Exponentially few RNA structures are designable

Hua-Ting Yao
LIX, UMR 7161, Ecole Polytechnique
Palaiseau, France

Cedric Chauve
Department of Mathematics, Simon Fraser University
Burnaby, Canada

Mireille Regnier
LIX, UMR 7161, Ecole Polytechnique
Palaiseau, France

Yann Ponty
yann.ponty@lix.polytechnique.fr
LIX, UMR 7161, Ecole Polytechnique
Palaiseau, France

ABSTRACT

The problem of RNA design attempts to construct RNA sequences that perform a predefined biological function, identified by several additional constraints. One of the foremost objective of RNA design is that the designed RNA sequence should adopt a predefined target secondary structure preferentially to any alternative structure, according to a given metrics and folding model. It was observed in several works that some secondary structures are undesignable, *i.e.* no RNA sequence can fold into the target structure while satisfying some criterion measuring how preferential this folding is compared to alternative conformations.

In this paper, we show that the proportion of designable secondary structures decreases exponentially with the size of the target secondary structure, for various popular combinations of energy models and design objectives. This exponential decay is, at least in part, due to the existence of undesignable motifs, which can be generically constructed, and jointly analyzed to yield asymptotic upper-bounds on the number of designable structures.

KEYWORDS

RNA Design, Inverse Folding, Pattern Matching, Analytic Combinatorics, Neutral Networks

ACM Reference Format:
Hua-Ting Yao, Cedric Chauve, Mireille Regnier, and Yann Ponty. 2019. Exponentially few RNA structures are designable. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7–10, 2019, Niagara Falls, NY, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3307339.3342163

1 INTRODUCTION

RiboNucleic Acids (RNAs) are ubiquitous biomolecules equipped with a capacity of performing a wide variety of functions, both as a messenger enabling gene synthesis (mRNAs), as a regulator of gene expression (miRNAs...) or as a direct performer of a large collection of enzymatic activities (ncRNAs) [28]. For a large subset of ncRNA family, the adoption of a predefined structure is instrumental to the function(s) of individual molecules [45], and even, at times, the survival of its hosts organisms [20]. Accordingly, the evolutionary pressure on RNA families induced by RNA structure, at the secondary structure level, is at the core of most approaches for the identification of novel ncRNA families [46]. Improved characterizations of this pressure for individual represents a key challenge of RNA Bioinformatics and the object of current work, for instance in the case of the elusive long non-coding RNAs (lncRNAs) [38].

This strong connection between RNA structure and function has motivated the continuous development of mature computational methods for structure prediction [24, 34, 49]. More recently, researchers have attempted to harness the success of folding prediction approaches, and tackled a *de novo* design of structured RNAs. In its historic setting [24], the RNA Design, or inverse folding, consists in designing a sequence of nucleotides, folding into a predefined structure according to a criterion which can computed using available computational methods. RNA design is now an established problem in RNA bioinformatics, motivated by applications ranging from synthetic biology [42] to RNA therapeutics [47] through systems biology [14] and nanotechnologies [21].

Computationally, RNA design is a hard problem [4], motivating the development of several design methodologies [6] relying on exact exponential algorithms (constraint-programming [19], SAT solving) on heuristics (local search [1, 2, 5, 24, 48], genetic algorithms [13, 32], sampling [37]...). While the former methods are limited in their scope of applications by their extreme computational demands, methods of the latter category have encountered numerous applied successes [23], and enjoy a growing popularity. Historic objectives of design include the adoption, by the produced sequence, of a structure having energy as close as possible to the Minimal Free-Energy (MFE) achieved by the sequence. Modern formulations also include the minimization of defects, properties of the thermodynamic equilibrium that indicate a notion of distance to the expectation of a perfect design [9]. Those include the *probability defect* [9], the probability of not folding into the target structure, or the *ensemble defect* [48], the expected base pair distance to the target at the thermodynamic equilibrium.

However, some secondary structures do not admit a solution to the design problem. This was first observed by Aguirre-Hernández *et al* [1], with the discovery of two *undesignable* motifs, motifs for which alternatives would always be preferred by the usual Turner energy models [43]. This claim was later generalized to simple base-pair based energy models by a study of a combinatorial version of RNA design [22], exhibiting motifs whose presence within a structure precludes its designability. However, the prevalence, in the folding space, of such undesignable motifs, and their impact on the overall combinatorics of designable structures, was never been assessed to date. Moreover, a characterization of undesignable structures would allow for sanity checks within design methods, avoiding the costly execution of a heuristics-based algorithm in contexts where no such solution exists.

Another motivation for this work pertains to theoretical evolutionary studies, where the RNA sequence to structure relationship

represents an attractive model of neutral network [18], and a crucial conceptual framework to quantify the evolvability of species [10]. Indeed, the sequence/structure relationship in RNA enables the existence of, possibly large and highly diverse, subsets of sequences (genotype) folding into the same structure (phenotype), thus achieving the same fitness level. Studies of RNA neutral networks [17, 40] often require an enumeration of accessible phenotypes, *i.e.* the number of RNA secondary structures of a given size which are adopted as the most stable structure for *some* sequence. Since no exact method is known to compute this quantity, studies rely on available asymptotic estimates for the number of all secondary structures [44]. Such an implicit assumption of universal designability may bias studies [27, 29] of the underlying evolutionary dynamics, by artificially inflating the cardinality of structural ensembles. It is thus crucial to provide more precise (approximate) expressions for the number of designable structures.

In this work, we show that the existence of small undesignable motifs, which we call *local obstructions*, constitutes an intrinsic feature of RNA design objectives. An enumeration of the secondary structure that avoid those motifs thus represents an upper bound on the number of designable structures. A direct consequence of this observation is that the proportion of designable secondary structures is typically negligible beyond a certain sequence sizes. Indeed, a tree motif perspective on the problem, coupled with classic results in analytic combinatorics [16] imply that the proportion of designable structures over n nucleotides scales like α^n, where $\alpha < 1$ can be numerically computed from any collection of local obstructions. As a side product of our automated method for computing local obstructions, we are also able to establish a list of likely candidate sequences for each motifs of a given size.

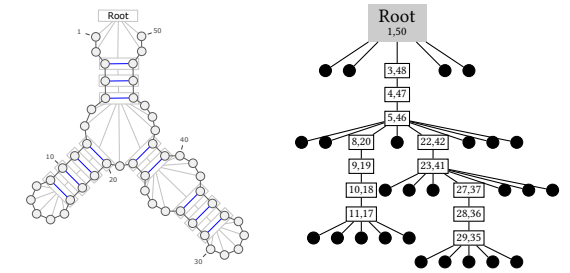
After dedicating Section 2 to formal definitions for the key concepts, and state our main result, we describe in Section 3 the application of our general strategy on a simple combinatorial version of RNA design. We then show in Section 4 how small local obstructions can be computed, for any combination of defect, tolerance and energy models. Section 5 introduces a generic specification for enumerating secondary structures that avoid a collection of local obstructions, and describes a simple numerical procedure to derive asymptotic equivalents for the number of such structures. Section 6 presents the results of our analysis of different design objectives, using the realistic Turner energy model.

2 BACKGROUND AND RESULTS OVERVIEW

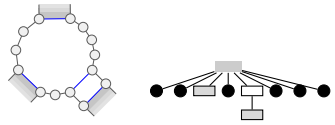
RNA secondary structure. An RNA can be abstracted as a sequence $w \in \Sigma^*, \Sigma := \{A, C, G, U\}$, of nucleotides, having length $n := |w|$. For a sequence w, a secondary structure is a set S of base pairs $(i, j), i < j \in [1, n]$, representing the interaction of nucleotides at positions i and j through hydrogen bonds, such that:

(1) Base pairs are pairwise non-crossing, *i.e.* $\nexists (i, j), (k, l) \in S$ such that $i < k < j < l$;
(2) A minimal distance of θ is ensured between interacting positions, *i.e.* $\forall (i, j) \in S, j - i > \theta$;
(3) Any position of $[1, n]$ is involved in at most one base pair.

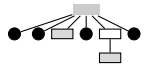
Positions of $[1, n]$ that are not involved in any base pair are called *unpaired*. In the following, we will denote by $\mathcal{S}$ the entire set of



(a) Secondary structure S, as a graph (left) or a tree (right)



(b) Motif $m_1 \in S$ (left–graph, right–tree)



(c) Motif $m_2 \notin S$

Figure 1: Graph and tree representations of an RNA secondary structure S of size 50 (1a). It features an occurrence of the motif m_1 (1b), having size 14 with 2 paired leaves, rooted at node $(5, 46)$. Although m_2 (1c) resembles m_1, it lacks two unpaired nodes and is thus does not strictly occurs at position $(5, 46)$.

secondary structures, and by $\mathcal{S}_n$ its restriction to structures of length n.

