RNA Triplet Repeats: Improved Algorithms for Structure Prediction and Interactions

Kimon Boehmer, Sarah J. Berkemer, Sebastian Will, Yann Ponty

RNA Triplet Repeats: Improved Algorithms for Structure Prediction and Interactions

Kimon Boehmer[1], Sarah J. Berkemer[1,2], Sebastian Will[1], Yann Ponty[1*]

[1*]Laboratoire d'Informatique de l'Ecole Polytechnique (LIX CNRS UMR 7161), France, Institut Polytechnique de Paris, 1 Rue Honoré d'Estienne d'Orves, Palaiseau, 91120, France.
[2]Earth-Life Science Institute, Institute of Science Tokyo, 2-12-1-I7E-318 Ookayama, Tokyo, 152-8550, Japan.

*Corresponding author(s). E-mail(s): yann.ponty@lix.polytechnique.fr; Contributing authors: kimon.bohmer@ens-paris-saclay.fr; berkemer@lix.polytechnique.fr; sebastian.will@polytechnique.edu;

Abstract

RNAs composed of Triplet Repeats (TR) have recently attracted much attention in the field of synthetic biology. We study the mimimum free energy (MFE) secondary structures of such RNAs and give improved algorithms to compute the MFE and the partition function. Furthermore, we study the interaction of multiple RNAs and design a new algorithm for computing MFE and partition function for RNA-RNA interactions, improving the previously known factorial running time to exponential. In the case of TR, we show computational hardness but still obtain a parameterized algorithm. Finally, we propose a polynomial-time algorithm for computing interactions from a base set of RNA strands and conduct experiments on the interaction of TR based on this algorithm. For instance, we study the probability that a base pair is formed between two strands with the same triplet pattern, allowing an assessment of a notion of orthogonality between TR.

Keywords: RNA folding, RNA interactions, triplet repeats, dynamic programming, NP-hardness

1 Introduction

RNAs composed of Triplet Repeats (TR) have attracted much attention, and harbour promises in the field of synthetic biology, due to their demonstrated capacity to self-assemble into droplets [1, 2]. Those can in turn be used to compartmentalize cellular processes, thereby creating a “clean room”, free of the natural cellular clutter, where synthetic circuits can be executed without interference. The exact process underlying this phenomena is still the object of ongoing investigations, but it is hypothesized that repetitive RNAs may induce Liquid-Liquid Phase separation mediated by unstable/transient structures. Repetitive RNAs are also found at the origin of severe Neurological Triplet Expansion Diseases (TED), including Friedreich attaxia [3] and Triplet Repeat Diseases (TRD) such as Huntington disease [4]. For multiple TEDs and TRDs, overly expanded RNAs have been observed to aggregate into RNA foci, leading to a sequestration of RNA binding proteins. Local secondary structures and interactions are impacted by the repeat, and generally believed to contribute to the pathogenicity and treatment efficiency. To study those phenomena *in silico*, and in particular the impact of the repeated motif and number of repeats on aggregates, one needs to predict the MFE structure of potentially large RNAs, and many-body interactions. Recently, coarse-grained simulations showed a disparity between odd or even numbers of triplet repeats [5] as well as extensions to quadruplet and non-redundant tandem repeats [6].

RNA folding by energy minimization is a classic algorithmic problem in Bioinformatics, historically solved in time $\Omega(n^3)$ using dynamic programming [7, 8]. Despite recent suggestions for heuristics [9], the best algorithm to date to solve energy minimization has runtime $\mathcal{O}(n^{2.8603})$ [10], and both its implementation and extension beyond a base-pair maximization setting represent considerable challenges. Prior works have also investigated conditional lower bounds, and found that the existence of a $\mathcal{O}(n^{2-\varepsilon})$ algorithm would refute the Strong Exponential Time Hypothesis (SETH) [10]. Meanwhile, an $\mathcal{O}(n^{\omega-\varepsilon})$ algorithm would disprove the k-clique conjecture, with $\omega < 2.373$ being the matrix multiplication exponent [10, 11].

RNA-RNA interaction prediction represents an equally relevant, yet computationally substantially more involved algorithmic problem. For a fixed number of interacting strands, polynomial-time algorithms have been proposed. For example, by excluding so-called zig-zag joint conformations, Alkan et al. [12] proposed a polynomial-time algorithm for the interaction of two strands, while also showing **NP**-hardness for the case where we include these conformations. In the unbounded case, Dirks et al. [13] gave a factorial-time algorithm for computing the partition function (PF) over multiple strands. Additionally, it was shown that energy minimization in this setting is **APX**-hard (and by that **NP**-hard) [14], even for a very simple energy model. The problem is very much the object of ongoing investigations at different level of granularities, with striking recent results [15].

In this work, we show that the repeated nature of RNA can be exploited to obtain substantially improved algorithms for several problems. First, we show that the MFE of a triplet-repeat RNA can be predicted in linear time, both with respect to base pair maximization and Turner energy model, and is realized by either the open chain or a single helix. By a change of algebra, the DP scheme can be used to calculate the partition function. We then consider the interaction of multiple triplet repeats and propose improved algorithms for the general (non-triplet) case as well as algorithms specifically for the interaction of TR. For the latter case, we show **NP**-hardness in a reasonable energy model. We then propose a polynomial-time algorithm for the setting where we are given a "soup" of strands instead of a fixed set, and, using this algorithm, conduct experiments on the probability that a base pair is folding, interacting with another identical sequence or interacting with a different sequence.

2 Definitions and Problem Statement

2.1 Definitions

RNA sequence and folding. An *RNA sequence* (or just *sequence*) is a word $s \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{U}\}^+$. The length of s is denoted by $|s|$ and the i-th position of s by s_i. A position on a sequence is also called a *base*. We associate to each base s_i its letter by $l(s_i)$. We define $P := \{\{\mathtt{C},\mathtt{G}\},\{\mathtt{A},\mathtt{U}\},\{\mathtt{G},\mathtt{U}\}\}$ as the set of *admissible base pairs*. A *(pseudoknot-free) secondary structure* S is a set of unordered pairs of bases, hereunder called *base pairs*, such that:

- each base pair is a Watson-Crick or Wobble pair, i.e. for all $\{s_i, s_j\} \in S$, $\{l(s_i), l(s_j)\} \in P$;
- each base is involved in at most one base pair, i.e. for all bases s_i, $|\{p \in S \mid s_i \in p\}| \leq 1$;
- S is *pseudoknot-free*, i.e. there are no $\{s_i, s_j\}, \{s_k, s_\ell\} \in S$ with $i < k < j < \ell$;
- each base pair encloses at least θ bases, i.e. if $\{s_i, s_j\} \in S$, then $j - i > \theta$. The *minimal base pair span* is usually denoted by θ, and we use $\theta := 3$ unless explicitly specified.

We denote by $\Omega(s)$, or in short Ω whenever clear from the context, the set of all pseudoknot-free secondary structures over sequence s.

We associate each secondary structure $S \in \Omega$ to a *free energy*, according to an *energy model* $E : \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{U}\}^+ \times \Omega \to \mathbb{R}$. For example, in the *base pair model* E_{bp}, we simply count the number of base pairs in S, hence set $E_{\text{bp}}(s, S) = -|S|$. More advanced energy models reason about the free energy introduced by motifs occurring in the secondary structure, such as the loops considered by the Turner nearest-neighbor model [16].

Interactions. A strand is an RNA sequence which is identified as a unique object in a set. In other words, in a set of strands R, we can have two strands $s \neq r$ that consist of the same sequences, that is $l(s_i) = l(r_i)$ for all $i \in \{1, ..., |s| = |r|\}$, but still are different objects. To describe the interaction of multiple strands, we are given a set R of strands, where $m := |R|$.

A *circular permutation* $\pi : R \to \{0, ..., m-1\}$ of a strand set R is a permutation of all elements in R except for one fixed strand s^*, which is fixed to position 0. Then, the bases are naturally ordered by $s_i <_\pi r_j \equiv s < r \vee (s = r \wedge i < j)$. We define O_π as the set of all tuples of bases $(s^1_{i_1}, ..., s^k_{i_k})$ such that there is a j with $s^j_{i_j} <_\pi s^{j+1}_{i_{j+1}} <_\pi ... <_\pi s^k_{i_k} <_\pi s^1_{i_1} <_\pi ... <_\pi s^{j-1}_{i_{j-1}}$.

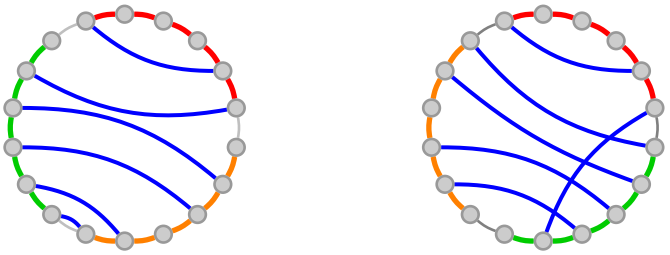


Fig. 1 The same secondary structure on a strand set with three strands drawn in two different circular permutations. The strands are depicted by the green, red and orange lines while blue lines indicate base pairs. Gray lines connect subsequent strands and depend on the strand permutation.

A *secondary structure* S of a strand set R is a set of base pairs $\{s_i, r_j\}$ from strands in $s, r \in R$ such that $\{l(s_i), l(r_j)\} \in P$, each base appears in at most one base pair and each intra-strand base pair encloses at least θ bases, i.e. $\{s_i, s_j\} \in S \to j - i > \theta$.

The *polymer graph* of a secondary structure S and a circular permutation π on R is a graph $G = (V, E)$ with $V := \{s_i \mid s \in R, 1 \leq i \leq |s|\}$ and $E := S \cup \{\{s_i, s_{i+1}\} \mid s \in R, 1 \leq i < |s|\} \cup C := \{\{s_{|s|}, r_1\} \mid (\pi(s) + 1) \bmod |R| = \pi(r)\}$. The edges $E - S$ are drawn in a cycle (naturally induced by the circular permutation), while the edges in S are drawn as straight lines between the bases. Examples for the polymer graphs of a single secondary structure under two different circular permutations can be found in fig. 1.

Two strands s, r are *connected* if there is a path from s_1 to r_1 that does not use edges from C. A secondary structure is connected if all of its strands are connected. Note that connectedness is independent of the circular permutation π.

A secondary structure S of a strand set R is called *pseudoknot-free* if there is a circular permutation π such that there are no crossing lines in the polymer graph, or formally, there are no two base pairs $\{s_i, t_k\}, \{u_\ell, r_j\} \in S$ with $(s_i, u_\ell, t_k, r_j) \in O_\pi$. The set of all pseudoknot-free secondary structures over a strand set R is denoted by $\Omega(R)$.

As for the folding, we associate to each $S \in \Omega(R)$ a free energy $E : 2^{\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{U}\}^*} \times \Omega \to \mathbb{R}$. In the base pair model, apart from the number of base pairs p of base pairs, we also add a strand association penalty K_{assoc} for each of the $(m - \ell)$ strand associations, where ℓ is the number of connected components (also called *complexes*) of S. Thus, the free energy of $S \in \Omega$ in this model is defined as $E(R, S) = -p + (m - \ell)K_{\text{assoc}}$.

2.2 Computational problems

For a single strand, two of the most classical problems in RNA bioinformatics are:

Minimum Free Energy (MFE) under Energy model E
Input: RNA sequence s
Output: Minimum free-energy $\min_{S \in \Omega(s)} E(s, S)$

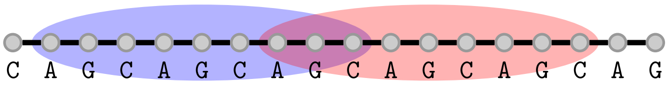


Fig. 2 The periodicity of triplet repeats RNAs (TR) induces multiple symmetries. leading to drastic simplifications of certain algorithmic problems.Here, the blue and red regions of the TR sequence are identical.

PARTITION FUNCTION UNDER ENERGY MODEL E
Input: RNA sequence s + positive temperature T in Kelvin (K)
Output: Partition function $\mathcal{Z}_s := \sum_{S\in\Omega(s)} \exp\left\{\frac{-E(s,S)}{kT}\right\}$

where $k = 1.987 \cdot 10^{-3}$kcal.mol^{-1}.K^{-1} is the Boltzmann constant.

In the multi-strand setting, we focus on energy minimization. In Dirks et al. [13], the authors adopt a thermodynamic perspective on the free energy of a secondary structure over multiple strands, such that potential rotational symmetries require an adjustment of the computed value. For the MFE, we focus on a more algorithmic perspective, where all rotationally symmetric structures are elements of a search space, and a simple base pair energy model. In our main algorithmic problem of interest, we are given a set of strands and are looking for the minimum free energy of the secondary structure over these strands:

MFE STRAND INTERACTION
Input: Set of strands R_0
Output: $\min_{S\in\Omega(R_0)} E(R_0, S)$

In a more applied setting in the field of synthetic biology, we consider a set of RNA strands composed out of triplet repeats to be present such that a sufficient subset thereof will self-assemble into droplets. Hence, we assume having a 'soup' of RNA strands in question as a basis for the formation of RNA droplets that will serve as point of departure.

Therefore, we consider here a slightly different setting, where the number of occurrences of each triplet/strand is unconstrained beyond the total number m of interacting strands. This allows to study situations where the strands concentrations are in excess, so that sequences can be locally seen as infinitely available often within a set (or "soup") R of strands. We then look for the best structure over m strands that all appear in R. Each sequence in the soup can appear zero or multiple times in a secondary structure. More formally:

MFE STRAND SOUP INTERACTION
Input: Set of *sequences* $R = \{r_1, ..., r_p\}$, $m \in \mathbb{N}$ encoded in unary
Output: $\min_{t_1\in R,\ldots,t_m\in R} \min_{S\in\Omega(\{t_1,\ldots,t_m\})} E(\{t_1, ..., t_m\}, S)$

2.3 Triplet repeats RNAs and their properties

Triplet repeat RNAs (TR). Of special interest to us are RNA sequences that are composed of *triplet repeats* (TR), that is, they have the form $(X \cdot Y \cdot Z)^k$ for

$X, Y, Z \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$ and $k \in \mathbb{N}^+$. We will describe how we can improve the general algorithms for the above computational problems in the case of TR.

An algorithmically consequential property of any region $[s_i, s_j]$ in a TR sequence is the following.

Observation 1. *For a triplet repeat sequence s and $1 \leq i \leq j \leq |s|$, one has*

$$[s_i, s_j] = [s_{i \bmod 3}, s_{j-(i-i \bmod 3)}].$$

In other words, we can shift any region three positions to the left or right, and in particular we can shift it to the beginning of the sequence, as visualized in fig. 2. That way, the index that usually denotes the beginning of the considered sequence in a dynamic programming (DP) algorithm can be restricted to values 1, 2 and 3. Hence, the length of the value range is constant and not linear anymore, which gives an easy linear improvement of running time and storage for MFE as well as PF computation.

We also note that TR sequences can be encoded exponentially more compact than general sequences. Each TR sequence is uniquely identified by its pattern $XYZ \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}^3$ and its number of repeats k. In other words, $6 + \lceil \log_2 k \rceil$ bits are enough to encode a TR sequence with k repeats. We will refer to this encoding as the *compact encoding*, while the *explicit encoding* consists of the complete sequence $s \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}^{3k}$. The latter can also be seen, equivalently in terms of asymptotic complexity, as a compact encoding where k is encoded in unary.

