HAL open science

Combinatorial RNA Design

Jozef Haleš, Alice Héliou, Ján Maňuch, Yann Ponty, Ladislav Stacho

To cite this version:

Jozef Haleš, Alice Héliou, Ján Maňuch, Yann Ponty, Ladislav Stacho. Combinatorial RNA Design. Algorithmica, 2017, 79 (3), pp.835–856. 10.1007/s00453-016-0196-x . hal-01285499v2

HAL Id: hal-01285499

https://inria.hal.science/hal-01285499v2

Submitted on 1 Aug 2016

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

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

Distributed under a Creative Commons Attribution 4.0 International License

Algorithmica manuscript No.
(will be inserted by the editor)

Combinatorial RNA Design

Designability and Structure-Approximating Algorithm in Watson-Crick and Nussinov-Jacobson Energy Models

Jozef Haleš · Alice Héliou · Ján Maňuch · Yann Ponty · Ladislav Stacho

Received: date / Accepted: date

Abstract We consider the *Combinatorial RNA Design problem*, a minimal instance of RNA design where one must produce an RNA sequence that adopts a given secondary structure as its minimal free-energy structure. We consider two free-energy models where the contributions of base pairs are additive and independent: the purely combinatorial *Watson-Crick model*, which only allows equally-contributing $\mathsf{A}-\mathsf{U}$ and $\mathsf{C}-\mathsf{G}$ base pairs, and the real-valued *Nussinov-Jacobson model*, which associates arbitrary energies to $\mathsf{A}-\mathsf{U}$, $\mathsf{C}-\mathsf{G}$ and $\mathsf{G}-\mathsf{U}$ base pairs.

We first provide a complete characterization of designable structures using restricted alphabets and, in the four-letter alphabet, provide a complete characterization for designable structures without unpaired bases. When unpaired bases are allowed, we characterize extensive classes of (non-)designable structures, and prove the closure of the set of designable structures under the stutter operation. Membership of a given structure to any of the classes can be tested in $\Theta(n)$ time, including the generation of a solution sequence for positive instances.

Finally, we consider a structure-approximating relaxation of the design, and provide a $\Theta(n)$ algorithm which, given a structure S that avoids two trivially non-designable motifs, transforms S into a designable structure constructively by adding at most one base-pair to each of its stems.

Keywords RNA structure · Inverse combinatorial optimization · String design

Acknowledgements The authors thank Cédric Chauve for fruitful discussions and constructive criticisms. YP is supported by the Pacific Institute for the Mathematical Sciences (PIMS), and the French Agence Nationale de la Recherche (ANR-14-CE34-0011).

Yann Ponty* and Alice Héliou
LIX (CNRS UMR 7161) Ecole Polytechnique & Inria Saclay, Palaiseau, France
∗ Correponding author
E-mail: yann.ponty@lix.polytechnique

Jozef Haleš and Ladislav Stacho
Department of Mathematics, Simon Fraser University, Canada

Ján Maňuch
Department of Mathematics, Simon Fraser University, Canada
Present address: Department of Computer Science, University of British Columbia, Canada

1 Introduction

RiboNucleic Acids (RNAs) are biomolecules which act in almost every aspect of cellular life, and can be abstracted as a sequence of nucleotides, i.e., a string over the alphabet $\{\mathsf{A}, \mathsf{U}, \mathsf{C}, \mathsf{G}\}$. Due to their versatility, and the specificity of their interactions, they are increasingly being used as therapeutic agents [24], and as building blocks for the emerging field of synthetic biology [18, 20]. A substantial proportion of the functional roles played by RNA rely on interactions with other molecules to activate/repress dynamical properties of some biological system, and ultimately require the adoption of a specific conformation. Accordingly, RNA bioinformatics has dedicated much effort to developing energy models [15, 22] and algorithms [16, 28] to predict the secondary structure of RNA, a combinatorial description of the conformation adopted by an RNA which only retains interacting positions, or base pairs. Historically, structure prediction has been addressed as an optimization problem, whose expected output is a secondary structure which minimizes some notion of free-energy [16, 28]. The performances of the RNA folding prediction problem have now reached a point where *in silico* predictions are generally considered reliable [15], allowing for large scale studies and fueling the discovery of an increasing number of functional families [9].

Due to the existence of expressive, yet tractable, energy models and conformational spaces, coupled with promising applications in multiple fields (pharmaceutical research, natural computing, biochemistry. . .), a wide array of computational methods [10, 4, 1, 5, 2, 21, 25, 13, 14, 8, 11, 17, 27, 3, 6] has been proposed to tackle the natural inverse version of the structure prediction, the RNA design problem. In this problem, one attempts to perform the *in silico* synthesis of artificial RNA sequences, performing a predefined biological function *in vitro* or *in vivo*. Given the prevalence of structure in the function of an RNA, one of the foremost goal of RNA design (sometimes named inverse folding in the literature) is to ensure that the designed sequence folds into a predefined secondary structure, preferentially to any alternative structure. In other words, the chosen conformation should not be challenged by alternative stable structures having similar or lower free-energy.

Despite a rich, fast-growing, body of literature dedicated to the problem, there is currently no exact polynomial-time algorithm for the problem. Moreover, the complexity of the problem remains open (see Section 5 for a discussion). It can be argued that this situation, quite exceptional in the field of computational biology, partly stems from the intricacies of the Turner free-energy model [22] which associates experimentally-determined energy contributions to $\sim 2.4 \times 10^4$ structure/sequence motifs. This motivates a reductionist approach, where one studies an idealized version of the RNA design problem, lending itself to algorithmic intuitions, while hopefully retaining the presumed difficulty of the original problem and provides intuitions for future studies of the problem under more sophisticated energy models.

In this work, we introduce the *Combinatorial RNA Design problem*, a *minimal* instance of the RNA design problem which aims at finding a sequence that admits the target structure as its unique base pair maximizing structure. After this short introduction, Section 2 states definitions and problems. In Section 3, we state our main results and prove them in Section 4, including an extended weighted version that allows additional types of base pairs. Finally, we conclude in Section 5 with some remarks, open problems and future extensions of this work.

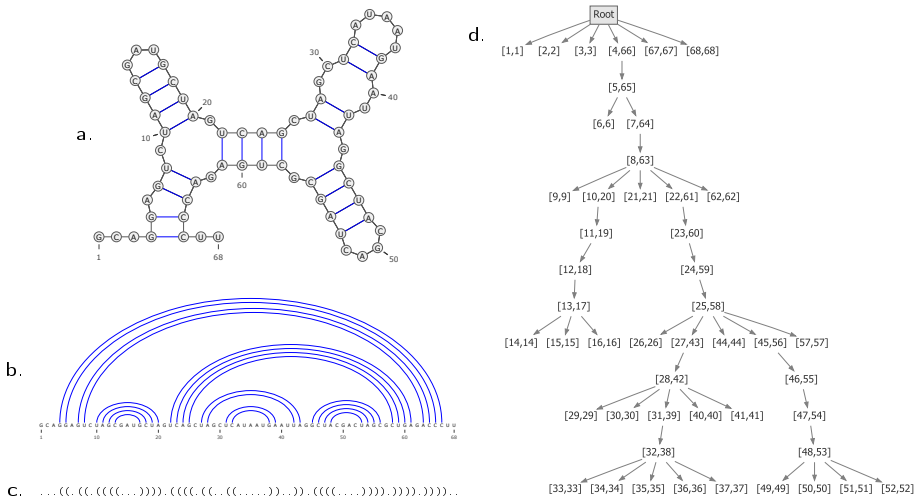

Fig. 1 Four equivalent representations for an RNA secondary structure of length 68, consisting of 20 base pairs forming 7 bands: outer-planar graph (**a.**), arc-annotated representation (**b.**), parenthesized expression (**c.**), and tree representation (**d**)

2 Definitions and notations

RNA secondary structure. An RNA can be encoded as a sequence of nucleotides, i.e., a string $w = w_1 \cdots w_{|w|} \in \{\mathsf{A}, \mathsf{U}, \mathsf{C}, \mathsf{G}\}^\star$. The prefix of w of length i is denoted as $w_{[1,i]}$ and $|w|_b$ denotes the number of occurrences of b in w. A (pseudoknot-free) secondary structure S on an RNA of length n is a pair (n, P), where P is a set of base pairs $\{(l_i, r_i)\}_{i=1}^{p} \subset [1, n]^2$ such that:

- $\forall i \in [1, p]$, $l_i < r_i$;
- Each position is involved in at most one base pair, *i.e.* $\forall i \neq j \in [1, p], l_i \neq l_j$, $l_i \neq r_j$, $r_i \neq r_j$;
- Base pairs are pairwise non-crossing,*i.e.* $\nexists i, j \in [1, p]$, $l_i < l_j < r_i < r_j$, .

The set of unpaired positions U_S of a secondary structure $S = (n, P)$ is the set of indices $k \in [1, n]$ that are not involved in any of the base pairs in S. A structure S is called saturated if and only if its positions are all paired, *i.e.* iff $U_S = \emptyset$. Conversely, a structure S is empty when none of its positions is paired, *i.e.* iff $S = (n, \emptyset)$.

