Impact Of The Energy Model On The Complexity Of RNA Folding With Pseudoknots

Saad Sheikh, Rolf Backofen, Yann Ponty

Impact Of The Energy Model On The Complexity Of RNA Folding With Pseudoknots

Saad Sheikh[a,d], Rolf Backofen[b], and Yann Ponty[c,d⋆]

[a] University of Florida, Gainesville, USA
[b] Albert Ludwigs University, Freiburg, Germany
[c] Ecole Polytechnique, CNRS UMR 7161, Palaiseau, France
[d] AMIB Team-Project, INRIA, Saclay, France

Abstract. Predicting the folding of an RNA sequence, while allowing general pseudoknots (PK), consists in finding a minimal free-energy matching of its n positions. Assuming independently contributing base-pairs, the problem can be solved in $\Theta(n^3)$-time using a variant of the maximal weighted matching. By contrast, the problem was previously proven NP-Hard in the more realistic nearest-neighbor energy model.
In this work, we consider an intermediate model, called the stacking-pairs energy model. We extend a result by Lyngsø, showing that RNA folding with PK is NP-Hard within a large class of parametrization for the model. We also show the approximability of the problem, by giving a practical $\Theta(n^3)$ algorithm that achieves at least a 5-approximation for any parametrization of the stacking model. This contrasts nicely with the nearest-neighbor version of the problem, which we prove cannot be approximated within any positive ratio, unless $P = NP$.

Keywords: RNA folding; General pseudoknots; Hardness; Inapproximability

1 Introduction

Ribonucleic Acid (RNA) is one of the key pieces to the puzzle of molecular biology. It plays a very large number of roles, not only by coding for proteins, but also through catalytic and regulatory functions. To play such roles, RNA folds into an intricate structure which is stabilized by the pairing, mediated by hydrogen bonds, of some of its positions. The conformations that arise from this folding process are instrumental to the function of an RNA. Consequently, the process of RNA folding has been extensively studied by molecular biology and biochemistry, and its *in silico* prediction has given rise to a wealth of computational approaches. Early work on the subject have focused on the secondary structure, a restriction of all admissible base-pairs that forbids crossing-interactions. Under

⋆ To whom correspondance should be addressed.

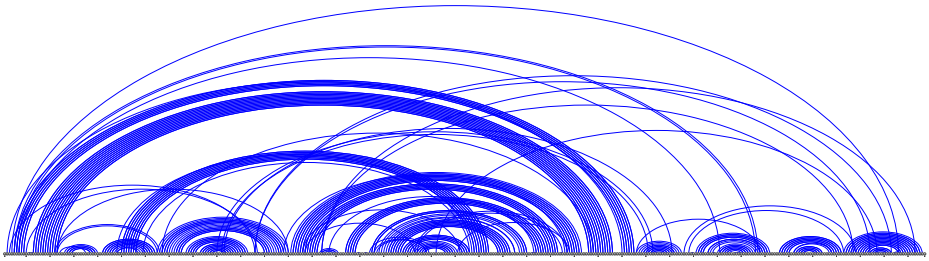


Fig. 1. Canonical base-pairs of the Oceanobacillus iheyensis group II intron, derived from a 3D model (PDB id: 3IGI) using `RNAView` [19].

the assumption of reasonable, additive, energy models, such a restriction implies an optimal-substructure property for computing the most stable conformation, i.e. the one having minimal free-energy. Polynomial-time algorithms, based on dynamic-programming (DP), have consequently been proposed for predicting the minimal free-energy *secondary structure* conformation of an RNA from its sequence. Unfortunately, the assumption of a non-crossing conformation may impede the quality of the prediction, since functionally essential crossing-interactions are found in many families of non-protein-coding RNAs (ncRNAs), as illustrated by Figure 1. For instance, **pseudoknots** (PK), can be found within the RFAM consensus [8] of at least 70 functional families of ncRNAs and are conserved throughout the evolution.

Taking general pseudoknots into account is known to turns RNA minimal-free energy prediction into a rather challenging problem. A pioneering work by Cary, Tabaska *et al* [6, 17] considered a simple additive model, associating independent energy contributions to each putative base-pair, and used a $\mathcal{O}(n^3)$ maximal-weighted matching algorithm to extract a minimal free-energy folding. Unfortunately, this energy model is regarded as unrealistic because of its incapacity to capture the interaction of consecutive – stacking – base-pairs, which constitute a primary stabilizing force in RNA folding. Such energy contributions are captured by the **nearest neighbor energy model**, in which the contribution of each base-pair depends on the base-pairing status and partners of its consecutive positions. The hardness of RNA folding assuming a nearest-neighbor energy model was independently established by Lyngsø and Pedersen [12], and Akutsu [1]. Subsequent efforts have therefore focused on providing either parameterized complexity algorithms [11, 20], heuristics [4, 18] or exact DP schemes for tractable subsets of pseudoknots [16, 15].