Looking into more structural properties of triplet repeats, we can observe that, since each base repeats after two other bases, there cannot be a base pair that encloses exactly 2 bases. Thus, requiring two ($\theta = 2$) or three ($\theta = 3$) enclosed bases between any base pair is equivalent:

Observation 2. *A secondary structure S for $(XYZ)^k$ fulfills minimum base pair span θ with $\theta \equiv_3 2$ if and only if it fulfills minimum base pair span $\theta + 1$.*

Finally, if we consider the graph $G = (\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}, P)$, where P is the set of allowed base pairs, we can see that it does not contain any triangles. From this we can observe:

Observation 3. *For any triplet sequence $(XYZ)^k$, there is a letter $V \in \{X, Y, Z\}$, that we call the* covering letter*, that is contained in all base pairs, i.e. $V \in p$ for all $p \in S$ and $S \in \Omega$.*

3 Single-Stranded Triplet Repeats

Our goal is to specify the exact MFE, and the corresponding secondary structure, when given a triplet pattern XYZ and length k of our TR sequence s, as well as the minimum base pair span θ. This will give us a very efficient way of computing the MFE in this simple setting.

3.1 Linear time solution for base pair maximization

We first consider the properties of the MFE structure for TR RNAs in a *base pair maximization model*, where the free energy E_{bp} of a secondary structure $S \in \Omega$ is such that $E_{\text{bp}}(s, S) = -|S|$.

We can first prove an upper bound on the number of base pairs in a TR sequence:

Lemma 1. *Consider a TR sequence $s := (XYZ)^k$ and a minimum number of enclosed bases $\theta \geq 0$, such that $\lfloor \frac{\theta+1}{3} \rfloor \leq k$. We have $E_{bp}(s, S) \leq k - \lfloor \frac{\theta+1}{3} \rfloor$ for any $S \in \Omega(s)$.*

Proof. Without loss of generality, let Z be the covering letter of s. Any non-empty secondary structure has an innermost base pair which must respect the minimum base pair span θ. For $\theta = 2$, which is equivalent to $\theta = 3$ by observation 2, as well as for $\theta = 4$, at least one Z base must remain unpaired, and increasing θ by 3 will result into one new unpairable Z base. Thus we know that at least $\lfloor \frac{\theta+1}{3} \rfloor$ Z bases will remain unpaired and at most $k - \lfloor \frac{\theta+1}{3} \rfloor$ Z-bases are pairable. Since every base pair must involve a Z base, we can conclude. □

We now show that this upper bound is almost always tight. To this end, first notice that for all triplet patterns XYZ such that $\{\{X, Y\}, \{X, Z\}, \{Y, Z\}\} \cap P = \varnothing$, no base pair can be built and thus the maximum value is trivially 0. We call TR sequences of such patterns *non-folding*, and all other TR sequences *folding*.

Lemma 2. *For $\theta \in \{0, 1\}$ and $k > 1$, we always have $E(s, S) = k$ for any secondary structure S over a folding sequence $s = (XYZ)^k$.*

Proof. If $\{X, Z\} \in P$, connect X and Z in each triplet. Else, connect the outermost pair (say without loss of generality $\{X, Y\}$). We obtain the inner sequence $(YZX)^{k-1}$ (with $k - 1 > 0$) and we can proceed as above since $\{Y, X\} \in P$. □

For the more natural case $\theta > 1$, the upper bound from lemma 1 is not always tight. The next lemma exactly specifies the MFE and its structure:

Lemma 3. *Let $\theta > 1$. The minimum MFE structure of a folding sequence $(XYZ)^k$ has value*

- $k - 1 - \frac{\theta-1}{3}$, *if* $(\{X, Z\} \notin P \wedge (\theta + 3k) \equiv_6 4) \vee (\{X, Y\}, \{Y, Z\} \notin P \wedge (\theta + 3k) \equiv_6 1)$
- $k - \lfloor \frac{\theta+1}{3} \rfloor$, *otherwise*

Furthermore, a minimum MFE structure is obtained by a single helix of base pairs of one letter pair p. If both $\{X, Z\} \in P$ and one of $\{X, Y\}$ and $\{Y, Z\} \in P$, we set $p := \{X, Z\}$ if $(\theta + 3k) \equiv_6 4$ and $p := \{X, Y\}$ (or $p := \{Y, Z\}$) if $(\theta + 3k) \equiv_6 1$; otherwise, we set p to the letters of an arbitrary pairable base pair.

Proof. We start by showing that the corresponding secondary structures achieve the claimed score. By observation 2, we only need to consider $\theta \equiv_3 0$ and $\theta \equiv_3 1$.

First assume $\{X, Z\} \in P$ and $\{X, Y\}, \{Y, Z\} \notin P$. We will derive the other cases from this one. Consider a large stacking of $X - Z$ bases. If $\theta = 3$, we only cannot match the $X - Z$ pair of the innermost repeat in the case $k \equiv_2 1$ and we only cannot match the $Z - X$ pair between the two innermost repeats in the case $k \equiv_2 0$. For all other pairs of repeats we obtain exactly two base pairs and hence we get $k - 1 = k - \lfloor \frac{\theta+1}{3} \rfloor$ base pairs. Inductively, let us show that we can obtain $k - \lfloor \frac{\theta'+1}{3} \rfloor$ base pairs for $\theta' := \theta + 3$. In other words, we only need to show that by increasing θ by 3, we get one base pair less. If the innermost base pair is $X - Z$, its enclosed region starts and ends with a Y and there are currently at least $\theta + 1$ free enclosed bases (because the region is of the form $Y(ZXY)^{\theta/3}$), and by deleting the $X-Z$ base pair, we obtain $XY(ZXY)^{\theta/3}Z$, that is $\theta+3$ enclosed bases. Else, for a $Z-X$ base pair, the region has the form $(XYZ)^{\theta/3}$. After deleting the innermost base pair $Z-X$, the new enclosed region starts and ends with a Y (the region is of the form $YZ(XYZ)^{\theta/3}XY$), so there are at least $\theta+4$ enclosed bases. Thus we can achieve $k-\lfloor\frac{\theta+1}{3}\rfloor$ base pairs.

If $\theta \equiv_3 1$, we distinguish two equivalence classes: In the first, k is even and $\theta \equiv_6 1$ *or* k is uneven and $\theta \equiv_6 4$, and in the second equivalence class, we have the other two cases.

For $\theta = 4$, for $k \equiv_2 1$, our lemma only claims $k-2$ base pairs. We can indeed leave the innermost repeat as well as the next $Z-X$ pair unpaired, and greedily create stackings outside of this region, obtaining $k-2$ base pairs. For $k \equiv_2 0$, We can proceed as for the even case in $\theta = 3$.

Consider $\theta+3$ now. We add an unpaired triplet in the middle of the sequence. Now, the number of base pairs is equal to the case $k-1$ (of opposite parity) with θ enclosed bases.

We thus established the lower bound for the $\{X,Z\} \in P$ case. For the "otherwise"-case, lemma 1 already gives us the required upper bound. Therefore, we only need to argue about the upper bound $k-1-\frac{\theta-1}{3}$ in the case that $\{X,Y\},\{Y,Z\} \notin P$ and $(\theta+3k) \equiv_6 1$. Assume a secondary structure that achieves more base pairs. Firstly, we cannot have any multiloops or exterior loops since that would imply two regions of unpaired enclosed bases, which then only allows $k-2\lfloor\frac{\theta+1}{3}\rfloor \leq k-1-\frac{\theta-1}{3}$ base pairs. Additionally, for each secondary structure S with $i < j'$ and $k > 0$ such that $\{i,j'\} \in S$ and the interval $[j'+1, j'+3k]$ only consists of unpaired bases, we can delete the base pair $\{i,j'\}$ and instead add base pair $\{i,j'+3k\}$ without reducing the number of base pairs. In other words, for any interval, it is always better to pair the leftmost base to the rightmost possible base than to any other interior base. We thus only need to consider the canonical structures of $X-Z/Z-X$-stackings.

Consider an odd k with all base pairs in the canonical way (for $\theta = 4$). The innermost triplet repeat bases X and Z have to stay unpaired, as well as the Z and X which are adjacent to that repeat. The innermost base pair $X-Z$ now has $7 = \theta+3$ enclosed bases. We thus have $k-2$ base pairs. Inductively, for $\theta' := \theta+6$, the next two innermost base pairs will have $\theta+3 < \theta'$ and $\theta+3+2 < \theta'$ enclosed bases, thus are both not available.

Consider an even k with all base pairs in the canonical way (for $\theta = 7$). The two innermost triplet repeats have to stay unpaired, as well as the Z and X which are adjacent to that repeat. The innermost base pair $X-Z$ now has $10 = \theta+3$ enclosed bases. The rest of the argument is exactly as above.

If $\{X,Z\} \notin P$, we can assume without loss of generality that $\{X,Y\} \in P$ (the arguments are symmetrical for $\{Y,Z\} \in P$, and we assumed to have a folding strand). We can reduce any such instance $(XYZ)^k$ to $(YZX)^{k-1}$ (by letting out the leftmost X and the rightmost Y and Z, and implicitly pairing these outermost X and Y, which is always optimal). Thus, all results can be directly obtained from the case $\{X,Z\} \in P$, by changing odd and even. The upper bound can also be derived by that. □

Setting $\theta = 3$, we get the following corollary:

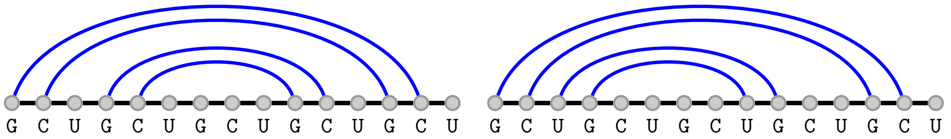


Fig. 3 Two distinct optimal secondary structures for $\texttt{GCU}_5$, for $\theta = 3$.

Corollary 1. *In the base pair maximization model, if $\theta = 3$, the MFE structure of any TR sequence $(XYZ)^k$ has $k-1$ base pairs.*

Determining the MFE is thus a simple calculation taking logarithmic time in the (explicit) size of the triplet repeat sequence. From this we can derive:

Theorem 1. *MFE prediction for compactly encoded TR in the base pair maximization model can be solved in linear time.*

The structure itself can then be computed by the following algorithm:

The algorithm clearly runs in linear time, when the input is an explicit (unary) encoding of the triplet repeats. Its correctness directly follows from lemma 3.

Remark 1. *The optimal secondary structure does not need to be unique. In particular, for a simple energy model, the number of optimal secondary structures for triplet repeats can even be exponential. For example, consider the sequence $(\texttt{GCU})^k$ as illustrated in fig. 3. When constructing the base pairs from outside to inside, in every step, we can choose between adding the base pairs* G-U*,* U-G*, or the base pairs* G-C*,* C-G*. This decision can be repeated $\lfloor \frac{k}{2} \rfloor - 1$ times (assuming $\theta = 3$), giving $\Omega(2^{k/2})$ different optimal secondary structures.*

3.2 Minimum Free-energy in the Turner model

For the Turner model, we will argue that the optimal structures obtained for BP maximization remains optimal for the Turner nearest neighbor model under reasonable assumptions, satisfied by current versions of the model [16]. We first show a helpful lemma:

Lemma 4. *Assume that a TR region of s where the covering letter appears k times has B branches. Then the number of base pairs is at most $k - B$.*

Proof. Let V be the covering letter of s. By observation 3, for each base pair $\{s_i, s_j\}$, either $l(s_i) = V$ or $l(s_j) = V$. Furthermore, each of the B branches contains one unpairable V-base (since $\theta = 3$). Thus, there are only $k - B$ pairable V-bases and we conclude. □

We show the absence of multiloops, *i.e.* structural motifs consisting of $B \geq 2$ branches, in the Turner MFE, with some simplifications. Their free energy contribution is composed of an initiation penalty α, a value β for each branch, and an asymmetry penalty γ. The overall contribution of a multiloop S is given by $E(s, S) = \alpha + \beta B + \gamma C + E_{\text{in}}$, where E_{in} is the MFE of the interior secondary structure of the branches. We will assume $N := \min_{V,W \in \{X,Y,Z\}:\{V,W\}\in P} E_{V,W}$ to be the best contribution of a

```
Input: String $r := (XYZ)^k$, minimum number of enclosed bases $\theta$
Output: Secondary structure $S \in \Omega$ with $E(S) \leq E(S')$ for all $S' \in \Omega$
$i \leftarrow 1, j \leftarrow 3k, S \leftarrow \varnothing$;
if No possible base pair patterns then
    return $S$
end
$p \leftarrow$ arbitrary $\{a, b\} \in P$ such that $a, b \in \{X, Y, Z\}$;
if $\{X, Z\} \in P$ and $k \mod 2 = 1$ then
    $p \leftarrow \{X, Z\}$;
end
if $\{Y, Z\} \in P$ and $k \mod 2 = 0$ then
    $p \leftarrow \{Y, Z\}$;
end
if $\{X, Y\} \in P$ and $k \mod 2 = 0$ then
    $p \leftarrow \{X, Y\}$;
end
while $j - i > \theta$ do
    if $\{r[i], r[j]\} = p$ then
        $S \leftarrow S \cup \{\{i, j\}\}$;
        $i \leftarrow i + 1$;
        $j \leftarrow j - 1$;
    end
    if $r[i] \notin p$ then
        $i \leftarrow i + 1$;
    end
    if $r[j] \notin p$ then
        $j \leftarrow j - 1$;
    end
end
return $S$
```

Algorithm 1: Computing the MFE of Triplet Repeat RNAs

single base pair appearing in a stacking in our triplet pattern, and we will not consider dangling ends etc.

Lemma 5. *Any Turner-MFE secondary structure S^* over a TR sequence does not contain any multiloop, assuming $\beta \geq N, \alpha > -\beta, \gamma \geq 0$.*

Proof. Let S be a multiloop structure on region s with k appearances of the covering letter and let S^* be a stacking on the same region. Their free energy values are related as follows:

$$\begin{aligned} E(s, S) &\geq \alpha + \beta B + \gamma C + (k - B)N && (1) \\ &> -\beta + \beta B + (k - B)N && (2) \\ &= (k - 1)N + (\beta - N)(B - 1) && (3) \\ &\geq (k - 1)N && (4) \end{aligned}$$

$$\geq E(s, S^*) \tag{5}$$

where (1) comes from our above observation and lemma 4, (2) from $\alpha > -\beta$ and $\gamma \geq 0$, (4) from $\beta \geq N$ and $B \geq 2$ (by definition of a multiloop). For inequality (5), first notice that S^* contains $k-1$ base pairs by corollary 1. As noticed in remark 1, we can choose which base pair is used in S^* without affecting the optimality. In particular, we can always choose the base pair consisting of the letters V, W that optimize their contribution, such that $E_{V,W} = N$. We get $E(s, S^*) \leq (k-1)N$. □

Remark 2. *The above assumptions are satisfied by the Turner 2004 energy model* ($\alpha = 9.25$*,* $\beta = -0.63$*,* $\gamma = 0.91$ *and* $N \leq -0.93$)*, as seen in the NNDB [16].*

lemma 4 also excludes secondary structures with multiple exterior faces. Thus, by the above two lemmata, we can conclude that the MFE in the Turner model is also of the canonical form described in the BP maximization setting.

