the Hamming distance distribution with respect to the natural sequence of the sampled sequences. The resulting distances are then normalized by sequence length n. We display in Figure 4 the distance distributions of three distinct sequence–structure pairs.*

These data show that almost all sampled sequences exhibit distances d to the reference sequence, where $0.6n \leq d \leq 0.9n$. *We observe mean distance,* μ, *where* $0.7n \leq \mu \leq 0.8n$. *In particular, none of the sampled sequences is within Hamming distance* 5 *to the reference sequence.*

4. COMPUTING IFR OF MUTANTS

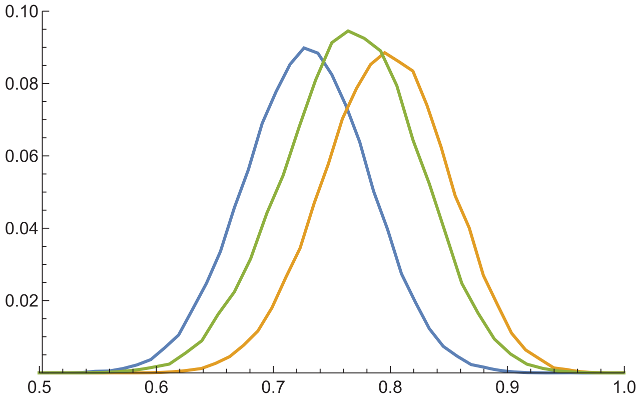
Our sampler facilitates the exploration of mutants in each respective Hamming distance. One key observable associated with such mutants is the IFR. The IFR is computed as follows. We associate an indicator variable to each sampled sequence that equals 1 if the sequence folds into the reference structure and 0 otherwise. By construction, the IFR is the mean of this random variable and we consider the IFR of a sequence–structure pair as a function of the Hamming distance, h, to the reference sequence, IFR(h).

Given a sequence–structure pair $(\overline{\sigma}, S)$, we sample 5×10^4 sequences from S having a fixed Hamming distance, h, where h ranges from 1 to 20. Then $\text{IFR}(h) = U/M$, where U is the number of sampled sequences folding back to S and M is the sample size.

We consider the microRNA let-7 family of three species: human (hum01-12), lizard (liz01-11), and drosophila (dro01-08), computing their IFRs, respectively. We display the mean[4] IFR of the three species in Figure 5.

Figure 5 shows that human has the highest mean IFR while drosophila has the lowest (except at $h = 1$). In addition, the mean IFR decreases for human is significantly slower than for drosophila. This can be interpreted to be a result of an increased robustness in higher-level organisms. To analyze robustness and dependencies of these findings, we compute the IFR of random sequence–structure pairs and to those of natural pairs. In the following, we restrict ourselves to the hum04 sequence–structure pair. We first consider random sequences compatible with the hum04-structure and thereby create new sequence–structure pairs. Then we compute IFR(5) of these pairs by sampling 5×10^4 sequences of Hamming distance 5. The IFR(5) is almost zero for these random sequences indicating that random compatible sequences have little or no connection with the hum04-structure. To identify sequences that are more closely related to the hum04-structure, we use our sampler, creating 100 sequence–structure pairs by sampling sequences from the

FIG. 4. Hamming distance distribution of sampled sequences for three sequence–structure pairs of the human microRNA let-7 family (hum01, hum09, and hum10). For each pair, we sample 5×10^4 sequences, using the unrestricted sampler in Barrett et al. (2017). We display the Hamming distance distribution of the sampled sequences to the natural sequence. The x-axis is the Hamming distance normalized by the sequence length, and the y-axis is frequency sampled sequences.



[4] Taken over the entire collection of let-7 micrRNAs of a given species in the database.

