



HAL
open science

RNA Triplet Repeats: Improved Algorithms for Structure Prediction and Interactions

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

► **To cite this version:**

Kimon Boehmer, Sarah J Berkemer, Sebastian Will, Yann Ponty. RNA Triplet Repeats: Improved Algorithms for Structure Prediction and Interactions. 2024. hal-04589903v1

HAL Id: hal-04589903

<https://hal.science/hal-04589903v1>

Preprint submitted on 27 May 2024 (v1), last revised 5 Jun 2024 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



RNA Triplet Repeats: Improved Algorithms for Structure Prediction and Interactions

Kimon Boehmer 

Laboratoire d'Informatique de l'Ecole Polytechnique (LIX UMR 7161), Institut Polytechnique de Paris, France

Sarah J. Berkemer  

Laboratoire d'Informatique de l'Ecole Polytechnique (LIX UMR 7161), Institut Polytechnique de Paris, France

Sebastian Will  

Laboratoire d'Informatique de l'Ecole Polytechnique (LIX UMR 7161), Institut Polytechnique de Paris, France

Yann Ponty¹  

Laboratoire d'Informatique de l'Ecole Polytechnique (LIX UMR 7161), Institut Polytechnique de Paris, France

Abstract

RNAs composed of Triplet Repeats (TR) have recently attracted much attention in the field of synthetic biology. We study the minimum 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 that avoids the previously-known factorial-time iteration over all permutations. 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 TRs 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 TRs.

2012 ACM Subject Classification [Replace ccsdesc macro with valid one](#)

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

Digital Object Identifier 10.4230/LIPIcs.WABI.2024.XXX

Funding *Kimon Boehmer*: Supported by ANR-funded SYNORG project (PI S.J. Berkemer)

¹ To whom correspondence should be addressed



1 Introduction

32 RNAs composed of Triplet Repeats (TR) have attracted much attention, and harbour
33 promises in the field of synthetic biology, due to their demonstrated capacity to self-assemble
34 into droplets [12, 9]. Those can in turn be used to compartmentalize cellular processes,
35 thereby creating a “clean room”, free of the natural cellular clutter, where synthetic circuits
36 can be executed without interference. The exact process underlying this phenomena is
37 still the object of ongoing investigations, but it is hypothesized that repetitive RNAs may
38 induce Liquid-Liquid Phase separation mediated by unstable/transient structures. Repetitive
39 RNAs are also found at the origin of severe Neurological Triplet Expansion Diseases (TED),
40 including Friedreich ataxia [20] and Triplet Repeat Diseases (TRD) such as Huntington
41 disease [13]. For multiple TEDs and TRDs, overly expanded RNAs have been observed
42 to aggregate into RNA foci, leading to a sequestration of RNA binding proteins. Local
43 secondary structures and interactions are impacted by the repeat, and generally believed to
44 contribute to the pathogenicity and treatment efficiency. To study those phenomena *in silico*,
45 and in particular the impact of the repeated motif and number of repeats on aggregates, one
46 needs to predict the MFE structure of potentially large RNAs, and many-body interactions.
47 Recently, coarse-grained simulations showed a disparity between odd or even numbers of
48 triplet repeats [16] as well as extensions to quadruplet and non-redundant tandem repeats [1].

49 RNA folding by energy minimization is a classic algorithmic problem in Bioinformatics,
50 historically solved in time $\Omega(n^3)$ using dynamic programming [18, 22]. Despite recent
51 misleading suggestions of linear-time alternatives [11], the best algorithm to date to solve
52 energy minimization has runtime $\mathcal{O}(n^{2.8603})$ [4], and both its implementation and extension
53 beyond a base-pair maximization setting represent considerable challenges. Prior works
54 have also investigated conditional lower bounds, and found that the existence of a $\mathcal{O}(n^{2-\epsilon})$
55 algorithm would refute the Strong Exponential Time Hypothesis (SETH) [4]. Meanwhile, an
56 $\mathcal{O}(n^{\omega-\epsilon})$ algorithm would disprove the k -clique conjecture, with $\omega < 2.373$ being the matrix
57 multiplication exponent [4, 5].

58 RNA-RNA interaction prediction represents an equally relevant, yet computationally
59 substantially more involved algorithmic problem. For a fixed number of interacting strands,
60 polynomial-time algorithms have been proposed. For example, by excluding so-called zig-zag
61 joint conformations, [2] proposed a polynomial-time algorithm for the interaction of two
62 strands, while also showing **NP**-hardness for the case where we include these conformations.
63 In the unbounded case, [8] gave a factorial-time algorithm for computing the partition
64 function over multiple strands. Additionally, it was shown that energy minimization in this
65 setting is **APX**-hard (and by that **NP**-hard) [6], even for a very simple energy model.

66 Contributions In this work, we show that the repeated nature of RNA can be exploited
67 to obtain substantially improved algorithms for several problems. First, we show that the
68 Minimum Free-Energy of a triplet-repeat RNA can be predicted in linear time (in the size
69 of the binary encoded triplet sequence), both with respect to base pair maximization and
70 Turner energy model, and are realized by either the open chain or a single helix.
71 We then consider the interaction of multiple triplet repeats and propose improved algorithms
72 for the general (non-triplet) case as well as algorithms specifically for the interaction of TR.
73 For the latter case, we show **NP**-hardness in a reasonable energy model.

74 2 Definitions and Problems Statement

75 2.1 Definitions

76 **RNA sequence and structure(s).** An RNA sequence (or just sequence) is a word
 77 $s \in \{A, C, G, U\}^+$. The length of s is denoted by $|s|$ and the i -th position of s by s_i . A
 78 position on a sequence is also called a base. We associate to each position s_i its letter by
 79 $l(s_i)$. We define $P := \{\{C, G\}, \{A, U\}, \{G, U\}\}$. A pseudoknot-free secondary structure S is a
 80 set of pairs of bases, hereunder called base pairs, such that:

- 81 ■ for all $\{s_i, s_j\} \in S$, $\{l(s_i), l(s_j)\} \in P$;
- 82 ■ each base is involved in at most one base pair, i.e. for all bases i , $|\{p \in S \mid i \in p\}| \leq 1$;
- 83 ■ S does not contain two base pairs $\{i, j\}, \{k, \ell\}$ with $i < k < j < \ell$;
- 84 ■ each base pair encloses at least θ bases, that is, if $\{i, j\} \in S$, then $j - i > \theta$. We usually
 85 call θ the minimal base pair span, and use $\theta = 3$ unless explicit specified.

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

88 We associate each secondary structure $S \in \Omega$ to a free energy, according to an *energy*
 89 *model* $E : \{A, C, G, U\}^+ \times \Omega \rightarrow \mathbb{R}$. In the base pair model E_{bp} , we simply count the number
 90 of base pairs in S , and set $E_{\text{bp}}(w, S) = -|S|$. More advanced energy models reason about
 91 the free energy introduced by motifs occurring in the secondary structure, such as the loops
 92 considered by the Turner nearest-neighbor model [21]. .

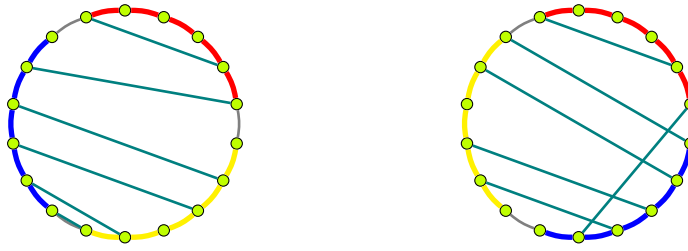
93 **Interactions.** A strand is an RNA sequence which is identified as a unique object in a set.
 94 In other words, in a set of strands R , we can have two strands $s \neq r$ that consist of the
 95 same sequences, that is $s_i = r_i$ for all $i \in \{1, \dots, |s| = |r|\}$, but still are different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

97 A *circular permutation* π of a strand set R is a permutation of $m - 1$ elements of R . We
 98 write Π_R (or just Π if clear from the context) for the set of all circular permutations over R .
 99 To express that strands s^1, \dots, s^m appear in that order in a circular permutation π , we write
 100 $(s^1, \dots, s^m) \in O_\pi$. Similarly, we write $(s_1^1, \dots, s_{i_m}^m) \in O_\pi$ to denote that the bases appear in
 101 this order in π . A *secondary structure* S of a strand set R is a set of base pairs $\{s_i, r_j\}$ from
 102 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
 103 and each intra-strand base pair encloses at least θ bases, i.e. $\{s_i, s_j\} \in S \rightarrow j - i > \theta$.