It is frequent that the complexity of solving any problem in computational biology optimally is tied to the chosen model (e.g. [3]). However, despite a significant amount of research focusing on predicting pseudoknots, the impact of a specific instantiation of the energy model on the computational complexity of RNA folding with pseudoknots has only been partly unexplored. In this extended abstract, we further study the influence of the energy model on the complexity and approximability of RNA folding with unconstrained pseudoknots. In addition to the base-pair and nearest-neighbor models, we consider the **stacking base-pairs energy model**, which captures the dependency between consecutive base-pairs. The computational complexity of RNA folding under this energy model was first studied by Ieong *et al* [9]. They were able to show the NP-hardness of maximizing the number of stacking-pair among the set of planar secondary structures, a restriction of general pseudoknots. This restriction was lifted by Lyngsø [13], who established the hardness of maximizing the number of base-pairs, allowing general types of pseudoknots. Approximation algorithms we also sought, leading to the current best 8/3-approximation $O(n^{10})$-time algorithm reported by Jiang [10]. However all of these works consider a purely combinatorial model, maximizing the number of base-stacking, while the contribution of stacking pairs to the free-energy may vary significantly. It is therefore a natural question to ask to what extent the hardness of folding with pseudoknots is affected by perturbations of the energy model. More generally, understanding what makes the problem hard, and just how hard, could be instrumental to the development of future algorithms, achieving better tradeoffs between sensibility and complexity.

This extended abstract is organized as follows. First we formally define in Section 2 our main problem, along with the different energy models considered. We discuss the NP-hardness of the stacking base-pairs in Section 3, and present an approximation in Section 4. In Section 5, we show the inapproximability of RNA folding with pseudoknots under the nearest-neighbor energy model. Finally Section 6 summarizes the contributions and describes futures lines of research.

2 Problem statement and free-energy models

Let $\omega \in \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}^*$ be an RNA sequence, and m be a partial matching of the positions in ω, i.e. a set of non overlapping pairs of positions in ω. An **energy model** is a real-valued function E_w that associates a **free-energy** to ω by summing over the contributions of local motifs in the

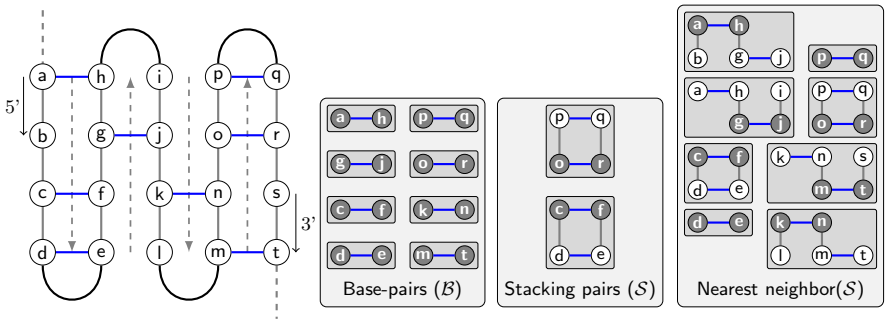


Fig. 2. Typical picture of a standard pseudoknot/matching (Left) and individual contributions of local motifs to the three energy models considered (Right). Dark nodes indicate the supporting base pair for each motif (i.e. (i, j) pairs in our definition below).

matching. The precise definition of local motifs will depend on the exact energy model.

A low free-energy indicates a stable folding. Furthermore, any free-energy contribution is usually determined up to an additive constant. Therefore one can assume that the contribution to the free-energy of any local motif is negative, with the exception of extremely unfrequent motifs which will be forbidden and assigned $+\infty$ contributions. Let us then rephrase the problem of as an optimization (minimization) problem.

RNA-PK-Fold(E) **problem**
Input: An RNA sequence w.
Ouput: A partial matching m over w, i.e. a set of pairwise disjoint pairs of positions in $[1, |w|]$, which minimizes $E_w(m)$.

The three reference energy models are usually considered:

- **Base-pairs model $\mathcal{B}$ [14, 6, 17]:** Here, local motifs are simply individual base pairs, independently contributing to the free-energy:

$$\mathcal{B}_w(m) = \sum_{(i,j)\in m} \Delta_{\mathcal{B}}(w_i, w_j)$$

  where $\Delta_{\mathcal{B}} : \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}^2 \to \mathbb{R}^- \cup \{+\infty\}$.
- **Stacking pairs model $\mathcal{S}$ [13, 5]:** This model only considers *consecutively nested* pairs as motifs, and disregards isolated pairs:

$$\mathcal{S}_w(m) = \sum_{(i,j),(i+1,j-1)\in m} \Delta_{\mathcal{S}}(w_i, w_j, w_{i+1}, w_{j-1})$$

  where $\Delta_{\mathcal{S}} : \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}^4 \to \mathbb{R}^- \cup \{+\infty\}$.