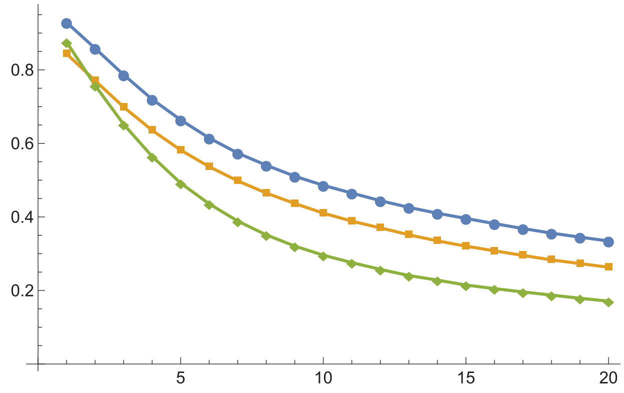


FIG. 5. Mean IFR of sequence–structure pairs of microRNA let-7 family of human (blue), lizard (yellow), and drosophila (green). The *x*-axis represents Hamming distance and the *y*-axis represents the IFR. IFR, inverse folding rate.

natural pairs of distance 5, 7, 10, and 20, respectively. Then, we combine the newly sampled sequences with the hum04-structure, creating new sequence–structure pairs. We compute their IFR(5) and sort the pairs by their IFR(5) in increasing order (Fig. 6). For reference purposes we display IFR(5) of the natural sequence–structure pair as a dashed line.

Figure 6 shows that IFR(5) of the natural sequence–structure pair is above the 95 percentile, that is, better than almost all of the newly created pairs. Furthermore, there exist very few pairs such that IFR(5) $\in [0.1, 0.3]$ holds. The proportion of sequence having high IFR(5), that is, IFR(5) > 0.3 drops when the sampled sequences have higher Hamming distance. This finding suggests that the natural pair is locally optimal.

5. CONSTRUCTING NEUTRAL PATHS

As discussed in Section 1, connectivity is of central importance in neutral networks. Combined with some form of density, it allows genotypes to explore, by means of point or pair mutations, extended portions of sequence space, while the phenotype (folded structure) is conserved. An exhaustive analysis of connectivity is not feasible even for relatively short sequence length, whence the explicit construction of specific paths within the neutral network is the best possible outcome. Neutral paths connecting two

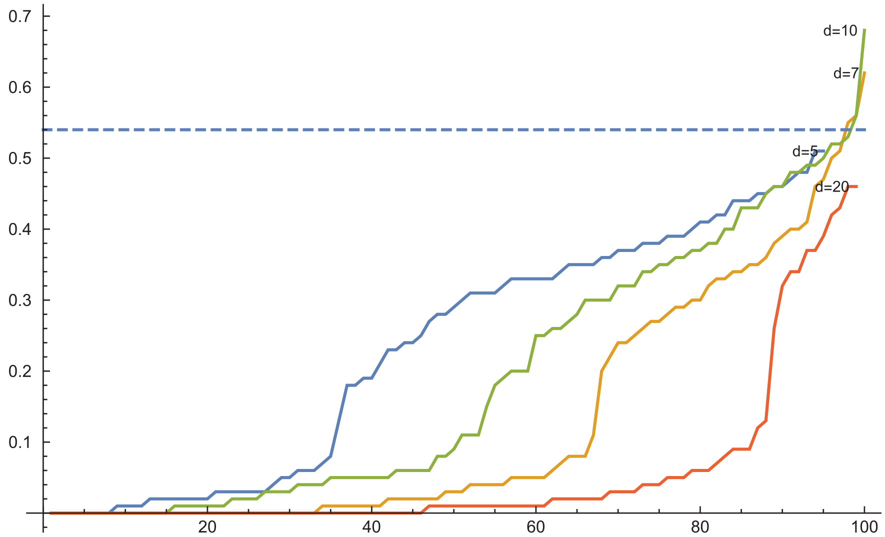


FIG. 6. IFR(5) of the natural sequence–structure pair versus adjacent sampled sequences. We compare the natural sequence–structure pair of hum04 and sample 100 sequences of distance 5, 7, 10, and 20, respectively. We display IFR(5) of the induced sequence–structure pairs sorted by their IFR(5) in increasing order for distance 5 (blue), 7 (yellow), 10 (green), and 20 (red). IFR(5) of the natural sequence–structure pair is displayed as the dashed line. Here, the *x*-axis is labeled by the sorted sequence–structure pairs and the *y*-axis represents the IFR.