3.3 Linear-time computation of the partition function

In the context of computing the partition function, one can write a weighted context-free grammar which, for any given pattern XYZ, simultaneously generates all TR sequences along with their associated set of secondary structures Ω.

As an example, a context-free grammar associated with the pattern `CAG` would be:

$$\begin{array}{lllll}
S^G_C \to & (\ {\cdot}_A\ S^C_G\ {\cdot}_A\ ) & \mid (\ {\cdot}_A\ S^C_G\ {\cdot}_A\ )\ S^G_C & \mid {\cdot}_C\ \ {\cdot}_A\ S^G_G & \mid {\cdot}_C\ \ {\cdot}_A\ \ {\cdot}_G \\
S^C_G \to & (\ S^G_C\ ) & \mid (\ S^G_C\ )\ {\cdot}_A\ S^C_G & \mid {\cdot}_G\ S^C_C & \mid {\cdot}_G\ \ {\cdot}_C \\
S^G_G \to & (\ S^G_C\ )\ {\cdot}_A\ \ {\cdot}_G & \mid (\ S^G_C\ )\ {\cdot}_A\ S^G_G & \mid {\cdot}_G\ S^G_C & \\
S^C_C \to & (\ {\cdot}_A\ S^C_G\ {\cdot}_A\ )\ {\cdot}_A & \mid (\ {\cdot}_A\ S^C_G\ {\cdot}_A\ )\ S^C_C & \mid {\cdot}_C\ \ {\cdot}_A\ S^C_G &
\end{array}$$

Here, the terminal S^G_C generates all secondary structures for the RNA sequence $(CAG)^k$ for all $k > 0$; S^C_G the structures of $(GCA)^k GC$ for $k \geq 0$; S^G_G the structure of $G(CAG)^k$ for $k > 0$; and S^C_C corresponds to the pattern $(CAG)^k C$ for some $k > 0$.

Following standard methodologies in enumerative/analytic combinatorics [17], such a grammar can be generically translated into a system of functional equations involving weighted generated functions for each non-terminal:

$$\begin{aligned}
S^G_C(z) &= \beta\, z^4\, S^C_G(z) + \beta\, z^4\, S^C_G(z)\, S^G_C(z) + z^2\, S^G_G(z) + z^3 \\
S^C_G(z) &= \beta\, z^2\, S^G_C(z) + \beta\, z^3\, S^G_C(z)\, S^C_G(z) + z\, S^C_C(z) + z^2 \\
S^G_G(z) &= \beta\, z^4\, S^G_C(z) + \beta\, z^3\, S^G_C(z)\, S^G_G(z) + z\, S^G_C(z) \\
S^C_C(z) &= \beta\, z^3\, S^C_G(z) + \beta\, z^2\, S^C_G(z)\, S^C_C(z) + z^2\, S^C_G(z)
\end{aligned}$$

where $\beta := e^{1/kT}$ is the Boltzmann weight associated to base pairs and, in particular:

$$S^G_C(z) = \sum_{s \in \mathcal{L}(S^G_C)} \beta^{\#\mathrm{BP}(s)}\, z^{|s|} = \sum_{k \geq 0} \sum_{\substack{s \in \mathcal{L}(S^G_C) \\ \text{such that } |s| = 3\,k}} e^{\frac{\#\mathrm{BP}(s)}{kT}}\, z^{3k} = \sum_{k \geq 0} \mathcal{Z}_{(CAG)^k}\, z^{3k}$$

The partition function of $\mathcal{Z}_{(CAG)^k}$ can then be obtained as $[z^{3k}]\, S_C^G(z)$, the coefficient of degree $3k$ in $S_C^G(z)$. Since the system of functional equations is algebraic, the coefficients of each generating function obey a linear recurrence with polynomial coefficients [18], which can be efficiently [19] and effectively computed [20]. We obtain an equation of the form:

$$\mathcal{Z}_{(CAG)^k} = P_1(k)\, \mathcal{Z}_{(CAG)^{k-1}} + P_2(k)\, \mathcal{Z}_{(CAG)^{k-2}} + \cdots + P_d(k)\, \mathcal{Z}_{(CAG)^{k-d}}$$

where each P_i is a polynomial in k, and d is a constant . $\mathcal{Z}_{(CAG)^k}$ can then be computed using a linear number of arithmetic operations. This property holds for other triplets and thus:

Theorem 2. *The partition function of a TR can be computed in $\Theta(k)$ arithmetic operations.*

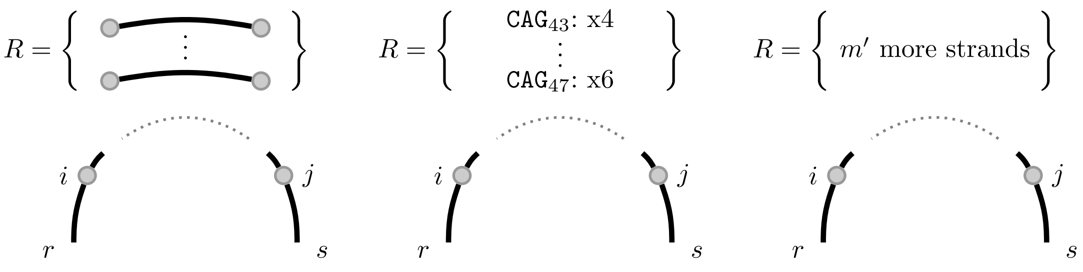
4 Interaction of Triplet Repeats

We now consider a set R_0 of triplet repeat strands. Our goal is to find the minimum free energy secondary structure for R_0. We defined the computational problem MFE Strand Interaction in section 2.2. In the base pair maximization model, this gives exactly the same definition as in [14], where the authors show that the problem is **APX**-hard (and by that **NP**-hard) for the general (non-triplet) case. On the other hand, Dirks et al [13] gave a factorial-time algorithm for computing the partition function over multiple strands. In this section, we improve both results: we show that the problem is **NP**-hard in a reasonable energy model even if restricted to triplet repeats of one pattern; we give an exponential-time, instead of factorial, algorithm for the problem.

However, our exponential-time algorithm is designed for solving the MFE from an algorithmic perspective, as discussed in section 2.2, and does not account for the rotational symmetries to the free energy described by Dirks et al. [13]. Consequently, the DP scheme will not necessarily compute the MFE value in this model, although on a practical level it is likely that the symmetry-corrected structure can be found by investigating a small number of suboptimal structures. Moreover, for the partition function, we can account for the algorithmic overcounting and additionally, if desired, for penalties associated with rotational symmetries.

4.1 General RNA-RNA interactions

The main difficulty of the problem lies in the fact that we need to consider all possible circular permutations of strands. Instead of trying all of these circular permutations one by one and applying a classical single-stranded folding algorithm, we build up the values for all possible circular permutations while exploring all possible joint secondary structures. More specifically, we will consider structures consisting of a leftmost strand and its position, a rightmost strand and its position, as well as a set of strands which have to appear in between the leftmost and rightmost strand (without specifying the ordering of these strands).



(a) (b) (c)

Fig. 4 Visualization of the structures used to compute the MFE in the (a) general setting, (b) TR setting and (c) strand soup setting.

We can formulate DP recurrences as follows: Let E_{s_i,r_j} be the minimum free energy induced by the base pair between the i-th base of strand s and the j-th base of strand r. In our DP equations, $R \subseteq R_0$ denotes the subset of still available strands, $s \in R$ the leftmost strand, $r \in R$ the rightmost strand, $1 \le i \le |s|$ the current position in s, $1 \le j \le |r|$ the current position in r, and $c \in \{0,1,2\}$ indicates whether s and r will be connected by a base pair (0: no base pair allowed, 1: at least one base pair required, 2: a base pair is not required; if the left and right strand are equal, then $c = 2$). The structures with which our algorithm works are visualized in fig. 4 (a). The main recurrences are as follows:

$$M_{R,s_i,r_j,c} = \min \begin{cases} \begin{cases} M_{R,s_{i+1},r_j,c} & \text{if } i+1 \le |s| \\ \min_{t\in R,c'\in\{0,1\}} M_{R-\{s\},t_1,r_j,c'} - \mathbb{1}_{c'=0}K_{\text{assoc}} & \text{if } i+1 > |s| \text{ and } c \neq 1 \\ +\infty & \text{else} \end{cases} \\ \begin{cases} E_{s_i,r_j} + \bar{M}_{R,s_i,r_j,2} & \text{if } c \neq 0 \\ +\infty & \text{if } c = 0 \end{cases} \\ \min_{R',t,k} E_{s_i,t_k} + \bar{M}_{R',s_i,t_k,2} + \bar{M}_{(R-R')\cup\{s\},t_k,r_{j+1},c} \end{cases}$$

where

$$\bar{M}_{R,s_i,r_j,c} = \begin{cases} M_{R,s_{i+1},r_{j-1},c} & \text{if } i+1 \le |s| \text{ and } j-1 \ge 1 \\ \min_{t\in R-\{s,r\},c'\in\{0,1\}} M_{R-\{s,r\},t_1,r_{j-1},c'} - \mathbb{1}_{c'=0}K_{\text{assoc}} & \text{if } i+1 > |s| \text{ and } j-1 \ge 1 \\ \min_{u\in R-\{s,r\},c'\in\{0,1\}} M_{R-\{s,r\},s_{i+1},u_{|u|},c'} - \mathbb{1}_{c'=0}K_{\text{assoc}} & \text{if } i+1 \le |s| \text{ and } j-1 < 1 \\ \min_{t,u\in R-\{s,r\},c'\in\{0,1\}} M_{R-\{s,r\},t_1,u_{|u|},c'} - \mathbb{1}_{c'=0}K_{\text{assoc}} & \text{else} \end{cases}$$

and $-K_{\text{assoc}}$ is a reward for an additional complex. We give this reward each time we "choose" a new strand from R and decide that it should not be connected to the other extremity of the interval ($c' = 0$). The $\bar{M}_{R,s_i,r_j,c}$ equation gives the MFE for the region $]s_i, r_j[$ (i.e. $[s_{i+1}, r_{j-1}]$ if $i+1 \le |s|$ and $j-1 \ge 1$, and introducing new strands in the other cases).

Choosing an arbitrary strand s, the minimum free energy can be finally computed by

$$E^*(R) = (m-1) \cdot K_{\text{assoc}} + \min_{r \in R - \{s\}, c \in \{0,1\}} M_{R, s_1, r_{|r|}, c}$$

and the optimal secondary structure can be obtained through backtracking.

For the initialization, we can set $M_{\{s\}, s_i, s_j} = 0$ for valid indices $j - i \leq \theta$ for any $s \in R$. We now prove that M_{R, s_i, r_j} is computed correctly. By slight abuse of notation, we write $s_i \in S$ for $s_i \in \bigcup_{P \in S} P$.

Definition 1. *An* interval *for this DP is denoted by $[R, s_i, r_j, c]$ where $s, r \in R$, $1 \leq i \leq |s|$, $1 \leq j \leq |r|$ and $c \in \{0, 1, 2\}$. An interval $[R', t_k, u_\ell, c']$ is* included *in interval $[R, s_i, r_j, c]$, written $[R', t_k, u_\ell, c'] \preccurlyeq [R, s_i, r_j, c]$, if one of the following holds:*

- *$R' \subset R$ and $|R'| < |R| - 1$*
- *$R' \subset R$, $|R'| = |R| - 1$ and $s = t \vee r = u$*
- *$R' = R$, $s = t$, $r = u$, $i \leq k$ and $\ell \leq j$.*

If we replace both *inequalities by strict inequalities in the last point, the interval is* strictly included *and we write $[R', t_k, u_\ell, c] \prec [R, s_i, r_j, c]$.*

Each such interval is associated to a minimum free energy as follows:

Definition 2. *Let $I := [R, s_i, r_j, c]$. $\Omega(I)$ is the set of all secondary structures that are valid for this interval, or more formally, a secondary structure S must fulfill:*

- *$S \in \Omega(R)$*
- *$s_k, r_\ell \notin S$ for any $k < i$ and $\ell > j$*
- *$c = 1$ implies the existance of a base pair between s and r (that is, $\{s_k, r_\ell\} \in S$ for some $i \leq k \leq |s|, 1 \leq \ell \leq j$) and $c = 0$ implies that there is no such base pair.*

The minimum free energy of I is defined as MFE$(I) := \min_{S \in \Omega(I)} E(R, S)$.

The minimum free energy of an open interval MFE$(]R, s_i, r_j, c[)$ is the minimum free energy over all secondary structures and all intervals $I' \prec I$ where c specifies the connectedness of s and r.

We also observe that an optimal structure is optimal for any substructure that includes all its base pairs:

Observation 4. *If $E(R, S) =$ MFE$([R, s_i, r_j, c])$ and S only contains base pairs in some interval $[R', t_k, u_\ell, c] \preccurlyeq [R, s_i, r_j, c]$, then $S =$ MFE$([R', t_k, u_\ell, c])$.*

We first show that our helper equation $\bar{M}$ is computed correctly:

Lemma 6. *Assuming that $M_{R', t_k, u_\ell, c'} =$ MFE$(I' := [R', t_k, u\ell, c'])$ for all $I' \preccurlyeq I := [R, s_i, r_j, c]$, we have $\bar{M}_{R, s_i, r_j, c} =$ MFE$(]R, s_i, r_j, c[)$.*

Proof. We distinguish four cases:

- **Case 1:** $i + 1 \leq |s|$ and $j - 1 \geq 1$. In that case, for any $I' \prec I$, we have $I' \preccurlyeq [R, s_{i+1}, r_{j-1}, c]$ and thus MFE$(I') \geq$ MFE$([R, s_{i+1}, r_{j-1}, c]) = \bar{M}_{R, s_i, r_j, c}$ by assumption. Thus MFE$(]R, s_i, r_j, c[) = \bar{M}_{R, s_i, r_j, c}$.

- **Case 2:** $i+1 > |s|$ and $j-1 \geq 1$. For any $I' \prec I$, there is a $t \in R - \{s\}$ and a $c' \in \{0,1\}$ with $I' \preccurlyeq [R - \{s\}, t_1, r_{j-1}, c']$. It thus suffices to minimize over the strands $R - \{s, r\}$ while taking into account a possible strand disconnection reward. We have $\min_{t \in R-\{s,r\}, c' \in \{0,1\}} M_{R-\{s\},t_1,r_{j-1},c'} - \mathbb{1}_{c'=0} K_{\text{assoc}} = \text{MFE}(]R, s_i, r_j, c[)$.
- **Case 3:** $i+1 \leq |s|$ and $j-1 < 1$. This case is completely symmetrical to Case 2.
- **Case 4:** $i+1 > |s|$ and $j-1 < 1$. For any $I' \prec I$, there are $t, u \in R - \{s, r\}$ with $I' \preccurlyeq [R - \{s, r\}, t_1, u_{|u|}, 2]$. It thus suffices to minimize twice over the strands $R - \{s, r\}$ while taking into account a possible strand disconnection reward. We have $\min_{t,u \in R-\{s,r\}, c' \in \{0,1\}} M_{R-\{s,r\},t_1,u_{|u|},c'} - \mathbb{1}_{c'=0} K_{\text{assoc}} = \text{MFE}(]R, s_i, r_j, c[)$.