104 The *polymer graph* of a secondary structure S and a circular permutation π on R is
 105 a graph $G = (V, E)$ with $V := \{s_i \mid s \in R, 1 \leq i \leq |s|\}$ and $E := S \cup \{\{s_i, s_{i+1}\} \mid s \in$
 106 $R, 1 \leq i < |s|\} \cup C := \{\{s_{|s|}, r_1\} \mid r \text{ follows } s \text{ in } \pi\}$. The edges $E - S$ are drawn in a cycle
 107 (naturally induced by the circular permutation), while the edges in S are drawn as straight
 108 lines between the bases. Two strands s, r are **connected** if there is a path from s_1 to r_1
 109 that does not use edges from C . A secondary structure is **connected** if all of its strands
 110 are connected. Examples for the polymer graphs of a single secondary structure under two
 111 different circular permutations can be found in Figure 1.

112 A secondary structure S is called *pseudoknot-free* if there is a circular permutation π
 113 such that there are no crossing lines in the polymer graph, or formally, there are no two base
 114 pairs $\{s_i, t_k\}, \{u_\ell, r_j\}$ with $(s_i, u_\ell, t_k, r_j) \in O_\pi$.

115 As for the folding, we associate to each $S \in \Omega(R)$ a free energy value. In the base pair
 116 model, additionally to counting the number p of base pairs, we also add a strand association
 117 penalty K_{assoc} for each of the $(m - \ell)$ strand associations, where ℓ is the number of connected
 118 components (*complexes*) in the polymer graph. Thus, the free energy of a secondary structure
 119 $S \in \Omega$ in this simple energy model is defined as $E(R, S) = -p + (m - \ell)K_{\text{assoc}}$. We may also
 120 require that all strands are connected; in that case, the strand association penalty is obsolete.



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

121 2.2 Computational problems

122 For a single strand, the two classical historical problems are:

MINIMUM FREE ENERGY (MFE) UNDER ENERGY MODEL E

Input: A sequence s

Output: Minimum free-energy $\min_{S \in \Omega(s)} E(s, S)$

PARTITION FUNCTION UNDER ENERGY MODEL E

Input: A sequence s and a positive temperature T

Output: Partition function $\mathcal{Z}_s := \sum_{S \in \Omega(s)} \exp\left\{\frac{-E(s, S)}{kT}\right\}$

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

124 In the multi-strand setting, we focus on energy minimization. In [8], the authors adopt a
 125 thermodynamic perspective on the free energy of a secondary structure over multiple strands,
 126 such that potential rotational symmetries require an adjustment of the computed value. We
 127 focus on a more algorithmic perspective, where all rotationally symmetric structures are
 128 elements of a search space, and a simple base pair energy model. In our main algorithmic
 129 problem of interest, we are given a set of strands and are looking for the minimum free energy
 130 of the secondary structure over these strands:

MFE STRAND INTERACTION

Input: Set of strands $R_0 = \{r_1, \dots, r_m\}$

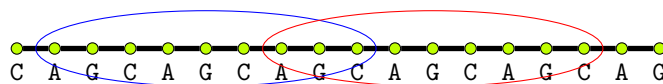
Output: $\min_{S \in \Omega(R_0)} E(R_0, S)$

131 We also consider a slightly different setting, where the number of occurrences of each
 132 triplet/strand is unconstrained beyond the total number m of interacting strands. This
 133 allows to studies situations where the strands concentrations are in excess, so that sequences
 134 can be locally seen as infinitely available often within a set (or “soup”) R of strands. We are
 135 then given the number of interacting strands m and look for the best secondary structure
 136 over m strands that all appear in R . More formally:

MFE STRAND SOUP INTERACTION

Input: Set of sequences $R = \{r_1, \dots, r_p\}$, $m \in \mathbb{N}$ encoded in unary

Output: $\min_{t_1, \dots, t_m: t_i \in R} \min_{S \in \Omega(\{t_1, \dots, t_m\})} E(\{t_1, \dots, t_m\}, S)$



■ **Figure 2** The blue and red region of the TR sequence are identical.

137 2.3 Triplet repeats RNAs and their properties

138 **Triplet repeat RNAs (TR).** Of special interest to us are RNA sequences that are composed
 139 of *triplet repeats* (TR), that is, they have the form $(X \cdot Y \cdot Z)^k$ for $X, Y, Z \in \{A, C, G, U\}$
 140 and $k \in \mathbb{N}^+$. We will describe how we can improve the general algorithms for the above
 141 computational problems in the case of TR.

142 An algorithmically convenient property about a region $[s_i, s_j]$ of a triplet repeat sequence
 143 is the following:

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

$$145 \quad [s_i, s_j] = [s_{i \bmod 3}, s_{j - (i - i \bmod 3)}].$$

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

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

158 Looking into more structural properties of triplet repeats, we can observe that, since each
 159 base repeats after two other bases, we cannot have exactly two enclosed base pairs at any
 160 point, otherwise we would have a base pair between two bases with the same label. Thus,
 161 requiring two ($\theta = 2$) or three ($\theta = 3$) enclosed bases between any base pair is equivalent:

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

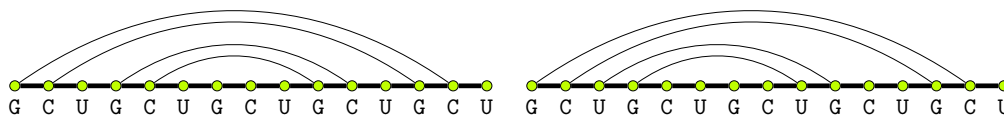
164 3 Single-Stranded Triplet Repeats

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

168 3.1 Linear time solution for base pair maximization

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

171 We can first prove an upper bound on the number of base pairs in a TR secondary
 172 structure:



■ **Figure 3** Two different optimal secondary structures for GCU_5 .

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

175 **Proof.** For any $X, Y, Z \in \{A, C, G, U\}$, there are $V, W \in \{X, Y, Z\}$ with $\{V, W\} \notin P$ (since
 176 the graph which represents the possible letter pairings does not contain triangles). Without
 177 loss of generality, let us assume these are X and Y . Each X and each Y must thus be paired
 178 to a Z or be unpaired. Due to the fact that any non-empty secondary structure has an
 179 innermost base pair which must respect the minimum base pair span θ , at least $\lfloor \frac{\theta+1}{3} \rfloor$ Z
 180 bases will remain unpaired (the $+1$ comes from Observation 2). It follows that there are
 181 exactly $k - \lfloor \frac{\theta+1}{3} \rfloor$ pairable Z -bases. Since every base pair must involve a Z base, the upper
 182 bound follows. ◀

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

187 ▶ **Lemma 4.** For $\theta \in \{0, 1\}$ and $k > 1$, we always have $E(s, S) = k$ for any secondary
 188 structure S over a folding sequence $s = (XYZ)^k$.

189 **Proof.** If $\{X, Z\} \in P$, connect X and Z in each triplet. Else, connect the outermost pair
 190 (say without loss of generality $\{X, Y\}$). We obtain the inner sequence $(YZX)^{k-1}$ (with
 191 $k-1 > 0$) and we can proceed as above since $\{Y, X\} \in P$. ◀

192 For the more natural case $\theta > 1$, the upper bound from Lemma 3 is not always tight. The
 193 next lemma exactly specifies the MFE and its structure:

194 ▶ **Lemma 5.** Let $\theta > 1$. The minimum MFE structure of a folding sequence $(XYZ)^k$ has
 195 value

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

197 ■ $k - \lfloor \frac{\theta+1}{3} \rfloor$, otherwise

198 Furthermore, a minimum MFE structure is obtained by choosing a letter pair and greedily
 199 stacking base pairs of this letter pair from the outermost to the innermost base. If both
 200 $\{X, Z\} \in P$ and one of $\{X, Y\}$ and $\{Y, Z\} \in P$, we choose $\{X, Z\}$ if $(\theta + 3k) \equiv_6 4$ and the
 201 letters of other base pair if $(\theta + 3k) \equiv_6 1$; otherwise, we choose the letters of an arbitrary
 202 base pair.

203 The proof of this lemma involves many case distinctions and can be found in the appendix.
 204 Setting $\theta = 3$, we get the following corollary:

205 ▶ **Corollary 6.** In the base pair maximization model, if $\theta = 3$, the MFE structure of any TR
 206 sequence $(XYZ)^k$ has $k - 1$ base pairs.

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

209 ► **Theorem 7.** *MFE prediction for compactly encoded TR in the base pair maximization*
 210 *model can be solved in linear time.*

211 ► **Remark 8.** The optimal secondary structure does not need to be unique. In particular, for
 212 a simple energy model, the number of optimal secondary structures for triplet repeats can
 213 even be exponential. For example, consider the sequence $(\text{GCU})^k$ as illustrated in Figure 3.
 214 When constructing the base pairs from outside to inside, in every step, we can choose whether
 215 we add the base pairs G-U, U-G or the base pairs G-C, C-G. This decision can be repeated
 216 $\lfloor \frac{k}{2} \rfloor - 1$ times (assuming $\theta = 3$), giving $\Omega(2^{k/2})$ different optimal secondary structures.

