An efficient dual sampling algorithm with Hamming distance filtration

Fenix W. Huang 1, Qijun He 1, Christopher Barrett 1,4 and Christian M. Reidys 1,2,3 ∗

1 Biocomplexity Institute of Virginia Tech, Blacksburg, VA, USA. 2 Department of Mathematics, Virginia Tech, Blacksburg, VA, USA. 3 Thermo Fisher Scientific Fellow in Advanced Systems for Information Biology. 4 Department of Computer Science, Virginia Tech, Blacksburg, VA, USA.

∗To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Recently, a framework considering RNA sequences and their RNA secondary structures as pairs, led to some 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.

Results: We present here the Hamming distance filtered, dual partition function, 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 being in the context of the evolution of natural sequence-structure pairs of microRNAs and the second constructing neutral paths. The former studies the inverse fold rate (IFR) of sequence-structure pairs, filtered by Hamming distance, observing that such pairs evolve towards higher levels of robustness, i.e., increasing IFR. The latter is an algorithm that construct neutral paths: given two sequences in a neutral network, we employ the sampler in order to construct short paths connecting them, consisting of sequences all contained in the neutral network.

Availability: The source code is freely available at http://staff.vbi.vt.edu/fenixh/HamSampler.zip

Supplementary information: Supplementary material containing additional data tables are available at Bioinformatics online.

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 conveyer of genetic information from its repository in DNA to the ribosomes. Over the last several decades, however, researchers have discovered an increasing number of important participate in processing of messenger RNAs, to help maintain the telomers of eukaryotic chromosomes, and to influence gene expression in multiple ways (Darnell [2011], Breaker [1996], Serganov and Patel [2007], Breaker and Joyce [1994]). 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 cuckoo can be a hurdle, as increasing number of researchers have...
are contact structures with noncrossing arcs when presented as a diagram, see Fig. 1.

The key feature of RNA secondary structures is that they can be inductively constructed. Stein and Everett (1978), Waterman et al. (1978), Nussinov et al. (1978) and Kleiman (1970) studied the combinatorics and folding of RNA secondary structures. The noncrossing arcs of RNA secondary structures allow for a recursive build: let $S_2(n)$ denotes the number of RNA secondary structures over $n$ nucleotides then we have Waterman (1978), $S_2(n) = S_2(n - 1) + \sum_{j=2}^{n-3} S_2(n - 2 - j)S_2(j)$, where $S_2(n) = 1$ for $0 \leq n \leq 2$. The recursion forms the basis for more than three decades of research resulting in what can be called 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. (Zaker and Stiegler 1981) Hofacker et al. (1994), employing the energy values derived in Mathews et al. (1999); Turner and Mathews (2010). The so-called inverse folding, i.e., identifying sequences that realize a given structure as MFE-structure, has been studied in Hofacker et al. (1994); Busch and Backofen (2006).

![Fig. 1. tRNA: secondary structure and diagram representation.](image)

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 via random graph theory (Reidys, 1997), in the context of the molecular quasispecies (Reidys et al. 1997) and by exhaustive enumeration (Grüner el al. 1990, Göbel 2000). A neutral network represents the set of all inverse folding solutions of a fixed structure. Graph properties, like for instance, size, density and connectivity are of crucial functionality in molecular evolution. Clearly, a vast, extended neutral network is more accessible than small, localized one and on a connected and dense neutral network, neutral evolution can easily be facilitated via point- and pair-mutations. On such a network, a population of RNA sequences can explore sequence space via gradual genotypic changes while maintaining its phenotype.

However, there is more to sequences and structures than MFE-folding: certain RNA sequences exhibit multiple, distinctly different, stable configurations (Baumstark et al. 1997; Schultz and Bartel 2000), as for example, riboswitches (Serganov and Patel 2007); Mandal and Breaker 2004. Recently Kezadegan et al. 2017 evolutionary trajectories, so called drift walks have been considered that are obtained by either neutral evolution or switching between a multiplicity of MFE-structures present at a fixed sequence. Such sequences indicate that is may not suffice to entire RNA energy landscape. Energy landscapes of sequences, i.e., 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. (Dill et al. 1997; Onuchic el al. 1997; Martinez 1984; Wolfsberg et al. 2004).