A shown in Fig. 1, a secondary structure S of length n can be unambiguously represented as a rooted ordered tree $T = (V := V_i \cup V_j, E)$, whose nodes are either intervals $[i, j] \in V_i$, $i < j$, representing base paired positions (i, j) in S, or singletons $\{i\} \in V_j$, representing an unpaired position i in S. Note that leaves of such tree represent only singletons. Any edge $(u \to v) \in E$ connects intervals such that $u \subset v$ and $\nexists v' \in V_i$ such that $u \subset v' \subset v$.

Energy model. An energy model assigns a free-energy value to each pair (w, S), where w is an RNA sequence and S is a secondary structure for w. Popular energy models for RNA folding prediction, such as the base pairs-based *Nussinov* and the *Turner* nearest-neighbors models, consider additive contributions associated with the *shallow subtrees, i.e.* subtrees of depth 1 occurring in S, and their respective nucleotides assignments. Hence, an *energy model* is a function $E : \Sigma^* \times \mathcal{S} \to \mathbb{R} \cup \{+\infty\}$ such that

$$E(w, S) = \sum_{T = \underset{a\ b\ c\ \cdots}{p} \in S} \Delta G\left(T, \{p \to w_p, a \to w_a, b \to w_b \ldots\}\right)$$

where $\Delta G(T, m)$ is the free-energy, expressed in kcal.mol^{-1} associated with the assignment m of concrete nucleotides from w to the (pairs of) positions in the subtree T. In practice, values taken by ΔG are tabulated or extrapolated from experimentally-measured values.

RNA Folding. RNA structure modeling aims, given a sequence w, to find one or several folding(s) of w into RNA secondary structure(s). Several paradigms exist, associated to different objective functions measuring the quality of a folding. In the energy minimization setting, a central algorithmic question is to compute the

minimum free energy (MFE) structure(s), optimizing the thermodynamic stability:

$$\mathsf{MFE}(w) = \left\{ S \in \mathcal{S}_{|w|} \mid E(w,S) = \min_{S' \in \mathcal{S}_{|w|}} E(w,S') \right\}.$$

A second, increasingly popular, paradigm strives to predict structures that are representative of the Boltzmann ensemble of low energy structures. Under the hypothesis of a Boltzmann equilibrium, statistical mechanics postulates that, for a given sequence w, the putative secondary structures follow a Boltzmann distribution

$$\mathbb{P}(S \mid w) = \frac{\mathcal{B}(w,S)}{\mathcal{Z}_w}$$

$$\text{with } \mathcal{B}(w,S) = e^{\frac{-E(w,S)}{RT}} \text{ and } \mathcal{Z}_w := \sum_{S' \in \mathcal{S}_{|w|}} \mathcal{B}(w,S')$$

where R is the Boltzmann constant, T is the temperature, $\mathcal{B}(w,S)$ is called the *Boltzmann factor* of w and S, and $\mathcal{Z}_w$ the *partition function* of w. Similarly, the probability of a base pair is defined as

$$p_w(i,j) = \sum_{\substack{S \in \mathcal{S}_n \\ (i,j) \in S}} \mathbb{P}(S \mid w)$$

and $p(i,i)$ represents the probability of i being left unpaired.

Note that, while an MFE structure has maximum probability in the Boltzmann ensemble, its probability can be arbitrary low, so achieving a high probability is not reducible to being an MFE. In fact, modern approaches typically elect structures that are, on average, maximally similar (MEA [31], centroids [8]) to random structures in the Boltzmann ensemble.

Defects and negative RNA design. Given a *target secondary structure* S^*, the *negative RNA design* problem, also called inverse folding, consists in producing one or several RNA sequences w that folds into S^* while avoiding alternative folds of similar quality for the chosen energy model.

The avoidance of alternative structures is captured by a notion of *defect*, defined as a function $\mathcal{D} : \Sigma^* \times \mathcal{S} \to \mathbb{R}$. RNA design methods usually consider one of the three following defects:

(1) The *Suboptimal Defect* $\mathcal{D}_S$ of a sequence w is defined as the energy distance to the first suboptimal, such that

$$\log \mathcal{D}_S(w,S^*) := -\min_{\substack{S \in \mathcal{S}_{|w|} \\ S \neq S^*}} E(w,S) - E(w,S^*);$$

(2) The *Probability Defect* $\mathcal{D}_P$ represents the probability of folding into any other structure than S^*:

$$\mathcal{D}_P(w,S^*) := \sum_{\substack{S \in \mathcal{S}_{|w|} \\ S \neq S^*}} \mathbb{P}(S \mid w) = 1 - \mathbb{P}(S^* \mid w);$$

(3) The *Ensemble Defect* $\mathcal{D}_E$ is the expected base pair distance between S^* and a random structure, generated with respect to the Boltzmann probability distribution:

$$\mathcal{D}_E(w,S^*) := \sum_{S \in \mathcal{S}_{|w|}} \mathbb{P}(S \mid w) \cdot |S \triangle S^*| = |w| - \sum_{(i,j) \in S^*} p_w(i,j)$$

with $|S \triangle S'|$ a shorthand for the set symmetric distance, also known as base pair distance.

Now we can define the *main objectives of negative RNA design.* Given a real-valued threshold $\varepsilon \geq 0$ and a defect $\mathcal{D}$, a sequence w is a *(negative)* $(\mathcal{D},\varepsilon)$*-design* for a structure S^* if and only if

$$\mathsf{MFE}(w) = \{S^*\} \quad \text{and} \quad \mathcal{D}(w,S^*) \leq \varepsilon. \tag{1}$$

Similarly, we call $(\mathcal{D},\varepsilon)$-designable a secondary structure that does admit at least a valid design. Note that the defect definition also depends on the chosen energy model, but we chose to make this dependency implicit for the sake of simplicity.

Motifs and local defect. A *motif* is a rooted ordered tree, similar to a secondary structure, but whose leaves may represent base paired positions. We say *a motif m occurs in a secondary structure S (resp. a motif m')* or *a secondary structure S (resp. a motif m') contains a motif m* if m is a subtree of S (resp. m'), rooted at any base paired node in S (resp. m') and obtained by deleting all the children for a subset of its base paired nodes. In other words, a node in m either has exactly all of its children within S (resp. m'), or none. See Fig. 1 for an example.

Consider a motif m, having a root base-pair (i,j) and paired leaves $(i_1,j_1),\ldots,(i_l,j_l)$, and let w, $|w| = n$, be an assignment of nucleotides to the positions of m. We define the *local defect* $\mathcal{D}^L(w,m)$ similarly as $\mathcal{D}$, by replacing $\mathcal{S}_n$ with

$$\mathcal{S}_m := \{S \in \mathcal{S}_n \mid (i,j) \in S \text{ and } (i_\ell,j_\ell) \in S, \forall \ell \in [1,l]\}$$

a restricted set of structures where both the root, and all the paired leaves, of m appear as base pairs. A crucial observation, of which we omit a formal proof in the interest of space, is stated in the following proposition.

Proposition 1. *For any defect $\mathcal{D} \in \{\mathcal{D}_S, \mathcal{D}_P, \mathcal{D}_E\}$, sequence w, $|w| = n$, and structure $S \in \mathcal{S}_n$, one has*

$$\mathcal{D}(w,S) \geq \mathcal{D}^L(w_m,m), \forall m \in S$$

where w_m is the restriction of w to the positions in m.

Corollary 1. *If there exists a motif $m \in S^*$ such that*

$$\mathcal{D}^L(w,m) \geq \varepsilon, \forall w \in \Sigma^{|m|},$$

then S^* cannot be $\mathcal{D}$-designed.

In other words, the presence in the target structure S^* of a motif that cannot be designed *locally* is sufficient to forbid the existence of a sequence w that would constitute a design for S^*.

Problem statement and results overview. In this work, we address the following question: *Given an energy model, a design criterion, how many secondary structures of a given length actually admit a negative design?* Our main result is summarized by the following theorem.

Theorem 1. *For any energy model, defect $\mathcal{D} \in \{\mathcal{D}_S, \mathcal{D}_P, \mathcal{D}_E\}$ and tolerance $\varepsilon \geq 0$, only an exponentially small fraction of the secondary structures in $\mathcal{S}_n$ are $(\mathcal{D},\varepsilon)$-designable.*

3 BASIC COMBINATORIAL DESIGN

Here, we consider the special case of the 0-dominance criterion in the simplest energy model, the Nussinov model, considered in a previous work [22]. In this setting, the design problem simplifies

into finding an RNA sequence admitting a unique folding maximizing the number of base pairs, such that this folding coincides with the given target secondary structure.