217 3.2 Minimum Free-energy in the Turner model

218 Let us now consider the Turner model. We will show that the optimal structures obtained for
 219 BP maximization remains optimal for the Turner nearest neighbor model under reasonable
 220 assumptions, satisfied by current versions of the model [21].

221 We first focus on showing the absence of multiloops, *i.e.* structural motifs consisting
 222 of $B \geq 2$ branches, in the Turner MFE. Their free energy contribution is composed of an
 223 initiation penalty α , a value β for each branch, and an asymmetry penalty γ . The overall
 224 contribution of a multiloop S is given by

$$225 \quad E(s, S) = \alpha + \beta B + \gamma C + E_{\text{in}}$$

226 where E_{in} is the MFE of the interior secondary structure of the branches. Let $N :=$
 227 $\min_{V,W \in \{X,Y,Z\}: \{V,W\} \in P} E_{V,W}$ be the best contribution of a single base pair appearing in
 228 our triplet pattern.

229 ► **Lemma 9.** *Any Turner-MFE secondary structure S^* over $(XYZ)^k$ does not contain any*
 230 *multiloops, assuming $\beta \geq N, \alpha > -\beta, \gamma \geq 0$.*

231 **Proof.** By Corollary 6, each branch of k' repeats will not contribute more than $k' - 1$ base
 232 pairs. Thus, the number of base pairs in the interior of the multiloop is at most $k - B$. Let S
 233 be a multiloop secondary structure on region s and let S^* be a stacking on the same region.
 234 Their free energy values are related as follows:

$$235 \quad E(s, S) \geq \alpha + \beta B + \gamma C + (k - B)N \tag{1}$$

$$236 \quad > -\beta + \beta B + (k - B)N \tag{2}$$

$$237 \quad = kN + \beta(B - 1) - NB \tag{3}$$

$$238 \quad = (k - 1)N + \beta(B - 1) - N(B - 1) \tag{4}$$

$$239 \quad = (k - 1)N + (\beta - N)(B - 1) \tag{5}$$

$$240 \quad \geq (k - 1)N \tag{6}$$

$$241 \quad \geq E(s, S^*) \tag{7}$$

242 where (1) comes from our above observation, (2) from $\alpha > -\beta$ and $\gamma \geq 0$, (6) from $\beta \geq N$
 243 and $B \geq 2$ (by definition of a multiloop). For inequality (7), first notice that S^* contains
 244 $k - 1$ base pairs. Corollary 6. As noticed in Remark 8, we can choose which base pair is
 245 used in S^* without affecting the optimality. In particular, we can always choose the base
 246 pair consisting of the letters V, W that optimize their contribution, such that $E_{V,W} = N$.
 247 We get $E(s, S^*) \leq (k - 1)N$. ◀

248 ► **Remark 10.** The above assumptions are satisfied by the Turner 2004 energy model ($\alpha = 9.25$,
 249 $\beta = -0.63$, $\gamma = 0.91$ and $N \leq -0.93$) [21].

250 ► **Lemma 11.** *The exterior face of an MFE secondary structure for $(XYZ)^k$ is restricted to*
 251 *a single outermost base pair.*

252 **Proof.** From Corollary 6, we know that such a structure will always achieve $k - 1$ stacked
 253 base pairs. Assume that there are two outermost helices, of k_1 and k_2 repeats (notice that
 254 if one helix consists of k' repeats and one or two additional bases, without completing the
 255 $k' + 1$ -st repeat, this does not increase the number of base pairs) with $k_1 + k_2 = k$. Since
 256 $\theta = 3$, the MFE structure for the two subregions has at most $k_1 - 1$ and $k_2 - 1$ base pairs, so
 257 the total number of base pairs will be at most $k - 2$. ◀

258 By the above two lemmata, we can conclude that the MFE in the Turner model is also of
 259 the canonical form described in the BP maximization setting.

260 3.3 Linear-time computation of the partition function

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

264 Below is the context-free grammar for the pattern **CAG**:

$$\begin{array}{llll}
 265 & S_C^G \rightarrow & (\cdot_A S_G^C \cdot_A) & | (\cdot_A S_G^C \cdot_A) S_C^G & | \cdot_C \cdot_A S_G^C & | \cdot_C \cdot_A \cdot_G \\
 266 & S_C^C \rightarrow & (S_G^C) & | (S_G^C) \cdot_A S_G^C & | \cdot_G S_G^C & | \cdot_G \cdot_C \\
 267 & S_G^C \rightarrow & (S_G^C) \cdot_A \cdot_G & | (S_G^C) \cdot_A S_G^C & | \cdot_G S_G^C & \\
 268 & S_C^C \rightarrow & (\cdot_A S_G^C \cdot_A) \cdot_A & | (\cdot_A S_G^C \cdot_A) S_C^C & | \cdot_C \cdot_A S_G^C &
 \end{array}$$

269 Namely, the terminal S_C^G generates all secondary structures for the RNA sequence $(CAG)^k$
 270 for all $k > 0$, S_C^C the structures of $(GCA)^k GC$ for $k \geq 0$, S_G^C the structure of $G(CAG)^k$ for
 271 $k > 0$, and S_C^C corresponds to the pattern $(CAG)^k C$ for some $k > 0$.

272 Following standard methodologies in enumerative/analytic combinatorics [7], such a
 273 grammar can be generically translated into a system of functional equations involving
 274 weighted generated functions for each non-terminal:

$$\begin{array}{l}
 275 \quad S_C^G(z) = \beta z^4 S_G^C(z) + \beta z^4 S_G^C(z) S_G^C(z) + z^2 S_G^C(z) + z^3 \\
 276 \quad S_C^C(z) = \beta z^2 S_G^C(z) + \beta z^3 S_G^C(z) S_G^C(z) + z S_G^C(z) + z^2 \\
 277 \quad S_G^C(z) = \beta z^4 S_G^C(z) + \beta z^3 S_G^C(z) S_G^C(z) + z S_G^C(z) \\
 278 \quad S_C^C(z) = \beta z^3 S_G^C(z) + \beta z^2 S_G^C(z) S_G^C(z) + z^2 S_G^C(z)
 \end{array}$$

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

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

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 [14], which
 can be efficiently [3] and effectively computed [19]. We obtain an equation of the form:

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

279 where each P_i is a polynomial in k , and d is a constant. $\mathcal{Z}_{(CAG)^k}$ can then be computed
 280 using a linear number of arithmetic operations. The same result holds for other triplets and
 281 we obtain:

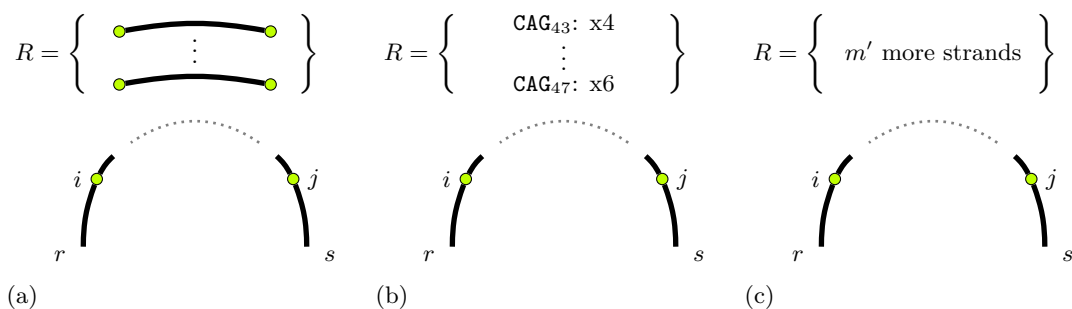


Figure 4 Visualization of the structures used to compute the MFE in the (a) general setting, (b) TR setting and (c) strand soup setting.

► **Theorem 12.** *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 [6], where the authors showed that the problem is **APX**-hard (and by that **NP**-hard) for the general (non-triplet) case. On the other hand, [8] gave a factorial-time algorithm for computing the partition function over multiple strands. In this section, we improve both results in the sense that on the one hand, we show that the problem is **NP**-hard in a reasonable energy model even if restricted to triplet repeats of one pattern, and on the other hand we give an exponential-time instead of factorial-time algorithm for the problem. However, notice that our exponential-time algorithm is designed for solving the MFE from an algorithmic perspective, as discussed in Section 2.2, and even though it can be translated to an algorithm for computing the partition function, adjustments for rotational symmetries must be made to obtain the same setting as Dirks *et al* [8].

4.1 General RNA-RNA interactions

The 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).

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, $1 \leq i \leq |s|$ the current position in s , $1 \leq j \leq |r|$ 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 Figure 4 (a). The main recurrences are as follows:

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

where

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

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 $i+1 \leq |s|$ and $j-1 \geq 1$, and introducing new strands in the other cases). The minimization requires some more detailed conditions which can be found in the appendix.