sequences are playing an important role in evolution, as suggested by Kimura (1968). To be clear, let us first specify the neutral path problem:

Given two sequences σ_1 and σ_2, both folding into the structure S, identify a path $\sigma_1 = \tau_0, \tau_1, \ldots, \tau_k = \sigma_2$, such that
(*) for all τ_i, $0 \leq i \leq k$, folds to S,
(**) τ_{i+1} is obtained from τ_i by either a compatible point or a base pair mutation.

The construction of such ''neutral paths'' has been studied in Göbel and Forst (2002) using a proof idea that facilitates the construction of neutral paths, for fixed, finite distance d, in random induced subgraphs. However, Göbel and Forst (2002) exhaustively checks whether such paths are neutral or not, irrespective of d, a task that becomes impracticable for large d. At present, there is no efficient way of finding neutral paths in neutral networks induced by folding algorithms, in particular, in case of the distance between the two sequences being large. Our sampling algorithm enables us to efficiently search for inverse folding solutions at a given Hamming distance. In the following, we shall employ our sampling algorithm to derive an efficient heuristic to solve the neutral path problem.

Certainly, given σ_1 and σ_2, both folding into S, one can always construct a path between them using the two above moves. By construction, this is a S-compatible path. Furthermore, there exists a minimum number of moves that have to be performed to traverse from σ_1 to σ_2. We refer to this as the S-compatible distance between σ_1 and σ_2, $d_S(\sigma_1, \sigma_2)$. Clearly, we have $\frac{1}{2}d \leq d_S \leq d$, for any S-compatible sequences. A neutral path, whose length equals the S-compatible distance, is called a shortest neutral path. In the context of the neutral path problem, we do not require the paths to be minimal in length.

Case 1: $d(\sigma_1, \sigma_2) \leq 5$. Here we exhaustively search all shortest S-compatible paths between σ_1 and σ_2 and check for neutrality. Note that we always have $d_S \leq d$, thus in the worst case, we need to check $5! = 120$ different paths and fold $2^5 = 32$ different sequences. This is feasible for sequence lengths shorter that 10^3 nucleotides, using standard secondary structure folding algorithms (Zuker and Stiegler, 1981; Hofacker et al., 1994).

Case 2: $d(\sigma_1, \sigma_2) > 5$. Suppose σ_1 and σ_2 have Hamming distance h. We sample m sequences from σ_1 with respect to S with distance filtration $h/2$. $m = 1000$ typically suffices, but higher sampling size can easily be realized if the IFR is too low. We then select such a sequence with minimum Hamming distance to σ_2, denoted by τ_s. We have $d(\sigma_1, \tau_s) = h/2 = h_1$ and $d(\tau_s, \sigma_2) = h_2$, where $h_1 + h_2 \geq h$. If $h_2 > h$, we claim the process fails and we conclude we cannot find a neutral path between σ_1 and σ_2. Otherwise, we repeat the process between σ_1 and τ_s, and between τ_s and σ_2, differentiating Case 1 and Case 2. We show the flow of the algorithm in Figure 7.

