RNA Triplet Repeats: Improved Algorithms for Structure Prediction and Interactions

Kimon Böhmer [1]

[1] 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 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 of the same TR pattern, allowing an assessment of a notion of orthogonality between TR.

Index Terms: RNA Secondary Structure Prediction, Dynamic Programming, NP-hardness.

1 Introduction

1.1 Conditions in the Lab

Before starting with the article, I want to briefly present the group I was working in and the general working environment during my internship.

My internship was with the AMIBIO group of the Laboratoire d'Informatique de l'École Polytechnique (LIX). The group focusses on computational and algorithmic methods for molecular and structural biology. The goal is to design and implement algorithms that have applications in synthetic biology, biotechnology and health. On the other hand, the theoretical-computer-science-environment of the LIX building shapes the group into a bioinformatics group with a very strong focus on mathematical rigor and an algorithmic viewpoint. Tools from parametrized complexity, graph theory and combinatorial analysis of algorithms are not only used but results are produced that are valuable per se for these domains, outside of the context of bioinformatics. For example, a recent graph-theoretical result about the treewidth was first shown in a paper by some members of the group (Marchand et al., 2023) before it was rediscovered by graph theorists (Chaplick et al., 2023). Examples for recent work in the group is the automated generation of dynamic programming schemes parametrized by the treewidth for prediction of RNA secondary structure with certain patterns of so-called pseudoknots (Marchand et al., 2023) or the design of an efficient algorithm for a broad subclass of the in general NP-hard problem of inverse folding (Boury et al., 2024).

My supervisors were the three permanent members of the AMIBIO group, namely Yann Ponty, Sarah Berkemer and Sebastian Will. Yann has a more algorithmic and theoretical background and was a great help for discussing combinatorial and algorithmic questions. With Sarah and Sebastian, I could better discuss the biology side of my project and the different motivation for the studied structures. I am very grateful to all three of them for their help. Together, we submitted a paper at the WABI 2024 conference, as described later.

In the first month of the internship, I became familiar with the topic and designed the first basic algorithms (as for sections sections 4.1, 4.3 and 4.5). In the second month, I started looking at the problems posed by pseudoknots on multiple strands. The corresponding results are not included in this report as they turned out to not fit well in the context of triplet repeats and the general research direction of the internship. I also wrote the NP-hardness proof for triplet repeats (section 4.4) and started looking at some more structural results for single triplet strands (section 3). The third month was mainly devoted to formalizing and writing down all achieved work to submit at the WABI 2024 conference. Additionally, I implemented the algorithm of section 4.5 and conducted biologically motivated experiments, as shortly described in section 5. The paper was accepted at the conference and is available (Boehmer et al., 2024). In contrast to the contents of this internship report, it contains some more experimental and structural results that I considered less interesting from an algorithmic perspective, and thus decided not to include in the report. On the other hand, section 4.2 is included in the report but not in the paper, since the main result was found and written down after the paper deadline, in the fourth and last month of the internship. In general, the fourth section contains the algorithmically most interesting and involved results.

The relevance of my project for synthetic biologists was something motivating throughout the internship. The internship is funded by the SYNORG project. In June, I presented my work to the group of synthetic biologists around Ariel Lindner from INSERM. Finding a common language between computer science and biology was challenging, but meeting the researchers that will use the results obtained during my internship was a very valuable experience.

1.2 Basics on RNA Secondary Structure Prediction

The aim of this section is to introduce basic knowledge and vocabulary about RNA secondary structure prediction to the computer scientist reader who might not be familiar with algorithmic bioinformatics.

A ribonucleid acid (RNA) consists of a ribose backbone chain of different *nucleotides* (also called *bases*), namely *adenine* (A), *cytosine* (C), *guanine* (G) and *uracil* (U). From a computer science perspective, this can be represented by a word $w \in \{A, C, G, U\}^+$, which is called the *primary structure* of an RNA.

However, additionally to the backbone, two nucleotides of an

RNA can be bound to each other by hydrogen bonds. In the most cases, these *base pairs* are between nucleotides of type C and G, A and U (Watson-Crick pairs) or G and U (Wobble pair). Due to physical/geometrical constraints, each base pair has to enclose at least θ bases (usually $\theta = 3$).

Additionally, when drawing the bases in a cycle and the base pairs as straight lines between the bases, two crossing lines form a *pseudoknot*. These pseudoknots do appear in practice, but are not very common and typically make the algorithmic problems much more difficult. Therefore, many works in algorithmic bioinformatics (including mine) ignore pseudoknots or first consider a simple setting where they are excluded.

The set of all base pairs of an RNA is called its *secondary stucture*. Predicting the secondary structure of an RNA given its primary structure is an important task in algorithmic bioinformatics.