□

Lemma 7. *The algorithm computes the table entries correctly, i.e.* $M_{R,s_i,r_j,c} = $ *MFE*$([R, s_i, r_j, c])$ *for all* $R \subseteq R_0$*,* $s_i, r_j \in R$ *and* $c \in \{0, 1, 2\}$*.*

Proof. We proceed by induction over the well-founded relation $\preccurlyeq$. Regarding the initialization, clearly no base pair can exist over an empty strand set, as well as over one strand where the number of enclosed base pairs between i and j is less than θ. Therefore, these table entries are correctly initialized by 0.

Let us assume that all $M_{R',t_k,u_\ell,c}$ with $[R', t_k, u_\ell, c] \preccurlyeq [R, s_i, r_j, c]$ except $M_{R,s_i,r_j,c}$ itself have been computed correctly.

- **Case 1:** $s_i \notin S$. If $i+1 \leq |s|$, we have $E(R, S) = \text{MFE}([R, s_{i+1}, r_j, c]) = M_{R,s_{i+1},r_j,c}$ by observation 4 and our induction hypothesis.
  Else, we first assume $c \neq 1$. Consider the strand t that follows s in the polymer graph of S and consider the value c' that specifies connectivity between t and r in S. Since i is unpaired, we again have $E(R, S) = \text{MFE}([R - \{s\}, t_1, r_j, c']) - \mathbb{1}_{c'=0} K_{\text{assoc}} = M_{R-\{s\},t_1,r_j,c} - \mathbb{1}_{c'=0} K_{\text{assoc}}$ as above.
  Finally, if $c = 1$, we look for the MFE of a structure in $[R, s_i, r_j, c]$ where s and r are connected by a base pair. Since there is only one base in s remaining and we leave it unpaired, there is no such structure and thus $\text{MFE}([R, s_i, r_j, 1]) = +\infty$.
- **Case 2:** $S = \{\{s_i, r_j\}\} \cup S'$, where S' is the best structure for any $I' \prec [R, s_i, r_j, c]$ with s and r arbitrarily connected (that is, $]R, s_i, r_j, 2[$). First assume $c \neq 0$. In this case, we have $E(R, S) = E_{s_i,r_j} + \text{MFE}(]R, s_i, r_j, 2[) = E_{s_i,r_j} + \bar{M}_{R,s_i,r_j,2}$, where we could apply lemma 6 because of the induction hypothesis.
  Now assume $c = 0$. We minimize over all structures such that s and r are not connected, but require $\{s_i, r_j\} \in S$. Thus $\text{MFE}([R, s_i, r_j, 0]) = +\infty$.
- **Case 3:** $S = \{s_i, t_k\} \cup S' \cup S''$ for some $t_k \neq r_j$, where S' (resp. S'') is the best structure for any $I' \prec [R', s_i, t_k, c]$ (resp. $I' \prec [R'', t_k, r_j, c]$), with R' being all strands between s and t in the polymer graph of S, and R'' being all strands between t and r.
  Note that s and t are connected, thus in S' the connectivity bit will be set to 2. On the other hand, the connectedness of t and r (for structure S'') is by transitivity of connectivity determined by the connectedness between s and r, that is, c. We then have $\text{MFE}([R, s_i, r_j, c]) = E_{s_i,t_k} + \text{MFE}(]R', s_i, t_k, 2[) + \text{MFE}(]R'', t_k, r_j, c[)$.

□

We now briefly discuss the running time. The number of table entries is bounded by $2^m \cdot n^2$, where $n := \sum_{r \in R} |r|$ is the size of the concatenated sequence. The last case of the DP equation dominates the running time for computing one entry. In the worst case, we iterate over $2^{|R|}$ subsets and n entries, which gives $\mathcal{O}(2^{|R|} \cdot n)$. Partitioned by subset size, we get

$$\sum_{t=0}^{m} \binom{m}{t} n^2 \cdot 2^t n = n^3 \cdot \sum_{t=0}^{m} \binom{m}{t} 2^t = n^3 \cdot \sum_{t=0}^{m} \binom{m}{t} 1^{m-t} 2^t = n^3 \cdot (1+2)^m = 3^m \cdot n^3$$

which bounds the total running time. Together with lemma 7, we conclude.

Detailed conditions and edge cases. When we minimize over all subsets, the following conditions must be respected:

$$\begin{aligned} &\{s,t\} \subseteq R' \subseteq R \wedge 1 \leq k \leq |t| \wedge (k = |t| \rightarrow c \neq 1) \\ \wedge &(s = t \rightarrow (k > i + \theta \wedge R' = \{s\})) \\ \wedge &(r \in R' \rightarrow (t = r \wedge k < j \wedge R' = R \wedge c \neq 0)) \end{aligned}$$

We minimize over all possible triples (R', t, k). A set R' must clearly include s and t to form a valid interval and k must be a valid position of t. If s_i is paired to $t_{|t|}$, s and j are disconnected $(c \neq 1)$. If $s = t$, we must respect θ and there is only one strand in R'. Finally, $r \in R'$ implies that s_i forms a base pair with some base of r (thus $t = r$ and $R' = R$), connectivity has to be allowed $(c \neq 0)$ and t_k must be in the interval $(k < j)$. These conditions are sufficient and match our algorithm.

When we minimize over two new inner strands (in the last case of $\bar{M}$), we clearly cannot choose the same strand for t and u, except if $|R| = 3$. Furthermore, we can clearly only minimize over new inner strands if such strands are still available. If $|R| \leq 3$, there may only be one available strand, or none at all, in which case the energy contribution is 0. We omit these edge cases in the presentation of the algorithm to maintain readability.

Theorem 3. MFE Strand Interaction *can be solved in time* $\mathcal{O}(3^m \cdot n^3)$.

4.2 Translation to Partition Function

For computing the partition function, we must take account of the arising rotational symmetries to avoid an "undercounting" of symmetrical secondary structures, before the canonical overcounting correction. We present an approach that allows to do that without iterating over all circular permutations, and can even incorporate an entropic symmetry correction as considered by [13].

We will always assume that secondary structures are connected and pseudoknot-free, and thus they have a unique pseudoknot-free permutation. In this context, we denote by $\{a,b\} \leq \{c,d\}$ that base pair $\{a,b\}$ includes base pair $\{c,d\}$, that is, c and d are not outside of the interval $[a,b]$.

Problem of over- and undercounting of symmetries. We introduce the notion of indistinguishability and will use the symbol $\sim$ to denote it. Two strands s, t are *indistinguishable* if their sequences are identical. Two sets R, R' of strands are *indistinguishable* if there is a bijection $f : R \to R'$ such that $s \sim f(s)$ for all $s \in R$. Two pairs $((s^k)_{k\in[m]}, S), ((s'^k)_{k\in[m]}, S')$ of a family of strands and a secondary structure are called *indistinguishable* if s^k and s'^k are indistinguishable for all $k \in [m]$ and $S' = \{\{s_i'^k, s_j'^\ell\} \mid \{s_i^k, s_j^\ell\} \in S\}$. An r*-symmetric* secondary structure is a secondary structure S with pseudoknot-free permutation $s^1, ..., s^m$ such that for all $i \in [r]$,

$$((s^k)_{k\in[m]}, S) \sim ((s^{(k+i\cdot m/r) \bmod m})_{k\in[m]}, S)$$

In other words, there are r cyclic shifts that the DP algorithm will not be able to distinguish. For a secondary structure which is not r-symmetric for any $r > 1$, our DP algorithm would count that structure m times (once for each "entry point" between two strands), because all cyclic shifts are distinguishable for the DP algorithm. However, in an r-symmetric structure, there are only m/r distinguishable entry points, and thus the secondary structure will only be counted m/r times. This danger of "undercounting" poses serious algorithmic challenges. We say that a secondary structure has *rotational symmetry* r if it is r-symmetric and that it has *maximum rotational symmetry* r if it has rotational symmetry r and for all $r' > r$, it is not r'-symmetric.

Let Ω be the set of all secondary structures, Ω_r be the set of all secondary structures with rotational symmetry r, and $\Omega_{\max=r}$ be the set of all secondary structures with maximum rotational symmetry r. We can first show a simple lemma which states that r-symmetric secondary structures have a multiple of r as their maximal rotational symmetry.

Lemma 8. *If $S \in \Omega_r$, then $S \in \Omega_{\max=t}$ with $t \bmod r = 0$.*

Proof. Any secondary structure has a maximum rotational symmetry. Assume for contradiction that for this maximum rotational symmetry t for S, $t \bmod r \neq 0$. We claim that S has rotational symmetry $s := \operatorname{lcm}(t, r) > t$.

First of all, we know that $m \bmod r = 0$ and $m \bmod t = 0$, since otherwise there could not be an r- (resp. t-) symmetry. Together with $m > \max(r, t)$, it follows that $m > s$ and that $m \bmod s = 0$. We will assume that $m = s$, and if m is a multiple of s, we can just consider $\frac{m}{s}$ strands to be one strand.

The repeat lengths s/r and s/t are coprime and s/r has a multiplicative inverse modulo s/t, i.e. there is some y such that $ys/r \bmod s/t = 1$. In particular, $iys/r \bmod s/t = i$. Consider two arbitrary strands a^1, a^d in the pseudoknot-free permutation $a^1, ..., a^s$. Take y such that $ys/r \equiv d \mod s/t$. The structures of a^1 and $a^{ys/r}$ have to be identical by r-symmetry, and by t-symmetry, $a^{ys/r}$ has to be identical to a^d. Thus the structures of all strands are identical and we have an s-symmetry. □

Another simple observation is that a higher symmetry implies a symmetry of its divisor:

Observation 5. *For $i, r \in \mathbb{N}^+$, if $S \in \Omega_{i\cdot r}$, then $S \in \Omega_r$.*

We now define some partition function values that we want to compute:

$$\mathcal{Z}_{\max=r} = \sum_{S\in\Omega_{\max=r}} \exp\{-E(S)/kT\}$$

$$\mathcal{Z}_r = \sum_{i=1}^{\lfloor \frac{m}{r} \rfloor} \frac{m}{i \cdot r} \cdot \sum_{S \in \Omega_{\max = i \cdot r}} \exp\{-E(S)/kT\}$$

Assume there is a set $R_{/r}$ of strands where each sequence appears exactly r times less than in R. There is at most one such distinguishable set, and it has size m/r. It suffices to compute a variant of the partition function over $R_{/r}$. Namely, an *extended secondary structure* is a pair of a secondary structure S' and a *marked* base pair $p \in S'$. Any pair of a secondary structure and a base pair is now part of the structure space $\bar{\Omega}(R_{/r})$. As for the standard secondary structure, we can restrict the space to structures of particular symmetries, e.g. $\bar{\Omega}_{\max=i}(R_{/r})$. The free energy of an extended secondary structure is defined as $\bar{E}((S', p)) = r \cdot E(S')$.

Intuitively, all predecessors of the marked base pair p (including itself), namely all $q \leq p$, will be flipped such that the first base of the pair is moved to the next symmetrical occurrence of this base. Therefore, we will sometimes refer to marked base pairs as all $q \leq p$, and not only p itself.

Cyclic shift operations on extended secondary structures. It will be convenient to talk about *cyclic shifts* of extended secondary structures. Each strand is moved one position to the right, and the last strand is moved to the front. Additionally, if the order of the bases of a base pair changes due to the cyclic shift, we change the markedness property (including parents) of the base pair. Two examples can be found in fig. 5.

An extended secondary structure (S', p) is *connected* if and only if:

- S' is connected and
- p is interior or $S' - \{q \in S' \mid q \leq p\}$ is connected.

The partition function over this space is defined as follows:

$$\bar{\mathcal{Z}}(R_{/r}) = \sum_{i=1}^{\lfloor \frac{m}{r} \rfloor} \frac{m}{i} \sum_{S \in \bar{\Omega}_{\max=i}(R_{/r})} \exp\{-\bar{E}(S)/kT\}$$

An i-rotational secondary structure is overcounted by a factor of $\frac{m}{i}$, which we account for in the above definition. Additionally, due to lemma 8, we only need to consider rotational symmetries which are multiples of the considered symmetry.

It is easy to extend the DP to capture this space: We add a *marked bit* b which is 1 if the marked base pair is in the region, and 0 else. An unpaired position does not change the marked bit. If $b = 1$ and we are in the case of a single stack, we add the two values for the case when the stack base pair is marked (in that case, the inner region has $b = 0$) or not. For $b = 0$, we do not change anything. If $b = 1$ and we are in the case of a multiloop, we add the value for the case where the first multiloop base pair is marked and the rest of the multiloop as well as the inner region has $b = 0$, to the value where the marked base pair is in the inner region and the value for the case where the rest of the multiloop has $b = 1$. In the end, we query the complete region with $b = 1$. These extensions do not increase the asymptotic complexity of our standard DP algorithm. We can thus derive:

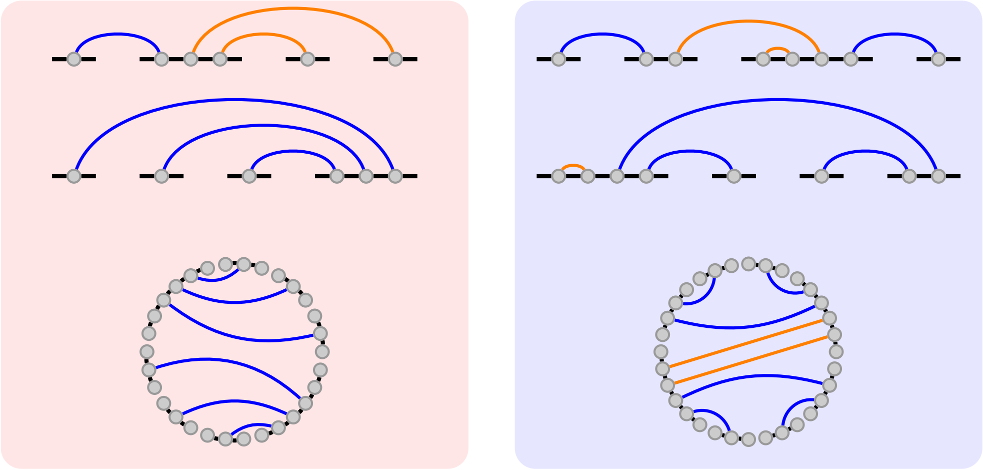


Fig. 5 Two extended secondary structures, where marked base pairs are colored orange. On the left side (red box), the innermost marked base pair is inter-strand (top), and therefore after the cyclic shift (middle), no base pair is marked, which results in an unconnected 2-symmetric structure (bottom). On the right side (blue box), the innermost marked base pair is intra-strand (top), and therefore after the cyclic shift (middle), it is still marked. This base pair ensures the connectedness of the 2-symmetric structure (bottom).

Lemma 9. *$\bar{\mathcal{Z}}(R_{/r})$ can be computed in time $\mathcal{O}(3^m \cdot n^3)$.*

We now describe a bijection between $\Omega_r(R)$ and $\bar{\Omega}(R_{/r})$. Fix an arbitrary strand $a \in R$. Consider $S \in \Omega_r(R)$. Relabel the strands to $s^0, ..., s^{m-1}$, where $s^0 = a$ and the other strands follow in the ordering of the unique pseudoknot-free permutation. Let p_S be the innermost base pair in S such that one base is on a strand between s^0 and $s^{m/r-1}$, and one is not. Our function $f : \Omega_r(R) \to \bar{\Omega}(R_{/r})$ is defined as

$$f(S) = (\{\{s_i^b, t_j^{c \bmod \frac{m}{r}}\} \mid \{s_i^b, t_j^c\} \in S \land 0 \le b < \frac{m}{r}\}, p_S)$$

Lemma 10. *$S \in \Omega_r(R)$ is connected if and only if $f(S)$ is connected.*

Proof. Assume S is connected. Consider $f(S) = (S', p)$ and assume that S' is not connected. Consider one connected component of strands in S' and call it C'. In S, by its r-symmetry, these strands repeat in components $C_0, ..., C_{r-1}$. The union of these components does not contain any outgoing base pairs, which gives a contradiction to the connectedness of S. Thus, S' is connected.