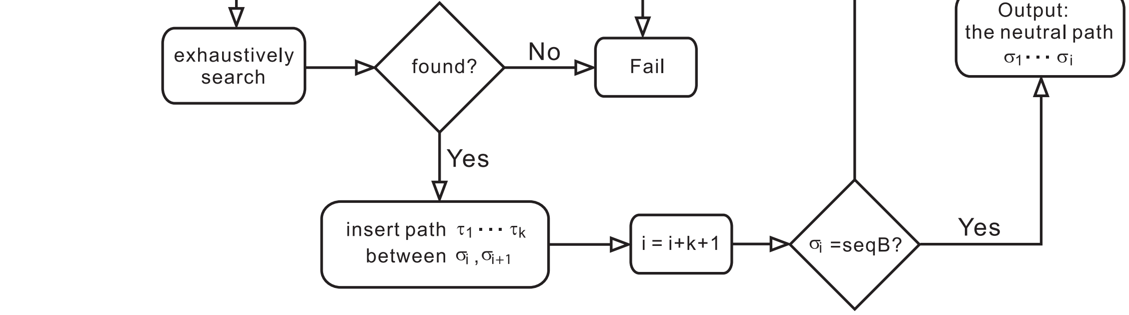


FIG. 7. Flowchart of the neutral path construction algorithm.