The set of all secondary structures is denoted by $\mathcal{S}$, and its restriction to structures of length n by $\mathcal{S}_n$. Secondary structures are typically expressed using a variety of equivalent representations, illustrated by Figure 1 and formally defined further in this section.

Given a sequence w and a structure $S = (|w|, P)$, let $u_i = \varepsilon$ if $i \in U_S$ and $u_i = w_i$, otherwise, where ε is the empty sequence. Define the S-paired restriction of w, denoted by $\mathsf{Paired}(w, S)$, as the subsequence of w consisting of the paired positions only, *i.e.* $\mathsf{Paired}(w, S) = u_1 \cdots u_{|w|}$. Similarly, define the paired restriction of $S = (n, P)$, denoted by $\mathsf{Paired}(S)$, as the substructure of S consisting of paired positions only, *i.e.* $\mathsf{Paired}(S) = (|\mathsf{Paired}(w, S)|, \{(|u_1 \cdots u_i|, |u_1 \cdots u_j|) \mid (i, j) \in P\})$, where w is any sequence of length n and u_i's are defined as above.

A maximal subset $B = \{(i,j),(i+1,j-1),\ldots,(i+\ell-1,j-\ell+1)\}$ of P for some integer i,j,ℓ is called a band (also referred to as helix or stem in related works) of size $\ell = |B|$, of $S = (n,P)$. Note that every base pair belongs to exactly one band. In other words, the base-pairs of a secondary structure can be unambiguously partitioned into a set of bands.

Dot-parentheses notation. A well-parenthesized sequence $s \in \{\texttt{(},\texttt{)},\texttt{.}\}^*$ can be used to represent a secondary structure. There exists a one-to-one correspondence between secondary structures and such well-parenthesized sequences: any base pair $(l,r) \in P$ becomes a pair of corresponding opening and closing parentheses in s at position l and r respectively ($s_l = \texttt{(}$ and $s_r = \texttt{)}$), and any unpaired position i corresponds to a dot ($s_i = \texttt{.}$). This representation is illustrated by Figure 1.c. A concatenation of two structures S and S', denoted by $S.S'$ or simply SS' wherever unambiguous, is the structure corresponding to the well-parenthesized sequence obtained by concatenating the well-parenthesized sequences of S and S'.

k-stutter. The k-stutter of a sequence s, denoted by $s^{[k]}$ is the result of an independent copy k-times of each of the characters in s. For instance, the 3-stutter of a sequence AUUC is AAAUUUUUUCCC. This operation also applies to an RNA structure S, and $S^{[k]}$ denotes the RNA structure obtained by applying the usual k-stutter to the dot-parentheses representation of S.

Tree representation. Alternatively, the tree representation, denoted by T_S, for $S = (n,P)$ is a rooted ordered tree whose vertex set V_S consists of intervals $[l,r]$ for any base pair $(l,r) \in P$, and $[k,k]$ for every $k \in U_S$. A virtual root $[0,n+1]$ is added for convenience. Any node labeled by $[k,k]$ is called unpaired, and any other node (including the virtual root) are considered as paired. The children of an interval $I \in V_S$ are the maximal proper subintervals $I' \in V_S$ of I ordered by the left points of the intervals. The degree of a vertex $I \in V_S$ is the total number of its paired neighbors, including its parent (if any). We denote by $D(S)$ the maximal degree of nodes in T_S. Figure 1.d shows the tree representation of a typical secondary structure.

Proper, greedy and separated coloring of the tree representation. Consider the tree representation T_S of structure S. A coloring of T_S associates a color, chosen among black, white, or gray, to each paired node of T_S that is different from the root. This coloring is called proper if:

i) Each node has at most one black child, at most one white child, and at most two grey children;
ii) Any c-colored node has at most one c-colored child;
iii) Black nodes shall not have a white child, and white nodes shall not have a black child.

A greedy coloring of T_S is the coloring obtained by recursive application of the following rule starting from the root and continuing towards leaves: if the node is black, color the first paired child black and the remaining paired children gray, if the node is white, color the first paired child white and the remaining paired children gray, otherwise (the gray node or the root), color the first paired child

black, second white and the remaining paired children gray. It is easy to check that if the degree of each node is at most four then the greedy coloring is a proper coloring.

Given a proper coloring of T_S, let the level of each node be the number of black nodes minus the number of white nodes on the path from this node to the root. A proper coloring is called separated if the two sets of levels, associated with gray and unpaired nodes respectively, do not overlap.

2.1 Statement of the generic RNA design problem

Consider an energy model $\mathcal{M}$, which associates a free-energy $E_{\mathcal{M}}(w, S) \in \mathbb{R}^- \cup \{+\infty\}$ to each secondary structure $S \in \mathcal{S}_{|w|}$ for a given RNA sequence w. The minimum free-energy (MFE) structure prediction problem is typically defined as follows:

RNA-FOLD$_{\mathcal{M}}$ problem
Input: RNA sequence w
Output: $S^\star_{\mathcal{M}}(w) := \operatorname{argmin}_{S' \in \mathcal{S}_{|w|}} E_{\mathcal{M}}(w, S')$,

where argmin returns a single structure $S^\star_{\mathcal{M}}(w)$, arbitrarily chosen amonst those having minimum free-energy.

The existence of alternative competing structures, *i.e.* one or several secondary structure(s) having (almost) minimal free-energy for a given RNA, impacts the efficacy of the folding process. The detection of such situations is therefore of interest, and can be rephrased as the following problem:

UNIQUE-FOLD$_{\mathcal{M}}$ problem
Input: Sequence w + Energy distance $\Delta > 0$
Output: True if, for every $S' \in \mathcal{S}_{|w|} \setminus \{S^\star_{\mathcal{M}}(w)\}$, one has:

$$E_{\mathcal{M}}(w, S') \geq E_{\mathcal{M}}(w, S^\star_{\mathcal{M}}(w)) + \Delta\,.$$

False otherwise.

We can now define the combinatorial RNA Design problem as:

RNA-DESIGN$_{\mathcal{M},\Sigma}$ problem
Input: Secondary structure S + Energy distance $\Delta > 0$
Output: RNA sequence $w \in \Sigma^\star$ – called an $(\mathcal{M}, \Sigma, \Delta)$-design for S – such that:

$$\text{RNA-FOLD}_{\mathcal{M}}(w) = S \quad \text{and} \quad \text{UNIQUE-FOLD}_{\mathcal{M}}(w, \Delta),$$

or $\varnothing$ if no such sequence exists.

The structures for which there exists an $(\mathcal{M}, \Sigma, \Delta)$-design are called $(\mathcal{M}, \Sigma, \Delta)$-designable. Let Designable$(\mathcal{M}, \Sigma, \Delta)$ be the set of all such structures. If it is clear from the context, we will usually drop $\mathcal{M}$, Σ and/or Δ.

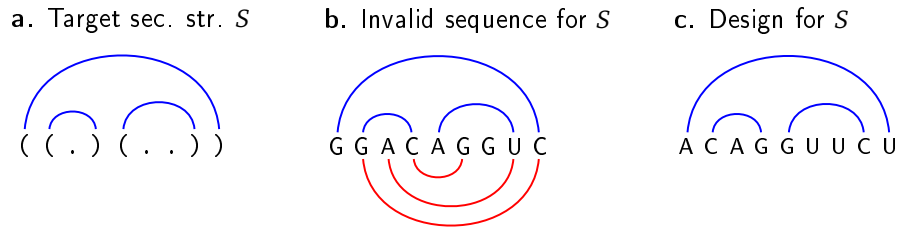

Fig. 2 The combinatorial RNA design problem: Starting from a secondary structure S (**a.**), our goal is to design an RNA sequence which uniquely folds, with maximum number of base pairs, into S. The sequence proposed in **b.** is invalid due to the existence of an alternative structure (lower half-plane, red) having the same number of base pairs as S. The right-most sequence (**c.**) is a design for S.

2.2 Base pair sum energy models

In this work, we will consider two types of base pair sum energy models, where the free-energy of a structure is simply obtained by sum, over all base pairs, individual independent contributions associated with each pair.

Definition 1 (Base pair sum energy model $\mathcal{M}$) *Let w be an RNA sequence and S a secondary structure in $\mathcal{S}_{|w|}$. Then*

$$E_{\mathcal{M}}(w, S) = \sum_{(l,r)\in S} E_{\mathcal{M}}(w_l, w_r)\,,$$

where $E_{\mathcal{M}}(x, y)$ is the energy induced by a base pair $x - y$.

To define a model of interest, it is sufficient to specify the energies of base pairs. First, we consider a minimal energy model, named Watson-Crick model due to its similarity with the DNA base-pairing rules. The model is purely combinatorial, as it associates a homogenous -1 energy contribution to each valid base-pair, and only allows $\mathsf{G} - \mathsf{C}$ and $\mathsf{A} - \mathsf{U}$ to pair.