The secondary structure essentialy determines the *free energy* of an RNA, where base pairs generally reduce the free energy to thus create a more stable physical state. In an oversimplified model, we can compute the free energy of a secondary structure as -1 times the number of base pairs. The first algorithmic problem of interest is then to minimize the free energy over all possible secondary structures, or in other words, to find the secondary structure with the maximum number of base pairs. This is achieved by the seminal Nussinov algorithm (Nussinov et al., 1980) in time $O(n^3)$ via dynamic programming (DP): We can compute the *minimum free energy* (MFE) for all subregions $[i, j]$ of our word. The DP then distinguishes three cases: The base at position i can be unpaired (then, the MFE of $[i, j]$ is equal to the MFE of region $[i + 1, j]$), it can be paired to j (then, the MFE is equal to the MFE of region $[i+1, j-1]$ minus the award of 1 for the created base pair $\{i, j\}$) or it can be paired to some k between i and j (then, the MFE is equal to the MFE of regions $[i+1, k-1]$ and $[k+1, j]$ minus the award of 1 for the created base pair $\{i, k\}$). This leads to following the following DP equation:

$$M_{i,j} = \min \begin{cases} M_{i+1,j} \\ M_{i+1,j-1} + E(i, j) \\ \min_{k=i+\theta+1}^{j-1} M_{i+1,k-1} + M_{k+1,j} + E(i, k) \end{cases}$$

where $E(i, j)$ is the free energy contribution of bases i and j, which in our energy model is -1 if the bases are C-G,A-U or G-U, and $+\infty$ otherwise. We have $O(n^2)$ table entries, and to compute one table entry, we need to minimize over linearly many positions where the first base can be paired to, giving a total running time of $O(n^3)$.

However, one can observe that the real secondary structure of an RNA sequence changes over time and the probability of observing a certain secondary structure S follows a Boltzmann distribution $\frac{\exp\{-E(S)/kT\}}{\sum_{S' \in \Omega} \exp\{-E(S')/kT\}}$, where T is the temperature in Kelvin and k is the Boltzmann constant $1.987 \cdot 10^{-3}$kcal.mol^{-1}.K^{-1}. Computing the numerator is easy, but the denominator, usually called the *partition function*, is a sum over exponentially many secondary structures and its computation is non-trivial. However, it is well-known that when we are given DP equations for energy minimization which are complete and unambiguous, we can obtain a DP for partition function computation by a simple change of algebra: We change minimizations to sums, sums to products and free energy values $E(S)$ to its Boltzmann value $\exp\{-E(S)/kT\}$. This results in the following DP equations for computing the partition function:

$$\mathcal{Z}_{i,j} = \sum \begin{cases} \mathcal{Z}_{i+1,j} \\ \mathcal{Z}_{i+1,j-1} \cdot \exp\{-E(i, j)/kt\} \\ \sum_{k=i+\theta+1}^{j-1} \mathcal{Z}_{i+1,k-1} \cdot \mathcal{Z}_{k+1,j} \cdot \exp\{-E(i, k)/kT\} \end{cases}$$

1.3 Remarks about authorship

This work has many common parts with the paper that I coauthored with my three supervisors. However, all results and the written text are my own work, but the few subsections that I mention in the following were not exclusively written by me: Section 1.4 about the biological motivation was largely written by Sarah Berkemer and Yann Ponty, section 1.5 about the algorithmic generalities on RNA-folding and interactions was jointly written by me and Yann Ponty and section 3.2 about the linear-time computation of the partition function was also jointly written by me and Yann Ponty. All other sections are exclusively my own work.

We now begin with the main part of the internship report.

1.4 Biological motivation for Triplet Repeats

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 (Isiktas et al., 2022; Guo et al., 2022). 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 (Srinivasan et al., 2023) and Triplet Repeat Diseases (TRD) such as Huntington disease (Kurokawa et al., 2023). 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 (Maity et al., 2023) as well as extensions to quadruplet and non-redundant tandem repeats (Aierken et al., 2023).

1.5 Algorithmic generalities on RNA folding and RNA interactions

RNA folding by energy minimization is a classic algorithmic problem in Bioinformatics, historically solved in time $\Omega(n^3)$ using dynamic programming (Nussinov et al., 1980; Zuker et al., 1981). Despite recent misleading suggestions of linear-time alternatives (Huang et al., 2019), the best algorithm to date to solve energy minimization has runtime $O(n^{2.8603})$ (Bringmann et al., 2019), 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 $o(n^2)$ algorithm would refute the Strong Exponential Time Hypothesis (SETH) (Bringmann et al., 2019). Meanwhile, a $o(n^{\omega})$ algorithm would disprove the k-clique conjecture, with $\omega < 2.373$ being the matrix multiplication exponent (Bringmann et al., 2019; Chang, 2019).