– **Nearest-neighbors model $\mathcal{N}$ [12, 16]:** This motif definition is even more expressive, allowing different contributions for each base-pair, depending on its bases, the base-pairing status of its consecutive neighbors and own their own partners:

$$\mathcal{N}_w(m) = \sum_{\substack{(i,j)\in m \\ i<j}} \Delta_{\mathcal{N}}(w_i, w_j, w_{i+1}, w_{j-1}, w_{m_{i+1}}, w_{m_{j-1}})$$

where $\Delta_{\mathcal{N}}$ is any function $\{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{U}\}^4 \times \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{U},\varnothing\}^2 \to \mathbb{R}^- \cup \{+\infty\}$, m_i denotes the partner of a position i in m (or $\varnothing$ if i is unpaired, while $w_\varnothing \equiv \varnothing$ by convention).

These three models induce different decompositions into motifs for any given structure, as illustrated by Figure 2.

3 NP-hardness of RNA-PK-Fold($\mathcal{S}$) in any non-degenerate stacking energy model

Consider the set of **canonical base-pairs** (A,U), (G,C) and (G,U). A **combinatorial stacking model** $\mathcal{S}^*$ specializes the stacking pairs model by assigning a -1.0 kcal.mol^{-1} contribution to any canonical stacking pair, and $+\infty$ to others. It was showed by Lyngsø [13] that the RNA-PK-Fold($\mathcal{S}^*$) problem is NP-complete, using a reduction from the BIN-PACKING problem.

Here we complement this result by showing its robustness, i.e. the NP-hardness of the problem under a wide class of stacking energy model.

Theorem 1 *Let $\mathcal{S}$ be a stacking energy model that allows (G,C) pairs, and forbids (A,C) and (A,G) pairs. Then RNA-PK-Fold($\mathcal{S}$) is NP-hard.*

Proof. In order to prove the hardness of RNA-PK-Fold($\mathcal{S}$), let us remind the statement of the 3-PARTITION problem:

3-PARTITION problem
Input: A multiset of integral values $X = \{x_i\}_{i=1}^n$ of cardinality $n = 3m$, such that $\sum_{i=1}^n x_i = m \cdot K$ for some $K \in \mathbb{N}$, and $\lfloor K/4 \rfloor < x < \lfloor K/2 \rfloor, \forall x \in X$.
Ouput: `True` if there exists a partition of X into m triplets $((x_{a_j}, x_{b_j}, x_{c_j}))_{j=1}^m$ such that

$$x_{a_j} + x_{b_j} + x_{c_j} = K, \forall j \in [1, m],$$

and `False` otherwise.

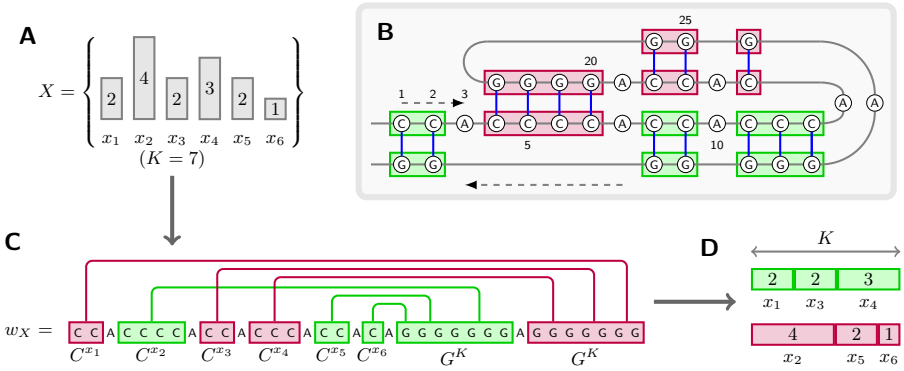


Fig. 3. Illustrating the reduction: Finding a 3-partition of a set of numbers X (**A**) is equivalent to finding a matching for w_X that produces a maximal number of stacking pairs (**C**), from which one easily deduces a set of equally summing triplets (**D**). Such a matching can be represented as a pseudoknotted secondary structure (**B**).

From Garey and Johnson [7], it is known that 3-PARTITION is **strongly NP-complete**, i.e. not only is the problem NP-hard, but it remains hard even when the elements of X are upper-bounded by some polynomial function $P(n)$.

Lemma 2 *Let X be a 3-PARTITION instance whose values are bounded by $P(n)$, and w_X be an RNA sequence such that:*

$$w_X = \mathsf{C}^{x_1}\mathsf{A}\mathsf{C}^{x_2}\mathsf{A}\mathsf{C}^{x_3}\mathsf{A}\cdots\mathsf{A}\mathsf{C}^{x_n}\underbrace{\mathsf{A}\mathsf{G}^K\mathsf{A}\mathsf{G}^K\mathsf{A}\cdots\mathsf{A}\mathsf{G}^K}_{m\ \text{times}}.$$

There exists a 3-partition of X into equally summing triplets if and only if there exists a solution to RNA-PK-Fold($\mathcal{S}$) over w_X having energy $k = \delta \cdot (K-3) \cdot m$ kcal.mol^{-1} with $\delta := \Delta_{\mathcal{S}}(\mathsf{C},\mathsf{G},\mathsf{C},\mathsf{G})$.