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 $s \in R$, and $M_{\emptyset,s_i,r_j} = 0$ for all s_i and r_j . The correctness of the algorithm and its running time are proven in the appendix. With n denoting the length of the longest strand sequence in R , we obtain:

► **Theorem 13.** *MFE STRAND INTERACTION can be solved in time $\mathcal{O}(3^m \cdot n^3)$.*

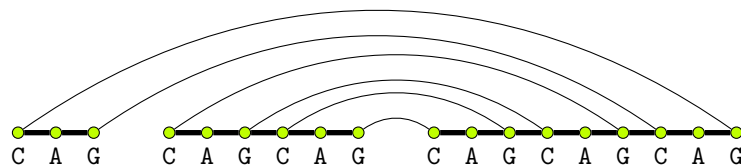
4.2 Strand interactions for triplet repeats

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 need to 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, \dots, a_p) where $a_i := |\{r \in R' \mid |r| = n_i\}|$ is the number of strands of size n_i in R . Then, each length has a number of occurrences s_i . An example is given in Figure 4 (b). Thus, the exponent will only depend on p , and using $n := \max_{r \in R} |r|$, we obtain the following result:

► **Theorem 14.** *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.



■ **Figure 5** Strands/optimal secondary structure corresponding to a valid summing triplet (1, 2, 3).

346 4.3 Computational hardness

347 In this subsection, we show that the parametrized approach seen before is the best we can
 348 hope for, and that, even for triplet repeats, the problem of deciding whether there is a
 349 secondary structure for R_0 with a free energy below a certain threshold t is **NP**-complete,
 350 for a reasonable energy model. Note that for the general (non-triplet) case, this has already
 351 been shown in [6]. Our result is surprising in the sense that the concatenation of TR strands
 352 always yields the same permutation, and the only additional difficulty compared to the
 353 single-stranded case arises from the fact that we do not know the indices of the strand
 354 borders.

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

364 Let us define the 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 $E(R, S) \leq t$?

365 Even if the following reduction does not work in the base pair maximization model, a DP
 366 algorithm for base pair maximization in this setting seems unlikely, as, under the assumption
 367 $\mathbf{P} \neq \mathbf{NP}$, one would not be able to generalize the algorithm to more complex energy models.
 368 We will show **NP**-hardness by reduction from the following problem:

SUMMING TRIPLETS

Input: list of distinct positive integers s_1, \dots, s_{3n} , encoded in unary

Output: Is there a partition of the input into triples (a_i, b_i, c_i) such that $a_i + b_i = c_i$?

369 This has been shown to be strongly NP-hard in [17]. We define $v := \sum_{i=1}^{3n} s_i$.
 370 The reduction is as follows: We create a strand $r_i := (CAG)^{s_i}$ for each integer s_i . Hence, we
 371 have $n = \frac{m}{3} = -E^{\text{bp}}$. We denote by R the set of strands. We set the target minimum free
 372 energy to $t := -(3v + 1)n$.
 373 Assume that there is a partition into summing triples. Our secondary structure is built such

374 that for each triple $a + b = c$, we add the base pairs

$$375 \quad (a_1, c_{|c|}), (a_3, c_{|c|-2}), (a_4, c_{|c|-3}), (a_6, c_{|c|-5}), \dots, (a_{|a|-2}, c_{|c|-|a|+3}), (a_{|a|}, c_{|c|-|a|+1}),$$

$$376 \quad (b_1, c_{|c|-|a|}), (b_3, c_{|c|-|a|-2}), \dots, (b_{|b|-2}, c_3), (b_{|b|}, c_1)$$

377 Note that all base pairs are labeled with $C - G$ or $G - C$. Figure 5 visualizes the secondary
378 structure for the exemplary triple $1 + 2 = 3$. We claim that S is unspseudoknotted for the
379 circular permutation $a_1 \cdot b_1 \cdot c_1 \dots a_n \cdot b_n \cdot c_n$ and that $E(R, S) = t$.

380 Since any two triples of strands are not connected, we have exactly n connected components.

381 Each connected component consists of one large stacked loop with innermost base pair
382 $(b_{|b|}, c_1)$ (i.e. we do not violate the constraint that every innermost base pair must include
383 three unpaired bases, because the base pair is inter-strand). Since $a + b = c$, the outermost
384 base pair is $(a_1, c_{|c|})$. There is no multiloop involved in S , so each triple (a_i, b_i, c_i) contributes
385 a free energy of $2|c| \cdot E^{\text{bp}} - K_{\text{assoc}} = -6n|c| - 1$. Since all triplets are correctly summing, we
386 have $\sum_{i=1}^n c_i = \frac{1}{2}v$. Thus indeed the minimum free energy is at most

$$387 \quad \sum_{i=1}^n -6n|c_i| - 1 = -6n \sum_{i=1}^n |c_i| - n = -6n \cdot \frac{1}{2}v - n = -3nv - n = t$$

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

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

390 **Proof.** First notice that in every valid secondary structure, all A bases remain unpaired
391 (since there are no U bases). There are $2v$ bases of C/G in total. Since we assumed that
392 one of them is unpaired, there can be at most $v - 1$ base pairs. We can have at most
393 $3n$ complexes, so the strand association penalty is reduced by at most $3n$. Thus we have
394 $E(R, S) \geq -3n(v - 1) - 3n = -3vn > -(3v + 1)n = t$. ◀

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

396 **Proof.** A hairpin loop must enclose at least three unpaired bases. Since in the CAG triplet
397 pattern any two consecutive bases involve at least one C or one G , we can apply Lemma 15
398 and conclude. ◀

399 Now assume for an arbitrary $S \in \Omega$ that $E(R, S) \leq t$. We first show that there must be
400 exactly n connected components, each with three strands. Assume that there is a connected
401 component with less than three strands. If it has only one strand, it must contain a hairpin
402 loop, and by Lemma 16, $E(R, S) > t$. If the complex contains two strands, first of all the
403 two strands have a different number of triplet repeats, since all s_i are distinct. This implies
404 that if the innermost loop is inter-strand (if it is intra-strand we again apply Lemma 16) and
405 has no multiloop, some G or C base must be unpaired (since base pairs can then only be
406 between the two strands, but one of the strands contains at least one G and one C base more
407 than the other). Then, by Lemma 15, $E(R, S) > t$. If it has a multiloop, there have to be
408 two innermost base pairs, one of which must be intra-strand, and we can apply Lemma 16.
409 Since we ruled out complexes of one or two strands and the total number of strand is divisible
410 by 3, we know that if there is a complex with four strands, our secondary structure will have
411 $< n$ connected components. Thus the best achievable score will be $-n + 1 - 3nv > t$. Hence,
412 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
413 with $|a_i| < |b_i| < |c_i|$. We claim that for all $i \in [n]$, $|a_i| + |b_i| = |c_i|$.

414 By contradiction, assume $|a_i| + |b_i| \neq |c_i|$ and first consider the case that there are no

415 multiloops. This implies that there is only one innermost base pair. If it is intra-strand, we
 416 obtain a contradiction to $E(R, S) \leq t$ by Lemma 16. If it is inter-strand, all remaining base
 417 pairs must be between one of two strands d, e on the one side and the third strand f on the
 418 other side. Since $|d| + |e| \neq |f|$ for any such partition, one of the two sides will be left with
 419 at least one unpaired G and one unpaired C , and we apply Lemma 15.

420 Now we consider the case of multiloops. Any multiloop where the cutpoint between the
 421 two recursive structures is on a strand border (and thus is not penalized) implies an
 422 innermost base pair in both recursive structures, and since by pigeonhole principle one of
 423 the two recursive structures is single-stranded, we have a hairpin loop and $E(R, S) > t$ by
 424 Lemma 16. In the other case, we have a multiloop penalty of $+1$. Thus we can lower bound
 425 $E(R, S) \geq -n - 3nv + 1 > t$.

426 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,
 427 each strand r corresponds to one integer $\frac{|r|}{3}$ in the set of integers of our original instance.
 428 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
 429 triples.