Now assume that $S' - \{q \in S' \mid q \le p\}$ is not connected. Consider a new extended secondary structure S'' that is obtained by cyclic shifts of S' until the strand of the second base of the innermost marked base pair is in front. Now, every marked inter-strand base pair is flipped in S'', and thus not marked anymore. Marked intra-strand base pairs however remain marked. If the innermost marked base pair is inter-strand, then in S'', no marked base pair exists, which corresponds to r separate connected components, a contradiction to the connectedness of S. Thus, the original innermost marked base pair must be intra-strand. We conclude that $f(S)$ is connected.

For the other direction, assume $f(S) = (S', p)$ is connected. If $S' - \{q \in S' \mid q \leq p\}$ is connected, each of the r symmetric repeats consist of at most one connected component. The marked base pair p then connects these connected structures with each other. Thus, S is connected.

If $S' - \{q \in S' \mid q \leq p\}$ is not connected, there is a marked interior base pair. By connectedness of S', there is a circular shift of the pe rmutation, where marked exterior base pairs are unmarked, such that the structure without the marked base pair is connected. We can then proceed as above. □

Lemma 11. *$S \in \Omega_{r\cdot i}(R)$ if and only if $f(S) \in \Omega_i(R_{/r})$.*

Proof. Assume $S \in \Omega_{r\cdot i}$. By observation 5, there are r symmetrical substructures and each of them is i-symmetric, and we thus get $f(S) \in \Omega_i(R_{/r})$. For the other way, f^{-1} replicates the (by assumption i-symmetric) structure r times, thus the resulting structure is $r \cdot i$-symmetric. □

Lemma 12. *The function $f : \Omega_r(R) \to \bar{\Omega}(R_{/r})$ is a bijection, preserves free energy and connectedness, and decreases the rotational symmetry by a factor of $\frac{1}{r}$.*

Proof. Connectedness follows from lemma 10, the decreasing of the rotational symmetry follows from lemma 11, and the preservation of the free energy is immediately clear since each base pair in $f(S)$ is weighted with a factor of r, and in S, it is replicated r times. It thus remains to show that f is a bijection.

We first show injectivity. Consider two different r-symmetric secondary structures $S, T \in \Omega_r(R)$. Consider $f(S) = (S', p_S)$ and $f(T) = (T', p_T)$. We order the strands with respect to the correct permutation of S, with $r^0 = a$. First we notice that if the unique pseudoknot-free permutations of S and T differ, so do the unique pseudoknot-free permutations of S' and T', which would imply $S' \neq T'$.

So we can assume that their pseudoknot-free permutation is the same. Because S and T are r-symmetric and different, there is a position s_i^p for $0 \leq p < \frac{m}{r}$ which is differently paired in S and T. Assume without loss of generality that $s_i^p \in S$. If $s_i^p \notin T$, we have $s_i^p \in S'$ but $s_i^p \notin T'$ and we are done. Else, if s_i^p is matched differently in S' and T', we are done again. We can thus assume that s_i^p is matched to the same s_j^q for $0 \leq q < \frac{m}{r}$ in S' and T'. Since s_i^p is differently matched in S and T, we must have $\{s_i^p, s_j^q\} \in S$ and $\{s_i^{p+\frac{m}{r}}, s_j^q\} \in T$, or the other way around. Thus, p_T has to be in the region enclosed by base pair $\{s_i^p, s_j^q\}$, but p_S cannot be in this region, because both endpoints of the base pair are between strands s^0 and $s^{\frac{m}{r}}$. Thus $f(S) \neq f(T)$.

We now show surjectivity. Consider an arbitrary $E \in \bar{\Omega}_r$, that is, a pair $E = (S', P)$. Now build a secondary structure S as follows: For easch base pair $\{s_i^p, s_j^q\}$, if it does not enclose P, add itself and its $r - 1$ symmetrical copies to S. For the case that it encloses P, assume wlog $p < q$. We add the base pair $\{s_j^q, s_i^{p+\frac{m}{r}}\}$ and its r symmetrical copies to S. It is easy to see that $f(S) = (S', P)$.

Finally notice that the function is total, i.e. it is defined for every $S \in \Omega_r(R)$. □

By lemma 12, we can now rewrite $\mathcal{Z}_r$ as follows:

$$\begin{aligned}
\mathcal{Z}_r &= \sum_{i=1}^{\lfloor \frac{m}{r} \rfloor} \frac{m}{i \cdot r} \cdot \sum_{S \in \Omega_{\max=i\cdot r}(R)} \exp\{-E(S)/kT\} \\
&= \sum_{i=1}^{\lfloor \frac{m}{r} \rfloor} \frac{m}{i \cdot r} \cdot \sum_{S \in \bar{\Omega}_{\max=i}(R_{/r})} \exp\{-\bar{E}(S)/kT\} \\
&= \bar{\mathcal{Z}}(R_{/r})
\end{aligned}$$

This quantity can be computed in time $\mathcal{O}(3^m \cdot n^3)$ by lemma 9. We can finally conclude:

Lemma 13. *$\mathcal{Z}_r$ can be computed in time $\mathcal{O}(3^m \cdot n^3)$.*

Using this result, we will proceed by showing that $\mathcal{Z}_{\max=r}$ can also be computed efficiently.

Lemma 14. *$\mathcal{Z}_{\max=r}$ can be computed in time $\mathcal{O}(3^m \cdot n^3 \cdot m)$, for any r.*

Proof. We can compute $\mathcal{Z}_{\max=r}(R)$ as follows. We create the following DP:

$$\mathcal{Z}[t] = \frac{t}{m} \cdot \left(\mathcal{Z}_t - \sum_{i=2}^{\lfloor \frac{m}{t} \rfloor} \frac{m}{i \cdot t} \cdot \mathcal{Z}[i \cdot t] \right)$$

Indeed, we can inductively verify the correctness of the equation, namely we can proof $\mathcal{Z}[t] = \mathcal{Z}_{\max=t}$, with respect to the energy contribution E', for all $t \in \{1, ..., m\}$. For the base case $t = m$, notice that

$$\mathcal{Z}[m] = \frac{m}{m} \cdot (\mathcal{Z}_m - 0) = \mathcal{Z}_m = \sum_{S \in \Omega_{\max=m}} \exp\{-E'(S)/kT\} = \mathcal{Z}_{\max=m}$$

Inductively, assume that $\mathcal{Z}[t']$ is correctly computed for all $m \geq t' > t$. We have

$$\begin{aligned}
\mathcal{Z}[t] &= \frac{t}{m} \cdot \left(\mathcal{Z}_t - \sum_{i=2}^{\lfloor \frac{m}{t} \rfloor} \frac{m}{i \cdot t} \cdot \mathcal{Z}[i \cdot t] \right) \\
&= \frac{t}{m} \cdot \left(\sum_{i=1}^{\lfloor \frac{m}{t} \rfloor} \frac{m}{i \cdot t} \cdot \sum_{S \in \Omega_{\max=i\cdot t}} \exp\{-E'(S)/kT\} - \sum_{i=2}^{\lfloor \frac{m}{t} \rfloor} \frac{m}{i \cdot t} \cdot \mathcal{Z}_{\max=i\cdot t} \right) \\
&= \frac{t}{m} \cdot \left(\begin{array}{l} \displaystyle\sum_{i=1}^{\lfloor \frac{m}{t} \rfloor} \frac{m}{i \cdot t} \cdot \sum_{S \in \Omega_{\max=i\cdot t}} \exp\{-E'(S)/kT\} \\ \displaystyle -\sum_{i=2}^{\lfloor \frac{m}{t} \rfloor} \frac{m}{i \cdot t} \cdot \sum_{S \in \Omega_{\max=i\cdot t}} \exp\{-E'(S)/kT\} \end{array} \right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{t}{m}\frac{m}{t} \sum_{S \in \Omega_{\max=t}} \exp\{-E'(S)/kT\} \\
&= \sum_{S \in \Omega_{\max=t}} \exp\{-E'(S)/kT\} \\
&= \mathcal{Z}_{\max=t}
\end{aligned}
$$

By lemma 13, each $\mathcal{Z}_t$ can be computed in time $\mathcal{O}(3^m \cdot n^3)$. This dominates the running time to compute one entry. Since we compute at most m entries, an overall running time in $\mathcal{O}(3^m \cdot n^3 \cdot m)$ follows. □

Now that we have the values for $\mathcal{Z}_{\max=r}$, we can compute the value of the partition function $\mathcal{Z}$:

$$
\mathcal{Z} = \sum_{S \in \Omega} \exp\{-E(S)/kT\} = \sum_{r=1}^{m} \sum_{S \in \Omega_{\max=r}} \exp\{-E(S)/kT\} = \sum_{r=1}^{m} \mathcal{Z}_{\max=r}
$$

By lemma 14, we can compute each $\mathcal{Z}_{\max=r}$ in time $\mathcal{O}(3^m \cdot n^3 \cdot m)$. Since we sum over m such entries, we finally obtain the following result:

Theorem 4. *The partition function $\mathcal{Z}$ over m strands can be computed in time $\mathcal{O}(3^m \cdot n^3 \cdot m^2)$.*

Remark 3. *It is easy to see that for each rotational symmetry r, we can add the symmetry correction $kT \log r$ as described by Dirks* et al *[13] to the DP equations, if desired. Thus, the above result also holds for this variant of the partition function.*

Remark 4. *The described technique directly translates to the other algorithms that we will present in the following sections. It can be applied to obtain the exact partition function for the triplet repeat setting (section 4.3) and the strand soup setting (section 4.5).*

4.3 Strand interactions for triplet repeats

We now consider the special case where all strands in our pool are triplet repeats. We call this restricted problem MFE Triplet Repeat Strand Interaction. Assume first that all strands have the same pattern and that we have a bounded number of different strand-lengths $p := |\{i \mid \exists r \in R : |r| = i\}|$. Regardless of the ordering of the strands, the resulting sequence of the concatenated strands is identical. We can therefore focus on the length of the strands and disregard their actual sequence.

We do not iterate over all subsets of R, since we only need to distinguish the number of strands of a certain length in the subset, in a count-sort-like manner. Thus we can represent a subset $R' \subseteq R$ by $(a_1, ..., a_p)$ where $a_i := |\{r \in R' \mid |r| = n_i\}|$ is the number of strands of size n_i in R. An example is given in fig. 4 (b). As argued in the following, the exponent only depends on p:

We need table entries for each possible configuration of remaining number of occurrences and for specifying the remaining number of bases on the leftmost and rightmost strand. Using $n := \max_{r \in R} |r|$, we bound the number of table entries by

$$n^2 \cdot \max_{s_1,\ldots,s_p : s_1+\ldots+s_p=m} \prod_{i=1}^{p} s_p \leq n^2 \cdot \left(\frac{m}{p}\right)^p$$

The running time for computing one table entry is dominated, as for the previous section, by the last case. We need to iterate over $\mathcal{O}((\frac{m}{p})^p)$ configurations to split our region into two strand sets, p lengths to determine the length of the strand on which we split and n positions for the index of the split. We finally obtain a running time of $\mathcal{O}((\frac{m}{p})^{2p} \cdot n^3 \cdot p)$, which is an XP algorithm parametrized by p.

Theorem 5. *There is an XP algorithm for* MFE Triplet Repeat Strand Interaction *parametrized by the number of different lengths p, running in $\mathcal{O}((\frac{m}{p})^{2p} \cdot n^3 \cdot p)$ time.*

Notice that this algorithm can be extended to the case where we have different triplet patterns; the parameter then becomes the number of non-identical strands.

4.4 Computational hardness

In this subsection, we show that the parametrized approach seen before is, in a sense and under standards assumptions, the best we can hope for. Moreover, even for triplet repeats, the problem of deciding whether there is a secondary structure for R_0 with a free energy below a certain threshold t remains **NP**-complete, for a reasonably-intricate energy model. Note that for the general (non-triplet) case, this has already been shown in [14]. Our result is surprising in the sense that the concatenation of TR strands always yields the same sequence, and the only additional difficulty compared to the single-stranded case arises from the fact that we do not know the indices of the strand borders.

Our reduction requires more than the naive base pair maximization model, but to keep the reduction simple, we will not use the full Turner energy model. Instead, we posit that each base pair gives a free energy reward of $E^{\mathrm{bp}} = -\frac{m}{3}$, where $m > 0$ is the number of interacting strands, while subdividing an interval into two intervals that are not strand-disjoint gives a multiloop penalty of $K_{\mathrm{multi}} = +1$. Furthermore, each connected component reduces the strand association penalty by $-K_{\mathrm{assoc}} := -1$. Finally, every hairpin loop must enclose at least three unpaired bases ($\theta = 3$). This model can be extended into the Turner model by setting equal energy values for interior and hairpin loops, and accounting for the multiloop penalty in the corresponding energy values.

Let us define our main decision problem:

Triplet Repeat Multi-Strand MFE
Input: A set R of explicitly encoded triplet repeat strands of the same pattern and a target free energy value t.
Output: Is there a secondary structure $S \in \Omega(R)$ with $E(R, S) \leq t$?

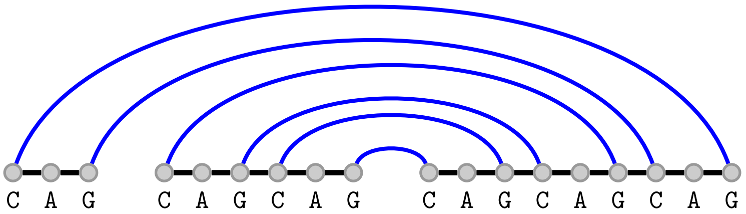


Fig. 6 Optimal secondary structure corresponding to a valid summing triple $(1, 2, 3)$. The $1 \cdot 3 + 2 \cdot 3$ bases on the left perfectly match to the $3 \cdot 3$ bases on the right.

Even if the following reduction does not work in the base pair maximization model, a DP algorithm for base pair maximization in this setting seems unlikely, as, under the assumption that $\mathbf{P} \neq \mathbf{NP}$, one would not be able to generalize the algorithm to more complex energy models.

We will show **NP**-hardness by reduction from the following problem:

Summing Triples
Input: List L of distinct positive integers $s_1, ..., s_{3n}$, encoded in unary
Output: Is there a partition of L into triples (a_i, b_i, c_i) such that $a_i + b_i = c_i$?

The problem was shown to be strongly NP-hard [21]. We define $v := \sum_{i=1}^{3n} s_i$.

The reduction is as follows: We create a strand $r_i := (CAG)^{s_i}$ for each integer s_i. Hence, we have $n = \frac{m}{3} = -E^{\text{bp}}$. We denote by R the set of strands. We set the target minimum free energy to $t := -(3v + 1)n$.