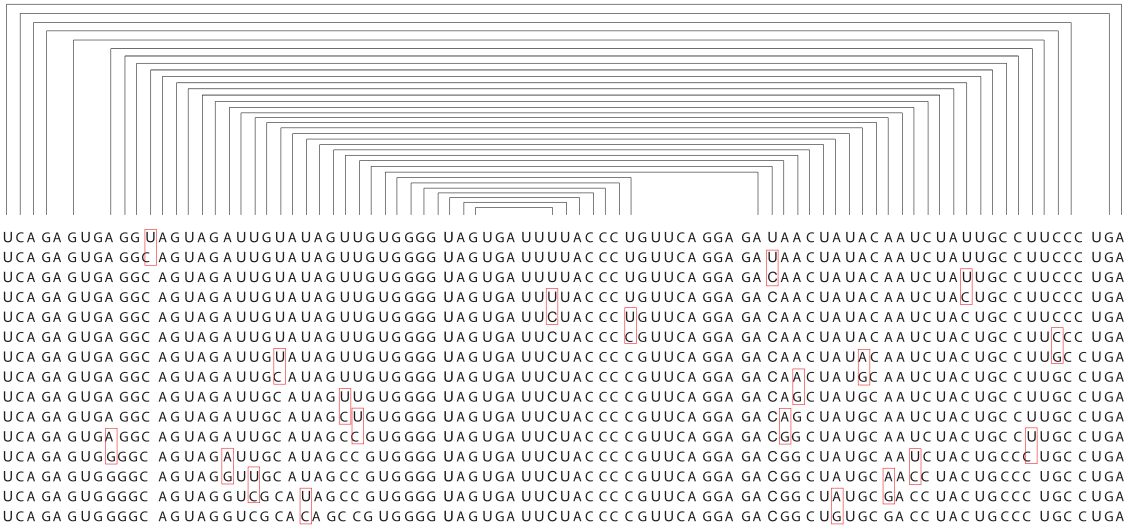

UCAGAGUGAGGUAGUAGAUUGUAUAGUUGUGGGUAGUGAUUUUACCCUGUUCAGGAGAUAACUAUACAAUCUAUUGCCUUCCCUGA
UCAGAGUGAGGCAGUAGAUUGUAUAGUUGUGGGUAGUGAUUUUACCCUGUUCAGGAGAUAACUAUACAAUCUAUUGCCUUCCCUGA
UCAGAGUGAGGCAGUAGAUUGUAUAGUUGUGGGUAGUGAUUUUACCCUGUUCAGGAGACAACUAUACAAUCUAUUGCCUUCCCUGA
UCAGAGUGAGGCAGUAGAUUGUAUAGUUGUGGGUAGUGAUUUUACCCUGUUCAGGAGACAACUAUACAAUCUACUGCCUUCCCUGA
UCAGAGUGAGGCAGUAGAUUGUAUAGUUGUGGGUAGUGAUUCUACCCUGUUCAGGAGACAACUAUACAAUCUACUGCCUUCCCUGA
UCAGAGUGAGGCAGUAGAUUGUAUAGUUGUGGGUAGUGAUUCUACCCCGUUCAGGAGACAACUAUACAAUCUACUGCCUUCCCUGA
UCAGAGUGAGGCAGUAGAUUGUAUAGUUGUGGGUAGUGAUUCUACCCCGUUCAGGAGACAACUAUACAAUCUACUGCCUUGCCUGA
UCAGAGUGAGGCAGUAGAUUGCAUAGUUGUGGGUAGUGAUUCUACCCCGUUCAGGAGACAACUAUGCAAUCUACUGCCUUGCCUGA
UCAGAGUGAGGCAGUAGAUUGCAUAGUUGUGGGUAGUGAUUCUACCCCGUUCAGGAGACAGCUAUGCAAUCUACUGCCUUGCCUGA
UCAGAGUGAGGCAGUAGAUUGCAUAGCUGUGGGUAGUGAUUCUACCCCGUUCAGGAGACAGCUAUGCAAUCUACUGCCUUGCCUGA
UCAGAGUGAGGCAGUAGAUUGCAUAGCCGUGGGUAGUGAUUCUACCCCGUUCAGGAGACGGCUAUGCAAUCUACUGCCUUGCCUGA
UCAGAGUGGGGCAGUAGAUUGCAUAGCCGUGGGUAGUGAUUCUACCCCGUUCAGGAGACGGCUAUGCAAUCUACUGCCCUGCCUGA
UCAGAGUGGGGCAGUAGGUUGCAUAGCCGUGGGUAGUGAUUCUACCCCGUUCAGGAGACGGCUAUGCAACCUACUGCCCUGCCUGA
UCAGAGUGGGGCAGUAGGUCGCAUAGCCGUGGGUAGUGAUUCUACCCCGUUCAGGAGACGGCUAUGCGACCUACUGCCCUGCCUGA
UCAGAGUGGGGCAGUAGGUCGCACAGCCGUGGGUAGUGAUUCUACCCCGUUCAGGAGACGGCUGUGCGACCUACUGCCCUGCCUGA

FIG. 8. A neutral path connecting the natural sequence of hum08 to a sequence having Hamming distance 20. All sequences along the path fold into the natural structure of hum08. This particular path has length 14 and consists of 8 point and 6 base pair mutants.

The process either fails at some point of the iteration or produces recursively a neutral path. We illustrate a particular neutral path, connecting the natural sequence of hum08 to a Hamming distance 20 sequence in Figure 8.

As for algorithmic performance, for hum04 we consider the natural sequence and structure pair and sample 100 sequences of Hamming distance 20, 19 of which being neutral. We pair each of these with the natural sequence and compute a neutral path. The algorithm succeeded 18 times and failed to produce a neutral path once. For hum08, we perform the same experiment for Hamming distances 20 and 40, respectively. In case of Hamming distance 20, we find neutral 85 sequences, for these the algorithm succeeds 83 times and fails twice. For distance 40, we find 53 neutral sequences: the algorithm succeeds 49 times and fails in four instances.

For a low-level organism microRNA, bra01, which is a Branchiostoma microRNA, at distance 20, we find 22 neutral sequences: 16 successes and 6 fails.

6. DISCUSSION

The problem of finding a sequence that folds into a given structure, *S*, has first been studied in Hofacker et al. (1994). The algorithm consists of two parts: first it constructs a random *S*-compatible sequence and second, it performs an adaptive walks of point mutant in the sequence such that it facilitates identifying a sequence that folds into *S*. In this process, neither an inverse fold solution is guaranteed nor the number of adaptive walks required is understood. Busch and Backofen (2006) show that such adaptive walks can be constructed much more easily, when proper care is taken where the process actually initiates, namely, choosing the *S*-compatible sequence such that it minimizes the free energy with respect to *S*. Levin et al. (2012) and Garcia-Martin et al. (2016) observe that Boltzmann sampled sequences exhibit a distinctively higher rate of folding again into *S*.