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

433 ► **Theorem 17.** *UNARY TRIPLET REPEAT MULTI-STRAND MFE is NP-complete.*

434 4.4 Strand soup interaction

435 We now consider the computational problem MFE STRAND SOUP INTERACTION as defined
 436 in Section 2.2. We can adapt the algorithm from above and fortunately we do not need
 437 to keep track of the (exponentially many) subsets anymore, yielding a polynomial-time
 438 algorithm. We do not charge any strand association penalty, since we require one single
 439 complex anyways. However, we still must enforce connectivity. To this end, we encode by
 440 $c = 1$ that s and r still need to be connected, and by $c = 2$ that they already are connected.
 441 Furthermore, instead of keeping track of a subset of remaining strands, we just need the
 442 number of remaining strands m , as seen in Figure 4 (c). We obtain the following DP equations:
 443

$$444 \quad M_{m,s_i,r_j,c} = \min \begin{cases} \begin{cases} M_{m,s_{i+1},r_j,c} & \text{if } i+1 \leq |s| \\ \min_{t \in R} M_{m-1,t_1,r_j,1} & \text{if } i+1 > |s| \text{ and } c \neq 1 \\ +\infty & \text{else} \end{cases} \\ E_{s_i,r_j} + \bar{M}_{m,s_i,r_j,2} \\ \min_{m',t,k \text{ s.t. } (*)} E_{s_i,t_k} + \bar{M}_{m',s_i,t_k,2} + \bar{M}_{m-m'+1,t_k,r_{j+1},c} \end{cases}$$

445 where

$$446 \quad \bar{M}_{m,s_i,r_j,c} = \begin{cases} M_{m,s_{i+1},r_{j-1},c} & \text{if } i+1 \leq |s| \text{ and } j-1 \geq 1 \\ \min_{t \in R} M_{m-1,t_1,r_{j-1},1} & \text{if } i+1 > |s| \text{ and } j-1 \geq 1 \\ \min_{u \in R} M_{m-1,s_{i+1},u_{|u|},1} & \text{if } i+1 \leq |r| \text{ and } j-1 < 1 \\ +\infty & \text{else} \end{cases}$$

447 The minimum free energy can be finally computed by

$$448 \quad E^*(R, m) = \min_{s,r \in R} M_{m,s_1,r_{|r|},1}$$

449 and the optimal secondary structure can be obtained through backtracking. We initialize
 450 $M_{1,s_i,s_j,2} = 0$ for all $j - i \leq \theta$.

451 The correctness mostly follows from Section 4.1, but we still have to argue that we correctly
 452 minimize over *connected* secondary structures only, which is done in the appendix.

453 Regarding the running time, the table size is bounded by $m \cdot p^2 \cdot n^2 \cdot 3$, where $n := \max_{s \in R} |s|$.

454 The running time to compute one table entry is dominated by the last case, where we
 455 minimize over $\mathcal{O}(m \cdot p \cdot n)$ triples and need $\mathcal{O}(p)$ time for each triple. In total, we obtain an
 456 algorithm with running time $\mathcal{O}(n^3 \cdot m^2 \cdot p^4)$. We can then conclude:

457 ► **Theorem 18.** *MFE UNLIMITED STRAND INTERACTION can be solved in time $\mathcal{O}(n^3 \cdot m^2 \cdot p^4)$.*

458 ► **Remark 19.** Additionally to restricting the number of interacting strands, one can extend
 459 the above algorithm to restrict the size of the concatenated sequence. This is possible by
 460 keeping track of the current size of the sub-interval in the DP tables, and updating these
 461 values whenever a new strand is introduced.

462 This might be useful if the sequences in the base set have different length, as the basic
 463 algorithm would favor larger sequences because they usually allow for more base pairs.

464 ► **Remark 20.** The case of triplet repeats gives a slight improvement to the running time.
 465 Since all strands look the same except for their length, we can use a table with entries of the
 466 form $M_{m,i,j,c}$, where i and j denote the remaining number of nucleotides in the leftmost and
 467 rightmost strand. This reduces the space complexity to $\mathcal{O}(m \cdot n^2)$, but the computation of
 468 one table entry still takes the same amount of time, giving an overall time complexity of
 469 $\mathcal{O}(n^3 \cdot m^2 \cdot p^2)$.

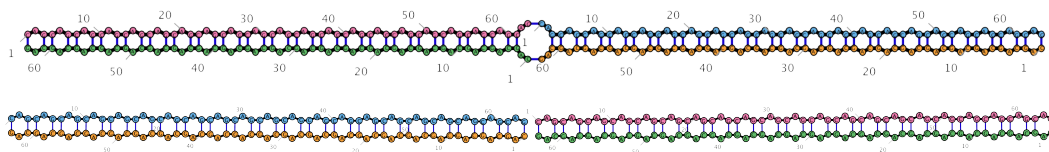
470 5 Empirical proof of concept

471 The goal of this section is to show how the algorithms described in the previous section can
 472 be used to answer biologically relevant questions regarding triplet repeats. We implemented
 473 the algorithm described in Section 4.4, which hereunder we call SoupFold, as well as its
 474 partition function equivalent, together with a (stochastic) backtracking procedure. Since we
 475 only limit the number of interacting strands but not their size, without further restrictions,
 476 the program would prefer large strands since they usually give more base pairs. To counteract
 477 this effect, we introduce a penalty on the length of a strand. Note that one could also set a
 478 maximum length of the concatenated sequence, as described in Remark 19. The source code
 479 is available at <https://github.com/kimonboehmer/soupfold/> and all experiments can be
 480 reproduced from its content.

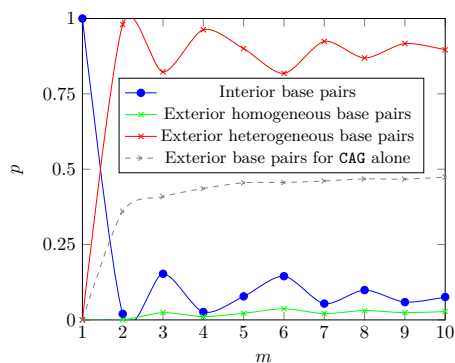
481 Regarding the stochastic backtracking, we must account for the overcounting of rota-
 482 tionally asymmetric secondary structures (since the algorithm uses normal permutations
 483 instead of circular permutations) as well as for the overcounting because of the positioning
 484 of different connected components. We address these two issues by rejection sampling. In
 485 theory, it is also necessary to adjust the overcounting correction for rotationally symmetric
 486 structures (because they are overcounted less often) but our experiments showed that the
 487 observed probability of encountering such rotational symmetries is 0 for triplets with 15
 488 repeats or more. Thus, for efficiency reasons, we do not include this case in our rejection
 489 sampling, arguing that the changes to the probability would be too small to observe.

490 5.1 Homogeneous triplet soup

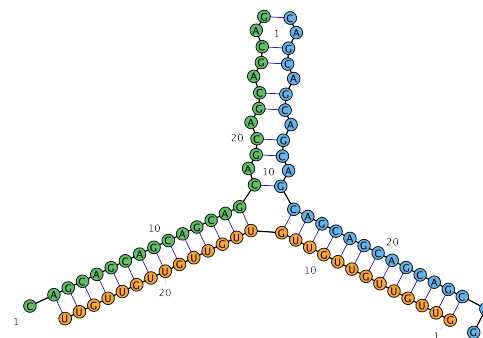
491 We exhibit some MFE secondary structures of interest for the interaction of multiple triplet
 492 repeat RNA strands. We first consider the case where all strands are of the same pattern.



■ **Figure 6** Connected and unconnected MFE structures for a four-strand CAG interaction, using RiboSketch [15]



■ **Figure 7** Probability p that a certain type of base pair is observed with increasing strand number m , for soup $\{\text{CAU}_{20}, \text{GGG}_{20}\}$. We also show the external base pair probability for a soup of just one pattern in dashed gray.



■ **Figure 8** Exemplary structure computed by SoupFold for strand pool $\{(\text{GUU})^9, (\text{CAG})^8, (\text{ACG})^{12}\}$ and $m = 3$, using RiboSketch [15]

493 The typical MFE structure will place the innermost base pair of a helix between two strands
 494 in order to avoid hairpin loops and the associated number of unpaired bases. Two examples of
 495 such secondary structures, one where we require connectedness and one where we do not, can
 496 be seen in Figure 6. The MFE of a soup of homogenous triplets behaves canonically, in the
 497 sense that all folding patterns behave almost identically (as can be expected, considering our
 498 results on single-strand triplets in Section 3). Furthermore, we observed that the number of
 499 base pairs increases canonically with the sequence length and with the number of interacting
 500 strands (except for the case of only one strand, where we loose one base pair due to a hairpin
 501 loop).

502 5.2 Heterogeneous triplet soups

503 Regarding the interactions of triplet repeat strands of different patterns, we can observe
 504 interactions of different triplet pattern strands in the MFE structure, which can even increase
 505 the number of base pairs compared to a homogeneous strand pool (see Figure 8).

506 In order to assess the capability of different strand soups to form droplets, we want to
 507 determine the probability of a base pair in the Boltzmann ensemble being between two
 508 strands (exterior) and not folding (interior). If the strand soup consists only of triplets of
 509 one pattern, all exterior base pairs will be homogeneous, as opposed to heterogeneous for an
 510 interaction of two strands of different patterns. In the homogeneous case, we can observe an
 511 increase of exterior base pairs for increasing number of interacting strands m , as presented
 512 by the gray line in Figure 7. The probabilities in a setting with strands of different pattern
 513 are much richer and less canonical, as can be seen at the example of the interaction of CAU

514 and GGG, presented by the other lines in Figure 7. These probabilities highly depend on the
515 number of strands, and only start to “converge” with quite high values of m .

516 To obtain a broader picture, we performed stochastic backtracking of our SoupFold
517 algorithm on all possible 4^6 pairs of triplet repeat patterns $\{TVW, XYZ\}$ as strand sets,
518 setting the number of interacting strands to $m \in \{2, 3, 4, 5\}$, and derived an estimated
519 probability of a base pair being interior, exterior-homogeneous or exterior-heterogeneous.
520 The probability of exterior homo- and heterogeneous base pairs for $m = 3$ and for all pairs of
521 TR patterns are exemplary visualized in Figure 9.