THEOREM 2. *Let d_n be the number of secondary structures that are designable in the Nussinov model with $\theta = 1$. Then*

$$d_n \in O\left(\frac{\alpha^n}{n\sqrt{n}}\right) \tag{2}$$

where $\alpha = 2.35\ldots$ is the smallest positive real root of

$$5z^8 - 14z^7 + 13z^6 - 8z^5 + 6z^4 - 2z^3 + 4z^2 - 4z + 1.$$

COROLLARY 2. *The probability that a uniform random secondary structure of length n, with $\theta = 1$, is designable in the Nussinov model for the 0-dominance criterion is in $O(\beta^n)$, where $\beta = \alpha(3-\sqrt{5})/2 < 1$.*

To prove Theorem 2, we rely on analytic combinatorics techniques, widely used in analysis of algorithms [15] and bioinformatics [41], exposed in [16]. Their application in this context involves the following steps:

(1) Identify a collection of secondary structure motifs $\mathcal{M}$ whose occurrence in S implies that S is not designable in the Nussinov model for the 0-dominance criterion;
(2) Design a grammar for the set $\mathcal{S}^{\overline{\mathcal{M}}}$ of all RNA secondary structures excluding this motif;
(3) Derive and solve a system of functional equations satisfied by the generating function $S^{\overline{\mathcal{M}}}(z) = \sum_{n\geq 0} s_n z^n$, where s_n is the number of structures of $\mathcal{S}^{\overline{\mathcal{M}}}$ of length n;
(4) Use singularity analysis to obtain an asymptotic equivalent for s_n, in particular the coefficient α of (2) in Theorem 2, called the *growth factor*, that drives the exponential growth of s_n as a function of n.

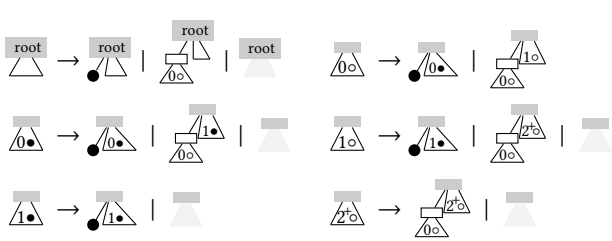
Corollary 2 follows from Theorem 2 and the fact that, when $\theta = 1$, the asymptotic number t_n of secondary structures of length n is such that

$$t_n \in \Theta\left(\frac{(2/(3-\sqrt{5}))^n}{n\sqrt{n}}\right). \tag{3}$$

The exponential decrease in the proportion of designable secondary structures follows from $2/\left(3-\sqrt{5}\right) = 2.62\ldots > \alpha = 2.35\ldots$.

We now turn to the proof of Theorem 2. For step (1) of the approach outlined above, we rely on the recent paper [22], where it was proved that a secondary structure cannot be designed if its tree representation includes an internal node whose children set contain ≥ 2 internal nodes and at least one leaf (the collection of motifs $\mathcal{M}$ discussed above).

The set of tree representations of the secondary structure avoiding this local motif can be generated by the context-free grammar given below.



Intuitively, this grammar keeps track of properties of the structural elements generated for the current internal node. Except for the root, each non-terminal is indexed by pairs taken from $(i, u) \in \{0, 1, 2^+\} \times \{\circ, \bullet\}$ where, within the current siblings, i represents the number of internal nodes/base pairs, and u expresses whether ($\bullet$) or not ($\circ$) a leaf has been generated. Notice that the grammar implicitly excludes structures having a three siblings composed of two internal nodes and one leaf ($(i, u) = (2^+, \bullet)$), i.e. the motif $\mathcal{M}$.

Following standard enumerative combinatorics techniques that links combinatorial specifications to the calculus of generating functions [16], one obtains that the ordinary generating function $S^{\overline{\mathcal{M}}}(z)$ is defined by the system of functional equations below.

$$\begin{aligned}
S^{\overline{\mathcal{M}}}(z) &= z \times S^{\overline{\mathcal{M}}}(z) + z^2 \times S_{0\circ}(z) \times S^{\overline{\mathcal{M}}}(z) + 1\\
S_{0\bullet}(z) &= z \times S_{0\bullet}(z) + z^2 \times S_{0\circ}(z) \times S_{1\bullet}(z) + 1\\
S_{1\bullet}(z) &= z \times S_{1\bullet}(z) + 1\\
S_{0\circ}(z) &= z \times S_{0\bullet}(z) + z^2 \times S_{0\circ}(z) \times S_{1\circ}(z)\\
S_{1\circ}(z) &= z \times S_{1\bullet}(z) + z^2 \times S_{0\circ}(z) \times S_{2^+\circ}(z) + 1\\
S_{2^+\circ}(z) &= z^2 \times S_{0\circ}(z) \times S_{2^+\circ}(z) + 1
\end{aligned}$$

Solving the system using algebraic elimination, followed by a careful choice of the right conjugate, one obtains a closed form for the generating function $S^{\overline{\mathcal{M}}}(z)$:

$$S^{\overline{\mathcal{M}}}(z) = -\frac{P(z)\sqrt{R(z)}}{Q(z)}$$

where

$$\begin{aligned}
P(z) &= 2z^5 - 7z^4 + 7z^3 - 6z^2 + 4z - 1\\
Q(z) &= 2z(z-1)(z^4 - 4z^3 + 3z^2 - 3z + 1)\\
R(z) &= 5z^8 - 14z^7 + 13z^6 - 8z^5 + 6z^4 - 2z^3 + 4z^2 - 4z + 1.
\end{aligned}$$

It follows from classic transfer theorems [15] that the singularity of this generating function is of square-root type, leading to the following asymptotic expansion for its coefficients,

$$s_n \in \Theta\left(\frac{\rho_{\mathcal{M}}{}^{-n}}{n\sqrt{n}}\right).$$

where $\rho_{\mathcal{M}}$ is the dominant singularity of $S_{\mathrm{r}}(z)$, *i.e.* the smallest root of $R(z)$, and can be numerically evaluated at $\rho_{\mathcal{M}} = 0.4262\ldots$.

Theorem 2 follows from $\alpha = 1/\rho_{\mathcal{M}}$ and the fact that $d_n \leq s_n$. Indeed, while it is necessary for a designable secondary structures to avoid $\mathcal{M}$, this condition is not sufficient.

4 LOCAL OBSTRUCTIONS

In this section, we describe an algorithm to compute local obstruction, motifs whose presence within a secondary structure forbids its design with respect to some predefined design objectives. Fig. 3 describes the main workflow of this study.

4.1 Emulating a local defect with constraints

The *minimal completion* of a motif m for a nucleotide assignment w is a pair (S_m, w_m) such that:

- S_m is the secondary structure obtained from m by adding θ unpaired nodes (leaves) under each paired leaf node;
- w_m is the sequence obtained by inserting θ occurrence of the letter A under paired leaves.

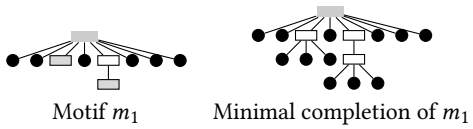


Figure 2: Minimal completion of the motif m_1 (Fig. 1b).

In this study, we set $\theta = 1$ for the *Nussinov* model and $\theta = 3$ for the *Turner* model. In other words, in the *Turner* model, the minimal completion is obtained by replacing any paired leaf by . Figure 2 illustrates the application of the minimal completion to the motif m_1 in Fig. 1b. A *trimming* operation is defined as the inverse of the completion, and allows to recover a motif/sequence pair from its completion.

Given a length k, a *folding constraint* C is a set consisting of positions from $[1, k]$ and pairs from $[1, k]^2$, respectively representing positions forced to remain unpaired and paired to a specific partner. The *constrained defect* $\mathcal{D}_C(w, S)$ can be defined by restricting the computation to structures compatible with the constraint C. Such constraints are supported by reference implementations of energy-minimization and partition-function algorithms, and can be easily enforced in simpler energy models.

Consider a motif m, and its minimal completion (S_m, w_m), we define the *induced constraint* C_m of m as consisting of:

- the root base pair of S_m;
- the base pairs in S_m stemming from the paired leaves in m;
- the unpaired positions introduced by the completion.

Intuitively, such a constraint will limit the alternative conformations, considered by the defect computation, to be consistent with the boundaries of the initial motif.