The high IFR of sampled sequences from a structural ensemble is not only useful in finding candidate sequences, for inverse folding problems reflect in some sense the robustness of the structure. High IFR in structural ensembles indicates that the structure is likely preserved within limited energy change and mutations on a sequence. This is quite subtle as competing structural configurations may offer a fixed sequence an even lower and thus more preferable free energy. The problem can therefore not be reduced to minimizing free energy of sequences for a fixed structure, it is context dependent.

However, sampled sequences from the structural ensemble are not conserved and differ vastly from each other. It is natural to bring evolutionary trajectories into the picture, necessitating the ability to study

Boltzmann sampled sequences having fixed Hamming distance to some reference sequence. This allows us to investigate local features and brings sequence information into the picture. By introducing the Hamming distance filtration, we can zoom into a specific sequence as well as its neighborhood in the structural ensemble. These sequences are not only sorted by the given structure but also evolutionary close to the reference sequence. This approach shifts focus to considering sequences and structures as pairs, as discussed in Barrett et al. (2017).

Levin et al. (2012) presents a Boltzmann sampler of sequences from a structural ensemble with Hamming distance restriction. The algorithm described in Levin et al. (2012) constitutes a constrained version of the algorithm described in Waldispühl et al. (2008), having a time complexity of $O(h^2n^3)$, where h is the Hamming distance. The partition function of sequences with distance filtration on all secondary structures is computed, requiring to consider all subintervals of $[1, n]$, as well as an additional for-loop index, induced by the concatenation of two substructures.

Our algorithm has a time complexity of $O(h^2n)$, a result of different recursions. We utilize the hierarchical organization, or equivalently the induced partial order of the arcs of a secondary structure, together with the fact that free energy is computed based on loops. This allows us to compute the partition function from the inside to the outside (bottom to top from the tree prospective). The routine is purely driven by the fixed structure, whence no redundant information is computed.

The dual sampler, that is, the Boltzmann sampler of sequences for a fixed structure, with Hamming distance filtration (enhanced sampler), brings sequence information into the picture. This enables us to study evolutionary questions with the enhanced dual sampler. IFRs and their Hamming distance dependence, but also questions as the structural diversity of the derived sequences, can be analyzed effectively with the enhanced sampler. These studies follow the generalized scheme of inferring information on any random variable over sequences partitioned into Hamming classes. Hamming classes in this sense can be viewed as blocks of a partition to which a random variable can be restricted to. Our analysis of IFR gives first indications that microRNAs of higher-level organisms exhibit higher robustness than those of organisms of lower level.

The enhanced sampler is furthermore useful for construction neutral paths. The naive approach to identifying neutral paths between two given sequences σ_1 and σ_2 (Göbel and Forst, 2002) is to exhaustively check all shortest compatible paths between them for neutrality. While this is feasible for small d_S, as d_S increases, the number of these shortest paths grows hyperexponential. In addition, a neutral path might still exist even when all shortest compatible paths are not neutral. The enhanced sampler shows that even at large Hamming distance, sampled sequence have a high IFR, provided reference sequence and structure are natural. This motivated the ‘‘divide and conquer’’ strategy employed to construct the neutral paths. We use the enhanced sampler to construct recursively ‘‘intermediate’’ sequences that are traversed by the neutral path. Iterating this process, we can reduce the Hamming distances to the point where exhaustive search becomes feasible (Fig. 7 in Section 5).