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., 2006 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., 2007 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) (Condon et al., 2021), even for a very simple energy model.

1.6 Contributions

In this work, we show that the repeated nature of TR can be exploited to obtain substantially improved algorithms for several problems. First, we show that the MFE of a TR RNA can be predicted in linear time and is realized by either the open chain or a single helix. 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 $s \in \{A, C, G, U\}^+$. 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 $P := \{\{C, G\}, \{A, U\}, \{G, U\}\}$. 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 $j - i > \theta$. We usually call θ the *minimal base pair* span, and use $\theta = 3$ unless explicitly specified.

We denote by $\Omega(s)$ (or just Ω), the set of all secondary structures over sequence s.

We associate each secondary structure $S \in \Omega$ to a *free energy*, according to an *energy model* $E : \{A, C, G, U\}^+ \times \Omega \to \mathbb{R}$. 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 (Turner et al., 2009).

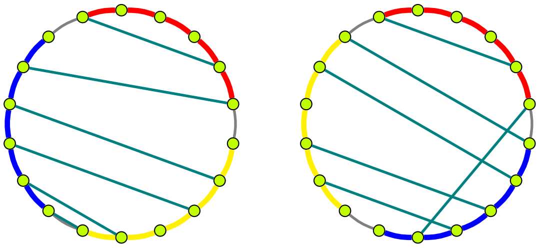

Figure 1. The same secondary structure on a strand set with three strands drawn in two different circular permutations. The strands are depicted by the blue, red and yellow lines while green lines indicate base pairs. Gray lines connect subsequent strands and depend on the strand permutation.

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 $i \in \{1, \ldots, |s| = |r|\}$, 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* $\pi : R \longrightarrow \{0, \ldots, m-1\}$ 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 $(s^1_{i_1}, \ldots, s^k_{i_k})$ such that there is a j with $s^j_{i_j} <_\pi s^{j+1}_{i_{j+1}} <_\pi \ldots <_\pi s^k_{i_k} <_\pi s^1_{i_1} <_\pi \ldots <_\pi s^{j-1}_{i_{j-1}}$.

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. $\{s_i, s_j\} \in S \longrightarrow j - i > \theta$.

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 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 $E : 2^{\{A,C,G,U\}^*} \times \Omega \to \mathbb{R}$. 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*) in the polymer graph. Thus, the free energy of $S \in \Omega$ in this model is defined as $E(R, S) = -p + (m - \ell)K_{\text{assoc}}$.

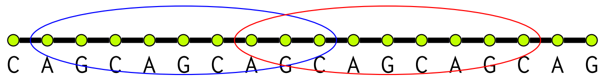

Figure 2. The blue and red region of the TR sequence are identical.

2.2 Computational problems

For a single strand, the two classical problems in RNA bioinformatics 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 in Kelvin (K)
Output: Partition function $\mathcal{Z}_s := \sum_{S\in\Omega(s)} \exp\{\frac{-E(s,S)}{kT}\}$

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

In the multi-strand setting, we first focus on energy minimization. In (Dirks et al., 2007), 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: $\min_{S\in\Omega(R_0)} E(R_0,S)$

However, we also show how the algorithm can be extended to the partition function setting.

We also consider 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. More formally:

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

2.3 Triplet repeat 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 $X,Y,Z \in \{A,C,G,U\}$ 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 convenient property about a region $[s_i, s_j]$ of a TR sequence is:

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

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

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 $XYZ \in \{A,C,G,U\}^3$ 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 $s \in \{A,C,G,U\}^{3k}$ (the latter can also be seen, asymptotically equivalent, 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 $G = (\{A,C,G,U\}, P)$, 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 $\theta \ge 0$, such that $\lfloor\frac{\theta+1}{3}\rfloor \le k$. We have $E_{bp}(s,S) \le k - \lfloor\frac{\theta+1}{3}\rfloor$ 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 $\{\{X,Y\},\{X,Z\},\{Y,Z\}\} \cap P = \emptyset$, 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.

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

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.*

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 $(\mathtt{GCU})^k$ as illustrated in fig. 3. When constructing the base pairs from outside to inside, in every step, we can choose whether we add 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.*

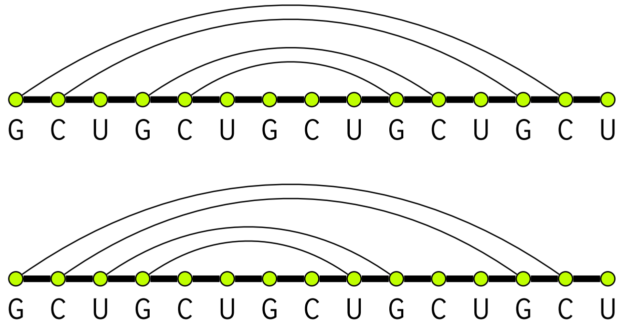

Figure 3. Two different optimal secondary structures for $\mathtt{GCU}_5$.