Proposition 2. *For every defect $\mathcal{D}$ and energy model considered in this work, one has $\mathcal{D}^L(w, m) = \mathcal{D}_{C_m}(w_m, S_m)$.*

In other words, the local defect of a motif can be practically computed by executing a constrained version of a, suitably constrained, global off-the-shelf algorithm (energy-minimization for $\mathcal{D}_S$, base-pair probability for $\mathcal{D}_P$ and $\mathcal{D}_E$) on the minimal completion of the motif. In particular, motifs that represent local obstructions to design, associated with large local defect, can be detected using this property as shown below.

4.2 Computing local obstructions

We now turn to the computation of a list of *local obstructions* over k nucleotides, motifs whose presence within any secondary structure implies that the overall defect exceeds a predefined *tolerance* $\varepsilon \geq 0$.

In principle, we could compute all possible motifs and nucleotides assignments, followed by an evaluation of the local defect, as described in the previous section. We could then simply consider as local obstructions any motif which, for all its compatible sequences, fails to satisfy the ε defect threshold. Indeed, Prop. 1 implies that any motif whose local defect exceeds ε cannot be part of a secondary structure having defect less than ε, thus representing a local obstruction. Since motifs are essentially structures over k nucleotides (for $\theta = 0$), the complexity of this approach would be in $O(3^k \cdot 4^k \cdot P(k))$, for $P(n)$ the complexity of the (constrained) direct MFE/defect algorithm.

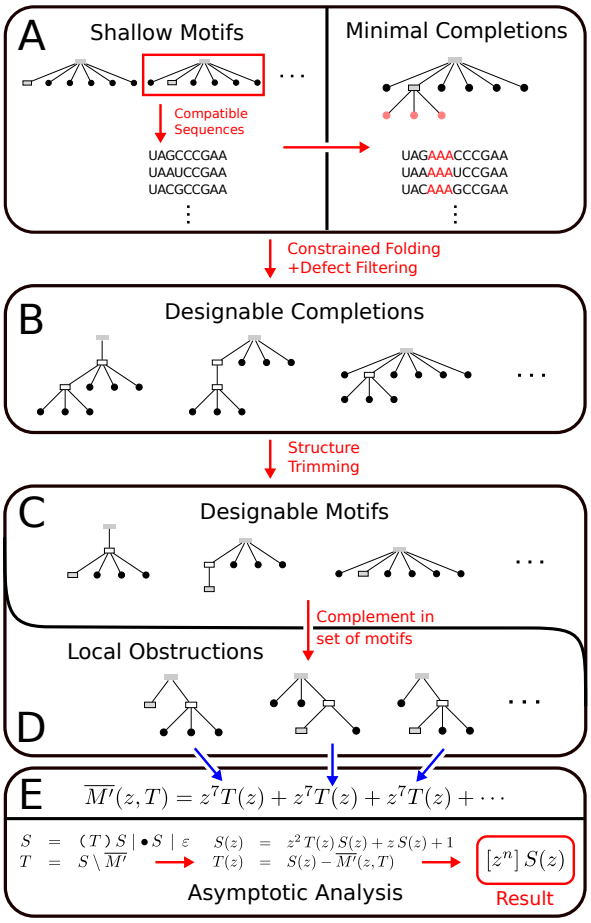


Figure 3: General workflow. Minimal completions are computed for all pairs of shallow motifs and compatible sequences (A). Constrained MFE predictions, filtered by defect, produce designs for refinements of the input motif (B). Designable completions are then trimmed into a set of designable motifs (C), itself complemented to obtain local obstructions (D). A grammar for secondary structures avoiding obstructions is built, and singularity analysis gives an asymptotic upper bound for designable structures (E).

This complexity can be further reduced by restricting the above computation to *shallow motifs*, motifs having tree height 1 consisting of paired and unpaired nodes underneath a root node. For any sequence assignment to such a motif, running a constrained energy minimization algorithm on the minimal completion of the sequence either returns one, or multiple co-optimal solutions. In the case of multiple solutions, the sequence admits several alternative local MFE folds. It is thus not suitable for any *refinement* of the shallow motif, *i.e.* any motif that includes the pairs of the shallow motif. Conversely, a unique MFE solution is provably a refinement of the shallow motif, and one concludes that the motif is designable. Since every motif over k nucleotides is a refinement of some shallow motif, this strategy produces the same output as the above-described

one. Its complexity, however, is reduced to $O(\varphi^k \cdot 4^k \cdot P(k))$, where $\varphi := (1+\sqrt{5})/2 \approx 1.62$ is the golden ratio, observing that shallow motifs are counted by the Fibonacci numbers.

Given a defect $\mathcal{D}$ restricted to a value ε, and a given motif size k, our algorithm execute the following steps:

- Enumerate all *shallow motifs* (of depth 1) of size k;
- For any such motif m°, consider any assignment w° consistent with the paired nodes in m°:
  - Build the minimal completion (S°_m, w°_m) of (m°, w°), and execute on w°_m a constrained MFE folding algorithm, using the induced constraint C_{m°;
  - If the MFE computation returns a unique structure $m^\star$, consider the motif m' obtained by trimming $m^\star$ (m' is a refinement of m°);
  - Evaluate the local defect and, if $\mathcal{D}^L(w^\circ, m') \leq \varepsilon$, add m' to the list $\mathcal{M}$ of designable motifs;
- Return $\overline{\mathcal{M}}$, the set of all motifs of size k not in $\mathcal{M}$.

A detailed version of the procedure is described in Alg. 1 and illustrated in Fig. 3.

Algorithm 1: Computing local obstructions of a given size

Input : A motif size k, an energy model E, a defect definition $\mathcal{D}$, and a tolerance $\varepsilon \geq 0$
Output: $\overline{\mathcal{M}}$ a, possibly empty, set of local obstructions of length k

```
M ← ∅;
Q_k ← Set of all shallow motifs of size k;
foreach m° ∈ Q_k do
    foreach w° ∈ Σ^k (compatible with m°) do
        C_{m°} ← Induced constraint of m°;
        (S°_m, w°_m) ← Minimal completion of (m°, w°);
        O ← MFE(w_m | C_{m°}) w.r.t. energy model E;
        if |O| = {S'} then
            if D_{C_{m°}}(w°, S') ≤ ε then
                m' ← trimming of S' ;
                M ← M ∪ {m'}
    R_k ← Complete set of all motifs of size k;
    return M̄ = R_k − M
```

Proposition 3. *Motifs returned by Alg. 1 are local obstructions.*

Proof. Fist, let us consider the properties of a motif m returned by the algorithm. Note that there exists only a single shallow motif m°, of which m is a refinement. Since $m \notin \mathcal{M}$ then, for each sequence w°, either a lower constrained MFE fold was found, or the local defect exceeded ε. In the latter case, Proposition 1 implies that any pair (S, w), where S features m, and sequence w having nucleotide assignment w° on the motif positions, has defect greater than ε, thus w is not a design for S. In the former case where an alternative motif m' is preferred to (or equally stable as) m for w°, then for any structure S containing m and sequence w, having nucleotide assignment w° on the positions of m, a competitor to S for w can be constructed by replacing m by m' in S. One concludes that, if $m \notin \mathcal{M}$, any structure S, $m \in S$, and sequence w does not represent a $(\mathcal{D}, \varepsilon)$-design. □

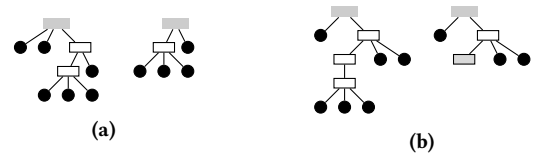


Figure 4: Two examples of pairs of redundant motifs. In both cases, the set of secondary structures rooted on the right motif strictly includes that of the right one, and we discard the left one from our computations.

The exhaustivity, for a given size k, of the list of motifs produced by Algorithm 1 remains unclear. Indeed, a motif is disregarded as a local obstruction as soon as its minimal completion folds correctly (with admissible defect) under suitable constraints for some sequence. Thus, there is no formal guarantee that a sequence would adopt this motif with an acceptable defect in the absence of constraints. However, we empirical observed that motifs not returned by the algorithm can overwhelmingly be included in design and, in particular, that the sequence of their minimal completion is a design for the completed structure. Moreover, the possible omission of some local obstructions is not overly critical, since our main goal is to provide upper bounds on the number of designable structures.