It is possible that the shortest possible neutral path has length strictly greater than the S-compatible distance, however, Case 1 does not consider any such paths. To validate the approach of Case 1, we consider sequence–structure pairs of the microRNA let-7 family across various species (human, cattle, lizard, and other low-level organism, 12 pairs for each class) as the origin. Then, for each sequence–structure pair, we identify inverse fold solutions by dual sampling 1×10^4 sequences of Hamming distance 5 and consider all neutral solutions[5] as the terminus. By exhaustive search, we observe that for all of these sequence pairs, there exists s neutral path, whose length is equal to the S-compatible distance.

ACKNOWLEDGMENTS

The authors want to thank Stanley Hefta and Peter Stadler for their input on this article. They gratefully acknowledge the help of Kevin Shinpaugh and the computational support team at BI, Mia Shu, Thomas Li, Henning Mortveit, Madhav Marathe, and Reza Rezazadegan for discussions. The fourth author is a Thermo Fisher Scientific Fellow in Advanced Systems for Information Biology and acknowledges their support of this work.

[5] On average 5×10^3 neutral sequences were found.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

HamSampler. Available at http://staff.vbi.vt.edu/fenixh/HamSampler.zip.

Barrett, C., Huang, F., and Reidys, C.M. 2017. Sequence-structure relations of biopolymers. *Bioinformatics* 33, 382–389.

Baumstark, T., Schrüder, A.R., and Riesner, D. 1997. Viroid processing: Switch from cleavage to ligation is driven by a change from a tetraloop to a loop e conformation. *EMBO J*. 16, 599–610.

Bernhart, S., Tafer, H., Mückstein, U., et al. 2006. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.* 1, 3.

Borenstein, E., and Ruppin, E. 2006. Direct evolution of genetic robustness in microRNA. *Proc. Natl. Acad. Sci*. 103, 6593–6598.

Breaker, R.R. 1996. Are engineered proteins getting competition from RNA? *Curr. Opin. Biotechnol*. 7, 442–448.

Breaker, R.R., and Joyce, G.F. 1994. Inventing and improving ribozyme function: Rational design versus iterative selection methods. *Trends Biotechnol*. 12, 268–275.

Busch, A., and Backofen, R. 2006. INFO-RNA—A fast approach to inverse RNA folding. *Bioinformatics* 22, 1823–1831.

Chen, S.-J., and Dill, K.A. 2000. RNA folding energy landscapes. *Proc. Natl. Acad. Sci*. 97, 646–651.

Darnell, J.E. 2011. *RNA: Life's Indispensable Molecule*. Cold Spring Harbor Laboratory Press, New York, NY.

Dill, K.A., and Chan, H.S. 1997. From levinthal to pathways to funnels. *Nat. Struct. Mol. Biol*. 4, 10.

Ding, Y., and Lawrence, C.E. 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.* 31, 7280–7301.

Freyhult, E., Moulton, V., and Clote, P. 2007. Boltzmann probability of RNA structural neighbors and riboswitch detection. *Bioinformatics* 23, 2054–2062.

Garcia-Martin, J.A., Bayegan, A.H., Dotu, I., et al. 2016. RNAdualPF: Software to compute the dual partition function with sample applications in molecular evolution theory. *BMC Bioinformatics* 17, 424.

Göbel, U. 2000. Networks of minimum free energy RNA secondary structures [PhD Thesis], University of Vienna.

Göbel, U., and Forst, C.V. 2002. RNA pathfinder–global properties of neutral networks. *Zeitschrift fuer physikalische Chemie* 216, 175.

Grüner, R., Giegerich, R., Strothmann, D., et al. 1996. Analysis of RNA sequence structure maps by exhaustive enumeration I. Structures of neutral networks and shape space covering. *Chem. Mon.* 127, 355–374.

Hofacker, I.L., Fontana, W., Stadler, P.F., et al. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* 125, 167–188.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* 217, 624–626.

Kleitman, D. 1970. Proportions of irreducible diagrams. *Stud. Appl. Math.* 49, 297–299.

Kozomara, A., and Griffiths-Jones, S. 2013. mirbase: Annotating high confidence micrornas using deep sequencing data. *Nucl. Acids Res*. 42, D68–D73.