In McCaskill (1990), McCaskill 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, as, 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; Bernardi et al. 2006; Rogers and Hesitsch 2014). Aside from global features, local features are being studied: for instance, local minima of the energy landscape, i.e., energy traps are crucial to the understanding of folding dynamics since they represent the metastable configurations (Chen and Dill 2000; Tinoco and Bustamante 1999). 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 that a fixed sequence is compatible with. This gives rise to consider the pairing (Barrett et al. 2017)

$$\eta: \mathcal{N}^\sigma \times S_\sigma \rightarrow \mathbb{R},$$

which maps a fixed sequence-structure pair into its free energy. Here $\mathcal{N}^\sigma$ and $S_\sigma$ denote the space of sequences, $\sigma$, and the space of secondary structures, $S$, respectively. The pairing illuminates the symmetry between sequences and structures, suggesting to consider the “dual” of RNA energy landscape, i.e., the spectrum of free energies of sequences with respect to 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 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 studying 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 paper 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, see Fig. 3. 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 Wadswitj et al. (2008), where the structure partition
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 here is based on the loop-decomposition of the fixed structure and has a time complexity of $O(h^2n)$.

The paper is organized as follows: in Section 2, we discuss our sampling algorithms. In Section 3, we study two application contexts of the dual Boltzmann sampler with the Hamming distance filtration. First we study the inverse fold rate as a function of Hamming distance and then we employ our dual sampler in order to explicitly construct neutral paths in neutral networks.

2 Methods

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

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. The straightforward approach would be to run a rejection sampler based on the sampler introduced in [Garcia-Martin et al., 2016]. Barrett et al. [2017]. However, as we shall prove in Section 3 this would result in a rather inefficient algorithm. Instead, we follow a different approach, introducing a new parameter $h$ associated to a substructure, representing the Hamming distance.

Let us first 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 presenting nucleotides and arcs are presenting base-pairs, see Fig. 1. Vertices are labeled by $n = \{1, 2, \ldots, n\}$ from left to right, indicating the orientation of the backbone from the 5′-end to the 3′-end. A base-pair, denoted by $(i,j)$ is an arc connecting vertices labeled by $i$ and $j$. Two arcs $(i,j)$ and $(r,s)$ are called crossing if for $i < r$, $i < r < j < s$, 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 from the energies of loops involving multiple base-pairs. A loop $L$ in a secondary structure is a sequence of intervals $[(i_1, b_1), (i_2, a_2), \ldots, (i_k, b_k)]$, $1 \leq i \leq k$, where $(a_1, b_1), (b_1, a_1+1), 1 \leq i \leq k-1$, are base-pairs. Since no crossing arcs are allowed, nucleotides in the interval $[a_1+1, b_1-1]$ 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 $k \geq 3$, a multiloop. Note that the arc $(a_1, b_1)$ is the maximal arc of the loop, i.e., $(b_1, a_1+1) \prec (a_1, b_1)$ for all $1 \leq i \leq k - 1$, whence $L$ can be represented by $(a_1, b_1)$. 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 (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, 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, i.e.,

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

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

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 denote such substructures by $X_{i,j}^S$. In case the interval $[i,j]$ contains no arcs, we simply refer to the substructure $X_{i,j}^S$ again as an interval. Given $S$, the concatenation of the two substructures $X_{i,q}^S \cup X_{j,r}^S$ is the substructure $X_{i,r}^S$. In the following we shall simply write $X_{i,r}^S$ instead of $X_{i,j}^S$.