Let us summarize the argument:

- A matching has minimal free-energy iff any block C^x is entirely paired to some contiguous substring of a single G^K block.
- A matching has minimal free-energy iff every position in every G^K block is connected.
- Any optimal matching thus gives us a mapping between the C^x and G^K blocks, which can be transformed in polynomial-time into a solution to the **3-Partition** problem.

Proof. X **is 3-partitionable** $\Rightarrow \exists m^*$ **such that** $\mathcal{S}_{w_X}(m^*) = \delta\cdot(K-3)\cdot m$**:**

If X is 3-partitionable, then there exists m disjoint triplets $((x_{a_j}, x_{b_j}, x_{c_j}))_{j=1}^{m}$ whose sum is identically K. If follows that the C^x blocks in w_X can be partitioned into triplets $(\mathsf{C}^{x_{a_j}}, \mathsf{C}^{x_{b_j}}, \mathsf{C}^{x_{c_j}})$ that can form a three-tier *perfect* helix with the j-th block $\mathsf{G}^K \equiv \mathsf{G}^{x_{a_j}}.\mathsf{G}^{x_{b_j}}.\mathsf{G}^{x_{c_j}}$. By creating x_{a_j} (resp. x_{b_j} and x_{c_j}) nested base-pairs between the block $\mathsf{C}^{x_{a_j}}$ ($\mathsf{C}^{x_{b_j}}$ and $\mathsf{C}^{x_{c_j}}$) and the beginning (resp. middle and ending) of the j-th block G^K, one obtains exactly $(x_{a_j})-1+(x_{b_j})-1+(x_{c_j}-1) = K-3$ stacking pairs. Repeating the operation for each triplet j yields a valid conformation with $(K-3)\cdot m$ stacking $(\mathsf{G},\mathsf{C})/(\mathsf{G},\mathsf{C})$ pairs, and the implication follows.

$\exists m^*$ such that $\mathcal{S}_{w_X}(m^*) = \delta\cdot(K-3)\cdot m \Rightarrow X$ is 3-partitionable:

Let us remark that the absence of U implies that the only admissible base-pairs are C/G or G/C, arising from interactions between C^x and G^K blocks respectively.

First, let us show that each G^K block contributes to at most $K-3$ stacking pairs, and that this upper-bound cannot be reached unless G^K creates K base-pairs with exactly 3 distinct C^x blocks. Indeed, it is easily seen that any G^K block, connected to b blocks $(\mathsf{C}^{x_{d_1}}, \ldots, \mathsf{C}^{x_{d_b}})$ by at least one base-pair, for a total number P of base-pairs, creates at most $P-b$ stacking pairs. This bound is reached when G^K is split into b portions, each forming a perfect helix with the corresponding $\mathsf{C}^{x_{d_i}}$ block. Noting that $x_i < \lfloor K/2 \rfloor$ is equivalent to $x_i \leq \lfloor K/2 \rfloor - 1$, one has that any connection of G^K with b blocks can therefore create at most $\min(K, b\cdot(\lfloor K/2 \rfloor - 1))$ base-pairs. It follows that, for $b=1$ and $b=2$, the maximum number of stackings involving G^K is bounded by $\lfloor K/2 \rfloor - 2$ and $2\lfloor K/2 \rfloor - 4 \leq K-4$ respectively. For $b \geq 3$, the number of base-pairs is potentially no longer limited by the lack of occurrences of C, but by the K occurrences of G in G^K. It follows that, when $b \geq 3$, the maximum number of stacking pairs is $K-3$, and is reached for $b=3$ when every position in G^K is paired.

Then let us assume the existence of a matching with $(K-3)\cdot m$ stacking pairs. Since $K-3$ is the upper-bound on the number of stacking pairs supported by a given G^K block, then each of the m G^K blocks must achieve this upper-bound. It follows that each G^K block must create a total of exactly K base-pairs with a triplet of blocks $(\mathsf{C}^x_a, \mathsf{C}^x_b, \mathsf{C}^x_c)$. A direct corollary is that every G and C in w_X must be paired.

We have now established that, within any matching having $m\cdot(K-3)$ stacking pairs, each G^K block creates exactly K base-pairs with a triplet of blocks $(\mathsf{C}^x_{a_i}, \mathsf{C}^x_{b_i}, \mathsf{C}^x_{c_i})$. To conclude on the implication, we need to show that each C^x block interacts with a single G^K block, i.e. that the $(\mathsf{C}^x_{a_i}, \mathsf{C}^x_{b_i}, \mathsf{C}^x_{c_i})$ triplets are mutually disjoint. Indeed, if a block C^{x_i} is found

in two distinct triplets, then there exists a block C^{x_j} that is not within any triplet (remember that there are $3K$ blocks C^x and K triplets). It follows that at most $m \cdot K - x_j$ base-pairs exist within this matching, which contradicts K base-pairs for every G^K block. Consequently, no C^{x_i} can be present in two distinct triplets, and the triplets are therefore disjoint.