Levin, A., Lis, M., Ponty, Y., et al. 2012. A global sampling approach to designing and reengineering RNA secondary structures. *Nucl. Acids Res.* 40, 10041–10052.

Lorenz, R., Flamm, C., and Hofacker, I.L. 2009. 2d projections of RNA folding landscapes. *GCB* 157, 21.

Mandal, M., and Breaker, R.R. 2004. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat. Struct. Mol. Biol.* 11, 29–36.

Martinez, H.M. 1984. An RNA folding rule. *Nucl. Acids Res.* 12(1 Pt 1), 323–334.

Mathews, D., Sabina, J., Zuker, M., et al. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288, 911–940.

McCaskill, J.S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29, 1105–1119.

Nebel, M.E., Scheid, A., and Weinberg, F. 2011. Random generation of RNA secondary structures according to native distributions. *Algorithms Mol. Biol.* 6:24.

Nussinov, R., Piecznik, G., Griggs, J.R., et al. 1978. Algorithms for loop matching. *SIAM J. Appl. Math.* 35, 68–82.

Onuchic, J.N., Luthey-Schulten, Z., and Wolynes, P.G. 1997. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem*. 48, 545–600.

Ponty, Y. 2008. Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy: The boustrophedon method. *J. Math. Biol.* 56, 107–127.

Reidys, C.M. 1997. Random induced subgraphs of general *n*-cubes. *Adv. Appl. Math.* 19, 360–377.

Reidys, C.M., Stadler, P.F., and Schuster, P. 1997. Generic properties of combinatory maps and neutral networks of RNA secondary structures? *Bull. Math. Biol.* 59, 339–397.

Rezazadegan, R., Barrett, C., and Reidys, C. 2018. Multiplicity of phenotypes and RNA evolution. *J. Thero. Biol.* 447, 139–146.

Rodrigo, G., and Fares, M.A. 2012. Describing the structural robustness landscape of bacterial small rnas. *BMC Evol. Biol*. 12, 52.

Rogers, E., and Heitsch, C.E. 2014. Profiling small RNA reveals multimodal substructural signals in a boltzmann ensemble. *Nucl. Acids Res.* 42, e171.

Schultes, E.A., and Bartel, D.P. 2000. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science* 289, 448–452.

Serganov, A., and Patel, D.J. 2007. Ribozymes, riboswitches and beyond: Regulation of gene expression without proteins. *Nat. Rev. Genet.* 8, 776–790.

Stein, P., and Everett, C. 1978. On a class of linked diagrams ii. asymptotics. *Discrete Math*. 21, 309–318.

Tacker, M., Stadler, P.F., Bornberg-Bauer, E.G., et al. 1996. Algorithm independent properties of RNA structure prediction. *Eur. Biophy. J.* 25, 115–130.

Tinoco, I., and Bustamante, C. 1999. How RNA folds. *J. Mol. Biol*. 293, 271–281.

Turner, D., and Mathews, D.H. 2010. NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucl. Acids Res.* 38(Database), 280–282.

Waldispühl, J., Devadas, S., Berger, B., et al. 2008. Efficient algorithms for probing the RNA mutation landscape. *PLoS Comput. Biol*. 4, e1000124.

Waterman, M.S. 1978. Secondary structure of single-stranded nucleic acids. *Adv. Math. (Suppl. Stud.)*. 1, 167–212.

Wolfinger, M.T., Svrcek-Seiler, W.A., Flamm, C., et al. 2004. Efficient computation of rna folding dynamics. *J. Phys. A* 37, 4731.

Zuker, M., and Stiegler, P. 1981. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* 9, 133–148.

Address correspondence to:
Prof. Christian M. Reidys
Biocomplexity Institute of Virginia Tech
1015 Life Science Circle
Blacksburg, VA 24061

E-mail: duck@santafe.edu