3.2 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.

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

$$\begin{array}{llll}
S_C^G \to \ (\ \cdot_A\ S_G^C\ \cdot_A\ ) & \mid (\ \cdot_A\ S_G^C\ \cdot_A\ )\, S_C^G & \mid \cdot_C\ \cdot_A\ S_G^G & \mid \cdot_C\ \cdot_A\ \cdot_G \\
S_G^C \to \ (\ S_C^G\ ) & \mid (\ S_C^G\ )\ \cdot_A\ S_G^C & \mid \cdot_G\ S_C^C & \mid \cdot_G\ \cdot_C \\
S_G^G \to \ (\ S_C^G\ )\ \cdot_A\ \cdot_G & \mid (\ S_C^G\ )\ \cdot_A\ S_G^G & \mid \cdot_G\ S_C^G & \\
S_C^C \to \ (\ \cdot_A\ S_G^C\ \cdot_A\ )\ \cdot_A & \mid (\ \cdot_A\ S_G^C\ \cdot_A\ )\, S_C^C & \mid \cdot_C\ \cdot_A\ S_G^C &
\end{array}$$

Namely, the terminal S_C^G generates all secondary structures for the RNA sequence $(CAG)^k$ for all $k > 0$, S_G^C 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 (Denise et al., 2010), such a grammar can be generically translated into a system of functional equations involving weighted generated functions for each non-terminal:

$$\begin{aligned}
S_C^G(z) &= \beta\, z^4\, S_G^C(z) + \beta\, z^4\, S_G^C(z)\, S_C^G(z) + z^2\, S_G^G(z) + z^3 \\
S_G^C(z) &= \beta\, z^2\, S_C^G(z) + \beta\, z^3\, S_C^G(z)\, S_G^C(z) + z\, S_C^C(z) + z^2 \\
S_G^G(z) &= \beta\, z^4\, S_C^G(z) + \beta\, z^3\, S_C^G(z)\, S_G^G(z) + z\, S_C^G(z) \\
S_C^C(z) &= \beta\, z^3\, S_G^C(z) + \beta\, z^2\, S_G^C(z)\, S_C^C(z) + z^2\, S_G^C(z)
\end{aligned}$$

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

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

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 (Lipshitz, 1989), which can be efficiently (Bostan et al., 2007) and effectively computed (Salvy et al., 1994). 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 also holds for other triplets and thus:

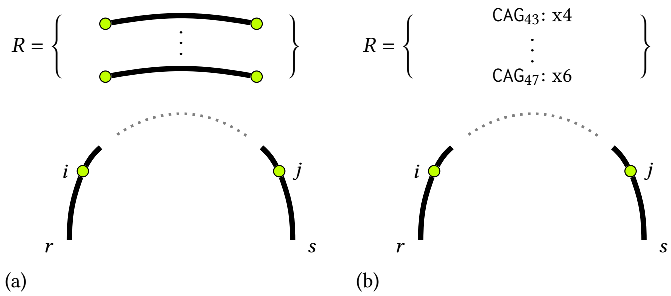

Figure 4. Visualization of the structures used to compute the MFE in the (a) general setting and (b) TR setting.

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 Condon et al., 2021, 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, Dirks et al., 2007 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. The exponential-time algorithm is shown for the minimization setting, but in section 4.2, we will show that we can extend it to the partition function setting, directly improving the result from Dirks et al., 2007.

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 fig. 4 (a). The main recurrences are as follows:

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

where

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

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 $s \in R$. The correctness of the algorithm and its running time are proven in the appendix. With n denoting the length of the concatenation of all strand sequences in R, we obtain:

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

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 Dirks et al., 2007.

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 $S' = \{\{s'^k_i, s'^\ell_j\} \mid \{s^k_i, s^\ell_j\} \in S\}$. An r-*symmetric* secondary structure is a secondary structure S with pseudoknot-free permutation $s^1, \ldots, s^m$ 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 4. *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 $a^1, \ldots, a^s$. 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 4. *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 seen in fig. 5.