5 ENUMERATING SECONDARY STRUCTURES WHILE AVOIDING LOCAL OBSTRUCTIONS

Next, we turn to the computation of asymptotic equivalent for the number of secondary structures that avoid a collection of local obstructions, computed using the algorithm outlined in the previous section. We first start by eliminating redundant motifs, *i.e.* motifs that merely extend another motif, as shown in Figure 4, which can be done by running a classic tree alignment algorithm [26] in a pairwise fashion.

5.1 Specification and generating function

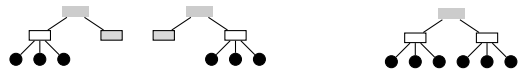
Our approach represents an instance of the symbolic method [16], and is similar in essence to the detailed example of Section 3.

We establish that the set of all secondary structures avoiding a set $\mathcal{M}$ of local obstructions is generated by the following specification:

$$\begin{array}{rcl} S & = & (\,T\,)\,S \mid \bullet S \mid \varepsilon \\ T & = & S \setminus \overline{M'} \end{array}$$

The first line essentially builds the set of all secondary structures ($\theta = 0$), and is highly reminiscent of Waterman's seminal decomposition [44]. The second line, however, subtracts the contributions of secondary structures which, when completed with a root, feature an occurrence of a local obstruction. In other words, M' denotes the set of enclosed forests built from the inner part of local obstructions

$$\overline{M'} := \{m' \mid \forall m \in \overline{\mathcal{M}}, m = (\,m'\,)\}.$$



(a) Two overlapping motifs (b) Common structure

Figure 5: Both motifs in (5a) represent local obstructions, each of them contributing $z^7\,T$ to the O.G.F.$\overline{M'}(z,T)$. However, those motifs are overlapping, and any structure in the intersection, such as (5b), will be subtracted twice. We work around this issue by including a correcting term z^{10} in $c(z,T)$.

Therefore, the system of generating functions can be written as

$$\begin{cases} S(z) &= z^2\,T(z)\,S(z) + z\,S(z) + 1 \\ T(z) &= S(z) - \overline{M'}(z,T) \end{cases} \tag{4}$$

where $\overline{M'}(z,T)$ denotes the Ordinary Generating Function (O.G.F.) of the set of enclosed structure of motifs, defined as

$$\overline{M'}(z,T) = \sum_{m' \in \overline{\mathcal{M}'}} z^{\gamma(m')}\,T^{\delta(m')} - c(z,T)$$

where $\gamma(m')$ (resp. $\delta(m')$) is the size (resp. number of paired leaves) of the motif m', and $c(z,T)$ is a correcting term to account for potential overlaps.

Indeed, in rare cases, some secondary structures may be counted in the O.G.F. associated to two or more motifs. Such structures would therefore be subtracted several times by the grammar, leading to an error while computing the singularity. Therefore, a correcting term $c(z,T)$ is introduced, as described in Figure 5 to counterbalance the overcounting in such (rare) situations. Given the scarcity of such situations, we computed those terms manually for each pair of motifs. A more systematic solution could be implemented, using ideas from Collet *et al* [7], but the lack of immediate needs led us to leave this for an extended version of this extended abstract.

5.2 Computing the dominant singularity

In general, one could use a symbolic calculus software (`Maple`) to solve the system (4), using some specialized package (`gfun` [39]) to extract the dominant singularity. However, in our case, such a approach turns out to scale poorly with the number of motifs and, more critically, paired leaves in the motifs (*i.e.* the degree of $T(z)$ in $\overline{M'}(z,T)$). Therefore, we consider an alternative approach which combines an elementary symbolic calculus with a numerical determination of the dominant singularity.

Indeed, rewriting the system shows that $T(z)$ is a solution of $G(z,y) = 0$ where

$$G(z,y) = z^2y^2 + y(z^2\overline{M'}(z,y) + z) + (z-1)\,\overline{M'}(z,y) + 1.$$

As $\overline{M'}(z,T)$ depends of $T^{\delta^\star}(z)$, for $\delta^\star = \max_{m' \in \overline{\mathcal{M}'}} \delta(m')$, the degree of G with respect to y might be greater than 2, and the problem is not directly amenable to the techniques developed in Section 3. Nevertheless, it follows from this smooth implicit-function schema that $T(z)$ is analytic. It is aperiodic and its dominant singularity, denoted ρ, is a non-zero root of $R(z)$ defined as the resultant of two polynomials in y, namely:

$$\begin{cases} P(z,y) &= G(z,y) - y \\ Q(z,y) &= \partial_y P(z,y) \end{cases}$$

The solution is easily derived by a numeric approach. The generating function $S(z)$ shares the same dominant singularity as $T(z)$. Thus, coefficients of $S(z)$ satisfy

$$[z^n]\,S(z) \in \Theta\left(\frac{\rho^{-n}}{n\sqrt{n}}\right).$$

Example. Let $\overline{\mathcal{M}}$ be restricted to the single motif (▭▭●), a special case of the local obstructions in the *Nussinov* Model described in [22]. Then, the O.G.F. of the set $\overline{\mathcal{M}'}$ with $\theta = 1$ is $1 + z^5T^2(z)$ and

$$G(z,y) = z + y(z+z^2) + y^2(z^2 - z^5 + z^6) + y^3z^7$$

Next, we compute the resultant of the polynomial $P(z,y)$ and its partial derivation on y

$$\begin{aligned} P(z,y) &= z^7y^3 + (z^6 - z^5 + z^2)y^2 + (z^2 + z - 1)y + z \\ Q(z,y) &= 3z^7y^2 + 2(z^6 - z^5 + z^2)y + (z^2 + z - 1) \end{aligned}$$

A numerical resolution of system locates the dominant singularity at $\rho = 0.3834$. We conclude that

$$[z^n]\,S(z) \in \Theta\left(\frac{2.6082^n}{n\sqrt{n}}\right).$$

6 RESULTS

We implemented Algorithm 1, and the numerical procedure to compute the dominant singularity described in Section 5, in `Python3` using the `pandas` library and `SymPy` [35], a Python library for symbolic computing. Our implementation is available at:
http://www.lix.polytechnique.fr/~ponty/?page=countingdesigns

6.1 Recovering the total number of secondary structures ($\theta = 3$)

As a first test, we ran Algorithm 1, using the suboptimal defect $\mathcal{D}_E$ as our objective and no tolerance for suboptimality ($\varepsilon := 0$), based on `RNAfold` [30] (version 2.4.12 with default parameters), in order to detect local obstructions of small sizes ($k \in [2,4]$). Unsurprisingly, but still reassuringly, our implementation returned three local obstructions, `()`, `(●)`, and `(●●)`, corresponding to the $\theta = 3$ minimal distance enforced by `RNAfold`.

Such local obstructions lead to a generating function

$$\overline{M'}(z) = 1 + z + z^2,$$

and our method produces the following asymptotic upper bound for the number of designable structures of size n:

$$[z^n]\,S(z) = s_n \in \Theta\left(\frac{2.289^n}{n\sqrt{n}}\right). \tag{5}$$

Those asymptotics match the ones reported by Hofacker *et al* [25].

6.2 Refined estimates for the phenotype space

We pushed our analysis further by using Algorithm 1 to compute the local obstructions of sizes up to 14 (hairpins) and up to 10 (internal loops and bulges). After removing the redundant motifs from the set, we manually computed the correcting terms $c(z,T)$ to avoid double-counting structures compatible with several obstructions. Then, we applied the methodology of Section 5 to produce the dominant term of the asymptotics, along with the first-order estimates for the proportion of designable structures reported in Table 1.

Defect	ε	#Local obstructions	ρ	Asymptotic equivalent	Equivalent	Proportion of designable structures* (upper bound) P_{10}(%)	P_{50} (%)	P_{100} (%)	P_{200} (%)	P_{500} (%)	P_{1000} (%)
$\mathcal{D}_E$	1	104	0.44917	$\Theta\left(\frac{2.226^n}{n\sqrt{n}}\right)$	0.973^n	76.1	25.4	6.48	$4.19 \cdot 10^{-1}$	$1.14 \cdot 10^{-4}$	$1.30 \cdot 10^{-10}$
$\mathcal{D}_P$	.5	117	0.44964	$\Theta\left(\frac{2.224^n}{n\sqrt{n}}\right)$	0.972^n	75.3	24.2	5.84	$3.41 \cdot 10^{-1}$	$6.81 \cdot 10^{-5}$	$4.64 \cdot 10^{-11}$
$\mathcal{D}_P$	.1	152	0.45967	$\Theta\left(\frac{2.176^n}{n\sqrt{n}}\right)$	0.95^n	59.9	7.69	0.59	$3.51 \cdot 10^{-3}$	$7.27 \cdot 10^{-10}$	$5.29 \cdot 10^{-21}$
$\mathcal{D}_P$	.01	174	0.48127	$\Theta\left(\frac{2.078^n}{n\sqrt{n}}\right)$	0.908^n	38.1	0.80	$6.44 \cdot 10^{-3}$	$4.14 \cdot 10^{-7}$	$1.10 \cdot 10^{-19}$	$1.22 \cdot 10^{-40}$

Table 1: Collections of local obstructions of size up to 12, and their consequences on the proportion of actually designable secondary structures. * Proportions of designable sequences computed using an assumption of equal constants for the asymptotic leading terms of the number of secondary structures, respectively allowing and forbidding local obstructions.