In particular, let $(i,j)$ be a base-pair, $L$ be the loop that is represented by $(i,j)$ and let $S_{i,j}$ be the substructure for which $(i,j)$ is the maximal arc. Suppose $(p_{i,q}), 1 \leq r \leq k$ are base-pairs in $L$, different from $(i,j)$, removing the arc $(i,j)$ produces a sequence of substructures $S_{p_{i,q}}, \ldots, S_{p_{i,q}}$, as well as a sequence of intervals $[q_{1} + 1, p_{1} - 1], [q_{1} + 1, p_{2} - 1], \ldots, [q_{k} + 1, j - 1]$.

Let $q_{0} = i$, concatenating the interval $[q_{0} + 1, j, p_{1} - 1]$ with $S_{p_{i,q}}$ produces a substructure, which we denote by $M_{q_{0} + 1, q_{1}}^r, 1 \leq r \leq k$. Let $R_{q_{0} + 1, q_{1}}$ be the substructure obtained by concatenating all $M_{q_{0} + 1, q_{1}}^r, 1 \leq r \leq k$, i.e., $\bigcup_{1 \leq r \leq k} M_{q_{0} + 1, q_{1}}^r$. By construction, removing $(i,j)$ from $S_{i,j}$ generates $R_{q_{0} + 1, q_{1}} \cup [q_{k} + 1, j - 1]$.

Note that the $R_{q_{0} + 1, q_{1}}$ can be obtained by concatenation recursively $M_{q_{0} + 1, q_{1}}^r, 1 \leq r \leq k$. We use the superscript $w$ to represent the intermediates (recursively concatenating from right to left):

$$R_{q_{w} + 1, q_{k}}^w = \bigcup_{w \leq r \leq k} M_{q_{r} + 1, q_{r+1}}^w \cup R_{q_{w} + 1, q_{k}}^w.$$

Clearly we have the following bipartition:

$$R_{q_{w} + 1, q_{k}}^w = M_{q_{w} + 1, q_{k}}^w \cup R_{q_{w} + 1, q_{k}}^w.$$

This decomposition of secondary structures allows us to compute the partition function efficiently.

Definition 1. Given a structure $S$ and a reference sequence $\sigma$, the partition function of $S$ with Hamming distance filtration $h$ to $\sigma$ is given by

$$Q_h^{S, \sigma} = \sum_{\sigma, d(\sigma, \sigma) = h} e^{-\frac{h}{R \beta}},$$

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

In the following we omit the explicit reference to $\sigma$ and simply write $Q_h^{S}$. The remaining task is now to calculate the probability of a sequence $\sigma$, given a structure $S$ and a Hamming distance $h$.

$$P_h^{\sigma, S} = \frac{Q_h^{S, \sigma}}{Z_h^{S}},$$

where $Z_h^{S}$ is the partition function of $S$ without Hamming distance filtration.
the partition functions of substructures $X_{a,b}$, $Q_h^{X_{a,b}}(N_a, N_b)$, where $X_{a,b} = S_{a,b}$, $R_{w,a}^w$, or $M_{w,a}^w$, whose left and right endpoints $a = N_a$ and $b = N_b$ are determined and contributes $h$ to Hamming distance. We consider the set of subsequences

$$\{\sigma_{a,b} \in S_{a,b}^h, d(\sigma_{a,b}, \mathcal{S}_{a,b}) = h, \sigma_{a} = N_a, \sigma_{b} = N_b\},$$

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

$$Q_h^{X_{a,b}}(N_a, N_b) = \sum_{\sigma_{a,b} \in S_{a,b}^h(N_a, N_b)} e^{-\eta(\sigma_{a,b}, X_{a,b})/T}$$

(3)

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

We next derive the recursion for $Q_h^{X_{a,b}}(N_a, N_b)$, computed from bottom to top.