An extended secondary structure (S', p) is *connected* 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 4, 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:

Lemma 5. *$\bar{\mathcal{Z}}(R_{/r})$ can be computed in time $O(3^m \cdot n^3)$.*

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 $s^0, \ldots, s^{m-1}$, 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

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

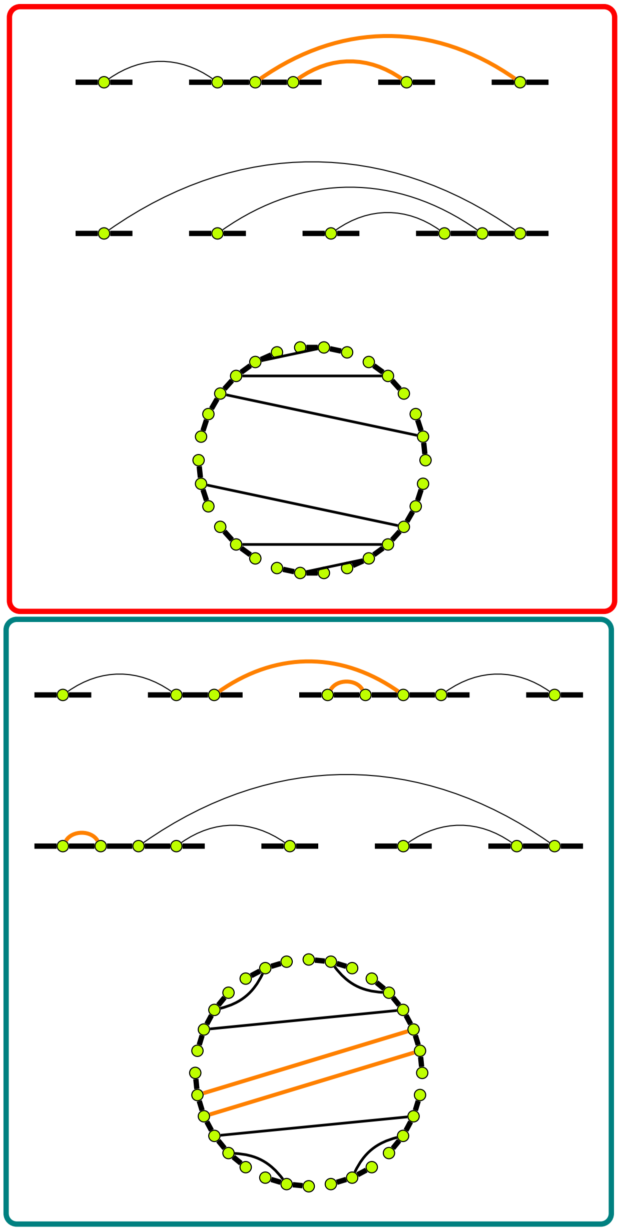

Figure 5. Two extended secondary structures, where marked base pairs are colored orange. In the red one, the innermost marked base pair is inter-strand (top), and therefore after the cyclic shift (middle), no base pair is marked, which results in an unconnected 2-symmetric structure (bottom). In the blue one, 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 6. *$S \in \Omega_r(R)$ 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 $C_0, \ldots, C_{r-1}$. 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 $S' - \{q \in S' \mid q \le p\}$ 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 $S' - \{q \in S' \mid q \le p\}$ is not connected, there is a marked interior base pair. By connectedness of S', there is a circular shift of the permutation, 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 7. *$S \in \Omega_{r\cdot i}(R)$ if and only if $f(S) \in \Omega_i(R_{/r})$.*

Proof. Assume $S \in \Omega_{r\cdot i}$. By observation 4, 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 8. *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 6, the decreasing of the rotational symmetry follows from lemma 7, 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 $0 \le p < \frac{m}{r}$ 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 not undefined for any $S \in \Omega_r(R)$. □

By lemma 8, 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 $O(3^m \cdot n^3)$ by lemma 5. We can finally conclude:

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

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

Lemma 10. *$\mathcal{Z}_{\max=r}$ can be computed in time $O(3^m \cdot n^3 \cdot m)$, for any r.*

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

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

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

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

By lemma 9, each $\mathcal{Z}_t$ can be computed in time $O(3^m \cdot n^3)$. This dominates the running time to compute one entry. Since we compute at most m entries, the running time 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 10, we can compute each $\mathcal{Z}_{\max=r}$ in time $O(3^m \cdot n^3 \cdot m)$. 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 $O(3^m \cdot n^3 \cdot m^2)$.*

Remark 2. *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. to the DP equations, if desired. Thus, the above result also holds for this variant of the partition function.*

Remark 3. *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). We briefly discuss the running time to show that we obtained a parametrized algorithm. 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 $O((\frac{m}{p})^p)$ 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 $O((\frac{m}{p})^{2p} \cdot n^3 \cdot p)$, which allows us to conclude:

Theorem 5. *There is an XP algorithm for MFE Triplet Repeat Strand Interaction parametrized by the number of different lengths p, running in $O((\frac{m}{p})^{2p} \cdot n^3 \cdot p)$ 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 the best we can hope for, and that, 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 is **NP**-complete, for a reasonable energy model. Note that for the general (non-triplet) case, this has already been shown in Condon et al., 2021. 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, each base pair gives a free energy reward of $E^{\text{bp}} = -\frac{m}{3}$, 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 $K_{\text{multi}} = +1$. Furthermore, each connected component reduces the strand association penalty by $-K_{\text{assoc}} := -1$. Finally, every hairpin loop must enclose at least three unpaired bases ($\theta = 3$). This model is extendable to the Turner model by setting equal energy values for interior and hairpin loops and account for the multiloop penalty in the corresponding energy values.