6.2.1 Inverse folding. In the classic setting of RNA design, the inverse folding, one attempts to design a sequence which admits a target structure as its unique MFE structure. This corresponds to choosing a suboptimal defect with $\varepsilon = 1$.

Our analysis reveal the existence of 104 motifs (after removal of redundant ones), an overwhelming majority of which contain isolated base pairs. Such motifs are expected, as they are heavily penalized, yet not explicitly forbidden (unless specified), by folding algorithms. Consecutive bulges, alternating on the 5' and 3' ends of an helix, also seem systematically suboptimal for the Turner model, a large interior loop being systematically favored as a candidate for the MFE. Finally, hairpin loops directly stemming from a multiloop are systematically discriminated, and a structure consisting of a larger unpaired stretch in the multiloop seem systematically favored.

Computing the dominant singularity yields $\rho = 0.44917$, which implies the following asymptotic upper bound on the number of secondary structures

$$[z^n]\, S(z) = s_n \in \Theta\left(\frac{2.22632^n}{n\sqrt{n}}\right). \tag{6}$$

The probability for a secondary structure of size n, taken uniformly at random, to be designable is upper-bounded by $P_n \in \Theta(0.973^n)$. Assuming the identity of constants involved in the leading terms of Equations (5) and (6), one concludes that, while about 3/4 of the structures of size 10 can be designed, this proportion quickly drops to less than 0.5% for RNAs of size 200, and reaches infinitesimal proportions (10^{-10}%) for very large RNAs of size 1 000.

6.2.2 Designing structures with large probabilities. Next, we analyze the probability defect $\mathcal{D}_P$, and investigate the impact of ε on the proportion of designable secondary structures. We consider 3 thresholds, $\varepsilon \in \{0.5, 0.1, 0.01\}$, associated with Boltzmann probabilities for the motif greater than 50%, 90% and 99% respectively. Executing Algorithm 1, followed by a removal of redundant motifs, led to the identification 117, 152 and 174 local obstructions respectively.

Interestingly, the $\varepsilon = 50\%$ case induces a dominant singularity of 0.44964, leading to a slightly slower asymptotic growth

$$[z^n]S(z) = s_n \in \Theta\left(\frac{2.22400^n}{n\sqrt{n}}\right)$$

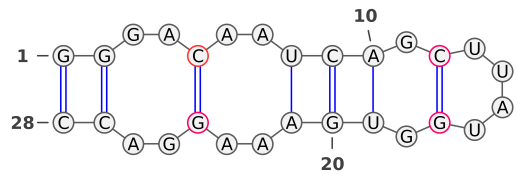


Figure 6: Example of secondary structure with isolated base pairs (red) in its MFE structure, obtained by connecting local solutions.

than for classic inverse folding. This is not entirely unexpected, since our definition of a valid design requires the target structure to be the sole MFE for the sequence. Thus, secondary structures satisfying some probability defect condition must also be solutions to the inverse folding problem. However, the observed divergence of the two singularities suggests that an exponentially small proportion (albeit with growth factor very close to 1) of MFE designs have Boltzmann probability greater than 50%.

For defect thresholds of 0.1 and 0.01 on the probability, the departure from the MFE design is much more pronounced, with respective singularities at 0.45967 and 0.48127 respectively, leading to asymptotic equivalents in

$$\Theta\left(\frac{2.17549^n}{n\sqrt{n}}\right) \quad \text{and} \quad \Theta\left(\frac{2.07783^n}{n\sqrt{n}}\right).$$

Again, assuming the equality of constants, we obtain proportions of designable structures bounded by $P_n = 0.95^n$ and $P_n = 0.908^n$ respectively. Those estimates support the notion of an extreme sparsity of designable structures in the folding space, with only three out of 10^{-5} (resp. 4 out of 10^{-9}) structures being designable for $\varepsilon = 0.1$ (resp. $\varepsilon = 0.01$). These abysmal proportions are consistent with the popular belief, which rigorously holds for the homopolymer model [12], that the Boltzmann probability of the MFE structure decreases exponentially with the sequence length in a random, uniformly distributed, RNA sequence.

7 CONCLUSION

In this work, we have addressed the designability of RNA structures for a variety of design paradigms, thresholds and energy models. We have described a procedure for computing a list of local motifs whose presence represents an obstruction to the design task.

This procedure is largely agnostic to the exact objectives of design, and holds for any design under mild assumptions (monotonicity of defects over loops). Using enumerative and analytic combinatorics techniques, we were able to automate the computation of asymptotic upper-bounds, revealing an overall sparsity of designable structures within the full conformational space, both in simple base-pair based and the complete Turner energy model. The number of designable structures still increases with the length of considered targets, but at a much slower rate than initially anticipated.

This work sets the stage for further analyses of designable structures, and unlocks a systematic way to address many further questions. For instance, the popular ensemble defect [48], could benefit from a more refined treatment using bivariate generating functions. Indeed, the ensemble defect is defined as an expectation, and is therefore fully additive on the Turner loops of the target secondary structures. One could therefore determine, through a trivial modification of Equation (4), the bivariate generating function $S(z,u) = \sum_{n,k\geq 0} s_{n,v} z^n u^k$, $s_{n,v}$ being an upper bound for the number of structures of size n having v ensemble defect. An application of the famous Drmota theorem [11] would then very likely provide sharper estimates, accounting for the accumulation of local defects rather than restricting to the worst one.

Enumerative aspects of this work could also easily be extended to secondary structures including algebraic types of pseudoknots. Indeed, multiple grammars have been shown to capture major pseudoknot classes while, at the same time, allowing for a characterization of generating functions [36]. An enumeration of designable structures would greatly help in the parametrization of free-energy models, a key aspect of pseudoknot prediction programs which has so far greatly hindered the development of predictive methods [33].

Regarding the complexity of our method for building local obstructions, we strongly believe its exponential nature may be intrinsic to the problem. More precisely, we believe that the list of local obstructions may generically grow exponentially with the length of investigated motifs. Indeed, since structures can be seen as motifs in our definitions, the existence of a polynomially-bounded list of local obstructions would imply a polynomial-time algorithm for decided whether an RNA can be designed. Unfortunately, the problem was recently been shown to be NP-hard [4], appearing to rule out a polynomial-time alternative to Algorithm 1.

On a more positive note, Algorithm 1 can easily be modified to keep the list of suitable candidate sequences for each and every designable motif. This allows to greatly restrict the search space of classic design algorithms, but also suggests a promising strategy for hard design instances. As an illustration, while investigating our database of local obstructions, we discovered that lonely base pairs appear in a few designable motifs, usually considered unstable in the *Turner* model and difficult to design for. For example, the structure `(((.(....).)))` is the MFE structure of the RNA sequence `UCAGCUUAUGGUGA`. We also found that the motif `((..(*)..))` could be designable for some collection of sequences. Combining sequences adopting these two motifs, we could verify that an RNA sequence

`GGGACAAUCAGCUUAUGGUGAAAGGACC`

is predicted by `RNAfold` to adopt its unique MFE structure of

`((..(..(((.(....).)))..)..))`

featuring two isolated base pairs, and a free-energy of -6.4kcal.mol^{-1}, a stability unmatched across several runs of `RNAinverse` [24] and `Nupack` [48]. While this observation remains anecdotal, it is supported by the success of recent approaches using (partial) libraries of local motifs [3].