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

527 For that, we can model the patterns as vertices of a graph and connect two patterns with
528 an edge if their heterogeneous base pair probability is high (we set the threshold to 0.13).
529 We are then looking for a maximum independent set in this graph, *i.e.* the largest number of
530 triplets that do not have a high probability of interacting pairwise with each other. We used
531 an exact solver [10] to obtain an independent set of size 4, namely AGU, CAG, GGC, UGG.
532 We then executed our algorithm on these triplet patterns as strand soup, and could indeed
533 observe that the probability of exterior heterogeneous base pairs remained quite low. In
534 particular, for $m = 3$, the total probability of heterogeneous external base pairs is around
535 0.17, while the probability of homogeneous external base pairs is considerably higher (0.28).

536 **6 Conclusion and Discussion**

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

547 For future work, it is desirable to extend the MFE STRAND INTERACTION algorithm to
548 the full thermodynamic setting considered by [8]. While the extension to the Turner model
549 does not pose any algorithmic challenges, it would be interesting to see how one can correct
550 symmetries and overcounting for the partition function during the dynamic programming
551 without iterating over each circular permutation separately, as well as implement a variant
552 of the inside/outside algorithm to compute exactly base-pairing probabilities and other
553 expected values of additive properties. Finally, the joint conformation space explored in this
554 work is heavily restricted by the existence of a non-crossing strand ordering. More complex
555 conformational spaces could be captured by using dynamic programming approaches akin to
556 the ones being used to include pseudoknots in RNA structure prediction.

557 — References —

- 558 1 Dilimulati Aierken and Jerelle A Joseph. Accelerated simulations of rna phase separation: a
559 systematic study of non-redundant tandem repeats. *bioRxiv*, pages 2023–12, 2023.
- 560 2 Can Alkan, Emre Karakoc, Joseph H Nadeau, S Cenk Sahinalp, and Kaizhong Zhang. Rna–rna
561 interaction prediction and antisense rna target search. *Journal of Computational Biology*,
562 13(2):267–282, 2006.
- 563 3 Alin Bostan, Frédéric Chyzak, Grégoire Lecerf, Bruno Salvy, and Éric Schost. Differential
564 equations for algebraic functions. In C. W. Brown, editor, *ISSAC’07: Proceedings of the 2007
565 international symposium on Symbolic and algebraic computation*, pages 25–32. ACM Press,
566 2007. doi:10.1145/1277548.1277553.
- 567 4 Karl Bringmann, Fabrizio Grandoni, Barna Saha, and Virginia Vassilevska Williams. Truly
568 subcubic algorithms for language edit distance and rna folding via fast bounded-difference
569 min-plus product. *SIAM Journal on Computing*, 48(2):481–512, 2019. arXiv:<https://doi.org/10.1137/17M112720X>, doi:10.1137/17M112720X.
- 570 5 Yi-Jun Chang. Hardness of rna folding problem with four symbols. *Theoretical Computer
571 Science*, 757:11–26, 2019. URL: <https://www.sciencedirect.com/science/article/pii/S0304397518304912>, doi:10.1016/j.tcs.2018.07.010.
- 572 6 Anne Condon, Monir Hajiaghayi, and Chris Thachuk. Predicting minimum free energy struc-
573 tures of multi-stranded nucleic acid complexes is apx-hard. In *27th International Conference
574 on DNA Computing and Molecular Programming (DNA 27)(2021)*. Schloss-Dagstuhl-Leibniz
575 Zentrum für Informatik, 2021.
- 576 7 A. Denise, Y. Ponty, and M. Termier. Controlled non-uniform random generation of decompos-
577 able structures. *Theoretical Computer Science*, 411(40):3527–3552, 2010. URL: <https://www.sciencedirect.com/science/article/pii/S0304397510002914>, doi:10.1016/j.tcs.2010.
578 05.010.
- 579 8 Robert M Dirks, Justin S Bois, Joseph M Schaeffer, Erik Winfree, and Niles A Pierce.
580 Thermodynamic analysis of interacting nucleic acid strands. *SIAM review*, 49(1):65–88, 2007.
- 581 9 Haotian Guo, Joseph C Ryan, Xiaohu Song, Adeline Mallet, Mengmeng Zhang, Victor Pabst,
582 Antoine L Decrulle, Paulina Ejsmont, Edwin H Wintermute, and Ariel B Lindner. Spatial
583 engineering of E. coli with addressable phase-separated RNAs. *Cell*, 185(20):3823–3837, 2022.
- 584 10 Fanny Hauser, Ferdinand Ermel, and Kimon Boehmer. Clique cover based vertex cover solver.
585 <https://github.com/f-erm/CliqueCoverBasedVertexCoverSolver>, 2024.
- 586 11 Liang Huang, He Zhang, Dezhong Deng, Kai Zhao, Kaibo Liu, David A Hendrix,
587 and David H Mathews. LinearFold: linear-time approximate RNA folding by
588 5’-to-3’ dynamic programming and beam search. *Bioinformatics*, 35(14):i295–i304,
589 07 2019. arXiv:[https://academic.oup.com/bioinformatics/article-pdf/35/14/i295/
590 50721438/bioinformatics_35_14_i295.pdf](https://academic.oup.com/bioinformatics/article-pdf/35/14/i295/50721438/bioinformatics_35_14_i295.pdf), doi:10.1093/bioinformatics/btz375.
- 591 12 Atagun U Isiktas, Aziz Eshov, Suzhou Yang, and Junjie U Guo. Systematic generation and
592 imaging of tandem repeats reveal base-pairing properties that promote RNA aggregation. *Cell
593 Reports Methods*, 2(11), 2022.
- 594 13 Ryo Kurokawa, Mariko Kurokawa, Akihiko Mitsutake, Moto Nakaya, Akira Baba, Yasuhiro
595 Nakata, Toshio Moritani, and Osamu Abe. Clinical and neuroimaging review of triplet repeat
596 diseases. *Japanese Journal of Radiology*, 41(2):115–130, 2023.
- 597 14 L. Lipshitz. D -finite power series. *Journal of Algebra*, 122(2):353–373, 1989.
- 598 15 Jacob S Lu, Eckart Bindewald, Wojciech K Kasprzak, and Bruce A Shapiro. RiboSketch:
599 versatile visualization of multi-stranded RNA and DNA secondary structure. *Bioinformatics*,
600 34(24):4297–4299, 06 2018. arXiv:[https://academic.oup.com/bioinformatics/
601 article-pdf/34/24/4297/48919841/bioinformatics_34_24_4297.pdf](https://academic.oup.com/bioinformatics/article-pdf/34/24/4297/48919841/bioinformatics_34_24_4297.pdf), doi:
602 10.1093/bioinformatics/bty468.
- 603 16 Hiranmay Maity, Hung T Nguyen, Naoto Hori, and D Thirumalai. Odd–even disparity in the
604 population of slipped hairpins in rna repeat sequences with implications for phase separation.
605 *Proceedings of the National Academy of Sciences*, 120(24):e2301409120, 2023.

- 609 17 Colin McDiarmid. Pattern minimisation in cutting stock problems. *Discrete applied mathematics*, 98(1-2):121–130, 1999.
- 610
- 611 18 R Nussinov and A B Jacobson. Fast algorithm for predicting the secondary structure of
612 single-stranded rna. *Proceedings of the National Academy of Sciences*, 77(11):6309–6313,
613 1980. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.77.11.6309>, arXiv:<https://www.pnas.org/doi/pdf/10.1073/pnas.77.11.6309>, doi:10.1073/pnas.77.11.6309.
- 614
- 615 19 B. Salvy and P. Zimmerman. GFUN: a Maple package for the manipulation of generating
616 and holonomic functions in one variable. *ACM Transactions on Mathematical Software*,
617 20(2):163–177, 1994.
- 618 20 Sharan R. Srinivasan, Claudio Melo de Gusmao, Joanna A. Korecka, and Vikram
619 Khurana. Chapter 18 - repeat expansion disorders. In Michael J. Zigmond,
620 Clayton A. Wiley, and Marie-Francoise Chesselet, editors, *Neurobiology of Brain Disorders*
621 (*Second Edition*), pages 293–312. Academic Press, second edition edition, 2023. URL:
622 <https://www.sciencedirect.com/science/article/pii/B9780323856546000484>, doi:10.
623 1016/B978-0-323-85654-6.00048-4.
- 624 21 Douglas H. Turner and David H. Mathews. NNDB: the nearest neighbor parameter data-
625 base for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*,
626 38(suppl_1):D280–D282, 10 2009. arXiv:[https://academic.oup.com/nar/article-pdf/38/
627 suppl_1/D280/11217894/gkp892.pdf](https://academic.oup.com/nar/article-pdf/38/suppl_1/D280/11217894/gkp892.pdf), doi:10.1093/nar/gkp892.
- 628 22 Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using
629 thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 01 1981.
630 arXiv:<https://academic.oup.com/nar/article-pdf/9/1/133/6201945/9-1-133.pdf>, doi:
631 10.1093/nar/9.1.133.