Definition 2 (Watson-Crick energy model $\mathcal{W}$)

$$E_{\mathcal{W}}(x, y) = \begin{cases} -1 & \textit{if } \{x, y\} = \{\mathsf{G}, \mathsf{C}\} \textit{ or } \{x, y\} = \{\mathsf{A}, \mathsf{U}\} \\ +\infty & \textit{otherwise.} \end{cases}$$

A more general model, named the Nussinov-Jacobson model, allows $\mathsf{G} - \mathsf{U}$ base-pairs to occur, and associates arbitrary weights to the base pairs depending on their content. It is named after the authors of the first polynomial-time algorithm for predicting the MFE under a similar energy model [16].

Definition 3 (Nussinov-Jacobson energy model $\mathcal{N}$)

$$E_{\mathcal{N}}(x, y) = \begin{cases} \alpha & \textit{if } \{x, y\} = \{\mathsf{G}, \mathsf{C}\} \\ \beta & \textit{if } \{x, y\} = \{\mathsf{A}, \mathsf{U}\} \\ \gamma & \textit{if } \{x, y\} = \{\mathsf{G}, \mathsf{U}\} \\ +\infty & \textit{otherwise.} \end{cases}$$

where $\alpha, \beta, \gamma \in \mathbb{R}^-$ and $\alpha, \beta < \gamma$.

Note that the last condition of the above definition is typically satisfied by empirical estimates of base pair energies. Namely, $\mathsf{G}-\mathsf{U}$ base pairs, also named Wobble base pairs, are much weaker than its alternatives. They are mediated by a single hydrogen bond, as opposed to 2 (resp. 3) bonds for $\mathsf{A}-\mathsf{U}$ (resp. $\mathsf{G}-\mathsf{C}$).

We say that the structure is compatible with a sequence w, according respect to an energy model $\mathcal{M}$, if $E_{\mathcal{M}}(w,S) < +\infty$.

It is worth noting that minimizing $E_{\mathcal{W}}(w,S)$ is equivalent to maximizing $|S|$, thus RNA-FOLD$_{\mathcal{W}}$ is a classic base pair maximization problem. Moreover, both RNA-FOLD$_{\mathcal{W}}$ and RNA-FOLD$_{\mathcal{N}}$ can be solved in polynomial time using dynamic programming, historically in $\mathcal{O}(n^3)$ complexity [16], or in $\mathcal{O}(n^3/\log(n))$ current best time complexity [7]. A backtracking procedure reconstructs the structure having minimal energy, and can be easily adapted to provide a $\Theta(n^3)$ algorithm for the UNIQUE-FOLD$_{\mathcal{M}}$ problem.

3 Statement of the results

In this section, we characterize sets of secondary structures which can or cannot be designed using (a subset of) $\{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{U}\}$, a (restricted) set of base pairs and a desired energy distance Δ within an energy model $\mathcal{M}$. The proofs of our statements are largely interconnected, and have been regrouped in Section 4.

First, let us remark that the empty secondary structures are the only ones that are designable for arbitrary large energy distances Δ.

Theorem 1 *For any $\mathcal{M} \in \{\mathcal{W},\mathcal{N}\}$, and any energy distance Δ such that*

$$\Delta > -E_{\mathcal{M}}(\mathsf{X},\mathsf{Y}), \quad \forall \mathsf{X},\mathsf{Y} \in \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{U}\}^2, \tag{1}$$

only the empty secondary structures are designable.

Proof For any non-empty secondary structure S having energy E on some sequence w, removing any base pair $\mathsf{X}-\mathsf{Y}$ from S yields an alternative structure S' whose energy is $E' = E - E_{\mathcal{M}}(\mathsf{X},\mathsf{Y}) < E + \Delta$. In other words, S' is a competing structure at distance less than Δ of S, *i.e.* w is not a valid Δ-design for S.

Moreover, any empty structure of length n is designable. Indeed, none of the models allows for pairs of the form $\mathsf{X}-\mathsf{X}, \mathsf{X} \in \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{U}\}$, so any sequence X^n admits the empty structure, having 0 energy, as its only secondary structure having finite free-energy, *i.e.* X^n is a design for the empty structure for any finite $\Delta > 0$. □

3.1 Watson-Crick model $\mathcal{W}$ ($\Delta = 1$)

We provide (partial) characterizations for the sets Designable(Σ) of designable structures over partial alphabets Σ in the $\mathcal{W}$ model. From Theorem 1, combined with the purely combinatorial nature of the energy model, we observe that non-trivial structures are designable only when $\Delta \in (0,1]$, and that the set of designable structures is unaffected by the precise value of Δ on the segment. Therefore we set $\Delta = 1$ without loss of generality, and consider the structures whose designed sequences lose at least one base-pair when forming alternative structures. We obtain the following meta-theorem.

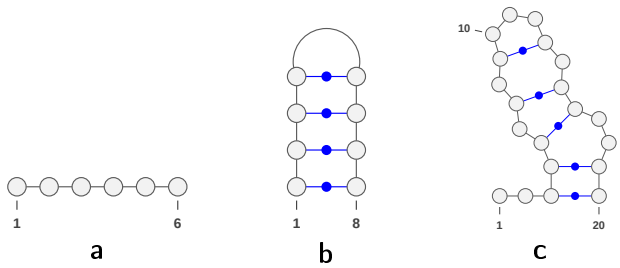

Fig. 3 Examples of structures satisfying conditions of **R1** (**a.**), **R2** (**b.**) and **R3** (**c.**).

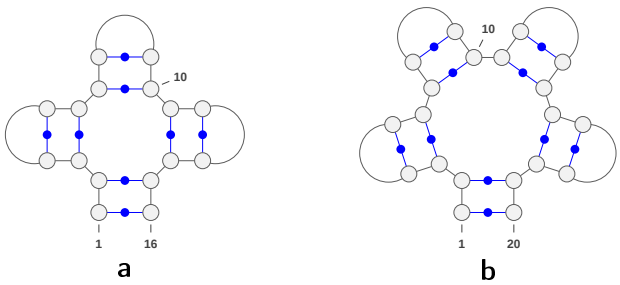

Fig. 4 Examples of two saturated structures, one satisfying conditions of **R4** (**a.**) and one that does not (**b.**). The structure on the left has pair degree 4 and is designable, and the structure on the right has pair degree 5 and is not designable.

Theorem 2 *In the Watson-Crick energy model $\mathcal{W}$, and assuming an energy distance of $\Delta = 1$, results* **R1** *through* **R8** *hold.*

Let $\Sigma_{c,u}$ be an alphabet with c pairs of complementary bases and u bases without a complementary base. Without loss of generality in the $\mathcal{W}$ model, we will assume that $\Sigma_{1,0} = \{\mathsf{G}, \mathsf{C}\}$ and $\Sigma_{1,1} = \{\mathsf{G}, \mathsf{C}, \mathsf{A}\}$.

3.1.1 Designability over restricted alphabets.

First, we provide a complete characterization of designable (secondary) structures using an alphabet Σ of restricted cardinality:

R1 For every $u \in \mathbb{N}^+$, $\mathsf{Designable}(\Sigma_{0,u}) = \{(n, \emptyset) \mid \forall n \in \mathbb{N}\}$;
R2 $\mathsf{Designable}(\Sigma_{1,0}) = \{S \in \mathcal{S} \mid S \text{ is saturated and } D(S) \leq 2\} \cup \{(n, \emptyset) \mid \forall n \in \mathbb{N}\}$;
R3 $\mathsf{Designable}(\Sigma_{1,1}) = \{S \in \mathcal{S} \mid D(S) \leq 2\}$.

Figure 3 shows examples of secondary structures satisfying conditions of these three results.

3.1.2 Designability over the complete alphabet $\Sigma_{2,0} = \{\mathsf{A}, \mathsf{U}, \mathsf{C}, \mathsf{G}\}$.

We first characterize the set of designable saturated structures, *i.e.* structures whose positions are all paired to some other position.

R4 $\{S \in \mathsf{Designable}(\Sigma_{2,0}) \mid S \text{ is saturated}\} = \{S \in \mathcal{S} \mid D(S) \leq 4 \text{ and } S \text{ is saturated}\}$.

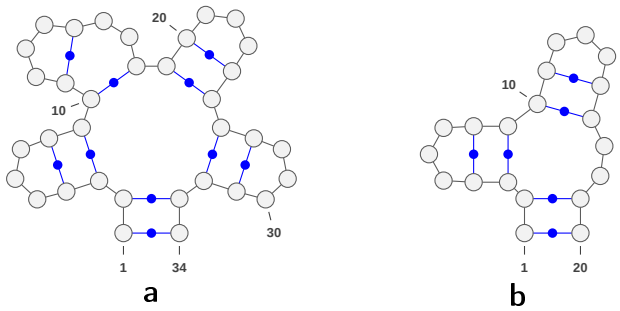

Fig. 5 Examples of two structures containing motifs m_5 (**a.**) and $m_{3\circ}$ (**b.**), respectively. By **R5** these two structures are not designable.

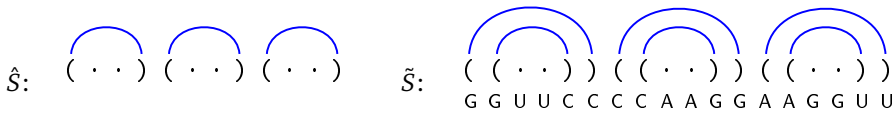

Fig. 6 An example of non-designable (left) and designable structure (right).