Therefore, the interacting blocks found in a matching having energy $\delta \cdot (K-3) \cdot m$ induce a partition of the $\{\mathsf{C}^{x_i}\}_{i=1}^{3m}$ blocks into m triplets. Furthermore, each $(\mathsf{C}_a^x, \mathsf{C}_b^x, \mathsf{C}_c^x)$ triplet must give rise to K base-pairs, and therefore $x_a + x_b + x_c \geq K$. Since the triplets are disjoint and partition a set of a total $m \cdot K$ occurrences of C, then any excess of C within a triplet implies a lack of C within another, so one has $x_a + x_b + x_c = K$. We conclude that any matching of w_X having energy $\delta \cdot (K-3) \cdot m$ induces the existence of a partition $\{\mathsf{C}^{x_i}\}_{i=1}^{3m}$ blocks into disjoint triplets $(\mathsf{C}^{x_a}, \mathsf{C}^{x_b}, \mathsf{C}^{x_c})$ such that $x_a + x_b + x_c = K$ which, in turn, implies the existence of a 3-partition for X. □

It follows from Lemma 2 that any algorithm for RNA-PK-Fold($\mathcal{S}$) gives an algorithm for the 3-PARTITION problem. Furthermore the length of w_X exactly equals $\sum_{i=1}^{n} x_i + K \cdot m + 2m - 1 = 2K \cdot m + 2m - 1 \in \mathcal{O}(n^2 \cdot P(n))$ where $P(n)$ is the polynomial upper bound on the value of each x_i. Therefore any polynomial algorithm for RNA-PK-Fold($\mathcal{S}$) gives a polynomial algorithm for the 3-PARTITION problem. Since 3-PARTITION is NP-hard, then so is RNA-PK-Fold($\mathcal{S}$) and Theorem **??** follows. □

4 Approximability of RNA-PK-Fold($\mathcal{S}$) in the stacking model

Since objective functions are usually derived experimentally or statistically, it is a natural question to ask whether hard problems can be efficiently approximated. Previous works on the subjetc only considered a combinatorial version of the problem, and the current best algorithm [10] produces a matching whose number of stacking pairs is guaranteed to be at least $3/8 \cdot OPT$, where OPT is the maximal number of stacking pairs in any matching. Unfortunately, this result does not hold for arbitrary-valued stacking energy models, as the free-energy of valid stacking pairs may greatly vary. For instance, the latest version of the Turner model reports a factor ~ 3.6 discrepancy between stacking canonical pairs, bringing the guaranteed approximation ratio down to $1/10$. By contrast, we show that RNA-PK-Fold($\mathcal{S}$) can be approximated in polynomial time up to a factor at least $1/5$, for any stacking model $\mathcal{S}$.

Input : An RNA sequence w
Output: A matching m of non-overlapping pairs of positions
$G = (V, E) \leftarrow ([1, |w| - 1], \varnothing)$;
$M \leftarrow \varnothing$;
foreach $u, v \in V$ **do**
  if w_u *base pairs with* w_{v+1} and w_{u+1} *base pairs with* w_v **then**
    `// Label each edge with its weight/energy`
    $E \leftarrow E \cup (u, v, -\Delta_{\mathcal{S}}(w_u, w_{v+1}, w_{u+1}, w_v))$;
  end
end
$m' \leftarrow$ `MaxWeightedMatching`(G);
foreach $(u, v) \in m'$ *sorted by increasing value* $\Delta_{\mathcal{S}}(w_u, w_{v+1}, w_{u+1}, w_v)$ **do**
  if $\forall (u', v') \in m, \{u', v'\} \cap \{u, v+1, u+1, v\} = \varnothing$ *or*
  $(u', v') \in \{(u, v+1), (u+1, v)\}$ **then**
    $m \leftarrow m \cup \{(u, v+1), (u+1, v)\}$;
  end
end
return m

Algorithm 1: A 5-approximation for any stacking energy model.

Theorem 3 *In any stacking energy model, RNA-PK-Fold*$(\mathcal{S}) \in APX$*, and can be approximated in polynomial time within a factor at least* $1/5$.

Proof. To prove the approximability of RNA-PK-Fold($\mathcal{S}$), let us consider Algorithm 1. This algorithm contracts consecutive positions in the RNA sequence as vertices, and adds an edge, weighted according to the energy function, between any pair of compatible positions. Computing a maximal weighted matching on this graph gives a set of stacking-pair which is not necessarily a valid matching, since distinct pairs of stacking pairs may induce more than a single partners for a given position. Therefore the algorithm considers the returned stacking pairs in decreasing order, and only retains the stacking pairs that do not conflict with the current selection of stacking-pairs.