**Case 1**: $(i, j)$ is $<$-minimal, i.e., $S_{i,j}$ is a hairpin loop ($k = 0$). By eq. (3), summing over all subsequence $\sigma_{i,j} \in S_{i,j}^h(N_i, N_j)$ we derive

$$Q_h^{S_{i,j}}(N_i, N_j) = \sum_{\sigma_{i,j} \in S_{i,j}^h(N_i, N_j)} e^{-\eta(\sigma_{i,j}, S_{i,j})/T}.$$

**Case 2**: $(i, j)$ is non-minimal and $k = 1$, i.e., $L$ is an interior 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 $d(\sigma_{i,j}, \sigma_{i,j}) = h$ and $d(\sigma_{p,q}, \sigma_{p,q}) = t$, where $0 \leq t \leq h$. 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

$$\sum_{t, t_1 + t_2 = h - t} \sum_{N_p, N_q} e^{-\eta(\sigma_{i,j}, S_{i,j})/T} Q_h^{S_{i,j}}(N_p, N_q),$$

where $t + t_1 + t_2 = h$, $N_p, N_q \in N, S_{i,j} \in S_{i,j}^{[t_1, t_2]}(N_i, N_q)$ and $\sigma_{i,j} \in S_{i,j}^{[t_1, t_2]}(N_p, N_q)$. Here $\delta_x = 1$ if $N_x = \sigma_x$, and $\delta_x = 0$, otherwise, for $x = p, q$.

**Case 3**: $(i, j)$ is non-minimal and $k \geq 2$, i.e., $L$ 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 $Q_h^{S_{i,j}}(N_i, N_j)$ via recursive bipartitioning. Removing $(i, j)$ produces $R_{q+1, k}^{q+1}$, as well as $[q+1, j-1]$, see Fig. 2(A). The energy $\eta(\sigma_{i,j}, S_{i,j})$ is then given by

$$\eta(\sigma, R_{q+1, q+1}) + \eta_{\text{mult}}((i, j)) + \eta_{\text{mult}}([q+1, j-1]),$$

where $\eta_{\text{mult}}$ is the energy contribution of forming a multiloop, $\eta_{\text{mult}}((i, j))$ is the energy contribution of base-pair $(i, j)$ in a multiloop, and $\eta_{\text{mult}}([q+1, j-1])$ is the energy contribution from the unpaired base interval in a multiloop. The sum of the latter three component is denoted by $\eta^p$.

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

$$\sum_t \sum_{N_{q+1}, N_{q+1}} e^{-\eta^p/T} Q_t^{R_{q+1, q+1}}(N_{q+1}, N_{q+1}),$$

where $0 \leq t \leq h$, $N_{q+1}, N_{q+1} \in N$, $\sigma_{q+1, k+1} \in S_{q+1, k+1}(N_i, N_q)$ and $\eta^p = \eta_{\text{mult}}((u, v)) + \eta_{\text{mult}}([r, s])$. Here $\eta_{\text{mult}}((u, v))$ is the energy contribution of base-pair $(u, v)$ in a multiloop and $\eta_{\text{mult}}([r, s])$ is the contribution of a segment of unpaired bases in a multiloop. We present the recursions in Fig. 2(B).

The introduction of the intermediate substructures $M_{w,a}^w$ and $R_{w,a}^w$ avoids processing concatenation of substructures simultaneously, which would result in a $O((k+1)^2)$ time complexity. The family of intermediate substructures $M_{w,a}^w$ and $R_{w,a}^w$ 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 bipartitions, resulting in a $(k - 1)(h)$ 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 $O(h^2n)$.

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 $h$ to $\sigma$ is given by $Q_h^S = Q_h^{S_{0,n+1}}(N_0, N_n)$, where $N_0$ and $N_n$ are "formal" nucleotides, discussed above.

Having computed the partition function $Q_h^S \sigma$, we implement the
An efficient dual sampling algorithm with Hamming distance filtration

3 Results

In this section, we first study the Hamming distance distribution of sequences generated via the unrestricted dual sampler [Barrett et al., 2017]. We perform this analysis for natural sequences as well as random sequences. The resulting distribution shows that a simple rejection sampler is rather inefficient and motivates the algorithm derived in Section 2.