Figure 4 shows two saturated secondary structures one with pair degree 4 and one with pair degree 5. By **R4**, the former is $\Sigma_{2,0}$-designable, while latter is not.

When unpaired positions are allowed in the target structure, our characterization is only partial:

R5 Let m_5 represent "a node having degree more than four", and $m_{3\circ}$ be "a node having one or more unpaired children, and degree greater than two" (cf. Figure 5), then

$$\mathsf{Designable}(\Sigma_{2,0}) \cap \{S \in \mathcal{S} \mid S \text{ contains } m_5 \text{ or } m_{3\circ}\} = \emptyset\,;$$

R6 Let Sep be the set of structures for which there exists a separated (proper) coloring of the tree representation, then $\mathsf{Sep} \subset \mathsf{Designable}(\Sigma_{2,0})$;

R7 The set of $\Sigma_{2,0}$-designable structures is closed under the k-stutter operations:

$$\forall S \in \mathcal{S}, \forall k \in \mathbb{N}^+ : \quad S \in \mathsf{Designable}(\Sigma_{2,0}) \implies S^{[k]} \in \mathsf{Designable}(\Sigma_{2,0})\,.$$

We note however that reverse implication is not true: $S^{[k]} \in \mathsf{Designable}(\Sigma_{2,0})$ does not imply that $S \in \mathsf{Designable}(\Sigma_{2,0})$. For instance, it is easily verified that $\hat{S}^{[2]}$ is $\Sigma_{2,0}$-designable, while $\hat{S}$ is not, as shown in Figure 6. Membership to the classes described in **R1-R5** can be tested by trivial linear-time algorithms. These algorithms can also be easily adapted into linear-time algorithms for the production of a concrete design, thereby offering partial solutions to the $\mathsf{RNA\text{-}DESIGN}_{\mathcal{M},\Sigma}$ problem.

3.1.3 Structure-approximating algorithm.

Unfortunately, avoiding m_5 and $m_{3\circ}$, while necessary, is generally not sufficient to ensure designability. For instance, consider $\hat{S}$ in Figure 6 clearly does not contain

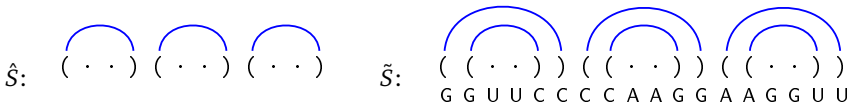

Fig. 7 Application of the structure-approximating algorithm to the non-designable structure $\hat{S}$ in Fig. 6: A base pair (circled black node) is inserted in the greedily colored tree, offsetting the levels of white and unpaired nodes (crosses) to even and odd levels respectively, so that the resulting tree is proper/separated, representing a designable structure.

m_5 or $m_{3\circ}$, yet it cannot be designed. In such cases, unwanted interactions can be somehow penalized by duplicating some base pairs. For instance, duplicating a single base pair in $\hat{S}$ yields a $\Sigma_{2,0}$-designable structure $\tilde{S}$, as shown by Figure 7.

R8 Any structure S avoiding m_5 and $m_{3\circ}$ can be transformed in $\Theta(n)$ time into a $\Sigma_{2,0}$-designable structure S'. This is done by duplicating a subset of the base pairs of S, at most one per band, such that the greedy coloring of the resulting structure is proper and separated, as illustrated by Figure 7.

3.2 Nussinov-Jacobson energy model $\mathcal{N}$ ($\Delta \leq \min(|\alpha|, |\beta|)$])

We consider the validity of the above results in the Nussinov-Jacobson model. Note that the consideration of $\mathsf{G}-\mathsf{U}$ base pairs, by loosening the notion of complementarity, forces us to abuse our notation for $\Sigma_{i,j}$. Namely, $\Sigma_{2,0}$ is taken to represent the full alphabet, even though it now strictly allows three types of base pairs.

Theorem 3 *For any $\Delta \in (0, \min(|\alpha|, |\beta|)]$, statements* **R1** *through* **R8** *hold in the Nussinov-Jacobson model.*

4 Proofs

4.1 Watson-Crick model ($\Delta = 1$)

R1 is trivial since, in the absence of complementary letters, empty structures are the only one whose energy is not infinite.

Theorem 4 ($\Rightarrow$ Result R2, R3 and R4) *A saturated secondary structure S is $\Sigma_{c,0}$-designable if and only if $D(S) \leq 2c$.*

Proof First, we will show that the degree condition is necessary. Assume to the contrary that $D(S) > 2c$ and S has a design w. Let $[a, b]$ be a vertex with degree $d \geq 2c + 1$ in T_S. Let $\{[l_i, r_i]\}_{i=1}^{d}$ be the (paired) children of $[a, b]$ and the node $[a, b]$ if $[a, b]$ is not the root. Let $L_i = l_i$ and $R_i = r_i$ if $[l_i, r_i]$ is a child of $[a, b]$, and $L_i = r_i$ and $R_i = l_i$ if it is $[a, b]$. Then among bases $w_{L_1}, \ldots, w_{L_d}$ must be a pair of repeated letters. Let $w_{L_i} = w_{L_j}$ be such a pair with $L_i < L_j$. It is easy to check that $S \setminus \{(l_i, r_i), (l_j, r_j)\} \cup \{(L_i, R_j), (R_i, L_j)\}$ is a structure compatible with

w with the same number of base pairs as S, a contradiction with the assumption that w is a design for S.

To show that the degree condition is also sufficient, we need further definitions and claims. First, we say that a sequence $w \in \Sigma^*$ is saturable if there is a saturated structure compatible with w. Note that the concatenation of two saturable sequences is also saturable. Then the following claim characterizes the cases when a saturable sequence can be split into saturable sequences.

Claim 4.1 *Let $w = uv$ be a saturable sequence of length k. If u is saturable, then so is v.*

Proof Consider a saturated structure S compatible with sequence w and a saturated structure S_u compatible with u. We will construct a saturated structure S_v compatible with v.

Consider a graph G with vertex set $\{1, \ldots, k\}$ and edge set defined by pairs in $S \cup S_u$. Obviously, this graph is a collection of alternating paths (alternating between pairs from S and from S_u, starting and ending with positions in v) and alternating cyclic paths, and it has a planar embedding such that all vertices lie on a line in their order: pairs in S are drawn as non-crossing arcs above the line and pairs in S_u as non-crossing arcs below the line. Note that every position in v is an end-point of exactly one path in the collection.

Define set of base pairs S_v by pairing the end-points of the paths in G, cf. Figure 8. We will show that S_v is a structure. Consider a graph G' constructed by adding pairs in S_v to G. This graph is a collection of cyclic paths. Consider an embedding of G' into plane that extends the planar embedding of G by adding arcs corresponding to the pairs in S_v below the line containing all the vertices. If two base pairs $b, b' \in S_v$ cross then the cyclic path containing b and the cyclic path containing b' intersect in exactly one point. By Jordan's curve theorem, this is a contradiction. It follows that S_v is a saturated structure, and hence v is also saturable. □

We define w to be an atomic saturable sequence if no proper prefix of w is saturable. Every saturated structure compatible with an atomic saturable sequence w contains the base pair $(1, |w|)$, since otherwise it contains the base pair $(1, j)$, with $j < |w|$, and consequently $w_{[1,j]}$ is a saturable proper prefix of w. On the other hand, by Claim 4.1, if every saturated structure compatible with w contains the pair $(1, |w|)$, then w is an atomic saturable sequence. A design w that is also an atomic saturable sequence will be called an atomic saturable design. A concatenation of two or more atomic saturable designs is obviously not an atomic saturable sequence and it is not necessarily a design. However, we have the following claims.

Claim 4.2 *The concatenation of t atomic saturable designs $w^1, \ldots, w^t$ for structures $S^1, \ldots, S^t$, such that $w_1^i \neq w_1^j, \forall 1 \leq i < j \leq t$, is a design for the concatenated (saturated) structure $S = S^1 \cdots S^t$.*

Proof Assume that $W := w^1 \cdots w^t$ is not a design, then there exist a saturated structure $S' \neq S$ for W. We show that positing such an alternative structure leads to a contradiction. Recall that each S^i is saturated and contains a pair $(1, |w^i|)$. If S' pairs the first and last letters in each w^i, $i \in [1, t]$, then $S' = S$ since each w^i

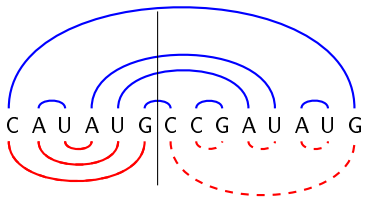

Fig. 8 Construction of the saturated structure compatible with the suffix v. The vertical line splits the sequence into a prefix u and a suffix v. Top (blue) and bottom (red) arcs depict saturated structures compatible with w and u respectively. Dashed arcs represent the induced saturated structure compatible with v, they are connecting end-points of the alternating bottom/top full path.