Assume that there is a partition into summing triples. Our secondary structure is built such that for each triple $a + b = c$, we add the base pairs

$$(a_1, c_{|c|}), (a_3, c_{|c|-2}), (a_4, c_{|c|-3}), (a_6, c_{|c|-5}), ..., (a_{|a|-2}, c_{|c|-|a|+3}), (a_{|a|}, c_{|c|-|a|+1}),$$
$$(b_1, c_{|c|-|a|}), (b_3, c_{|c|-|a|-2}), ..., (b_{|b|-2}, c_3), (b_{|b|}, c_1)$$

Note that all base pairs are labeled with $C-G$ or $G-C$. fig. 6 visualizes the secondary structure for the exemplary triple $1 + 2 = 3$. We claim that S is unpseudoknotted for the circular permutation $a_1 \cdot b_1 \cdot c_1 \ldots a_n \cdot b_n \cdot c_n$ and that $E(R, S) = t$.

Since any two triples of strands are not connected, we have exactly n connected components. Each connected component consists of one large stacked loop with innermost base pair $(b_{|b|}, c_1)$ (i.e. we do not violate the constraint that every innermost base pair must include three unpaired bases, because the base pair is inter-strand). Since $a + b = c$, the outermost base pair is $(a_1, c_{|c|})$. There is no multiloop involved in S, so each triple (a_i, b_i, c_i) contributes a free energy of $2|c| \cdot E^{\text{bp}} - K_{\text{assoc}} = -6n|c| - 1$. Since all triples are correctly summing, we have $\sum_{i=1}^{n} c_i = \frac{1}{2}v$. Thus indeed the minimum free energy is at most

$$\sum_{i=1}^{n} -6n|c_i| - 1 = -6n \sum_{i=1}^{n} |c_i| - n = -6n \cdot \frac{1}{2}v - n = -3nv - n = t$$

Before showing the opposite direction, we introduce the following simple lemmata:

Lemma 15. *If some C or G base remains unpaired in a secondary structure S, $E(R, S) > t$.*

Proof. First notice that in every valid secondary structure, all A bases remain unpaired (since there are no U bases). There are $2v$ bases of C/G in total. Since we assumed that one of them is unpaired, there can be at most $v - 1$ base pairs. We can have at most $3n$ complexes, so the strand association penalty is reduced by at most $3n$. Thus we have $E(R, S) \geq -3n(v - 1) - 3n = -3vn > -(3v + 1)n = t$. □

Lemma 16. *If S contains a hairpin loop, $E(R, S) > t$.*

Proof. A hairpin loop must enclose at least three bases. Since in the `CAG` triplet pattern any two consecutive bases involve at least one C or G, we can apply lemma 15 and conclude. □

Now assume for an arbitrary $S \in \Omega$ that $E(R, S) \leq t$. We first show that there must be exactly n connected components, each with three strands. Assume that there is a connected component with less than three strands. If it has only one strand, it must contain a hairpin loop, and by lemma 16, $E(R, S) > t$. If the complex contains two strands, first of all the two strands have a different number of triplet repeats, since all s_i are distinct. This implies that if the innermost loop is inter-strand (if it is intra-strand we again apply lemma 16) and has no multiloop, some G or C base must be unpaired (since base pairs can then only be between the two strands, but one of the strands contains at least one G and one C base more than the other). Then, by lemma 15, $E(R, S) > t$. If it has a multiloop, there have to be two innermost base pairs, one of which must be intra-strand, and we can apply lemma 16.

Since we ruled out complexes of one or two strands and the total number of strand is divisible by 3, we know that if there is a complex with four strands, our secondary structure will have $< n$ connected components. Thus the best achievable score will be $-n + 1 - 3nv > t$. Hence, any $S \in \Omega$ with $E(R, S) \leq t$ consists of n complexes, each consisting of three strands a_i, b_i, c_i with $|a_i| < |b_i| < |c_i|$. We claim that for all $i \in [n]$, $|a_i| + |b_i| = |c_i|$.

By contradiction, assume $|a_i| + |b_i| \neq |c_i|$ and first consider the case that there are no multiloops. This implies that there is only one innermost base pair. If it is intra-strand, we obtain a contradiction to $E(R, S) \leq t$ by lemma 16. If it is inter-strand, all remaining base pairs must be between one of two strands d, e on the one side and the third strand f on the other side. Since $|d| + |e| \neq |f|$ for any such partition, one of the two sides will be left with at least one unpaired G and one unpaired C, and we apply lemma 15.

Now we consider the case of multiloops. Any multiloop where the cutpoint between the two recursive structures is on a strand border (and thus is not penalized) implies an innermost base pair in both recursive structures, and since by pigeonhole principle one of the two recursive structures is single-stranded, we have a hairpin loop and $E(R, S) > t$ by lemma 16. In the other case, we have a multiloop penalty of $+1$. Thus we can lower bound $E(R, S) \geq -n - 3nv + 1 > t$.

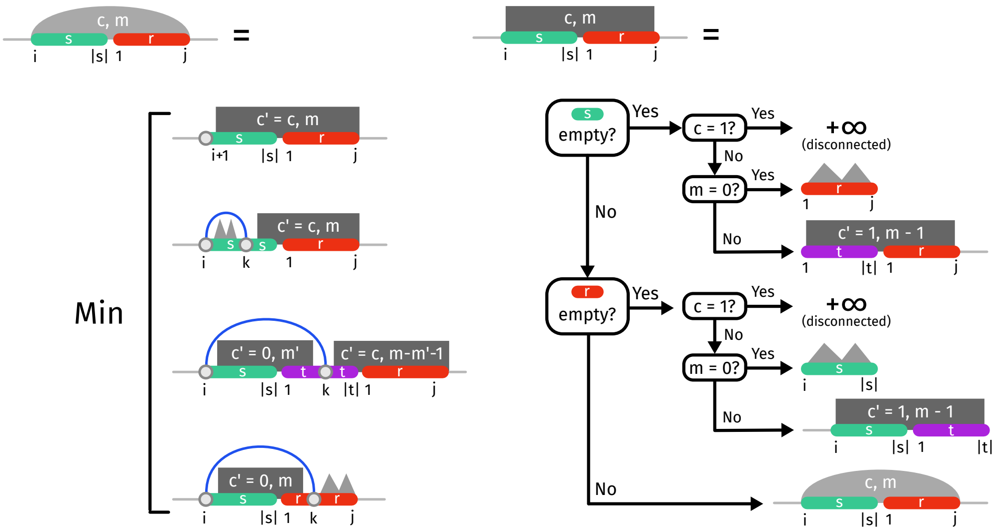


Fig. 7 Schematic illustration of the dynamic programming scheme for the Strand Soup Interaction model (base pair-based energy model). The main case/matrix (Left) minimizes free-energy over all secondary structures for a sequence over m strands, beginning with a $[i, |s|]$ suffix of the strand s, and ending with a prefix $[1, j]$ of a strand r. Akin to the Nussinov recursions, the decomposition partitions the space of secondary structure based on the base-pairing status (paired/unpaired) and partner (if relevant) of the first nucleotide s_i, creating further base pairs through recursive calls. An auxiliary matrix (Right) uniformly handles cases where r or s may be fully depleted and one or several new strand(s) need to be inserted. For both matrices, m strands remain to be inserted. This number is updated and distributed over recursive calls to reflect the addition of a strand. A connectivity bit c, if set to true, forces the connection of r and s (either directly or transitively), ultimately ensuring that the strands end up forming a connected graph. Not shown here is the classic $\Theta(n^3)$ PD scheme to obtain the MFE of single-strands, visually indicated by *mountains* above.

This finishes the proof that $|a_i| + |b_i| = |c_i|$, and we get $\frac{|a_i|}{3} + \frac{|b_i|}{3} = \frac{|c_i|}{3}$. By the construction, each strand r corresponds to one integer $\frac{|r|}{3}$ in the set of integers of our original instance. Thus, $(\frac{|a_i|}{3}, \frac{|b_i|}{3}, \frac{|c_i|}{3})$ for all complexes $\{a_i, b_i, c_i\}$ for $1 \leq i \leq n$ is a valid set of summing triples.

The reduction is polynomial-time, since in the Summing Triples problem, the integers are encoded in unary. Membership in **NP** follows by the fact that we can evaluate the energy given a secondary structure and its unpseudoknotted circular permutation.

Theorem 6. Unary Triplet Repeat Multi-Strand MFE *is* ***NP****-complete.*

4.5 Predicting *strand soup* interactions

We now consider the computational problem MFE Strand Soup Interaction as defined in section 2.2. In comparison to above, we no longer need to keep track of the (exponentially many) subsets, or consider any strand association penalty since we require one single complex, but must enforce global connectivity of the strands set. Towards that goal, we introduce a *connectivity bit* c such as $c = 1$ indicate that the two outer strands have to be (transitively) connected in the corresponding interval, and $c = 0$ if they do not need to be connected (but can still be).

Let E_{s_i,r_j} be the energy contribution of pairing the i-th base of strand s to the j-th base of strand j, and let $M_s[i,j]$ be the classical single-stranded minimum free energy in the interval from i to j in strand s. Then the minimum free-energy (BP energy model) of a secondary structure over m strands subject to c, flanked by a suffix $[i,|s|]$ of a strand s and a prefix $[1,j]$ of j, obeys:

$$M_{s_i,r_j,m,c} = \min \begin{cases} \overline{M}_{s_{i+1},r_j,m,c} \\ \min_k E_{s_i,s_k} + M_s[i+1,k-1] + \overline{M}_{s_{k+1},r_j,m,c} \\ \min\limits_{\substack{t\in R \\ 1\leq k\leq |t| \\ m'+m''=m-1}} E_{s_i,t_k} + \overline{M}_{s_{i+1},t_{k-1},m',0} + \overline{M}_{t_{k+1},r_j,m'',c} \\ \min_k E_{s_i,r_k} + \overline{M}_{s_{i+1},r_{k-1},m,c} + M_r[k+1,j] \end{cases} \tag{6}$$

with the following auxiliary table, responsible for the introduction of new strands whenever those identified by s and r have been entirely *consumed*:

$$\overline{M}_{s_i,r_j,m,c} = \begin{cases} \left.\begin{cases} \min_{t\in R} \overline{M}_{t_1,r_j,m-1,1} & \text{if } c=0 \text{ and } m>0 \\ M_r[1,j] & \text{if } c=0 \text{ and } m=0 \\ \infty & \text{if } c=1 \end{cases}\right| & \text{if } i > |s| \\ \left.\begin{cases} \min_{t\in R} \overline{M}_{s_i,t_{|t|},m-1,1} & \text{if } c=0 \text{ and } m>0 \\ M_s[i,|s|] & \text{if } c=0 \text{ and } m=0 \\ \infty & \text{if } c=1 \end{cases}\right| & \text{else if } j<1 \\ M_{s_i,r_j,m,c} & \text{otherwise} \end{cases} \tag{7}$$

The minimum free energy can be finally computed by

$$E^*(R,m) = \min_{s,r\in R} M_{m-2,s_1,r_{|r|},1} \tag{8}$$

and the optimal secondary structure can be obtained through backtracking. We initialize $M_{1,s_i,s_j,2} = 0$ for all $j - i \leq \theta$.

4.5.1 Proof of correctness

Let us first establish the correctness of the DP scheme which, despite similarities with that of section 4.1, implements a different decomposition strategy.

Theorem 7. *For any strand set R and number m of strands, the value found in $E^*(R,m)$, following its computation through Equations (6), (7) and (8), is the MFE of a connected secondary structure over m strands.*

Proof. We proceed by induction. We consider tuples of the form $(m,s,-i,r,j,c)$, and introduce an increasing lexicographic relation $\prec$ over tuples, defining a total order over such tuples. We then hypothesize that, for any tuple $(m',s',-i',r',j',c')$ such that $(m',s',-i',r',j',c') \prec (m,s,-i,r,j,c)$, the following properties holds:

- $\overline{M}_{m',{s'}_{i'},{r'}_{j'},c'}$ contains the MFE of $m - \mathbb{1}_{i>|s|} - \mathbb{1}_{j<1}$ connected strands, starting and ending with s_i and r_j respectively, both of which may be empty. Additionally, if $c = 1$, the MFE is restricted to structures (transitively) connecting s_i to r_j;
- $M_{m',{s'}_{i'},{r'}_{j'},c'}$ contains the MFE of m connected strands, starting and ending with non-empty s_i and r_j respectively. Additionally, if $c = 1$, the MFE is restricted to structures (transitively) connecting s_i to r_j.

In both of the above cases, a $+\infty$ value is expected whenever constraints are overall unsatisfiable.

First, let us observe that base cases only concern $\overline{M}$, and let us discuss the correction of the equation in this context:

- If $i > |s|$ (s_i is actually empty), and $c = 1$ (s needs to be connected to r, but no such base pair has been formed at this stage), then $\overline{M}$ represents the MFE over an empty set of secondary structure, and should be set to ∞ as seen in the DP equation. The same reasoning holds when $j < 1$ (r_j empty), and $c = 1$;
- If $i > |s|$ (s_i empty), $m = 0$ (no strand left to insert) and $c = 0$ (no obligation to pair s_i to r_j), then the only energy of a structure stems from the MFE over r_j (if non-empty), found in $M_r[1, j]$ as correctly returned by the equation;
- If $j < 1$ (r_j empty), $m = 0$ (no strand left to insert) and $c = 0$ (no obligation to pair s_i to r_j), then the only energy of a structure stems from the MFE over s_i (if non-empty), found in $M_s[i, |s|]$ as correctly returned by the equation.

Next, we use the two induction hypotheses to show the correctness of the value found in $M_{m,s_i,r_j,c}$ and, following that, the correctness of $\overline{M}_{m,s_i,r_j,c}$. To show that $M_{m,s_i,r_j,c}$ is correct, let us observe that, within any secondary structure, the first position (i in strand s) is either unpaired, or paired to some position k within a strand:

- If i is left unpaired, the MFE is that of the (possibly empty) $[i + 1, |s|]$ suffix of s. From the induction hypothesis, such an energy can be found in $\overline{M}_{m,s_{i+1},r_j,c}$, correctly computed since $(m, s, -(i + 1), r, j, c) \prec (m, s, i, r, j, c)$;
- If i is paired to some k in the interval $[i + \theta + 1, |s|]$ of s, then the MFE of any such structure includes the contribution E_{s_i,s_k} of the base pair (i, k) in s, the MFE of the structure enclosed by the base pair ($\rightarrow M_s[i + 1, k - 1]$) and the MFE over m strands beginning with the remainder of s ($\rightarrow \overline{M}_{m,s_{k+1},r_j,c}$, correctly computed since $i < k + 1$);
- If i is paired to some k in a non-flanking strand t, then the MFE is obtained as the sum of the BP energy E_{s_i,t_k}, and the MFE contributions of two structures, being assigned m' and m'' strands such that $m' + m'' + 1 = m$, respectively enclosed ($\rightarrow \overline{M}_{m',s_{i+1},t_{k-1},0}$; $c = 0$ since a BP (i, k) already connects s and t) and preceded ($\rightarrow \overline{M}_{m'',t_{k+1},r_j,c}$; c is propagated since t may already be connected to r through s) by (s_i, t_k);
- If i is paired to some k in r, then the MFE consists in the energy E_{s_i,r_k} of the BP (s_i, r_k), augmented by the MFE over the remainder of r and s *under* (s_i, r_k) ($\rightarrow \overline{M}_{s_{i+1},r_{k-1},m,c}$) and the independent folding of the portion $[k+1, j]$ of r following r_k ($\rightarrow M_r[k + 1, j]$).