Here we apply the refined Boltzmann sampler in order to gain deeper insight into IFR and neutral paths. First, given a sequence-structure pair, \((\sigma, S)\) we study the rate at which sampled sequences, filtered by Hamming distance, fold into \(S\). Secondly, we apply the sampler in order to develop an efficient heuristic that constructs paths within neutral networks, i.e., given two sequences, both of which folding into a fixed structure \(S\), we identify a path consisting of sequences all of which folding into \(S\), such that two consecutive sequences on the path differ only by a point- or pair-mutations.

Hamming distance distribution: 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 \(5 \times 10^4\) sequences using the unrestricted sequences sampler in [Barrett et al., 2017]. Then we compute the Hamming distance distribution to the natural sequence of the sampled sequences. The distances are normalized by sequence length. We display in Fig. [3] the distance distribution of three distinguished sequence-structure pairs, whose mean distance is in some sense minimal, typical, and maximal, respectively. The full spectrum of these distributions is presented in the SM, Fig. 1.

The data show that the sampled sequences have distances between 60% and 90% of the sequence length to the reference sequence. The mean distance is 70% to 80% of the sequence length, indicating that the unrestricted sampler in [Barrett et al. 2017] does not produce the full sampled sequences still have a Hamming distance of 60% to 90% of the sequence length. However, we observe some variations of the distribution, whence we have not only a dependence on structure, but also on the reference sequence. This implies the sampled sequences are not uniformly distributed. More interestingly, we observe that the Hamming distance distribution corresponds to the natural sequence is the least concentrated around the mean.

Inverse fold rate: we study now the inverse fold rate (IFR) of the sampled sequences with respect to different sequence-structure pairs, for different Hamming distances. We associate an indicator variable to each sampled sequence: taking the state 1 if the sequence actually 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 \((\sigma, S)\), we sample \(5 \times 10^4\) sequences from \(S\) having a fixed Hamming distance, \(h\), where \(h\) is ranging from 1 to 20. Then 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 (hun01-12), lizard (liz01-11) and drosophila (dro01-08), computing their IFRs respectively. We display the mean IFR of the three species in Fig. [4] and the IFR distributions of individual pairs within the three species in the SM, Fig. 4, Fig. 5 and Fig. 6, respectively.

The Fig. [4] shows that human has the highest, mean IFR while drosophila has the lowest. In addition, the mean IFR decreases for human significantly smaller than for drosophila. 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 hun04-structure-pair. We first consider random sequences compatible with the hun04-structure and thereby create new sequence-structure pairs. Then we compute IFR(\(h\)) of these pairs by sampling \(5 \times 10^4\) sequences of Hamming Distance 5. The IFR(\(h\)) is almost zero for these random sequences indicating that random compatible sequences have little or no connection with the hun04-structure. To identify sequences that are closer related to the hun04-structure, we use our sampler, creating 100 sequence-structure pairs by sampling sequences from the natural pairs of distance 5, 7, 10 and 20, respectively. Then we combine the newly sampled sequences with the hun04-structure, creating

Fig. 4. 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.
by their IFR(5) in increasing order, see Fig. 5. For reference purposes we display IFR(5) of the natural sequence-structure pair as a dashed line.

Fig. 5 shows that IFR(5) of the natural sequence-structure pair is above the 95 percentile, i.e., better than almost all of the newly created pairs. Furthermore, there exists very few pairs such that IFR(5) ∈ [0.1, 0.3] holds. The proportion of sequence having high IFR(5), i.e., IFR(5) > 0.3 drops when the sampled sequence have higher Hamming distance. This finding suggests that the natural pair is locally optimal.

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. 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. To be clear, let us first specify the neutral path problem:

Given two sequences σ₁ and σ₂, both folding into the structure S, identify a path σ₁ = τ₀, τ₁, . . . , τₖ = σ₂, such that
(* for all τᵢ, 0 ≤ i ≤ k, folds to S,
(**) τᵢ₊₁ is obtained from τᵢ by either a compatible point- or a base-pair mutation.