is a design, a contradiction. Let w_i be the leftmost sequence such that w_1^i is not paired with $w_{|w^i|}^i$ in S'. Since S' must be also saturated, w_1^i must be paired. Let w_k^j, $j \geq i$, be the partner of w_1^i in S', and let $u := w^i \cdots w^{j-1} w_{[1,k]}^j$. If $k = |w^j|$, then $j > i$ and, by complementarity, $w_1^i = w_1^j$ which contradicts the preconditions. Hence, we can assume that $k < |w^j|$. Since u and each of the $w^i, \ldots, w^{j-1}$ are saturable, by iterated application of Claim 4.1, we conclude that $v = w_{[1,k]}^j$ is saturable as well. This contradicts the precondition that w^j is an atomic saturable design, since v is a proper prefix of w^j. We conclude that no alternative saturated folding exists for W, i.e., W is a design for S. □

Claim 4.3 *Consider t atomic saturable designs $w^1 = w_1^1 \cdots w_{|w^1|}^1$, $\ldots$, $w^t = w_1^t \cdots w_{|w^t|}^t$ and a pair a, b of complementary letters such that $w_1^i \neq b$ for every $1 \leq i \leq t$ and $w_1^i \neq w_1^j$ for every $1 \leq i < j \leq t$. Then $W = aw^1 \cdots w^t b$ is an atomic saturable design.*

Proof We will first show that W is an atomic saturable sequence. Assume to the contrary that there is a proper prefix of W that is saturable. Consider the shortest such prefix $aw^1 \cdots w^i w_{[1,j]}^{i+1}$ with $1 \leq j < |w^{i+1}|$ and $1 \leq i < t$. Obviously, a has to be paired with w_j^{i+1}, otherwise we can find a shorter saturable prefix. This implies that $b = w_j^{i+1}$ and that $w^1 \cdots w^i w_{[1,j-1]}^{i+1}$ is saturable as well. By repeated application of Claim 4.1, we have that $w_{[1,j-1]}^{i+1}$ is saturable. Since it is a prefix of atomic saturable sequence w^{i+1}, it must be the empty sequence, i.e., $j = 1$. Therefore, $b = w_1^{i+1}$, a contradiction with the assumptions of the claim. Thus, W is an atomic saturable sequence.

Now we will show that W is a design. Consider any MFE (saturated) structure S for W. Since W is atomic saturable, a is paired with b in S. By Claim 4.2, $w^1 \cdots w^t$ is a design. It follows that W is a design as well. □

To prove the sufficiency of the degree condition, consider the following algorithm, which takes as input a saturated structure S with $D(S) \leq 2c$, and returns a design w for S:

- Let $\{[l_i, r_i]\}_{i=1}^d$ be the children of the root. Assign to each w_{l_i}, w_{r_i} complementary bases such that $\forall 1 \leq i < j \leq d: \; w_{l_i} \neq w_{l_j}$.

– While there exists an unprocessed internal node $[a, b]$ whose parent has been processed (if there is no such node, stop and return w). Let $\{[l_i, r_i]\}_{i=1}^{d}$ be the children of $[a, b]$. Assign to each w_{l_i}, w_{r_i} complementary bases such that $\forall 1 \leq i \leq d: \ w_{l_i} \neq w_b$ and $\forall 1 \leq i < j \leq d: \ w_{l_i} \neq w_{l_j}$.

Note that since the alphabet contains c pairs of complementary bases, the assignment at each step of the algorithm is possible. We will show that the returned sequence w is a design for S. We will show by tree induction on the size subtrees that $w_i \cdots w_j$ is an atomic saturable design for every internal node $[i, j]$. It is easy to check that this is satisfied at the leaves. Consider an internal node u. By the induction hypothesis, sequences for each child subtree of u are atomic saturable designs. Furthermore, by the choice of bases at children nodes of u, all assumptions of Claim 4.3 are satisfied, hence, the sequence for node u is also an atomic saturable design. The claim holds. Finally, we can apply Claim 4.2 at the root, which yields that w is a design. □

Corollary 5 (Result R2) *A structure S is $\Sigma_{1,0}$-designable if and only if it does not contain any base pairs, or it is saturated and $D(S) \leq 2$.*

Proof If S contains a base pair and an unpaired position, then it can be easily checked that S is not $\Sigma_{1,0}$-designable. Hence, any $\Sigma_{1,0}$-designable structure is either empty, and trivially designable using a single letter, or saturated. In the latter case, by Theorem 4, we know that designable structures are exactly those that are saturated, and such that $D(S) \leq 2$. The claim follows. □

Corollary 6 (Result R3) *A structure S is $\Sigma_{1,1}$-designable if and only if*

$$D(S) \leq 2.$$

Proof First, suppose S is $\Sigma_{1,1}$-designable and let w be a design for S. Then $\mathsf{Paired}(w, S)$ is a design for the paired restriction $\mathsf{Paired}(S)$ of S. Since $\mathsf{Paired}(S)$ is saturated, $\mathsf{Paired}(w, S)$ is over alphabet $\Sigma_{1,0} \subset \Sigma_{1,1}$, and by Theorem 4, $D(\mathsf{Paired}(S)) \leq 2$. Hence, $D(S) = D(\mathsf{Paired}(S)) \leq 2$.

Conversely, suppose that $D(S) \leq 2$. Construct a design for S as follows. Since $\mathsf{Paired}(S)$ is saturated, by Theorem 4, there is a design $\bar{w}$ for $\mathsf{Paired}(S)$ over $\Sigma_{1,0} \subset \Sigma_{1,1}$. Construct w from $\bar{w}$ by inserting the base without a complementary base at every unpaired position of S. Let S' be an MFE structure for w. Obviously, all unpaired positions in S are also unpaired in S'. We must have $\mathsf{Paired}(S') = \mathsf{Paired}(S)$, otherwise we have an alternative structure for $\bar{w}$, a contradiction. Hence, $S' = S$, i.e., w is a design for S. □

Result **R4** follows immediately from Theorem 4 by taking $c = 2$.

Lemma 7 (Result R5) *Any structure that contains m_5 or $m_{3\circ}$ is not $\Sigma_{2,0}$-designable.*

Proof Assume that S is $\Sigma_{2,0}$-designable and let w be a design for S. Then $\mathsf{Paired}(w, S)$ is a design for $\mathsf{Paired}(S)$. Since $\mathsf{Paired}(S)$ is saturated, by Theorem 4, $D(S) = D(\mathsf{Paired}(S)) \leq 4$, hence, S cannot contain motif m_5.

Now, assume that S contain motif $m_{3\circ}$ appearing at node $[a, b]$ of T_S. Let $\{[l_i, r_i]\}_{i=1}^{3}$ be some paired children of $[a, b]$ and the node $[a, b]$ if $[a, b]$ is not the root, and $[u, u]$ an unpaired child of $[a, b]$. Let $L_i = l_i$ and $R_i = r_i$ if $[l_i, r_i]$ is a child

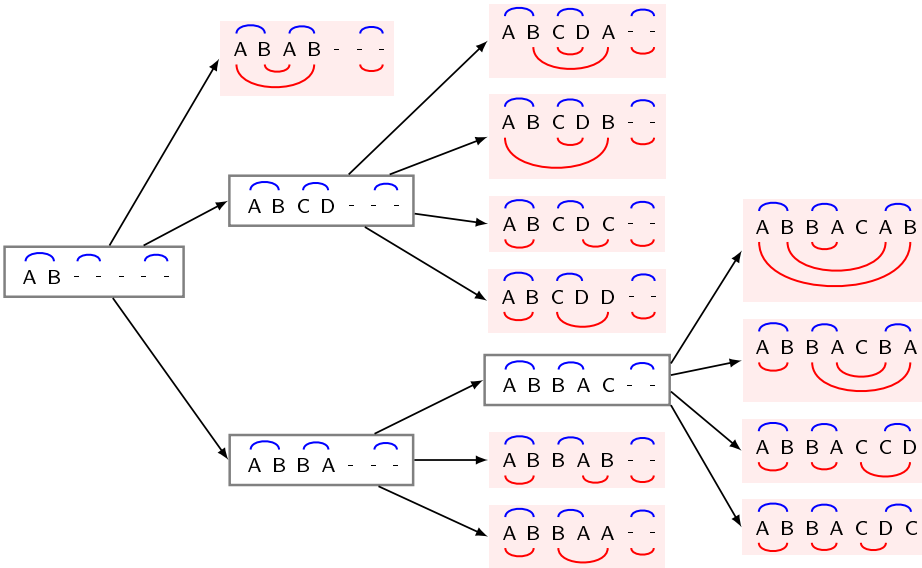

Fig. 9 Exhaustive search and systematic counter-examples for the design of $m_{3\circ}$. $\mathsf{A}-\mathsf{B}$ and $\mathsf{C}-\mathsf{D}$ respectively represent the first and second pairs of letters found in the design in prefix order, allowing the factorization of trivial symmetries.

of $[a, b]$, and $L_i = r_i$ and $R_i = l_i$ if it is $[a, b]$. If among bases $w_{L_1}, \ldots, w_{L_3}$ there is a pair of repeated letters, then we can construct an alternative MFE structure for w (see the first paragraph in the proof of Theorem 4). Assume that these three bases are different. Then for some $i = 1, 2, 3$, w_u equals either w_{l_i} or w_{r_i}, say it equals w_{l_i}. Then $S \setminus \{(l_i, r_i)\} \cup \{(u, r_i)\}$ is an MFE structure for S, a contradiction with the assumption that w is a design for S. □

Theorem 8 (Result R6) *If the tree representation of a structure S admits a separated coloring then S is $\Sigma_{2,0}$-designable.*

Proof Given a sequence w, we define the level $L(i)$ of position i as $L(i) = |w_{[1,i]}|_{\mathsf{G}} - |w_{[1,i]}|_{\mathsf{C}}$.