Now let m^* be the optimal matching for the given RNA string w, m' be the maximal matching over G, and m be the matching finally returned by the algorithm. Let us remark that m' induces a set of matched pairs over w that does not strictly constitutes a matching, as some position may be matched twice. Nevertheless let us write $\mathcal{S}_w(m')$ as a shorthand for the total energy of m', obtained by summing over the stacking pairs induced by m'. Any matching can be decomposed as a set of stacking pairs (leaving a set of isolated, non-contributive, base-pairs), hence one has $\mathcal{S}_w(m') \leq \mathcal{S}_w(m^*) \leq 0$. Any edge (i, j) in m' may conflict with at most 4 other, adjacent, stacking-pairs. Furthermore, the algorithm con-

siders the edges in m' by decreasing contribution, so the stacking pairs induced by any edge (i, j) in m' may only conflict with four stacking pairs having (negative) contribution of smaller absolute value. Discarding these competitors guarantees that at least $\frac{1}{5}$ of the total energy of m' is retained in m, i.e. $\mathcal{S}_w(m) \leq \frac{1}{5}\mathcal{S}_w(m') \leq 0$, and one therefore concludes that $\frac{\mathcal{S}_w(m)}{\mathcal{S}_w(m^*)} \geq \frac{1}{5}$. □

Remark that the actual approximation ratio achieved by Algorithm 1 might be better than 1/5, even in the worst-case scenario. However, this crude upper-bound already establishes the approximability of the problem, nicely contrasting with our upcoming inaproximability result for the nearest-neighbor version of the problem.

5 Inapproximability of RNA-PK-Fold($\mathcal{N}$) in the nearest-neighbor energy model

The stacking model, considered in the above sections, makes the prediction of RNA structure NP-Hard, yet remains approximable in general. By contrast, let us show that RNA-PK-Fold($\mathcal{N}$), the nearest-neighbor version of the problem, is non-approximable. More precisely, let us show the stronger property that, unless $P = NP$, there is no polynomial-time algorithm that guarantees to find a matching whose free-energy approximates that of the optimal matching up to a strictly positive factor $r(n)$.

Theorem 4 *There exists instances of the nearest-neighbor model such that RNA-PK-Fold($\mathcal{N}$) $\notin APX$.*

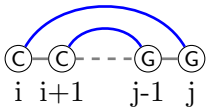
Proof. Let us briefly outline our proof strategy. We encode any set of numbers X as a string w, having length polynomial on the sum of values in X, and whose matchings are either forbidden ($+\infty$ free-energy), empty (0 free-energy), or have negative energy. Focusing on the latter category, we show that any negative energy matching can be turned, in polynomial-time, into a solution to the 3-PARTITION problem. It follows that any polynomial-time algorithm that guarantees a positive-ratio approximation, thereby producing a matching having negative free-energy anytime such a matching exists, immediately yields a polynomial-time algorithm for the 3-PARTITION problem. The NP-hardness of this problem allows us to conclude on the hardness of approximating RNA-PK-Fold($\mathcal{N}^*$), within any positive ratio, in the nearest-neighbor energy problem.

Let us consider the 3-PARTITION problem, fully defined in Section 3. For any instance $X = \{x_i\}_{i=1}^{3m}$ of the problem, let us consider the following

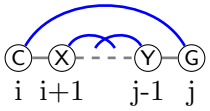
RNA sequence:

$$w = \mathsf{C}^{x_1}\mathsf{AC}^{x_2}\mathsf{A}\cdots\mathsf{AC}^{x_{3m}}\mathsf{A}\underbrace{\mathsf{G}^K\mathsf{UG}^K\mathsf{U}\cdots\mathsf{G}^K\mathsf{U}}_{m \text{ times}}\mathsf{U}^{2m}$$

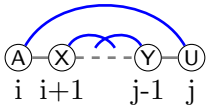
Moreover let us consider a nearest-neighbor energy model $\mathcal{N}^*$, defined by a function $\Delta^*_{\mathcal{N}}$ such that:

$\Delta^*_{\mathcal{N}}$: (A) 

$\longrightarrow -1, \quad \forall i < j,$

(B) 

$\longrightarrow -1, \quad \forall i < j, \forall X \neq C, \forall Y,$
($i+1$ and $j-1$ **must both** base-pair somewhere, possibly together)

(C) 

$\longrightarrow -1, \quad \forall i < j, \forall (X, Y),$
($i+1$ and $j-1$ **must both** base-pair somewhere, possibly together)

(D) Any other motif $\longrightarrow +\infty, \ \forall i < j.$

Lemma 5 *Let X be a 3-PARTITION instance whose values are bounded by $P(n)$. Then the following statements are equivalent:*

- *There exists a 3-partition of X into m triplets of equal sum.*
- *There exists a matching of strictly negative energy over w under $\mathcal{N}^*$.*