The construction of such “neutral paths” has been studied in (Gobel and Forst, 2002) using a proof idea that facilitates the construction of neutral paths, for fixed, finite distance d, in random induced subgraphs. However (Gobel 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 a neutral networks induced by folding algorithms, in particular in case of the distance between two sequences being large. In the following we shall employ our sampling algorithm in order to derive an efficient heuristic to solve the neutral path problem.

Certainly, given σ₁ and σ₂, 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 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(σ₁, σ₂) ≤ 5. Here we exhaustively search all shortest S-compatible paths between σ₁ and σ₂ and check for neutrality. Note that we always have dₛ ≤ d, thus in the worst case, we need to check 5! = 120 different paths and fold 2^{10} = 32 different sequences. This is feasible for sequence lengths shorter that 10³ nucleotides, using standard secondary structure folding algorithms (Zuker and Stiegler, 1981; Hofacker et al., 1994).

Case 2: d(σ₁, σ₂) > 5. Suppose σ₁ and σ₂ have Hamming distance h. We sample m sequences from σ₁ 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 σ₂, denoted by τᵢ. We have d(σ₁, τᵢ) = h/2 = h₁ and d(τᵢ, σ₂) = h₂, where h₁ + h₂ ≥ h. If h₂ > h we claim the process fails and we conclude we can not find a neutral path between σ₁ and σ₂. Otherwise, we repeat the process between σ₁ and τᵢ, and between τᵢ and σ₂, differentiating Case 1 and Case 2. We show the flow of the algorithm in Fig. 6.

Fig. 6. The algorithm.

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 Fig. 7.

Fig. 7. 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.

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
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.

4 Discussion

The problem of finding a sequence that folds into a given structure, $S$, has first been studied in [1993]. The algorithm consists of two parts: first it constructs a random $S$-compatible sequence and secondly it performs an adaptive walks of point mutant in the sequence such that facilitates identifying a sequence that folds into $S$. In this process, neither the inverse fold solution is guaranteed nor the number of adaptive walks required is understood. [2006] shows 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 $S$.

By exhaustive search, we can zoom into a specific sequence as well as its neighborhood in the structural ensemble. These sequences are conserved and differ vastly from each other. It is natural to bring some reference sequence. This allows us to investigate local features and consider all neutral solutions for-loop index, induced by the concatenation of two substructures. 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.

5 Acknowledgments

We want to thank Stanley Hetta and Peter Stadler for their input on this manuscript. We gratefully acknowledge the help of Kevin Shipaugh 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.

References

Barrett, C., Huang, F., and Reidys, C. M. (2017). Sequence-structure relations of biopolymers. Bioinformatics, 33(3), 382–389.

Baumstark, T., Schröder, A. R., and Riemser, D. (1997). Viroid processing: switch from cleavage to ligation is driven by a change from a tetraloop to a loop e conformation. The EMBO journal, 16(3), 599–610.

Bernhart, S., Tafer, H., Mückstein, U., Flamm, C., Stadler, P., and Hofacker, I. (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. Proceedings of the National Academy of Sciences, 103(17), 6593–6598.

Breaker, R. R. (1996). Are engineered proteins getting competition from ma? Current Opinion in Biotechnology, 7(4), 442–448.

Breaker, R. R. and Joyce, G. F. (1994). Inventing and improving ribozyme function:
Busch, A. and Backofen, R. (2006). INFO-RNA—a fast approach to inverse rna folding. *Bioinformatics*, 22(15), 1823–31.

Chen, S.-J. and Dill, K. A. (2000). Rna folding energy landscapes. *Proceedings of the National Academy of Sciences*, 97(2), 646–651.

Darnell, J. E. (2011). *RNA: life’s indispensable molecule*. Cold Spring Harbor Laboratory Press.