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 $E(R, S) \leq t$?

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 **P** ≠ **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:

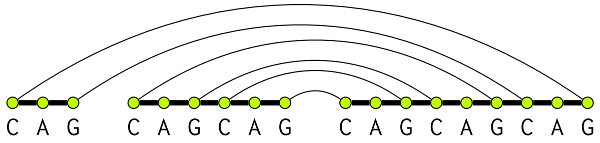

Figure 6. Optimal secondary structure corresponding to a valid summing triple $(1, 2, 3)$.

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

This has been shown to be strongly NP-hard (McDiarmid, 1999). 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

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

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

Proof. A hairpin loop must enclose at least three unpaired bases. Since in the *CAG* triplet pattern any two consecutive bases involve at least one C or one G, we can apply lemma 11 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 12, $E(R,S) > t$. 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 12) 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 11, $E(R,S) > t$. 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 12.

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 12. 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 11.

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 12. In the other case, we have a multiloop penalty of +1. Thus we can lower bound $E(R,S) \geq -n - 3nv + 1 > t$.

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 Strand soup interaction

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

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

where

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

The minimum free energy can be finally computed by

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

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

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

Regarding the running time, the table size is bounded by $m \cdot p^2 \cdot n^2 \cdot 3$, where $n := \max_{s\in R} |s|$. The running time to compute one table entry is dominated by the last case, where we minimize over $O(m \cdot p \cdot n)$ cutpoints and need $O(p)$ time for each new strand. In total, we obtain an algorithm with running time $O(n^3 \cdot m^2 \cdot p^4)$. We can then conclude:

Theorem 7. *MFE Unlimited Strand Interaction can be solved in time $O(n^3 \cdot m^2 \cdot p^4)$.*

Remark 4. *Additionally 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 favor larger sequences because they usually allow for more base pairs.

Remark 5. *The case of triplet repeats gives a slight improvement to 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 $O(m \cdot n^2)$, but the computation of one table entry still takes the same time, giving an overall time complexity of $O(n^3 \cdot m^2 \cdot p^2)$.*

5 Empirical proof of concept

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. 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 4. The source code to reproduce analyses is available at:

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

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 is also 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 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. 8).

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. 7. 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. 7. 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. A visualization and a small discussion can be found in the appendix. 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.

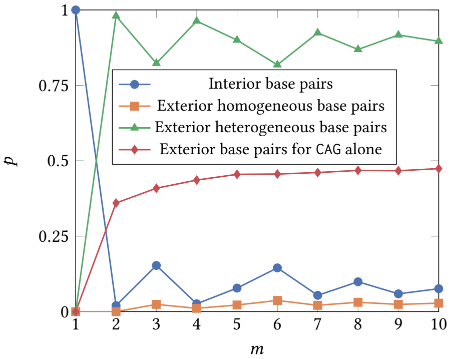

Figure 7. Probability p that a certain type of base pair is observed for increasing #strands m, either in a soup $\{\mathtt{CAU}_{20}, \mathtt{GGG}_{20}\}$, or for $\mathtt{CAG}_{20}$.

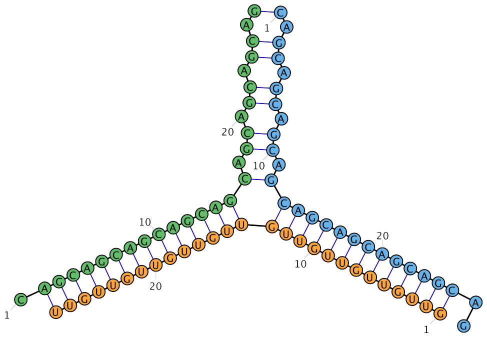

Figure 8. Exemplary MFE structure for strand pool $\{(\mathtt{GUU})^9, (\mathtt{CAG})^9, (\mathtt{ACG})^9\}$ computed by SoupFold with $m = 3$ (RiboSketch (Lu et al., 2018)).

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 (Hauser et al., 2024) 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.

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 Dirks et al., 2007. 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. One could extend the notion of "pseudoknot types", like kissing hairpins or H-type pseudoknots, by a specification of the strands each helix end comes from, and incorporate this extension to various algorithms for pseudoknots. For instance, during the internship, I briefly studied the incorporation of multiple strands to Auto-DP (Marchand et al., 2023). An overcounting of pseudoknotted structure in the partition function can be avoided by restricting the allowed pseudoknots by enforcing that for each crossing of two base pairs, at least one base pair must remain on one strand. Formalizing and extending these insights is also a desirable candidate for future work.