632 **A** Appendix for Section 3

633 **A.1** Proof for Lemma 5

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

636 First assume $\{X, Z\} \in P$ and $\{X, Y\}, \{Y, Z\} \notin P$. We will derive the other cases from this
637 one. Consider a large stacking of $X - Z$ bases. If $\theta = 3$, we only cannot match the $X - Z$
638 pair of the innermost repeat in the case $k \equiv_2 1$ and we only cannot match the $Z - X$ pair
639 between the two innermost repeats in the case $k \equiv_2 0$. For all other pairs of repeats we
640 obtain exactly two base pairs and hence we get $k - 1 = k - \lfloor \frac{\theta+1}{3} \rfloor$ base pairs. Inductively,
641 let us show that we can obtain $k - \lfloor \frac{\theta'+1}{3} \rfloor$ base pairs for $\theta' := \theta + 3$. In other words, we
642 only need to show that by increasing θ by 3, we get one base pair less. If the innermost
643 base pair is $X - Z$, its enclosed region starts and ends with a Y and there are currently
644 at least $\theta + 1$ free enclosed bases (because the region is of the form $Y(ZXY)^{\theta/3}$), and by
645 deleting the $X - Z$ base pair, we obtain $XY(ZXY)^{\theta/3}Z$, that is $\theta + 3$ enclosed bases. Else,
646 for a $Z - X$ base pair, the region has the form $(XYZ)^{\theta/3}$. After deleting the innermost base
647 pair $Z - X$, the new enclosed region starts and ends with a Y (the region is of the form
648 $YZ(XYZ)^{\theta/3}XY$), so there are at least $\theta + 4$ enclosed bases. Thus we can achieve $k - \lfloor \frac{\theta+1}{3} \rfloor$
649 base pairs.

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

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

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

658 We thus established the lower bound for the $\{X, Z\} \in P$ case. For the “otherwise”-case,
659 Lemma 3 already gives us the required upper bound. Therefore, we only need to argue about
660 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
661 a secondary structure that achieves more base pairs. Firstly, we cannot have any multiloops
662 or exterior loops since that would imply two regions of unpaired enclosed bases, which then
663 only allows $k - 2 \lfloor \frac{\theta+1}{3} \rfloor \leq k - 1 - \frac{\theta-1}{3}$ base pairs. Additionally, for each secondary structure
664 S with $i < j'$ and $k > 0$ such that $\{i, j'\} \in S$ and the interval $[j' + 1, j' + 3k]$ only consists
665 of unpaired bases, we can delete the base pair $\{i, j'\}$ and instead add base pair $\{i, j' + 3k\}$
666 without reducing the number of base pairs. In other words, for any interval, it is always
667 better to pair the leftmost base to the rightmost possible base than to any other interior
668 base. We thus only need to consider the canonical structures of $X - Z/Z - X$ -stackings.

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

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

678 If $\{X, Z\} \notin P$, we can assume without loss of generality that $\{X, Y\} \in P$ (the arguments

679 are symmetrical for $\{Y, Z\} \in P$, and we assumed to have a folding strand). We can reduce
 680 any such instance $(XYZ)^k$ to $(YZX)^{k-1}$ (by letting out the leftmost X and the rightmost
 681 Y and Z , and implicitly pairing these outermost X and Y , which is always optimal). Thus,
 682 all results can be directly obtained from the case $\{X, Z\} \in P$, by changing odd and even.
 683 The upper bound can also be derived by that. \blacktriangleleft

684 **B** Appendix for Section 4

685 **B.1** Proof of correctness for the exponential-time algorithm

686 We now prove that M_{R, s_i, r_j} is computed correctly. By slight abuse of notation, we write
 687 $s_i \in S$ for $s_i \in \bigcup_{P \in S} P$.

688 **► Definition 21.** An interval for this DP is denoted by $[R, s_i, r_j, c]$ where $s, r \in R$, $1 \leq i \leq |s|$,
 689 $1 \leq j \leq |r|$ and $c \in \{0, 1, 2\}$. An interval $[R', t_k, u_\ell, c']$ is included in interval $[R, s_i, r_j, c]$,
 690 written $[R', t_k, u_\ell, c'] \preceq [R, s_i, r_j, c]$, if one of the following holds:

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

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

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

697 **► Definition 22.** Let $I := [R, s_i, r_j, c]$. $\Omega(I)$ is the set of all secondary structures that are
 698 valid for this interval, or more formally, a secondary structure S must fulfill:

- 699 \blacksquare $S \in \Omega(R)$
- 700 \blacksquare $s_k, r_\ell \notin S$ for any $k < i$ and $\ell > j$
- 701 \blacksquare $c = 1$ implies the existence of a base pair between s and r (that is, $\{s_k, r_\ell\} \in S$ for some
 702 $i \leq k \leq |s|, 1 \leq \ell \leq j$) and $c = 0$ implies that there is no such base pair.

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

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

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

709 **► Observation 23.** If $E(R, S) = MFE([R, s_i, r_j, c])$ and S only contains base pairs in some
 710 interval $[R', t_k, u_\ell, c] \preceq [R, s_i, r_j, c]$, then $S = MFE([R', t_k, u_\ell, c])$.

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

712 **► Lemma 24.** Assuming that $M_{R', t_k, u_\ell, c'} = MFE(I' := [R', t_k, u_\ell, c'])$ for all $I' \preceq I :=$
 713 $[R, s_i, r_j, c]$, we have $\bar{M}_{R, s_i, r_j, c} = MFE([R, s_i, r_j, c])$.

714 **Proof.** We distinguish four cases:

- 715 \blacksquare **Case 1:** $i + 1 \leq |s|$ and $j - 1 \geq 1$. In that case, for any $I' \prec I$, we have $I' \preceq$
 716 $[R, s_{i+1}, r_{j-1}, c]$ and thus $MFE(I') \geq MFE([R, s_{i+1}, r_{j-1}, c]) = \bar{M}_{R, s_i, r_j, c}$ by assumption.
 717 Thus $MFE([R, s_i, r_j, c]) = \bar{M}_{R, s_i, r_j, c}$.

XXX:20 RNA Triplet Repeats: Improved Algorithms for Structure Prediction and Interactions

718 ■ **Case 2:** $i + 1 > |s|$ and $j - 1 \geq 1$. For any $I' \prec I$, there is a $t \in R - \{s\}$ and a
 719 $c' \in \{0, 1\}$ with $I' \preceq [R - \{s\}, t_1, r_{j-1}, c']$. It thus suffices to minimize over the strands
 720 $R - \{s, r\}$ while taking into account a possible strand disconnection reward. We have
 721 $\min_{t \in R - \{s, r\}, c' \in \{0, 1\}} M_{R - \{s\}, t_1, r_{j-1}, c'} - \mathbb{1}_{c'=0} K_{\text{assoc}} = \text{MFE}(\lrcorner R, s_i, r_j, c)$.

722 ■ **Case 3:** $i + 1 \leq |s|$ and $j - 1 < 1$. This case is completely symmetrical to Case 2.

723 ■ **Case 4:** $i + 1 > |s|$ and $j - 1 < 1$. For any $I' \prec I$, there are $t, u \in R - \{s, r\}$
 724 with $I' \preceq [R - \{s, r\}, t_1, u_{|u|}, 2]$. It thus suffices to minimize twice over the strands
 725 $R - \{s, r\}$ while taking into account a possible strand disconnection reward. We have
 726 $\min_{t, u \in R - \{s, r\}, c' \in \{0, 1\}} M_{R - \{s, r\}, t_1, u_{|u|}, c'} - \mathbb{1}_{c'=0} K_{\text{assoc}} = \text{MFE}(\lrcorner R, s_i, r_j, c)$.

727 ◀

728 ► **Lemma 25.** *The algorithm computes the table entries correctly, i.e. $M_{R, s_i, r_j, c} = \text{MFE}(\lrcorner R, s_i, r_j, c)$*
 729 *for all $R \subseteq R_0$, $s_i, r_j \in R$ and $c \in \{0, 1, 2\}$.*

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

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

736 ■ **Case 1:** $s_i \notin S$. If $i + 1 \leq |s|$, we have $E(R, S) = \text{MFE}(\lrcorner R, s_{i+1}, r_j, c) = M_{R, s_{i+1}, r_j, c}$ by
 737 Observation 23 and our induction hypothesis.