REFERENCES

[1] Rosalía Aguirre-Hernández, Holger H Hoos, and Anne Condon. 2007. Computational RNA secondary structure design: empirical complexity and improved methods. *BMC Bioinformatics* 8 (2007), 34. https://doi.org/10.1186/1471-2105-8-34

[2] Assaf Avihoo, Alexander Churkin, and Danny Barash. 2011. RNAexinv: An Extended Inverse RNA Folding from Shape and Physical Attributes to Sequences. *BMC Bioinformatics* 12, 1 (2011), 319. https://doi.org/10.1186/1471-2105-12-319

[3] Stanislav Bellaousov, Mohammad Kayedkhordeh, Raymond J Peterson, and David H Mathews. 2018. Accelerated RNA secondary structure design using preselected sequences for helices and loops. *RNA (New York, N.Y.)* 24 (Nov. 2018), 1555–1567. Issue 11. https://doi.org/10.1261/rna.066324.118

[4] Édouard Bonnet, Paweł Rzążewski, and Florian Sikora. 2018. Designing RNA Secondary Structures Is Hard. In *Research in Computational Molecular Biology - 22nd Annual International Conference, RECOMB 2018 (Lecture Notes in Computer Science)*, Benjamin J. Raphael (Ed.), Vol. 10812. Springer, Paris, 248–250.

[5] Anke Busch and Rolf Backofen. 2006. INFO-RNA—a fast approach to inverse RNA folding. *Bioinformatics* 22, 15 (2006), 1823–31. https://doi.org/10.1093/bioinformatics/btl194

[6] Alexander Churkin, Matan Drory Retwitzer, Vladimir Reinharz, Yann Ponty, Jérôme Waldispühl, and Danny Barash. 2018. Design of RNAs: comparing programs for inverse RNA folding. *Briefings in bioinformatics* 19 (March 2018), 350–358. Issue 2. https://doi.org/10.1093/bib/bbw120

[7] Gwendal Collet, Julien David, and Alice Jacquot. 2018. Random Sampling of Ordered Trees according to the Number of Occurrences of a Pattern. (2018). Submitted.

[8] Ye Ding, Chi Yu Chan, and Charles E Lawrence. 2005. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA (New York, N.Y.)* 11 (Aug. 2005), 1157–1166. Issue 8. https://doi.org/10.1261/rna.2500605

[9] Robert M. Dirks, Milo Lin, Erik Winfree, and Niles A. Pierce. 2004. Paradigms for computational nucleic acid design. *Nucleic Acids Research* 32, 4 (2004), 1392–1403. https://doi.org/10.1093/nar/gkh291

[10] Jeremy A Draghi, Todd L Parsons, Günter P Wagner, and Joshua B Plotkin. 2010. Mutational robustness can facilitate adaptation. *Nature* 463 (Jan. 2010), 353–355. Issue 7279. https://doi.org/10.1038/nature08694

[11] Michael Drmota. 1997. Systems of functional equations. *Random Structures and Algorithms* 10, 1-2 (1997), 103–124.

[12] Jérémie Du Boisberranger, Danièle Gardy, and Yann Ponty. 2012. The weighted words collector, In International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AOFA 2012), France) Nicolas, Broutin (INRIA and Canada) Luc, Devroye (McGill (Eds.). *Discrete Mathematics & Theoretical Computer Science* AQ, 243–264.

[13] Ali Esmaili-Taheri and Mohammad Ganjtabesh. 2015. ERD: a fast and reliable tool for RNA design including constraints. *BMC bioinformatics* 16 (Jan. 2015), 20. https://doi.org/10.1186/s12859-014-0444-5

[14] Sven Findeiß, Maja Etzel, Sebastian Will, Mario Mörl, and Peter F Stadler. 2017. Design of Artificial Riboswitches as Biosensors. *Sensors (Basel, Switzerland)* 17, 9 (Aug. 2017), E1990. Issue 9. https://doi.org/10.3390/s17091990

[15] Philippe Flajolet and Andrew M. Odlyzko. 1990. Singularity Analysis of Generating Functions. *SIAM J. Discrete Math.* 3, 2 (1990), 216–240. https://doi.org/10.1137/0403019

[16] Philippe Flajolet and Robert Sedgewick. 2009. *Analytic Combinatorics* (1 ed.). Cambridge University Press, New York, NY, USA.

[17] Fontana, Stadler, Bornberg-Bauer, Griesmacher, Hofacker, Tacker, Tarazona, Weinberger, and Schuster. 1993. RNA folding and combinatory landscapes. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics* 47 (March 1993), 2083–2099. Issue 3.

[18] W Fontana and P Schuster. 1987. A computer model of evolutionary optimization. *Biophysical chemistry* 26 (May 1987), 123–147. Issue 2-3.

[19] Juan Antonio Garcia-Martin, Peter Clote, and Ivan Dotu. 2013. RNAiFOLD: a constraint programming algorithm for RNA inverse folding and molecular design. *Journal of Bioinformatics and Computational Biology* 11, 2 (2013), 1350001. https://doi.org/10.1142/S0219720013500017

[20] José Vicente Gomes-Filho and Lennart Randau. 2019. RNA stabilization in hyperthermophilic archaea. *Annals of the New York Academy of Sciences* (April 2019), 14060. https://doi.org/10.1111/nyas.14060

[21] Stephan Grabbe, Heinrich Haas, Mustafa Diken, Lena M Kranz, Peter Langguth, and Ugur Sahin. 2016. Translating nanoparticulate-personalized cancer vaccines into clinical applications: case study with RNA-lipoplexes for the treatment of melanoma. *Nanomedicine (London, England)* 11 (Oct. 2016), 2723–2734. Issue 20. https://doi.org/10.2217/nnm-2016-0275

[22] Jozef Haleš, Alice Héliou, Ján Maňuch, Yann Ponty, and Ladislav Stacho. 2017. Combinatorial RNA Design: Designability and Structure-Approximating Algorithm in Watson-Crick and Nussinov-Jacobson Energy Models. *Algorithmica* 79, 3 (2017), 835–856. https://doi.org/10.1007/s00453-016-0196-x

[23] Stefan Hammer, Christian Günzel, Mario Mörl, and Sven Findeiß. 2019. Evolving methods for rational de novo design of functional RNA molecules. *Methods (San Diego, Calif.)* (May 2019). https://doi.org/10.1016/j.ymeth.2019.04.022

[24] I. L. Hofacker, W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker, and P. Schuster. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly* 125, 2 (1994), 167–188. https://doi.org/10.1007/BF00818163

[25] Ivo L Hofacker, Peter Schuster, and Peter F Stadler. 1998. Combinatorics of RNA secondary structures. *Discrete Applied Mathematics* 88, 1-3 (1998), 207–237.

[26] Tao Jiang, Lusheng Wang, and Kaizhong Zhang. 1995. Alignment of trees—an alternative to tree edit. *Theoretical Computer Science* 143, 1 (1995), 137–148.

[27] Thomas Jörg, Olivier C Martin, and Andreas Wagner. 2008. Neutral network sizes of biological RNA molecules can be computed and are not atypically small. *BMC bioinformatics* 9 (Oct. 2008), 464. https://doi.org/10.1186/1471-2105-9-464

[28] Ioanna Kalvari, Joanna Argasinska, Natalia Quinones-Olvera, Eric P Nawrocki, Elena Rivas, Sean R Eddy, Alex Bateman, Robert D Finn, and Anton I Petrov. 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic acids research* 46 (Jan. 2018), D335–D342. Issue D1. https://doi.org/10.1093/nar/gkx1038

[29] Ryan Kennedy, Manuel E Lladser, Zhiyuan Wu, Chen Zhang, Michael Yarus, Hans De Sterck, and Rob Knight. 2010. Natural and artificial RNAs occupy the same restricted region of sequence space. *RNA (New York, N.Y.)* 16 (Feb. 2010), 280–289. Issue 2. https://doi.org/10.1261/rna.1923210

[30] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. 2011. ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB* 6 (Nov. 2011), 26. https://doi.org/10.1186/1748-7188-6-26

[31] Zhi John Lu, Jason W Gloor, and David H Mathews. 2009. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA (New York, N.Y.)* 15 (Oct. 2009), 1805–1813. Issue 10. https://doi.org/10.1261/rna.1643609

[32] Rune B Lyngsø, James Wj Anderson, Elena Sizikova, Amarendra Badugu, Tomas Hyland, and Jotun Hein. 2012. FRNAkenstein: multiple target inverse RNA folding. *BMC Bioinformatics* 13 (2012), 260. https://doi.org/10.1186/1471-2105-13-260

[33] Chi H Mak and Ethan N H Phan. 2018. Topological Constraints and Their Conformational Entropic Penalties on RNA Folds. *Biophysical journal* 114 (May 2018), 2059–2071. Issue 9. https://doi.org/10.1016/j.bpj.2018.03.035

[34] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology* 288, 5 (1999), 911–940. https://doi.org/10.1006/jmbi.1999.2700

[35] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, Amit Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. 2017. SymPy: symbolic computing in Python. *PeerJ Computer Science* 3 (Jan. 2017), e103. https://doi.org/10.7717/peerj-cs.103