References

Aierken, Dilimulati and Jerelle A Joseph (2023). "Accelerated simulations of RNA phase separation: a systematic study of non-redundant tandem repeats". In: *bioRxiv*, pp. 2023–12.

Alkan, Can et al. (2006). "RNA–RNA interaction prediction and antisense RNA target search". In: *Journal of Computational Biology* 13.2, pp. 267–282.

Boehmer, Kimon et al. (2024). "RNA Triplet Repeats: Improved Algorithms for Structure Prediction and Interactions". In.

Bostan, Alin et al. (2007). "Differential equations for algebraic functions". In: *ISSAC'07: Proceedings of the 2007 international symposium on Symbolic and algebraic computation*. Ed. by C. W. Brown. ACM Press, pp. 25–32. DOI: 10.1145/1277548.1277553.

Boury, Théo, Laurent Bulteau, and Yann Ponty (2024). "RNA inverse folding can be solved in linear time for structures without isolated stacks or base pairs". In.

Bringmann, Karl et al. (2019). "Truly Subcubic Algorithms for Language Edit Distance and RNA Folding via Fast Bounded-Difference Min-Plus Product". In: *SIAM Journal on Computing* 48.2, pp. 481–512. DOI: 10.1137/17M112720X. eprint: https://doi.org/10.1137/17M112720X. URL: https://doi.org/10.1137/17M112720X.

Chang, Yi-Jun (2019). "Hardness of RNA folding problem with four symbols". In: *Theoretical Computer Science* 757, pp. 11–26. ISSN: 0304-3975. DOI: https://doi.org/10.1016/j.tcs.2018.07.010. URL: https://www.sciencedirect.com/science/article/pii/S0304397518304912.

Chaplick, Steven et al. (2023). "Snakes and ladders: a treewidth story". In: *International Workshop on Graph-Theoretic Concepts in Computer Science*. Springer, pp. 187–200.

Condon, Anne, Monir Hajiaghayi, and Chris Thachuk (2021). "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)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik.

Denise, A., Y. Ponty, and M. Termier (2010). "Controlled non-uniform random generation of decomposable structures". In: *Theoretical Computer Science* 411.40, pp. 3527–3552. ISSN: 0304-3975. DOI: https://doi.org/10.1016/j.tcs.2010.05.010. URL: https://www.sciencedirect.com/science/article/pii/S0304397510002914.

Dirks, Robert M et al. (2007). "Thermodynamic analysis of interacting nucleic acid strands". In: *SIAM review* 49.1, pp. 65–88.

Guo, Haotian et al. (2022). "Spatial engineering of E. coli with addressable phase-separated RNAs". In: *Cell* 185.20, pp. 3823–3837.

Hauser, Fanny, Ferdinand Ermel, and Kimon Boehmer (2024). *Clique Cover Based Vertex Cover Solver*. https://github.com/f-erm/CliqueCoverBasedVertexCoverSolver.

Huang, Liang et al. (July 2019). "LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search". In: *Bioinformatics* 35.14, pp. i295–i304. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz375. eprint: https://academic.oup.com/bioinformatics/article-pdf/35/14/i295/50721438/bioinformatics_35_14_i295.pdf. URL: https://doi.org/10.1093/bioinformatics/btz375.

Isiktas, Atagun U et al. (2022). "Systematic generation and imaging of tandem repeats reveal base-pairing properties that promote RNA aggregation". In: *Cell Reports Methods* 2.11.

Kurokawa, Ryo et al. (2023). "Clinical and neuroimaging review of triplet repeat diseases". In: *Japanese Journal of Radiology* 41.2, pp. 115–130.

Lipshitz, L. (1989). "D-Finite Power Series". In: *Journal of Algebra* 122.2, pp. 353–373.

Lu, Jacob S et al. (June 2018). "RiboSketch: versatile visualization of multi-stranded RNA and DNA secondary structure". In: *Bioinformatics* 34.24, pp. 4297–4299. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty468. eprint: https://academic.oup.com/bioinformatics/article-pdf/34/24/4297/48919841/bioinformatics_34_24_4297.pdf. URL: https://doi.org/10.1093/bioinformatics/bty468.

Maity, Hiranmay et al. (2023). "Odd–even disparity in the population of slipped hairpins in RNA repeat sequences with implications for phase separation". In: *Proceedings of the National Academy of Sciences* 120.24, e2301409120.

Marchand, Bertrand et al. (2023). "Automated design of dynamic programming schemes for RNA folding with pseudoknots". In: *Algorithms for Molecular Biology* 18.1, p. 18.