Claim 8.1 *Consider any structure S compatible with sequence w that contains some $\mathsf{A}-\mathsf{U}$ base pair between positions at different levels, then there exists a position G or C that is left unpaired in S.*

Proof Assume that the $\mathsf{A}-\mathsf{U}$ base pair occurs at position (a, b), and note that the bases of the substring $w_{[a+1,b-1]}$ can only base pair among themselves without introducing crossings. We will show that G's and C's are not balanced in this substring. Since $w_b \in \{\mathsf{A}, \mathsf{U}\}$, $L(b) = L(b-1)$. Hence, by the definition of L, we have that

$$|w_{[a+1,b-1]}|_{\mathsf{G}} - |w_{[a+1,b-1]}|_{\mathsf{C}} = L(b-1) - L(a) = L(b) - L(a) \neq 0\,.$$

Therefore, at least one G or C in the substring remains unpaired in this structure. □

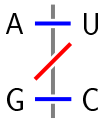

Fig. 10 The compatibility graphs of the Watson-Crick (blue edges) and the Nussinov-Jacobson (blue and red edges) energy models are bipartite.

Consider a separated coloring of the tree representation of S. We will use this coloring to construct a design w for S, by specifying a nucleotide at each position of w. First, for each unpaired position i, set $w_i = \mathsf{A}$. Second, apply a modified version of the algorithm described in the proof of the Theorem 4 to set the bases of paired positions in which black nodes are assigned to base pair $\mathsf{G}-\mathsf{C}$, white nodes to $\mathsf{C}-\mathsf{G}$ and grey nodes to $\mathsf{A}-\mathsf{U}$ or $\mathsf{U}-\mathsf{A}$. The algorithm ignores unpaired nodes in the tree representation of S. Since the coloring is proper such assignment is always possible at every step of the algorithm. We claim that for any node $[i, j]$ (paired or unpaired), the level of position i is the same as the level of the node $[i, j]$. To verify this, observe that the substring of w corresponding to any subtree has the same number of G's and C's. Hence, for any node $[i, j]$, the level of position i depends only on nodes on the path from this node to the root. It is easy to check that the level of i is equal to the level of the node. Note that if $[i, j]$ is a grey node then the level of position j is the same as the level of i, i.e., the same as the level of $[i, j]$.

We will show that the constructed w is a design for S. Since all C's and U's of w are paired in S, S is an MFE structure for w. We need to show that it is the only MFE structure for w. Consider an MFE structure S' for w different from S. Since w has the same number of G's and C's, S' must pair all G's, C's and U's of w. We will show that all unpaired positions in S are also unpaired in S'. Assume to the contrary that position i is unpaired in S, but it is paired to j in S'. We must have $w_i = \mathsf{A}$ and $w_j = \mathsf{U}$. Since the coloring is separated, the unpaired node $[i, i]$ has a different level than the grey node containing j, and hence, the level of i is different from the level of j. It follows by Claim 8.1 that some G or C is unpaired in S', a contradiction. Consider the paired restrictions of S, S' and w. Both $\mathsf{Paired}(S)$ and $\mathsf{Paired}(S')$ are saturated and compatible with $\mathsf{Paired}(w, S)$ and they are different since S and S' are different and agree on the unpaired positions. Furthermore, $\mathsf{Paired}(w, S)$ can be produced by the algorithm described in the proof of Theorem 4 for the input structure $\mathsf{Paired}(S)$, and hence, by Theorem 4, $\mathsf{Paired}(w, S)$ is a design for $\mathsf{Paired}(S)$, which contradicts the existence of $\mathsf{Paired}(S')$. Hence, w is a design for S. □

Next, we show the closure of the set of designable structures under the k-stutter operation. To that purpose, we introduce the compatibility graph of an energy model $\mathcal{M}$, whose vertices are the four nucleotides $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$, and whose edges correspond to valid base-pairs in $\mathcal{M}$, i.e. having finitely-valued contributions.

Definition 4 (Bipartite energy model) *An energy model $\mathcal{M}$ is* bipartite *if and only if its compatibility graph is bipartite.*

The Watson-Crick energy model $\mathcal{W}$ and the Nussinov-Jacobson energy model $\mathcal{N}$ are bipartite, as can be seen in Figure 10.

Theorem 9 (Result R7) *For any bipartite energy model $\mathcal{M}$ and any energy distance Δ, if w is a Δ-design for a structure S, then for any integer $k \geq 1$, $w^{[k]}$ is also a Δ-design for $S^{[k]}$.*

Proof Consider a designable structure S and let $w = w_1 \cdots w_n$ be a design for S. We will show that $w^{[k]}$ is a design for $S^{[k]}$. Let us use the ith block in $S^{[k]}$ (resp. $w^{[k]}$) as a shorthand for the subset $[1 + i \cdot k, 1 + (i+1) \cdot k)$ of its positions. Note that the positions involved in the ith block of $S^{[k]}$ correspond to the ith position in S (resp. w).

Consider a valid structure $S' \neq S^{[k]}$ for $w^{[k]}$. Define an interaction multigraph $I(S') = (V_{I(S')}, E_{I(S')})$ of S' as follows: the vertex set $V_{I(S')}$ is the set of positions $\{1, \ldots, n\}$ in w, and there are as many edges between i and j in $I(S')$ as there are base-pairs in S' between the ith and j^{th} blocks. Clearly, $I(S')$ is a multigraph of maximal degree k. Moreover any edge between the ith and j^{th} block in $I(S')$ corresponds to a valid base-pair (i, j) for w. Therefore the sequence of nucleotides read along any path in $I(S')$ must constitute a valid path in the compatibility graph. Since the energy model is bipartite, then any cyclic path cannot have odd length, and so $I(S')$ is also bipartite.

Since $I(S')$ is a bipartite multigraph of maximal degree k, then by König's theorem [12] it is k edge-colorable (see [23, page 52] for an English version of the proof). In other words, we can color the base-pairs of S', using less than k colors, such that each block in S' is involved in at most one base-pair of each color. Therefore we can partition the base-pairs of S' into k structures $S'_1, S'_2, \cdots, S'_k$ that are compatible with w. Note that the base-pairs of S'_i are pairwise non-crossing since S' itself is non-crossing.

The sequence w is a design for S, thus one has

$$E_{\mathcal{W}}(w, S'_i) \geq E_{\mathcal{W}}(w, S), \text{ for every } 1 \leq i < k.$$

Moreover, one has $S' \neq S^{[k]}$ so there exists a structure S'_j such that $S'_j \neq S$, and therefore $E_{\mathcal{W}}(w, S'_j) \geq E_{\mathcal{W}}(w, S) + \Delta$. It follows that

$$E_{\mathcal{W}}(w^{[k]}, S') = \sum_{i=0}^{k-1} E_{\mathcal{W}}(w, S'_i) \geq k \cdot E_{\mathcal{W}}(w, S) + \Delta = E_{\mathcal{W}}(w^{[k]}, S^{[k]}) + \Delta.$$

We conclude that $S^{[k]}$ is the sole MFE structure for $w^{[k]}$, and has energy at least Δ less than its foremost competitor, so $w^{[k]}$ is a Δ-design for $S^{[k]}$.

□

Result **R7** immediately follows from Theorem 9, by reminding the bipartite nature of the Watson-Crick energy model.

Theorem 10 (Result R8) *Each structure S without m_5 and $m_{3\circ}$ can be transformed into a $\Sigma_{2,0}$-designable structure S' by inflating a subset of its base pairs (at most one per band). Furthermore, this transformation can be done in $\Theta(n)$ time.*

Proof We start with the greedy coloring of T_S. Since S does not contain m_5 and $m_{3\circ}$, it is a proper coloring and there is no node having both a grey child and an unpaired child. We will insert base pairs within S so that the grey nodes and any unpaired node end up at levels of different parities. If the root has a grey child,

assign even parity to the grey nodes, otherwise (if the root has an unpaired child, or no grey and no unpaired children), assign even parity to the unpaired nodes.