Proof. X **is 3-PARTITIONABLE** $\Rightarrow \exists m^*$ **such that** $\mathcal{N}^*_w(m^*) < 0$**:** Since X is 3-PARTITIONABLE, then there exists a partition of X into m disjoint triplets $((x_{a_j}, x_{b_j}, x_{c_j}))_{j=1}^m$ whose sum are identically K. Consider the matching that pairs each G^K block with one of the triplet of blocks $\mathsf{C}^{x_{a_j}}$, $\mathsf{C}^{x_{b_j}}$ and $\mathsf{C}^{x_{c_j}}$, creating nested sequences of base-pairs, and completed with $3 \cdot m$ (A, U) unconstrained base-pairs over the remaining positions. Clearly, all the positions are involved in a base-pair, and consecutive $\mathsf{CC}\cdots\mathsf{GG}$ are nested as required by energy rule (A). Therefore, any base-pair falls within the scope of energy rules (A), (B) or (C), and the final energy of the matching is $\mathcal{N}^*_w(m^*) = -m \cdot (K + 3) < 0$.
$\exists m^*$ **such that** $\mathcal{N}^*_w(m^*) < 0 \Rightarrow$ X **is 3-PARTITIONABLE:** Let us start by proving that, within an energy model $\mathcal{N}^*$, any matching of w having negative energy is a perfect matching, i.e. every position in the matching is paired. Since $\mathcal{N}^*$ only allows (C, G) and (A, U) pairs, therefore any valid (finite, negative contribution) base-pair (i, j) must involve a

position in the left half of w (C or A) and a position in its right half (G or U), i.e. such that $i \leq m \cdot (K+3) < j$. In order to be valid, (i,j) must also be in a context where $i+1$ (resp. $j-1$) is paired to $j' \geq m \cdot (K+3)$ (resp. $i' < m \cdot (K+3)$). The same argument applies to $(i+1, j')$ and $(i', j-1)$, and one easily shows by induction that any matching featuring a base-pair (i,j) has infinite energy unless every position in $[i,j]$ is paired. It follows that any matching having negative energy is perfect on some interval $[a,b]$, $a \leq m \cdot (K+3) < b$, and leaves the remaining positions unpaired.

Now let us consider which bounds for the interval $[a,b]$ are compatible with a negative energy. Let us denote by $w_{[a,b]}$ the $[a,b]$ factor in a sequence w, and by $|w|_t$ the number of occurrences of some letter t in w, then one has $|w_{[a,b]}|_A = |w_{[a,b]}|_U$ and $|w_{[a,b]}|_\mathsf{C} = |w_{[a,b]}|_\mathsf{G}$. Observe that, since $x_i < K/2$, one has $\frac{|w_{[a,b]}|_A}{|w_{[a,b]}|_C} \leq \frac{1}{1+K/2}$. Furthermore, if b falls before the final run U^{2m}, then $b < m \cdot (2K+4)$ and one has $\frac{1}{1+K} \leq \frac{|w_{[a,b]}|_U}{|w_{[a,b]}|_G}$. It follows that $\frac{|w_{[a,b]}|_\mathsf{A}}{|w_{[a,b]}|_\mathsf{C}} < \frac{|w_{[a,b]}|_\mathsf{U}}{|w_{[a,b]}|_\mathsf{G}}$, i.e. the matching cannot be perfect on $[a,b]$, and its energy cannot be negative. We are then left to consider the case where $m \cdot (2K+4) \leq b \leq |w_X|$. In such a case, one has $|w_{[a,b]}|_\mathsf{G} = m \cdot K$ and one has $a = 1$. Indeed any greater value a would lead to less than $\sum x_i = m \cdot K$ copies of C, and some G would be left alone. Remark that $|w_{[1,b]}|_\mathsf{A} = 3m$, so one must have $b = |w|$, from which we conclude that any matching having negative energy is perfect, i.e. base-pairs every position.

Let us finally show that a 3-PARTITION of X can be retrieved from a matching having negative energy. Remind that energy rule (A) forces two consecutive occurrences of C to pair with consecutive occurrences of G. This property extends transitively, and any C^{x_i} block in w must therefore be entirely connected to a single G^K block. Since a matching of negative energy is perfect, then all the positions in a G^K block must be base-paired. Two C^{x_i} blocks are not sufficient ($x_i < K/2$) to saturate a G^K block, and four blocks would be too large ($x_i > K/4$), violating the constraint that each block must be entirely paired to a single G^K block. Therefore a triplet $(\mathsf{C}^{x_{a_i}}, \mathsf{C}^{x_{b_i}}, \mathsf{C}^{x_{c_i}})$ blocks is in total interaction with each G^K block, and the corresponding values (a_i, b_i, c_i) constitute a 3-PARTITION of X. □

From Lemma 5, one knows that the existence of a 3-PARTITION for X can be derived from the existence of a matching of w having negative energy under $\mathcal{N}^*$. Now assume there exists a polynomial-time algorithm $\mathcal{A}$ that guarantees an $r(n) > 0$ approximation ratio. Then $\mathcal{A}$

would produce a matching m such that $\mathcal{N}_w^*(m) = \mathcal{N}_w^*(m^*)/r(n)$, for m^* the optimal matching. In particular, $\mathcal{A}$ would produce a matching having negative energy anytime such a matching exists. One could then decide, in polynomial time, the 3-partitionability of any set X. Since the decision version of 3-PARTITION is NP-Hard, then there is no such algorithm unless $P = NP$. □

6 Conclusion/perspectives

We considered the influence of the energy model on the computational complexity of RNA folding with general pseudoknots. In the simplest base-pair model, the problem is exactly equivalent to finding a maximal weighted matching in the graph of compatible positions, and can be solved in $\Theta(n^3)$ [17]. By contrast, it was previously established that the more expressive nearest-neighbor model made the problem NP-Hard [1, 12]. We completed this result by showing that this problem is actually inapproximable within any ratio. Turning to a less expressive – yet realistic – stacking energy model, we have showed that, although NP-hard, the problem could be approximated in polynomial time, at least up to a $\frac{1}{5}$ approximation ratio.