McDiarmid, Colin (1999). "Pattern minimisation in cutting stock problems". In: *Discrete applied mathematics* 98.1-2, pp. 121–130.

Nussinov, R and A B Jacobson (1980). "Fast algorithm for predicting the secondary structure of single-stranded RNA." In: *Proceedings of the National Academy of Sciences* 77.11, pp. 6309–6313. DOI: 10.1073/pnas.77.11.6309. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.77.11.6309. URL: https://www.pnas.org/doi/abs/10.1073/pnas.77.11.6309.

Salvy, B. and P. Zimmerman (1994). "GFUN: a Maple package for the manipulation of generating and holonomic functions in one variable". In: *ACM Transactions on Mathematical Software* 20.2, pp. 163–177. ISSN: 0098-3500.

Srinivasan, Sharan R. et al. (2023). "Chapter 18 - Repeat expansion disorders". In: *Neurobiology of Brain Disorders (Second Edition)*. Ed. by Michael J. Zigmond, Clayton A. Wiley, and Marie-Francoise Chesselet. Second Edition. Academic Press, pp. 293–312. ISBN: 978-0-323-85654-6. DOI: https://doi.org/10.1016/B978-0-323-85654-6.00048-4. URL: https://www.sciencedirect.com/science/article/pii/B9780323856546000484.

Turner, Douglas H. and David H. Mathews (Oct. 2009). "NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure". In: *Nucleic Acids Research* 38.suppl_1, pp. D280–D282. ISSN: 0305-1048. DOI: 10.1093/nar/gkp892. eprint: https://academic.oup.com/nar/article-pdf/38/suppl_1/D280/11217894/gkp892.pdf. URL: https://doi.org/10.1093/nar/gkp892.

Zuker, Michael and Patrick Stiegler (Jan. 1981). "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information". In: *Nucleic Acids Research* 9.1, pp. 133–148. ISSN: 0305-1048. DOI: 10.1093/nar/9.1.133. eprint: https://academic.oup.com/nar/article-pdf/9/1/133/6201945/9-1-133.pdf. URL: https://doi.org/10.1093/nar/9.1.133.

A Appendix for Section 3

A.1 Proof for Lemma 5

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. □

B Appendix for Section 4

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

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 $s = t \lor r = u$*
- *$R' = R$, $s = t$, $r = u$, $i \leq k$ 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 5. *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 13. *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 $\text{MFE}(I') \geq \text{MFE}([R, s_{i+1}, r_{j-1}, c]) = \bar{M}_{R,s_i,r_j,c}$ by assumption. Thus $\text{MFE}(]R, s_i, r_j, c[) = \bar{M}_{R,s_i,r_j,c}$.
- **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 14. *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 5 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 13 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 $O(2^{|R|} \cdot n)$. 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 14, we conclude.

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

$$\{s, t\} \subseteq R' \subseteq R \land 1 \leq k \leq |t| \land (k = |t| \to c \neq 1)$$
$$\land (s = t \to (k > i + \theta \land R' = \{s\}))$$
$$\land (r \in R' \to (t = r \land k < j \land R' = R \land c \neq 0))$$

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.

B.2 Proof for the connectivity in Section 4.5

Analogous to section 4.1, we define an interval $[m, s_i, r_j, c]$ and a relation $[m', t_k, u_\ell, c'] \preccurlyeq [m, s_i, r_j, c]$ if and only if $m' < m - 1$ or $m' = m - 1 \land (s = t \lor r = u)$ or $m' = m \land s = t \land r = u \land i \leq k \land \ell \leq j$. Since we just change the representation of our set R to an integer m, the correctness of the algorithm can be shown by the same arguments as for the exponential algorithm. We only show here that the connectivity specifier $c \in \{1, 2\}$ actually enforces connectivity. For this, we introduce the following notation: $\gamma(m, s_i, r_j)$ means that 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 MFE structure computed by $M_{m,s_i,r_j,2}$ is either connected or consists of two connected components, one containing s and one containing r. In other words, adding a base pair between s and r to such a structure will make it connected. Let $[m] := \{1, \ldots, m\}$.

Lemma 15. *$\gamma(m', s_i, r_j) \land \bar{\gamma}(m', s_i, r_j)$ for all $m' \in [m]$, $s, r \in R$, $i \leq |s|$ and $j \leq |r|$.*

Proof. Clearly, a secondary structure over an interval with $m = 1$ is always connected, 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 $\preccurlyeq$, assume that $\gamma(m', t_k, u_\ell)$ and $\bar{\gamma}(m', t_k, u_\ell)$ for any $[m', t_k, u_\ell, c'] \preccurlyeq [m, s_i, r_j, c]$. We show $\gamma(m, s_i, r_j)$ and $\bar{\gamma}(m, s_i, r_j)$. By case distinction:

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

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

□

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

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

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