Now we proceed from the children of the root towards leaves adjusting parity level for grey and unpaired nodes to keep one type even and the other one odd. We repeatedly apply the following simple operation on T_S: If the node N does not match its intended parity level. Denote N_P the parent of N (N_P is not the root as all children of the root already have the correct parity level) and N_{PP} the parent of N_P. Insert a new paired node N_N between N_{PP} and N_P, assign it with the color of N_P, and apply the greedy algorithm on N_N. Observe that N_P always takes either black or white color changing the parity level of all its descendants (including N). Note that the children of N_P may get recolored, we can even get one more grey child but after this operation the parity levels of all children of N are correct and we do not change parity levels outside the subtree rooted at N. After fixing all nodes, we get a separated proper coloring (which is actually the greedy coloring) of $T_{S'}$. Hence, by Theorem 8, S' is designable. Figure 7 illustrates this process. □

4.2 Nussinov-Jacobson energy model $\mathcal{N}$

We will show that, for a value of $\Delta \in (0, \min(|\alpha|, |\beta|)]$ our results for the Watson-Crick model transpose to the more general Nussinov-Jacobson model.

First, we will establish that **R1**, **R2**, **R3** and **R7** hold in the $\mathcal{N}$ model.

R1 concerns an alphabet that does not allow base pairs to occur, so their weighting is unconsequential. **R2** is equally trivial since the uniform weighting of every occurrence of a base pair type does not affect the relative order of structures.

R3 requires a clarification, as the introduction of two partners A and G for U somehow assign an unambiguous semantics to $\Sigma_{1,1}$ notation, leading to a disjunctive discussion. Let us assume that $\Sigma_{1,1}$ represents $\{\mathsf{A}, \mathsf{C}, \mathsf{G}\}$ (resp. $\{\mathsf{A}, \mathsf{C}, \mathsf{U}\}$), then the argument used for **R2** holds, since A (resp. C) cannot form base pairs. **R7** is a direct corollary of Theorem 9 which is applicable for any bipartite energy model. Since the Nussinov-Jacobson model is bipartite, as shown in Figure 10, then **R7** also holds in the $\mathcal{N}$ model.

Definition 5 *Let $X \subseteq \{\mathsf{C}, \mathsf{G}, \mathsf{A}, \mathsf{U}\}$. A design w for a structure S is X-*unpaired *if and only if the bases of w found at unpaired positions in S belong to X. If $X = \{b\}$ is a singleton, the notation is shortened to b-*unpaired*.*

Let $n_{\mathsf{GC}}(w, S)$ (resp. $n_{\mathsf{AU}}(w, S)$, $n_{\mathsf{GU}}(w, S)$) be the number of $\mathsf{G}-\mathsf{C}$ (resp. $\mathsf{A}-\mathsf{U}$, $\mathsf{G}-\mathsf{U}$) base pairs of S on w. Note that

$$E_{\mathcal{W}}(w, S) = -n_{\mathsf{GC}}(w, S) - n_{\mathsf{AU}}(w, S)$$

and

$$E_{\mathcal{N}}(w, S) = \alpha.n_{\mathsf{GC}}(w, S) + \beta.n_{\mathsf{AU}}(w, S) + \gamma.n_{\mathsf{GU}}(w, S).$$

Proposition 11 *Let $\Delta_{\mathcal{W}} = 1$ and $0 < \Delta_{\mathcal{N}} \leq \min(|\alpha|, |\beta|)$, if a structure is* A-*unpaired and $(\mathcal{W}, \Sigma_{2,0}, \Delta_{\mathcal{W}})$-designable then it is also $(\mathcal{N}, \Sigma_{2,0}, \Delta_{\mathcal{N}})$-designable.*

Proof Let w be an A-unpaired $(\mathcal{W}, \Sigma_{2,0}, \Delta_{\mathcal{W}})$-design for S. Since in $\mathcal{W}$ there are no $\mathsf{G}-\mathsf{U}$ base pairs, we have

$$n_{\mathsf{GC}}(w,S) = |w|_{\mathsf{G}} = |w|_{\mathsf{C}}\,, \qquad n_{\mathsf{AU}}(w,S) = |w|_{\mathsf{U}}\,,$$

and for any other structure $S' \in \mathcal{S}_{|w|}$,

$$n_{\mathsf{GC}}(w,S') \leq n_{\mathsf{GC}}(w,S) \qquad n_{\mathsf{AU}}(w,S') \leq n_{\mathsf{AU}}(w,S)\,,$$

with at least one of the inequalities being strict.

We will show that w is also a $(\mathcal{N}, \Sigma_{2,0}, \Delta_{\mathcal{N}})$-design for S. Consider any alternative structure $S' \in \mathcal{S}_{|w|}$. If $n_{\mathsf{GU}}(w,S') = 0$, then

$$\begin{aligned} E_{\mathcal{N}}(w,S') &= \alpha\, n_{\mathsf{GC}}(w,S') + \beta\, n_{\mathsf{AU}}(w,S') \\ &\geq \alpha n_{\mathsf{GC}}(w,S) + \beta\, n_{\mathsf{AU}}(w,S) + \min\{|\alpha|,|\beta|\} \geq E_{\mathcal{N}}(w,S) + \Delta_{\mathcal{N}}\,, \end{aligned}$$

as required. Otherwise ($n_{\mathsf{GU}}(w,S') > 0$), first observe that

$$n_{\mathsf{GC}}(w,S') + n_{\mathsf{GU}}(w,S') \leq |w|_{\mathsf{G}} \quad \text{and} \quad n_{\mathsf{AU}}(w,S') + n_{\mathsf{GU}}(w,S') \leq |w|_{\mathsf{U}}\,. \tag{2}$$

Now, we have

$$\begin{aligned} E_{\mathcal{N}}(w,S') &= \alpha\, n_{\mathsf{GC}}(w,S') + \beta\, n_{\mathsf{AU}}(w,S') + \gamma\, n_{\mathsf{GU}}(w,S') \\ &\underset{(2)}{\geq} \alpha\, |w|_{\mathsf{G}} - \alpha\, n_{\mathsf{GU}}(w,S') + \beta\, |w|_{\mathsf{U}} - \beta\, n_{\mathsf{GU}}(w,S') + \gamma\, n_{\mathsf{GU}}(w,S') \\ &= \alpha\, |w|_{\mathsf{G}} + \beta\, |w|_{\mathsf{U}} + (\gamma - \alpha - \beta) n_{\mathsf{GU}}(w,S') \\ &\geq E_{\mathcal{N}}(w,S) + \gamma - \alpha - \beta. \end{aligned}$$

Moreover, since $\alpha, \beta, \gamma < 0$ and $\alpha, \beta < \gamma$, then

$$\gamma - \alpha - \beta \geq \max(|\alpha|,|\beta|) \geq \min(|\alpha|,|\beta|) \geq \Delta_{\mathcal{N}}.$$

We conclude that

$$E_{\mathcal{N}}(w,S') \geq E_{\mathcal{N}}(w,S) + \Delta_{\mathcal{N}}$$

as required. □

Corollary 12 *Results* **R4**, **R5**, **R6** *and* **R8** *hold in any Nussinov-Jacobson energy model* $\mathcal{N}$ *with* $0 < \Delta_{\mathcal{N}} \leq \min(|\alpha|,|\beta|)$.

Proof The validity of Results **R6** and **R8** in $\mathcal{N}$ follows directly from a close inspection of the constructive proofs in the Watson-Crick energy model, both establishing the existence of A-unpaired designs. Proposition 11 therefore applies, and extends the validity of those designs to any suitable $\Delta_{\mathcal{N}}$.

R5 follows from the fact that, in the proof of Theorem 7, our counterexamples 'locally' trade one base pair in S for another in the alternative structure, and that the two base pairs are of the same type. Therefore, both structures have the same energy in the Nussinov-Jacobson energy model.

From this, we conclude on the validity of **R4**. Indeed, any failure to the degree condition $D(S) \leq 4$ implies the existence of m_5 in S, *i.e.* such structures are undesignable and the degree condition is therefore necessary.

□

5 Conclusion, discussion and perspectives

In this work, we introduced the *Combinatorial RNA Design problem*, a *minimal* instance of the RNA design problem which aims at finding a sequence that admits the target structure as its unique base pair maximizing structure using $\mathsf{A}-\mathsf{U}$ and $\mathsf{G}-\mathsf{C}$ base pairs. First, we provided complete characterizations for the structures that can be designed using restricted alphabets. Then we considered the RNA design under a four-letters alphabet, and provide a complete characterization of designable saturated structures, i.e., free of unpaired positions. Turning to those target structures that contain unpaired positions, we provided partial characterizations for classes of designable/undesignable structures, and showed that the set of designable structures is closed under the stutter operation. Finally, we introduced structure-approximating version of the problem and, assuming that the input structure avoids two motifs, provided a structure approximating algorithm of ratio 2 for general structures. We showed that our results also hold in the more realistic Nussinov-Jacobson energy model, which allows $\mathsf{G}-\mathsf{U}$ base pairs to occur, and associates arbitrary negative free-energy to each base pair type ($\mathsf{G}-\mathsf{U}$ being the weakest).

An important question that is left open by this work is the computational complexity of the RNA design problem. Schnall-Levin *et al.* [19] established the NP-hardness of a more general problem, called the inverse Viterbi algorithm, which takes as input a stochastic grammar (representing the energy model) and a targeted parse tree (representing the structure), and outputs a sequence (design) whose most probable parsing should match the target. However this result does not settle the complexity of the RNA design, essentially because the proposed reduction relies critically on an encoding of 3-SAT instances within the input grammar. While the hypothetical *perfect* grammar/energy model for RNA folding probably differs from the currently accepted Turner model, it should ultimately reflect the laws of physics and should certainly not depend on the instance. As the reduction [19] requires a different grammar (i.e., energy model) for each instance, it does not seem easily adaptable into a proof that holds for a fixed energy model. Consequently, despite two decades of work on the subject, the computational tractability of RNA design is still open, either in its general instance and in our purely combinatorial version.