[36] Markus E Nebel and Frank Weinberg. 2012. Algebraic and combinatorial properties of common RNA pseudoknot classes with applications. *Journal of computational biology : a journal of computational molecular cell biology* 19 (Oct. 2012), 1134–1150. Issue 10. https://doi.org/10.1089/cmb.2011.0094

[37] Vladimir Reinharz, Yann Ponty, and Jérôme Waldispühl. 2013. A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotide distribution. *Bioinformatics* 29, 13 (2013), i308–i315. https://doi.org/10.1093/bioinformatics/btt217

[38] Elena Rivas, Jody Clements, and Sean R Eddy. 2017. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature methods* 14 (Jan. 2017), 45–48. Issue 1. https://doi.org/10.1038/nmeth.4066

[39] Bruno Salvy and Paul Zimmermann. 1994. GFUN: A Maple Package for the Manipulation of Generating and Holonomic Functions in One Variable. *ACM Trans. Math. Softw.* 20, 2 (June 1994), 163–177. https://doi.org/10.1145/178365.178368

[40] P Schuster, W Fontana, P F Stadler, and I L Hofacker. 1994. From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings. Biological sciences* 255 (March 1994), 279–284. Issue 1344. https://doi.org/10.1098/rspb.1994.0040

[41] Defne Surujon, Yann Ponty, and Peter Clote. 2019. Small-World Networks and RNA Secondary Structures. *Journal of Computational Biology* 26, 1 (2019), 16–26. https://doi.org/10.1089/cmb.2018.0125

[42] Melissa K. Takahashi and Julius B. Lucks. 2013. A modular strategy for engineering orthogonal chimeric RNA transcription regulators. *Nucleic Acids Research* 41, 15 (2013), 7577–7588. https://doi.org/10.1093/nar/gkt452

[43] Douglas H. Turner and David H. Mathews. 2010. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research* 38, Database issue (2010), D280–D282. https://doi.org/10.1093/nar/gkp892

[44] Michael Waterman. 1978. Secondary Structure of Single-Stranded Nucleic Acids. *Advances in Mathematics: Supplementary Studies* 1 (1978), 167–212.

[45] Zasha Weinberg, Christina E Lünse, Keith A Corbino, Tyler D Ames, James W Nelson, Adam Roth, Kevin R Perkins, Madeline E Sherlock, and Ronald R Breaker. 2017. Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic regions. *Nucleic acids research* 45 (Oct. 2017), 10811–10823. Issue 18. https://doi.org/10.1093/nar/gkx699

[46] Sebastian Will, Christina Otto, Milad Miladi, Mathias Möhl, and Rolf Backofen. 2015. SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics (Oxford, England)* 31 (Aug. 2015), 2489–2496. Issue 15. https://doi.org/10.1093/bioinformatics/btv185

[47] Sherry Y. Wu, Gabriel Lopez-Berestein, George A. Calin, and Anil K. Sood. 2014. RNAi Therapies: Drugging the Undruggable. *Science Translational Medicine* 6, 240 (2014), 240ps7. https://doi.org/10.1126/scitranslmed.3008362

[48] Joseph N Zadeh, Brian R Wolfe, and Niles A Pierce. 2011. Nucleic acid sequence design via efficient ensemble defect optimization. *Journal of Computational Chemistry* 32, 3 (2011), 439–52. https://doi.org/10.1002/jcc.21633

[49] M. Zuker and P. Stiegler. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research* 9 (1981), 133–148.

A PROOF OF PROPOSITION 1

Proposition. *For any defect $\mathcal{D} \in \{\mathcal{D}_S, \mathcal{D}_P, \mathcal{D}_E\}$, sequence w, $|w| = n$, and structure $S \in \mathcal{S}_n$, one has*

$$\mathcal{D}(w, S) \geq \mathcal{D}^L(w_{[|m|]}, m), \forall m \in S$$

where $w_{[|m|]}$ is the restriction of w to the positions in m.

Proof. Let constraint C be the set of all (un)paired positions of $S \setminus m$ plus the paired leaves and the root of m and $\mathcal{S}_C \subset \mathcal{S}$ denotes the set of secondary structures of size n that are compatible with the constraint C. Then, we have the follow inequality,

$$\begin{array}{rcl} \mathbb{P}(S \mid w) & = & \frac{\mathcal{B}(w,S)}{\sum_{S' \in \mathcal{S}_n} \mathcal{B}(w,S')} \\ & \leq & \frac{\mathcal{B}(w,S)}{\sum_{S' \in \mathcal{S}_C} \mathcal{B}(w,S')} \\ & = & \frac{\mathcal{B}(w_{[|m|]},m)}{\sum_{m' \in \mathcal{S}_m} \mathcal{B}(w_{[|m|]},m')} \\ & = & \mathbb{P}(m \mid w_{[|m|]}) \end{array}$$

Therefore, $\mathcal{D}_P(w, S) \geq \mathcal{D}_P^L(w_{[|m|]}, m)$.

Similarly,

$$\begin{array}{rcl} \sum_{S' \in \mathcal{S}_n} \mathbb{P}(S' \mid w) \cdot |S' \triangle S| & \geq & \sum_{S' \in \mathcal{S}_C} \mathbb{P}(S' \mid w) \cdot |S' \triangle S| \\ & = & \sum_{m' \in \mathcal{S}_m} \mathbb{P}(m' \mid w_{[|m|]}) \cdot |m' \triangle m| \end{array}$$

Thus, the inequality for $\mathcal{D} = \mathcal{D}_E$

For the case $\mathcal{D} = \mathcal{D}_S$, we consider m' such that,

$$m' := \underset{\substack{x \in \mathcal{S}_m \\ x \neq m}}{\operatorname{argmin}} E(w_{[|m|]}, x) - E(w_{[|m|]}, m)$$

A such m' exists since the set $\mathcal{S}_m$ is finite. Let $S' \in \mathcal{S}_C$ be the secondary structure containing m' at the position of m. We have,

$$E(w_{[|m|]}, m') - E(w_{[|m|]}, m) = E(w, S') - E(w, S)$$

In addition, we have, by definition, $\mathcal{D}_S(w, S) \geq e^{-(E(w,S') - E(w,S))}$, which implies the inequality $\mathcal{D}_S(w, S) \geq \mathcal{D}_S^L(w_{[|m|]}, m)$ □

B PROOF OF PROPOSITION 2

Proposition. *For every defect $\mathcal{D}$ and energy model considered in this work, one has $\mathcal{D}^L(w, m) = \mathcal{D}_{C_m}(w_m, S_m)$.*

Proof. The schema of proof is similar to the above one. Let $\mathcal{S}_{C_m}$ be the set of secondary structures of length $|S_m|$ that are compatible with the folding constraint C_m. One can observe that the way to make the completion structure is a bijection from $\mathcal{S}_m$ to $\mathcal{S}_{C_m}$. Let m' be a motif equivalent to m and $S_{m'} \in \mathcal{S}_{C_m}$ be its minimum completion structure. The energy of the structure S_m (resp. $S_{m'}$) is the sum of the energy contribution with the motif m (resp. m') and with the constrained part, which is the same for both structures. The later one is a constant for a given sequence. For the reason of simplicity, we denote it by E_C. Then, for a given sequence w and its minimum completion w_m, we have

$$E(w_m, S_{m'}) - E(w_m, S_m) = E(w, m') - E(w, m)$$

This proves the proposition for the case where $\mathcal{D} = \mathcal{D}_S$

For the case of the Probability Defect $\mathcal{D}_P$,

$$\begin{array}{rcl} \mathbb{P}(S_m \mid w_m) & = & \frac{e^{-E(w_m,S_m)/RT}}{\sum_{S_{m'} \in \mathcal{S}_C} e^{-E(w_m,S_{m'})/RT}} \\ & = & \frac{e^{-E(w,m)/RT} e^{-E_C/RT}}{\sum_{m' \in \mathcal{S}_m} e^{-E(w,m')/RT} e^{-E_C/RT}} \\ & = & \frac{e^{-E(w,m)/RT}}{\sum_{m' \in \mathcal{S}_m} e^{-E(w,m')/RT}} \\ & = & \mathbb{P}(m \mid w) \end{array}$$

Thus, the equality.

Furthermore, for any motif $m' \in \mathcal{S}_m$ and its minimal completion $S_{m'}$, the base pair distance are equal between motifs m' and m and between their minimal completion structures, $|m' \triangle m| = |S_{m'} \triangle S_m|$, because the completion part is same for both motifs. Therefore, the equality for the case of the ensemble defect $\mathcal{D} = \mathcal{D}_E$ □