These 5 cases can be verified to match the contributions in Equation (6). Moreover, the decomposition is complete, as it covers all possible outcomes for s_i. We conclude that minimizing over those yields the correct value, *i.e.* the $M_{m,s_i,r_j,c}$, and that the correctness of M and $\overline{M}$ for each $(m', s', -i', r', j', c') \prec (m, s, -i, r, j, c)$ implies the correctness of M for $(m, s, -i, r, j, c)$.

We are then only left to establish the correctness of $\overline{M}_{m,s_i,r_j,c}$, noting that its only difference in comparison with $M_{m,s_i,r_j,c}$ is its support for empty flanking regions. We focus on cases where one or several additional strands needs to be inserted and can be $(m > 0)$:

- If s is fully depleted $(i > |s|)$, then a strand t needs to be introduced to replace the leftmost flanking s. Since t is new it needs to be connected to r and, transitively, to each of the $m-1$ other strands that will be inserted by subsequent calls ($\rightarrow c = 1$). The MFE of such a structure can therefore be found in $\overline{M}_{t_1,r_j,m-1,1}$, correctly computed by the equation since $(m-1, t, -1, r, j, c) \prec (m, s, -i, r, j, c)$;
- Similarly, if r is fully depleted $(j < 1)$, then a strand t needs to be introduced as the new rightmost flanking strand r. Since t is new it needs to be connected to s and the other $m-1$ other strands inserted by subsequent calls ($\rightarrow c = 1$). The MFE of such a structure is found in $\overline{M}_{s_i,t_{|t|},m-1,1}$, which correctly computed by the equation since $(m-1, s, -i, t, |t|, c) \prec (m, s, -i, r, j, c)$;
- If neither s nor r is depleted, then we are left to consider the structures over m connected strands, flanked by s_i and r_j, with connectivity bit c. The MFE of such a structure is found in $M_{m,s_i,r_j,c}$, correctly computed as established above.

Having already discussed base/terminal cases, we establish the completeness of the case decomposition and, in turn, on the correctness of the value found in $\overline{M}_{m,s_i,r_j,c}$. The induction step follows, and we finally conclude that the content of both $M_{m,s_i,r_j,c}$ and $\overline{M}_{m,s_i,r_j,c}$ match their specification.

The correctness of the MFE found in $E^*(R, m)$ follows from the choice of two globally-flanking strands s and r, left fully available (s_1 and $r_{|r|}$) while ensuring their connectivity $(c = 1)$. For each (s, r) pair, the MFE of such structures can be found in $M_{m-2,s_1,r_{|r|},1}$, so the minimization computed in Equation (8), concluding our proof. □

4.5.2 Complexity

Regarding the running time the number of entries, needed to compute the DP tables in Equations (6) and (7), is bounded by $\mathcal{O}(m \cdot p^2 \cdot n^2)$ entries, with $n := \max_{s \in R} |s|$. The running time to compute one table entry is dominated by the repeated evaluation of the third line in Equation (6), where we minimize over $\mathcal{O}(m \cdot p \cdot n)$ different ways to introduce a new strand t, figure out a base pairing partner $k \in [1, |t|]$ for s_i, and split m into (m', m'') such that $m' + m'' = m - 1$. In total, we obtain an algorithm with running time $\mathcal{O}(n^3 \cdot m^2 \cdot p^3)$. We can then conclude:

Theorem 8. MFE Strand Soup Interaction *over m strands, taken from a collection of p sequences, can be solved in time $\mathcal{O}(n^3 \cdot m^2 \cdot p^3)$.*

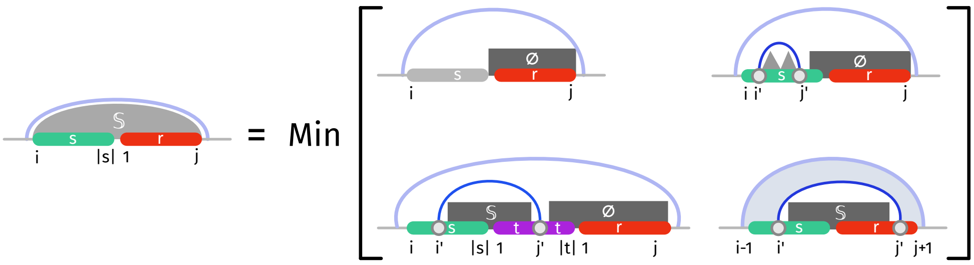


Fig. 8 Partial illustration of our extended DP scheme, capturing a simple Turner-style energy model in the strand soup paradigm. The decomposition for filling the $M^{\mathbb{S}}$ matrix postulates the existence of an enclosing base pair $(i-1, j+1)$, and investigates the first paired position $i' \geq i$ in the $[i, |s|]$ region of s: Such a position may not exist ($s[i, |s|]$ fully unpaired), or be paired to j'' located in s, in a new strand t, or in r. In the latter case, the existence of two consecutive nested base pairs between s and r represents an interaction loop (stacked pairs, bulges or interior loop), usually associated with an energy bonus.

Remark 5. *In addition to restricting the number of interacting strands, one can extend the above algorithm to restrict the size of the concatenated sequence. This is possible by keeping track of the current size of the sub-interval in the DP tables, and updating these values whenever a new strand is introduced. This might be useful if the sequences in the base set have different length, as the basic algorithm would otherwise favor larger sequences because they usually allow for more base pairs.*

Remark 6. *The case of triplet repeats induces a slight improvement of the running time. Since all strands look the same except for their length, we can use a table with entries of the form $M_{m,i,j,c}$, where i and j denote the remaining number of bases in the leftmost and rightmost strand. This reduces the space complexity to $\mathcal{O}(m \cdot n^2)$, but the computation of one table entry still requires the same time, giving an overall time complexity of $\mathcal{O}(n^3 \cdot m^2 \cdot p)$.*

4.5.3 Incorporating a simple nearest-neighbor energy model

The DP scheme underlying Equations (6), (7) and (8) can be modified to capture a simplified nearest neighbor model, where the energy of a secondary structure S over m strands. It crucially requires the definition of an *interaction loop* between two strands s and r delimited by two base pairs (s_i, r_j) and $(s_{i'}, r_{j'})$ such that $1 \leq i < i' \leq |s|$ and $1 \leq j' < j \leq |r|$. We denote by $\Delta G(s_i, s'_i, r'_j, r_j)$ the energy of an interaction loop (s_i, s'_i, r'_j, r_j), accessible via table lookup in popular libraries such as the Vienna package [22]. We can then define the simplified Turner energy of a secondary structure S as:

$$E(S; s_1, \ldots, s_m) = \sum_{\text{strand } s_i} E^{\mathcal{T}}(S|_{s_i}; s_i) + \sum_{\substack{\textit{Interaction loop } (s_i, s'_i, r'_j, r_j) \\ \text{(stack, bulge, or interior loop)}}} \Delta G(s_i, s'_i, r'_j, r_j)$$

where $S|_{s_i}$ represents the restriction of the (multi-strand) secondary structure S to closed substructures occurring within the strand s.

The MFE in such a model can be obtained through a minor update of the dynamic programming scheme described in the previous section, where:

- Individual base pairs no longer contribute to the free-energy individually, yet still contribute as constitutive of loops;
- Since interaction loops are characterized by a pair of directly nested base pairs, each involving two strands, we duplicate both matrices M and $\overline{M}$ to indicate the presence ($M^{\mathbb{S}}$ and $\overline{M}^{\mathbb{S}}$) or absence ($M^{\varnothing}$ and $\overline{M}^{\varnothing}$) of an enclosing base pair (s_{i-1}, r_{j+1});
- The rules of $M^{\mathbb{S}}$ need to be adapted to explicitly make the first base pair (i', j') such that $i \leq i'$ and $j' \leq j$, available for scoring. If no such base pair exist, then s needs to be entirely consumed, and a new strand be allocated by subsequent calls;
- $\overline{M}^{\mathbb{S}}$ and $\overline{M}^{\varnothing}$ essentially act as above, but also respectively propagate the presence/absence of an enclosing base pair;
- The MFE contribution of a substructure formed within a region of a strand s is now denoted as $M^{\mathbb{S}}_s[i,j]$ if enclosed by a base pair (s_i, s_j), or $M^{\varnothing}_s[i,j]$ otherwise. In the Turner energy model, those values are classically computed in time $\mathcal{O}(|s|^4)$, *e.g.* using the Zuker/Stiegler DP scheme [8].

$$M^{\varnothing}_{s_i,r_j,m,c} = \min \begin{cases} \overline{M}^{\varnothing}_{s_{i+1},r_j,m,c} \\ \min_k M^{\mathbb{S}}_s[i,k] + \overline{M}^{\varnothing}_{s_{k+1},r_j,m,c} \\ \min\limits_{\substack{t\in R\\ 1\leq k\leq |t|\\ m'+m''=m-1}} \overline{M}^{\mathbb{S}}_{s_{i+1},t_{k-1},m',0} + \overline{M}^{\varnothing}_{t_{k+1},r_j,m'',c} \\ \min_k \overline{M}^{\mathbb{S}}_{s_{i+1},r_{k-1},m,c} + M^{\varnothing}_r[k,j] \end{cases} \tag{9}$$

$$M^{\mathbb{S}}_{s_i,r_j,m,c} = \min \begin{cases} \overline{M}^{\varnothing}_{s_{|s|+1},r_j,m,c} \\ \min_{i\leq i'<j'\leq |s|} M^{\mathbb{S}}_s[i',j'] + \overline{M}_{s_{j'+1},r_j,m,c} \\ \min\limits_{\substack{t\in R\\ i\leq i'\leq |s|, 1\leq j'\leq |t|\\ m'+m''=m-1}} \overline{M}^{\mathbb{S}}_{s_{i'+1},t_{j'-1},m',0} + \overline{M}^{\varnothing}_{t_{j'+1},r_j,m'',c} \\ \min_{\substack{i\leq i'\leq |s|\\ 1\leq j'\leq j}} \Delta G(s_{i-1}, s_{i'}, r_{j'}, r_{j+1}) + \overline{M}^{\mathbb{S}}_{s_{i'+1},r_{j'-1},m,c} + M^{\varnothing}_r[j'+1,j] \end{cases} \tag{10}$$

$$\overline{M}^{\xi\in\{\varnothing,\mathbb{S}\}}_{s_i,r_j,m,c} = \begin{cases} \left.\begin{cases} \min_{t\in R} \overline{M}^{\xi}_{t_1,r_j,m-1,1} & \text{if } c=0 \text{ and } m>0 \\ M^{\xi}_r[1,j] & \text{if } c=0 \text{ and } m=0 \\ \infty & \text{if } c=1 \end{cases}\right| & \text{if } i>|s| \\ \left.\begin{cases} \min_{t\in R} \overline{M}^{\xi}_{s_i,t_{|t|},m-1,1} & \text{if } c=0 \text{ and } m>0 \\ M^{\xi}_s[i,|s|] & \text{if } c=0 \text{ and } m=0 \\ \infty & \text{if } c=1 \end{cases}\right| & \text{else if } j<1 \\ M^{\xi}_{s_i,r_j,m,c} & \text{otherwise} \end{cases} \tag{11}$$

$$E^*(R,m) = \min_{s,r\in R} M^{\varnothing}_{m-2,s_1,r_{|r|},1} \tag{12}$$

The complexity of this algorithm is increased to $\mathcal{O}(n^4 \cdot m^2 \cdot p^3)$ with $n := \max_{s\in R} |s|$.

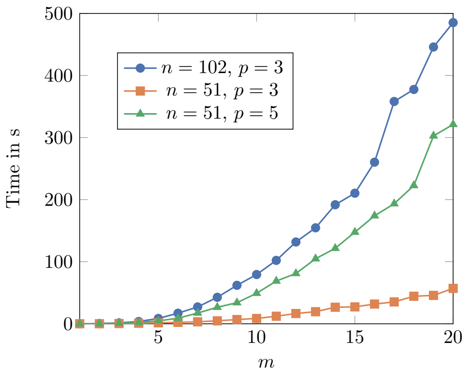


Fig. 9 Empirical running time for increasing m and various values for n and p.

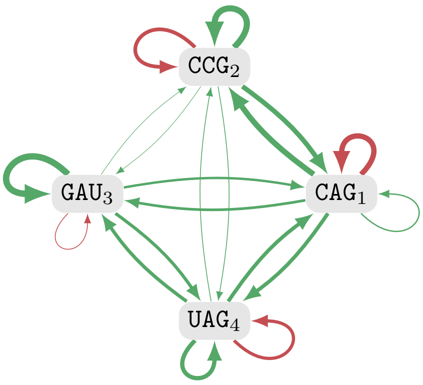


Fig. 10 Affinity of triplet repeats (%BPs in soup model, coded by line thickness) for external (green) and internal (red) interactions.

5 Empirical studies

The goal of this section is to show how the algorithms described in the previous section can be used to answer biologically relevant questions regarding triplet repeats. We implemented the algorithm described in section 4.5, which hereunder we call SoupFold, as well as its partition function equivalent, together with a (stochastic) backtracking procedure. The source code to reproduce our analyses is available at:

https://github.com/kimonboehmer/soupfold/

Since we only limit the number of interacting strands but not their size, without further restrictions, the program would prefer large strands since they usually give more base pairs. To counteract this effect, we introduce a penalty on the length of a strand. Note that one could also set a maximum length of the concatenated sequence, as described in remark 5. The empirically observed running time matches the theoretical running time well, as can be seen in fig. 9.

Regarding the stochastic backtracking, we must account for the overcounting of rotationally asymmetric secondary structures as well as for the overcounting because of the positioning of different connected components. We address these two issues by rejection sampling.

In theory, it would also be necessary to adjust the overcounting correction for rotationally symmetric structures (because they are overcounted less often) but our experiments showed that the observed probability of encountering such rotational symmetries is 0 for triplets with 15 repeats or more. Thus, for efficiency reasons, we do not include this case in our rejection sampling, arguing that the changes to the probability would be too small to observe.

5.1 Homogeneous triplet soup

We first consider the case where all strands are of the same pattern. The MFE of a soup of homogenous triplets behaves canonically, in the sense that all folding patterns

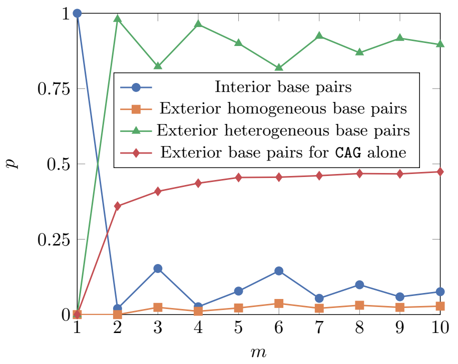


Fig. 11 Probability p that a certain type of base pair is observed for increasing #strands m, either in a soup $\{\texttt{CAU}_{20}, \texttt{GGG}_{20}\}$, or for $\texttt{CAG}_{20}$.