Dill, K. A., Chan, H. S., *et al.* (1997). From Levinthal to pathways to funnels. *Nature structural biology*, 4(1), 10–19.

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

Freyhalt, E., Moulton, V., and Clote, P. (2007). Boltzmann probability of rna structural neighbors and riboswitch detection. *Bioinformatics*.

García-Martín, J. A., Bayegan, A. H., Dotu, I., and Clote, P. (2016). RNADualPF: software to compute the dual partition function with sample applications in molecular evolution theory. *BMC Bioinformatics*, 17(1), 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(2/2002), 175.

Grüner, R., Giegerich, R., Stadler, P. F., Bornberg-Bauer, E., Hofacker, I. L., Stadler, P. F., and Schuster, P. (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., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (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.

Kleinman, D. (1970). Proportions of irreducible diagrams. *Studies in Appl. Math.*, 49, 297–299.

Kozomara, A. and Griffiths-Jones, S. (2013). mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, 42(D1), D68–D73.

Levin, A., Lis, M., Ponty, Y., O'Donnell, C. W., Devadas, S., Berger, B., and Waldispühl, J. (2012). A global sampling approach to designing and reengineering rna secondary structures. *Nucleic acids research*, 40(20), 10041–10052.

Lorenz, R., Flamm, C., and Hofacker, I. L. (2009). 2d projections of rna folding landscapes. *GCB*, 157(11-20), 21.

Mandall, M. and Breaker, R. R. (2004). Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nature Structural and Molecular Biology*, 11(1), 29–36.

Martinez, H. M. (1984). An rna folding rule. *Mathews, D., Sabina, J., Zuker, M., and Turner, D. (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), doi:10.1186/1748–7188–6–24.

Nussinov, R., Piecznik, G., Griggs, J. R., and Kleitman, D. J. (1978). Algorithms for loop matching. *SIAM J. Appl. Math.*, 35(1), 68–82.

Otuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry*, 48(1), 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(1-2), 107–27.

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

Reidy, C. M., Stadler, P. F., and Schuster, P. (1997). Generic properties of combinatorial maps and neutral networks of RNA secondary structures. *Bull. Math. Biol.*, 59(2), 339–397.

Rezazadegan, R., Barrett, C., and Reidy, C. (2017). Neutrality and continuity in rna evolution. to appear in *Bioinformatics*.

Rodrigo, G. and Fares, M. A. (2012). Describing the structural robustness landscape of bacterial small mRNAs. *BMC evolutionary biology*, 12(1), 52.

Rogers, E. and Heitsch, C. E. (2014). Profiling small RNA reveals multimodal substructural signals in a boltzmann ensemble. *Nucl. Acids Res.*, 42(22), e171.

Schultes, E. A. and Bartel, D. P. (2000). One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, 290(5478), 448–452.

Serganov, A. and Patel, D. J. (2007a). Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nature reviews. Genetics*, 8(10), 776.

Serganov, A. and Patel, D. J. (2007b). Riboswitches, riboswitches and beyond: regulation of gene expression without proteins. *Nat. Rev. Genet.*, 10, 776–790.

Stein, P. and Everett, C. (1978). On a class of linked diagrams ii. asymptotics. *Discrete Mathematics*, 21(3), 309–318.

Tacker, M., Stadler, P. F., Bornberg-Bauer, E. G., Hofacker, I. L., and Schuster, P. (1996). Algorithms independent properties of RNA structure prediction. *Eur. Biophy. J.*, 25, 115–130.

Timoco, I. and Bastamante, C. (1999). How rna folds. *Journal of molecular biology*, 293(2), 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., and Clote, P. (2008). Efficient algorithms for probing the rna mutation landscape. *PLoS computational biology*, 4(8), e1000124.

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

Wolfinger, M. T., Svercek-Seiler, W. A., Flamm, C., Hofacker, I. L., and Stadler, P. F. (2004). Efficient computation of rna folding dynamics. *Journal of Physics A: Mathematical and General*, 37(17), 4731.

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