Quite nicely, a similar approach could be used to refine the computational complexity of RNA-RNA interaction prediction. Already proven NP-complete by Alkan *et al* [2], it can be verified that our approximation algorithm achieves the same ratio for RNA-RNA interactions. Furthermore, our NP-hardness and inapproximability results consider bi-partite strings, for which an algorithm for the RNA-RNA interaction problem, suitably parameterized, would yield the same matching as an algorithm for RNA folding with general pseudoknots.

These results show a difference in essence between the nearest-neighbor and the stacking models, which could serve as a starting point for a design of practical (approximation) algorithms for the stacking version of the problem. To that purpose, we plan to complement this study by investigating the existence of a polynomial-time approximation scheme for the problem. Another direction for complementing this study would consider the impact of the energy model on the parameterized-complexity of the problem.

Acknowledgement

This work was supported by the French ANR MAGNUM `ANR 2010 BLAN 0204` grant (YP) and by an INRIA Postdoc Program (SS).

References

1. Akutsu, T.: Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. Discrete Appl. Math. 104(1-3), 45–62 (2000)
2. Alkan, C., Karakoç, E., Nadeau, J.H., Şahinalp, S.C., Zhang, K.: Rna-rna interaction prediction and antisense rna target search. In: Proceedings of RECOMB'05. pp. 152–171. Springer-Verlag, Berlin, Heidelberg (2005)
3. Ashley, M.V., Berger-Wolf, T.Y., Chaovalitwongse, W., Dasgupta, B., Khokhar, A., Sheikh, S.: On approximating an implicit cover problem in biology. In: Proceedings of AAIM '09. pp. 43–54. AAIM '09 (2009)
4. Bindewald, E., Kluth, T., Shapiro, B.A.: Cylofold: secondary structure prediction including pseudoknots. Nucleic Acids Research 38(suppl 2), W368–W372 (2010)
5. Bon, M.: Prédiction de structures secondaires d'ARN avec pseudo-noeuds. Ph.D. thesis, Ecole Polytechnique (September 2009)
6. Cary, R.B., Stormo, G.D.: Graph-theoretic approach to RNA modeling using comparative data. Proceedings International Conference on Intelligent Systems for Molecular Biology 3, 75–80 (1995)
7. Garey, M.R., Johnson, D.S.: Complexity results for multiprocessor scheduling under resource constraints. SIAM Journal on Computing 4(4), 397–411 (1975)
8. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R.: Rfam: an RNA family database. Nucleic Acids Research 31(1), 439–441 (2003)
9. Ieong, S., Kao, M.Y., Lam, T.W., Sung, W.K., Yiu, S.M.: Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs. Journal of Computational Biology 10(6), 981–995 (2003)
10. Jiang, M.: Approximation algorithms for predicting RNA secondary structures with arbitrary pseudoknots. IEEE/ACM Trans. Comput. Biology Bioinform. 7(2), 323–332 (2010)
11. Liu, C., Song, Y., Shapiro, L.: RNA folding including pseudoknots: A new parameterized algorithm and improved upper bound. In: Algorithms in Bioinformatics. Lecture Notes in Computer Science, vol. 4645, pp. 310–322 (2007)
12. Lyngsø, R.B., Pedersen, C.N.: RNA pseudoknot prediction in energy-based models. J Comput Biol 7(3-4), 409–427 (2000)
13. Lyngsø, R.B.: Complexity of pseudoknot prediction in simple models. In: ICALP. pp. 919–931 (2004)
14. 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)
15. Reidys, C.M., Huang, F.W.D., Andersen, J.E., Penner, R.C., Stadler, P.F., Nebel, M.E.: Topology and prediction of RNA pseudoknots. Bioinformatics 27(8), 1076–1085 (Apr 2011)
16. Rivas, E., Eddy, S.: A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol 285, 2053–2068 (1999)
17. Tabaska, J.E., Cary, R.B., Gabow, H.N., Stormo, G.D.: An RNA folding method capable of identifying pseudoknots and base triples. Bioinformatics 14(8), 691–699 (1998)
18. Theis, C., Janssen, S., Giegerich, R.: Prediction of RNA secondary structure including kissing hairpin motifs. In: Proceedings of WABI 2010. pp. 52–64 (2010)
19. Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H., Westhof, E.: Tools for the automatic identification and classification of rna base pairs. Nucleic Acids Research 31(13), 4250–4263 (2003)
20. Zhao, J., Malmberg, R., Cai, L.: Rapid ab initio prediction of RNA pseudoknots via graph tree decomposition. Journal of Mathematical Biology 56, 145–159 (2008)