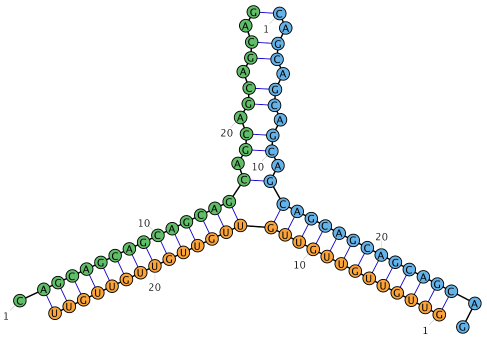


Fig. 12 Exemplary MFE structure for strand pool $\{(\texttt{GUU})^9, (\texttt{CAG})^9, (\texttt{ACG})^9\}$ computed by SoupFold with $m = 3$ (RiboSketch [23]).

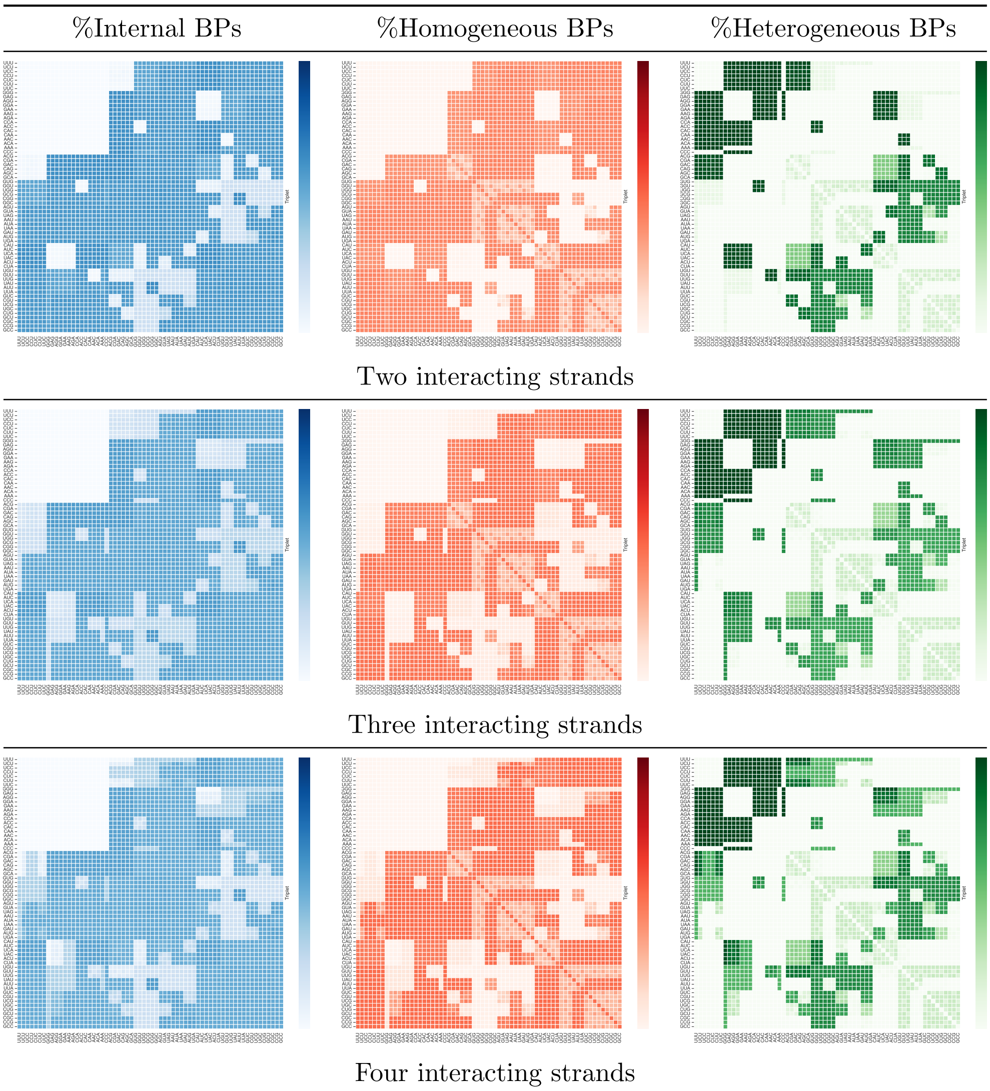
have almost identical MFE structures (as can be expected, considering our results on single-strand TR in section 3). Furthermore, we observed that the number of base pairs increases canonically with the sequence length and with the number of interacting strands (except for the case of only one strand, where we loose one base pair due to a hairpin loop).

5.2 Heterogeneous triplet soup

More interesting observations can be made in a heterogeneous pool. We can observe that different TR pattern strands can achieve more base pairs than the theoretical upper bound for a homogeneous strand pool (see fig. 12).

In order to assess the capability of different strand soups to form droplets, we want to determine the probability of a base pair in the Boltzmann ensemble being between two strands (*exterior*) as opposed to folding (*interior*). If the strand soup consists only of triplets of one pattern, all exterior base pairs will be *homogeneous*, as opposed to *heterogeneous* for an interaction of two strands of different patterns. In the homogeneous case, we can observe an increase of exterior base pairs for increasing number of interacting strands m, as presented by the red line in fig. 11. The probabilities in a setting with strands of different patterns are much richer and less canonical, as can be seen at the example of the interaction of CAU and GGG, presented by the other lines in fig. 11. These probabilities highly depend on the number of strands, and only start to “converge” with quite high values of m.

To obtain a broader picture, we performed stochastic backtracking on all possible 4^6 pairs of triplet repeat patterns $\{TVW, XYZ\}$ as strand sets, with m between 2 and 5, and computed the probability of a base pair being interior, exterior-homogeneous or exterior-heterogeneous. fig. 13 shows the probability of interior, exterior-homogeneous and exterior-heterogeneous base pairs for all pairs of TR, from $m = 2$ to 4. We can observe that the probabilities vary a lot and highly depend on the interacting triplets. Some pairs of triplets do not form base pairs at all, in which case all three corresponding tables have a blank entry. Usually, internal and exterior-homogeneous base pairs behave similarly. One can also see that the probability of heterogeneous



%Internal BPs

%Homogeneous BPs

%Heterogeneous BPs

Two interacting strands

Three interacting strands

Four interacting strands

Fig. 13 Interaction profiles for pairs of triplets in the heterogeneous soup model.

base pairs slightly increases with increasing m. On the other hand, the probability of observing interior base pairs is slightly decreasing.

From a synthetic biology perspective, some triplet repeats aggregate and form a Liquid-Liquid Phase Separation, which can be used to isolate subprocesses, thereby implementing a notion of orthogonality. In order to maximize the number of independent tasks being performed by a modified bacteria, it would then be desirable to find a large number of triplet repeat patterns such that the probability of heterogeneous base pairs is low.

For that, we can model the patterns as vertices of a graph and draw an edge if the heterogeneous base pair probability between two patterns for $m = 5$ is high (we set the threshold to 0.175). We then want to determine a maximum independent set (MIS), i.e. the largest set of triplets that do not have a high probability of interacting pairwise with each other. We used an exact solver [24] to obtain a MIS of size 4, namely `CAG, CCG, GAU, UAG`.

We then executed our algorithm on these triplet patterns as strand soup, and could indeed observe that the probability of exterior heterogeneous base pairs is clearly below 0.2 for values of m between 1 and 10. In fig. 10, we depict the number of base pairs that are between two types of strands, for $m = 5$ and our four independent TR patterns as strand soup. We added a bonus to the appearance of strands, to ensure that all strands of the soup appear equally often in the constructed structures. We observe that for three of the four triplets, for exterior base pairs, the most likely interacting strand is of the same type.

6 Conclusion and discussion

In this work, we investigated the algorithmic aspects of folding and interactions of triplet repeat RNA sequences, while also revisiting the general (non-triplet) setting in the interaction setting. For the folding of individual triplets, we found that their repetitive structure allows us to immediately characterize the MFE and partition function value, without the need of a more time-consuming DP approach. For interactions of RNA sequences, we exhibited a new algorithm with improved running time that avoids the factorial-time iteration over all permutations and acts as a foundation for the design of specialized algorithms, as the XP algorithm for triplet repeats. For the "strand soup" setting, we derived a polynomial-time algorithm and demonstrated possible uses for experiments regarding triplet repeats.

For future work, it is desirable to describe in detail how to extend the MFE Strand Interaction algorithm to the full thermodynamic setting considered in [13]. While the extension to the Turner model does not pose any algorithmic challenges, it would be interesting to implement a variant of the inside/outside algorithm to compute exactly base-pairing probabilities and other expected values of additive properties. Finally, the joint conformation space explored in this work is heavily restricted by the existence of a non-crossing strand ordering. More complex conformational spaces could be captured by using DP approaches akin to those used to include pseudoknots in RNA structure prediction.

Declarations

- Funding: ANR-funded SYNORG project (ANR-23-CE44-0027) and European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant (agreement no. 101029676)

- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use): Not applicable
- Ethics approval and consent to participate: Not applicable
- Consent for publication: Not applicable

- Data availability: Not applicable
- Materials availability: Not applicable
- Code availability: https://github.com/kimonboehmer/soupfold/
- Author contribution: All authors designed the study. KB wrote the proofs and contributed the proof of concept implementation. KB and YP developed the DP equations in the manuscript and the formal grammar for the linear time partition function computation. All authors contributed to writing and rereading of the manuscript.

References

[1] Isiktas, A.U., Eshov, A., Yang, S., Guo, J.U.: Systematic generation and imaging of tandem repeats reveal base-pairing properties that promote RNA aggregation. Cell Reports Methods **2**(11) (2022)

[2] Guo, H., Ryan, J.C., Song, X., Mallet, A., Zhang, M., Pabst, V., Decrulle, A.L., Ejsmont, P., Wintermute, E.H., Lindner, A.B.: Spatial engineering of E. coli with addressable phase-separated RNAs. Cell **185**(20), 3823–3837 (2022)

[3] Srinivasan, S.R., Melo de Gusmao, C., Korecka, J.A., Khurana, V.: Chapter 18 - repeat expansion disorders. In: Zigmond, M.J., Wiley, C.A., Chesselet, M.-F. (eds.) Neurobiology of Brain Disorders (Second Edition), Second edition edn., pp. 293–312. Academic Press, ??? (2023). https://doi.org/10.1016/B978-0-323-85654-6.00048-4 . https://www.sciencedirect.com/science/article/pii/B9780323856546000484

[4] Kurokawa, R., Kurokawa, M., Mitsutake, A., Nakaya, M., Baba, A., Nakata, Y., Moritani, T., Abe, O.: Clinical and neuroimaging review of triplet repeat diseases. Japanese Journal of Radiology **41**(2), 115–130 (2023)

[5] Maity, H., Nguyen, H.T., Hori, N., Thirumalai, D.: Odd–even disparity in the population of slipped hairpins in rna repeat sequences with implications for phase separation. Proceedings of the National Academy of Sciences **120**(24), 2301409120 (2023)

[6] Aierken, D., Joseph, J.A.: Accelerated simulations of rna phase separation: a systematic study of non-redundant tandem repeats. bioRxiv, 2023–12 (2023)

[7] Nussinov, R., Jacobson, A.B.: Fast algorithm for predicting the secondary structure of single-stranded rna. Proceedings of the National Academy of Sciences **77**(11), 6309–6313 (1980) https://doi.org/10.1073/pnas.77.11.6309 https://www.pnas.org/doi/pdf/10.1073/pnas.77.11.6309

[8] Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research **9**(1), 133–148 (1981) https://doi.org/10.1093/nar/9.1.133 https://academic.oup.com/nar/article-pdf/9/1/133/6201945/9-1-133.pdf

[9] Huang, L., Zhang, H., Deng, D., Zhao, K., Liu, K., Hendrix, D.A., Mathews, D.H.: LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. Bioinformatics **35**(14), 295–304 (2019) https://doi.org/10.1093/bioinformatics/btz375 https://academic.oup.com/bioinformatics/article-pdf/35/14/i295/50721438/bioinformatics_35_14_i295.pdf

[10] Bringmann, K., Grandoni, F., Saha, B., Williams, V.V.: Truly subcubic algorithms for language edit distance and rna folding via fast bounded-difference min-plus product. SIAM Journal on Computing **48**(2), 481–512 (2019) https://doi.org/10.1137/17M112720X https://doi.org/10.1137/17M112720X

[11] Chang, Y.-J.: Hardness of rna folding problem with four symbols. Theoretical Computer Science **757**, 11–26 (2019) https://doi.org/10.1016/j.tcs.2018.07.010

[12] Alkan, C., Karakoc, E., Nadeau, J.H., Sahinalp, S.C., Zhang, K.: Rna–rna interaction prediction and antisense rna target search. Journal of Computational Biology **13**(2), 267–282 (2006)

[13] Dirks, R.M., Bois, J.S., Schaeffer, J.M., Winfree, E., Pierce, N.A.: Thermodynamic analysis of interacting nucleic acid strands. SIAM review **49**(1), 65–88 (2007)

[14] Condon, A., Hajiaghayi, M., Thachuk, C.: Predicting minimum free energy structures of multi-stranded nucleic acid complexes is apx-hard. In: 27th International Conference on DNA Computing and Molecular Programming (DNA 27)(2021) (2021). Schloss-Dagstuhl-Leibniz Zentrum für Informatik

[15] Demaine, E.D., Gomez, T., Grizzell, E., Hecher, M., Lynch, J., Schweller, R., Shalaby, A., Woods, D.: Domain-based nucleic-acid minimum free energy: Algorithmic hardness and parameterized bounds. In: 30th International Conference on DNA Computing and Molecular Programming (DNA 30)(2024) (2024). Schloss Dagstuhl–Leibniz-Zentrum für Informatik

[16] Turner, D.H., Mathews, D.H.: NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Research **38**(suppl_1), 280–282 (2009) https://doi.org/10.1093/nar/gkp892

[17] Denise, A., Ponty, Y., Termier, M.: Controlled non-uniform random generation of decomposable structures. Theoretical Computer Science **411**(40), 3527–3552 (2010) https://doi.org/10.1016/j.tcs.2010.05.010

[18] Lipshitz, L.: D-finite power series. Journal of Algebra **122**(2), 353–373 (1989)

[19] Bostan, A., Chyzak, F., Lecerf, G.e., Salvy, B., Schost, E.: Differential equations for algebraic functions. In: Brown, C.W. (ed.) ISSAC'07: Proceedings of the 2007 International Symposium on Symbolic and Algebraic Computation, pp. 25–32. ACM Press, ??? (2007). https://doi.org/10.1145/1277548.1277553

[20] Salvy, B., Zimmerman, P.: GFUN: a Maple package for the manipulation of generating and holonomic functions in one variable. ACM Transactions on Mathematical Software **20**(2), 163–177 (1994)

[21] McDiarmid, C.: Pattern minimisation in cutting stock problems. Discrete applied mathematics **98**(1-2), 121–130 (1999)

[22] Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L.: ViennaRNA package 2.0. Algorithms for molecular biology : AMB **6**, 26 (2011) https://doi.org/10.1186/1748-7188-6-26

[23] Lu, J.S., Bindewald, E., Kasprzak, W.K., Shapiro, B.A.: RiboSketch: versatile visualization of multi-stranded RNA and DNA secondary structure. Bioinformatics **34**(24), 4297–4299 (2018) https://doi.org/10.1093/bioinformatics/bty468 https://academic.oup.com/bioinformatics/article-pdf/34/24/4297/48919841/bioinformatics_34_24_4297.pdf

[24] Hauser, F., Ermel, F., Boehmer, K.: Clique Cover Based Vertex Cover Solver. GitHub (2024)