In our opinion, this exceptional resistance of the RNA design problem to any attempt so far at characterizing its computational complexity can be attributed to two main reasons:

- The inverse nature of the problem: While polynomial, the direct computation MFE folding for an RNA requires dynamic programming, and runs in $\Theta(n^3)$ time (up to polylogarithmic factors, see [26] for a complete state-of-the-art). Unfortunately, the *optimal-substructure* property of the direct problem does not transpose to the inverse problem.
  Therefore, solving the RNA design problem somehow requires inverting a non-trivial – yet polynomially computable – function. It is tempting here to draw a parallel with some areas of cryptography, where multiple protocols are based on a – sometimes difficult to establish – disymmetry between the complexities of the direct and inverse computation.

– The intricacies of the objective function: Current state-of-the-art implementations of MFE folding prediction algorithms rely on a sophisticated energy model, the Turner model [22]. This model associates energy contributions to as much as 24 000 different types of structure/sequence motifs, and vastly increases the complexity the characterization of the space, and energies, of competing structures.
  On the other hand, oversimplified statements for the problem, as would result from a relaxation of the uniqueness condition, can be trivially solved in linear time. Such problems are not only largely unrealistic from a biological perspective, but they also probably do not retain the potential difficulty of the general problem.

Besides complexity issues, natural extensions of this work may include the consideration of more sophisticated energy models such as those based on stacking pairs or, ultimately, to the full Turner energy model [22]). One could also consider incorporating additional constraints, expressed as the presence/avoidance of motifs [27], GC-content [17]... or the design under other objectives, such as the Boltzmann probability [25]. In the Nussinov-Jacobson model, our result could be completed by the consideration of more liberal values for Δ, although it should be noted that considering larger such values would only gradually deplete the sets of designable structures until it becomes empty when the conditions of Theorem 1 are met. More precise bounds for the ratio of the structure-approximating could also be established. Finally, the structure-approximation problem could be revisited in an optimization setting, in which one would attempt to minimize the number of modifications made to the structure, so that a given structure becomes designable (or, more modestly, belongs to an identified class of designable structures). We plan to address some of these questions in future works.

References

1. Aguirre-Hernández, R., Hoos, H.H., Condon, A.: Computational RNA secondary structure design: empirical complexity and improved methods. BMC Bioinformatics **8**, 34 (2007). DOI 10.1186/1471-2105-8-34
2. Avihoo, A., Churkin, A., Barash, D.: RNAexinv: An extended inverse RNA folding from shape and physical attributes to sequences. BMC Bioinformatics **12**(1), 319 (2011). DOI 10.1186/1471-2105-12-319
3. Bau, A., Waldmann, J., Will, S.: RNA design by program inversion via SAT solving. In: A.D. Palu, A. Dovier (eds.) Proceedings of Workshop on Constraint Based Methods for Bioinformatics (WCB 2013), pp. 85–94 (2013)
4. Busch, A., Backofen, R.: INFO-RNA—a fast approach to inverse RNA folding. Bioinformatics **22**(15), 1823–31 (2006). DOI 10.1093/bioinformatics/btl194
5. Dai, D.C., Tsang, H.H., Wiese, K.C.: RNADesign: Local search for RNA secondary structure design. In: IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (2009)
6. Esmaili-Taheri, A., Ganjtabesh, M., Mohammad-Noori, M.: Evolutionary solution for the RNA design problem. Bioinformatics **30**(9), 1250–1258 (2014). DOI 10.1093/bioinformatics/btu001. URL `http://dx.doi.org/10.1093/bioinformatics/btu001`
7. Frid, Y., Gusfield, D.: A simple, practical and complete $o(n^3/\log n)$-time algorithm for RNA folding using the Four-Russians speedup. Algorithms Mol Biol **5**, 13 (2010). DOI 10.1186/1748-7188-5-13. URL `http://dx.doi.org/10.1186/1748-7188-5-13`
8. Garcia-Martin, J.A., Clote, P., Dotu, I.: RNAiFOLD: a constraint programming algorithm for RNA inverse folding and molecular design. J Bioinform Comput Biol **11**(2), 1350,001 (2013). DOI 10.1142/S0219720013500017. URL `http://dx.doi.org/10.1142/S0219720013500017`

9. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R.: RFAM: an RNA family database. Nucleic Acids Res **31**(1), 439–441 (2003)
10. Hofacker, I.L., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M., Schuster, P.: Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie / Chemical Monthly **125**(2), 167–188 (1994). DOI 10.1007/BF00818163
11. Höner Zu Siederdissen, C., Hammer, S., Abfalter, I., Hofacker, I.L., Flamm, C., Stadler, P.F.: Computational design of RNAs with complex energy landscapes. Biopolymers **99**(12), 1124–1136 (2013). DOI 10.1002/bip.22337. URL `http://dx.doi.org/10.1002/bip.22337`
12. Kőnig, D.: Gráfok és alkalmazásuk a determinánsok és a halmazok elméletére. Matematikai és Természettudományi Értesítő **34**, 104–119 (1916)
13. Levin, A., Lis, M., Ponty, Y., O'Donnell, C.W., Devadas, S., Berger, B., Waldispühl, J.: A global sampling approach to designing and reengineering RNA secondary structures. Nuc Acids Res **40**(20), 10,041–52 (2012). DOI 10.1093/nar/gks768
14. Lyngsø, R.B., Anderson, J.W., Sizikova, E., Badugu, A., Hyland, T., Hein, J.: FRNAkenstein: multiple target inverse RNA folding. BMC Bioinformatics **13**, 260 (2012). DOI 10.1186/1471-2105-13-260
15. Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H.: Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol **288**(5), 911–940 (1999)
16. Nussinov, R., Jacobson, A.: Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci U S A **77**, 6903–13 (1980)
17. Reinharz, V., Ponty, Y., Waldispühl, J.: A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotide distribution. Bioinformatics **29**(13), i308–i315 (2013). DOI 10.1093/bioinformatics/btt217. URL `http://dx.doi.org/10.1093/bioinformatics/btt217`
18. Rodrigo, G., Landrain, T.E., Majer, E., Daròs, J.A., Jaramillo, A.: Full design automation of multi-state RNA devices to program gene expression using energy-based optimization. PLoS Comput Biol **9**(8), e1003,172 (2013). DOI 10.1371/journal.pcbi.1003172. URL `http://dx.doi.org/10.1371%2Fjournal.pcbi.1003172`
19. Schnall-Levin, M., Chindelevitch, L., Berger, B.: Inverting the Viterbi algorithm: an abstract framework for structure design. In: Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008, pp. 904–911 (2008). DOI 10.1145/1390156.1390270. URL `http://doi.acm.org/10.1145/1390156.1390270`
20. Takahashi, M.K., Lucks, J.B.: A modular strategy for engineering orthogonal chimeric RNA transcription regulators. Nucleic Acids Res **41**(15), 7577–7588 (2013). DOI 10.1093/nar/gkt452. URL `http://nar.oxfordjournals.org/content/41/15/7577.abstract`
21. Taneda, A.: MODENA: a multi-objective RNA inverse folding. Adv Appl Bioinform Chem **4**, 1–12 (2011)
22. Turner, D.H., Mathews, D.H.: NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res **38**(Database issue), D280–D282 (2010). DOI 10.1093/nar/gkp892. URL `http://dx.doi.org/10.1093/nar/gkp892`
23. Wilson, R.A.: Graphs, Colourings and the Four-colour Theorem. Oxford University Press (2002)
24. Wu, S.Y., Lopez-Berestein, G., Calin, G.A., Sood, A.K.: RNAi therapies: Drugging the undruggable. Science Translational Medicine **6**(240), 240ps7 (2014). DOI 10.1126/scitranslmed.3008362. URL `http://stm.sciencemag.org/content/6/240/240ps7.abstract`
25. Zadeh, J.N., Wolfe, B.R., Pierce, N.A.: Nucleic acid sequence design via efficient ensemble defect optimization. J Comput Chem **32**(3), 439–52 (2011). DOI 10.1002/jcc.21633
26. Zakov, S., Tsur, D., Ziv-Ukelson, M.: Reducing the worst case running times of a family of RNA and cfg problems, using valiant's approach. Algorithms Mol Biol **6**(1), 20 (2011). DOI 10.1186/1748-7188-6-20. URL `http://dx.doi.org/10.1186/1748-7188-6-20`
27. Zhou, Y., Ponty, Y., Vialette, S., Waldispuhl, J., Zhang, Y., Denise, A.: Flexible RNA design under structure and sequence constraints using formal languages. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics (ACM-BCB), BCB'13, pp. 229–238. ACM (2013). DOI 10.1145/2506583.2506623. URL `http://doi.acm.org/10.1145/2506583.2506623`
28. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res **9**, 133–148 (1981)