738 Else, we first assume $c \neq 1$. Consider the strand t that follows s in the polymer graph
 739 representation of S and consider the value c' that specifies connectivity between t and r in
 740 S . Since i is unpaired, we again have $E(R, S) = \text{MFE}(\lrcorner R - \{s\}, t_1, r_j, c') - \mathbb{1}_{c'=0} K_{\text{assoc}} =$
 741 $M_{R - \{s\}, t_1, r_j, c'} - \mathbb{1}_{c'=0} K_{\text{assoc}}$ as above.

742 Finally, if $c = 1$, we look for the MFE of a structure in $\lrcorner R, s_i, r_j, c$ where s and r are
 743 connected by a base pair. Since there is only one base in s remaining and we leave it
 744 unpaired, there is no such structure and thus $\text{MFE}(\lrcorner R, s_i, r_j, 1) = +\infty$.

745 ■ **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
 746 s and r arbitrarily connected (that is, $\lrcorner R, s_i, r_j, 2$). First assume $c \neq 0$. In this case, we
 747 have $E(R, S) = E_{s_i, r_j} + \text{MFE}(\lrcorner R, s_i, r_j, 2) = E_{s_i, r_j} + \bar{M}_{R, s_i, r_j, 2}$, where we could apply
 748 Lemma 24 because of the induction hypothesis.

749 Now assume $c = 0$. We minimize over all structures such that s and r are not connected,
 750 but require $\{s_i, r_j\} \in S$. Thus $\text{MFE}(\lrcorner R, s_i, r_j, 0) = +\infty$.

751 ■ **Case 3:** $S = \{s_i, t_k\} \cup S' + S''$ for some $t_k \neq r_j$, where S' (resp. S'') is the best structure
 752 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
 753 and t in the polymer graph representation of S , and R'' being all strands between t and
 754 r .

755 Note that s and t are connected, thus in S' the connectivity bit will be set to 2. On
 756 the other hand, the connectedness of t and r (for structure S'') is by transitivity of
 757 connectivity determined by the connectedness between s and r , that is, c . We then have
 758 $\text{MFE}(\lrcorner R, s_i, r_j, c) = E_{s_i, t_k} + \text{MFE}(\lrcorner R', s_i, t_k, 2) + \text{MFE}(\lrcorner R'', t_k, r_j, c)$.

759 ◀

760 We now briefly discuss the running time. The number of table entries is bounded by $2^m \cdot n^2$,
 761 where $n := \max_{r \in R} m$ is the maximum size of the concatenated sequence. Clearly, the last
 762 case of the DP equation dominates the running time for computing one entry. In the worst

763 case, we iterate over $2^{|R|}$ subsets and n entries, which gives $\mathcal{O}(2^{|R|} \cdot n)$. Partitioned by subset
764 size, we get

$$765 \quad \sum_{t=0}^m \binom{m}{t} n^2 \cdot 2^t n = n^3 \cdot \sum_{t=0}^m \binom{m}{t} 2^t = n^3 \cdot \sum_{t=0}^m \binom{m}{t} 1^{m-t} 2^t = n^3 \cdot (1+2)^m = 3^m \cdot n^3$$

766 which bounds the total running time. Together with Lemma 25, we conclude.

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

$$769 \quad \{s, t\} \subseteq R' \subseteq R \wedge 1 \leq k \leq |t| \wedge (k = |t| \rightarrow c \neq 1)$$

$$770 \quad \wedge (s = t \rightarrow (k > i + \theta \wedge R' = \{s\}))$$

$$771 \quad \wedge (r \in R' \rightarrow (t = r \wedge k < j \wedge R' = R \wedge c \neq 0))$$

772 We minimize over all possible triples (R', t, k) . A set R' must clearly include s and t to
773 form a valid interval and k must be a valid position of t . If s_i is paired to $t_{|t|}$, s and j
774 are disconnected ($c \neq 1$). If $s = t$, we must respect θ and there is only one strand in
775 R' . Finally, $r \in R'$ implies that s_i forms a base pair with some base of r (thus $t = r$ and
776 $R' = R$), connectivity has to be allowed ($c \neq 0$) and t_k must be in the interval ($k < j$). These
777 conditions are sufficient and match our algorithm.

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

783 B.2 Running time for Section 4.2

784 We need table entries for each possible configuration of remaining number of occurrences and
785 for specifying the remaining number of bases on the leftmost and rightmost strand. Using
786 $n := \max_{r \in R} |r|$, we bound the number of table entries by

$$787 \quad n^2 \cdot \max_{s_1, \dots, s_p: s_1 + \dots + s_p = m} \prod_{i=1}^p s_p \leq n^2 \cdot \left(\frac{m}{p}\right)^p$$

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

793 B.3 Proof for the connectivity in Section 4.4

794 Analogous to Section 4.1, we define an interval $[m, s_i, r_j, c]$ and a relation $[m', t_k, u_\ell, c'] \preceq$
795 $[m, s_i, r_j, c]$ if and only if $m' < m - 1$ or $m' = m - 1 \wedge (s = t \vee r = u)$ or $m' = m \wedge s =$
796 $t \wedge r = u \wedge i \leq k \wedge \ell \leq j$. Since we just change the representation of our set R to an
797 integer m , the correctness of the algorithm can be shown by the same arguments as for the
798 exponential algorithm. We only show here that the connectivity specifier $c \in \{1, 2\}$ actually
799 enforces connectivity. For this, we introduce the following notation: $\gamma(m, s_i, r_j)$ means that
800 the MFE structure computed by $M_{m, s_i, r_j, 1}$ is connected, and $\bar{\gamma}(m, s_i, r_j)$ means that the

801 MFE structure computed by $M_{m,s_i,r_j,2}$ is either connected or consists of two connected
 802 components, one containing s and one containing r . In other words, adding a base pair
 803 between s and r to such a structure will make it connected.

804 ► **Lemma 26.** $\gamma(m, s_i, r_j)$ and $\bar{\gamma}(m, s_i, r_j)$ hold.

805 **Proof.** Clearly, a secondary structure over an interval with $m = 1$ is always connected,
 806 i.e. $\gamma(1, t_k, t_\ell)$ and $\bar{\gamma}(1, t_k, t_\ell)$ hold for any valid t, k, ℓ . By induction over \preceq , assume that
 807 $\gamma(m', t_k, u_\ell)$ and $\bar{\gamma}(m', t_k, u_\ell)$ for any $[m', t_k, u_\ell, c'] \preceq [m, s_i, r_j, c]$. We show $\gamma(m, s_i, r_j)$ and
 808 $\bar{\gamma}(m, s_i, r_j)$. By case distinction:

809 ■ **Case 1:** $s_i \notin S$. If $i + 1 \leq |s|$, the structure is connected by assumption. Else, if $c = 2$,
 810 we need that a connection between s and r would make the structure connected. Indeed,
 811 by assumption, $[m - 1, t_1, r_j, 1]$ is connected, and together with a base pair between s and
 812 r , all strands are in one connected component. If $c = 1$, s and r are not yet connected
 813 and we do not connect them with the last possible base $s_{|s|}$, thus no connected secondary
 814 structure with these constraints exists.

815 ■ **Case 2:** $\{s_i, r_j\} \in S$. By hypothesis, the structure for $]m, s_i, r_j, 2[$ would be connected
 816 together with a base pair between s and r , thus the structure for $[m, s_i, r_j, c]$ is connected.

817 ■ **Case 3:** $\{s_i, t_k\} \in S$ for some t_k in the region. By assumption and the base pair $\{s_i, t_k\}$,
 818 the strands from s to t are connected. If $c = 1$, then by assumption $]m - m' + 1, t_k, r_{j+1}, 1[$
 819 is connected and thus all the structure is connected. For $c = 2$, assume a connection
 820 between s and r . Now by the fact that s is connected to t and transitivity, r is connected
 821 to t . We can apply our induction hypothesis to conclude that the substructure for the
 822 strands from t to r is connected, and by that, the complete structure is connected. ◀

824 We now argue (somewhat informally) why there cannot be a better connected secondary
 825 structure that the algorithm ignores. Assume that the last case of the \bar{M} equation is defined
 826 as for 4.1, that is, we minimize over the two next inner strands. Any structure that uses
 827 this case cannot be connected (as the component including s and r has no way of being
 828 connected to the component including the inner strands).

829 Assume also that when minimizing over strands, we lift the connectivity requirement ($c = 1$).
 830 In any secondary structure than can be obtained by at some point (at interval $[m, s_i, r_j, 2]$)
 831 minimizing over a strand with $c = 2$ but not with $c = 1$, we know that the chosen inner
 832 strand (say t) is not connected to r in the constructed secondary structure restricted to the
 833 region from s_i to r_j . Since the outer region before s_i and after r_j does not contain any base
 834 of strand t , strand t will not be connected to r in the complete structure.

835 So, after applying these changes to the DP, we cannot achieve a better connected secondary
 836 structure than before. The DP is now almost equivalent to the DP in Section 4.1, with
 837 representing the set R by a natural number m . We can thus repeat the correctness proof
 838 of section Section 4.1 to show that any (connected) secondary structure is covered by the
 839 equations, and thus the output of our DP is optimal.

