NP-completeness of the direct energy barrier problem without pseudoknots

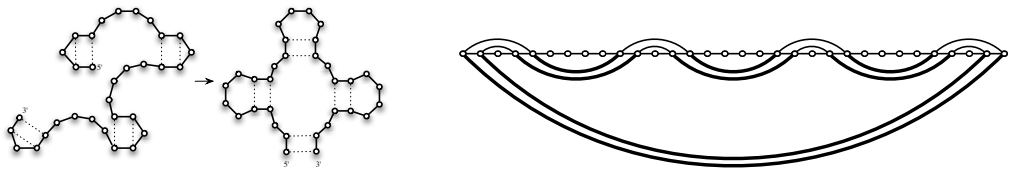
Ján Maňuch[1], Chris Thachuk[2], Ladislav Stacho[1], Anne Condon[2]

[1] Simon Fraser University
[2] University of British Columbia

Abstract. Knowledge of energy barriers between pairs of secondary structures for a given DNA or RNA molecule is useful, both in understanding RNA function in biological settings and in design of programmed molecular systems. Current heuristics are not guaranteed to find the exact energy barrier, raising the question whether the energy barrier can be calculated efficiently. In this paper, we study the computational complexity of a simple formulation of the energy barrier problem, in which each base pair contributes an energy of -1 and only base pairs in the initial and final structures may be used on a folding pathway from initial to final structure. We show that this problem is NP-complete.

1 Introduction.

We study the computational complexity of the energy barrier problem for nucleic acids: what energy barrier must be overcome for a DNA or RNA molecule to adopt a given final secondary structure, starting from a given initial secondary structure? We first provide some motivation for studying the energy barrier problem, then describe a simple formulation of the problem and summarize our results.



(a) (b)

Fig. 1: (a) An initial secondary structure (left) and a final secondary structure (right) for a given RNA strand. (b) A corresponding arc diagram. The top of the diagram denotes the base pairs in the initial structure while the bottom of the diagram, shown with bold arcs, denotes the base pairs of the final structure.

Motivation. Methods for calculating energy barriers are useful, in both rational design of programmed nucleic acid systems and in understanding the mechanisms of RNA function in the cell. This is because, both in the design and biological contexts, secondary structure and folding pathways are central to function. Many designed nucleic acid systems rely critically on the premise that the constituent molecules will follow certain folding pathways and avoid others [1–7]. Designs of such systems typically ensure that the desired folding pathway has a low energy barrier, compared with alternatives. While this property can be straightforward to establish for simple molecular systems, a method for energy barrier calculation would be useful when verifying that a system of large or even moderate scale has the desired behaviour [8]. In the biological context, knowledge of energy barriers between intermediate structures on the pathway from the open chain to the folded configuration of biological molecules is useful in determining folding efficiency and structure [9–12].

Methods for simulation of DNA or RNA folding pathways often estimate energy barriers between secondary structures, in order to calculate probabilities of transitioning between structures which are included in maps of the energy folding landscape [13–16] Heuristics for energy barrier calculation are also used to construct barrier trees, which are helpful in visualizing energy landscapes [17], and to elucidate properties of

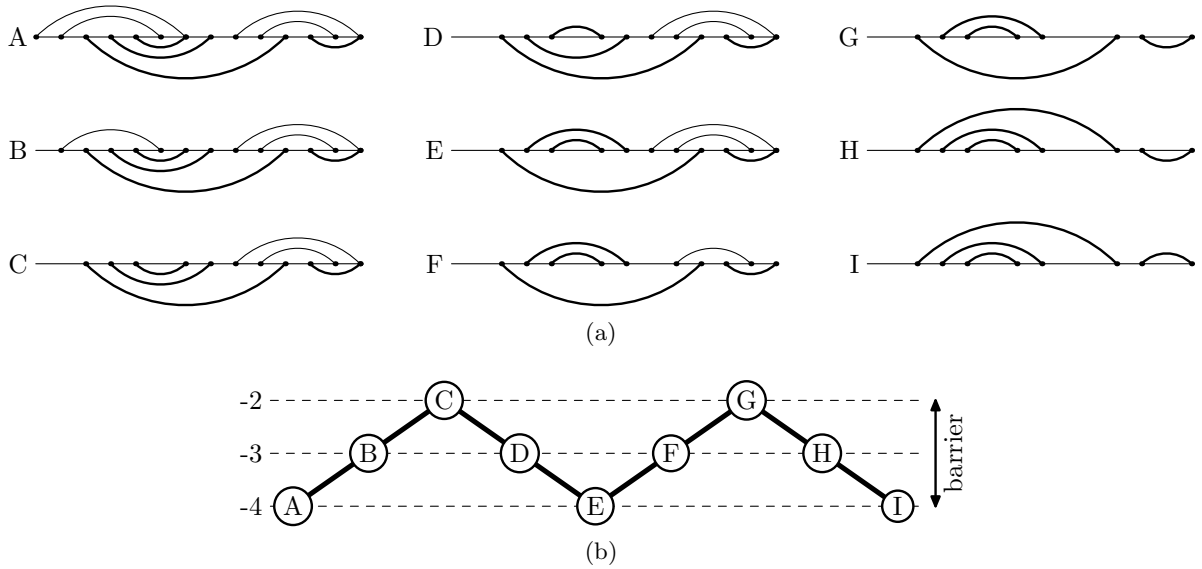


Fig. 2: (a) A possible folding pathway is shown for an initial structure A transitioning through intermediate structures $(B, C, \ldots)$ until the final structure I is reached. For a particular position in the pathway, the top of the arc diagram denotes the current base pairs of the structure. Each structure along the pathway differs from its neighbours by at most one arc. (b) The corresponding energy plot. The barrier in this example is two.

disordered systems in statistical physics [18]. They showed that the mean barrier between MFE structures of random sequences scales with the square root of the sequence length.

In light of its importance, it's natural to ask: what is the computational complexity of the energy barrier problem for nucleic acid secondary structures? An efficient algorithm for the problem might be a valuable replacement of currently-used heuristics in the applications mentioned above. On the other hand, intractability of the barrier problem would suggest that the result of a complex computation might be determined from observations of nucleic acid folding pathways.

The Energy Barrier Problem. We formulate the problem as follows. A *secondary structure* $\mathcal{T}$ for an RNA molecule of length n is a set of base pairs $i.j$, with $1 \leq i < j \leq n$, such that (i) each base index i or j appears in at most one base pair and (ii) the bases at indices i and j form a Watson-Crick (i.e., C-G, A-U, or A-T) base pair. Since we represent secondary structures using arc diagrams, we use the word *arc* interchangeably with base pair (see Fig. 1). Our main results pertain to pseudoknot free secondary structures, that is, structures with no crossing arcs — see Section 2 for definitions. We assume a very simple energy model for secondary structures in which each arc contributes an energy of -1. Thus, as is roughly consistent with more realistic energy models, the more base pairs in a structure the lower its energy. We denote the energy of secondary structure $\mathcal{T}$ by $E(\mathcal{T})$.

A *folding pathway* is a sequence of pseudoknot free secondary structures for a given molecule, each of which differs from its predecessor by the addition or removal of one arc (see Fig. 2). The *energy barrier* of a folding pathway $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_r$ is the maximum of $E(\mathcal{T}_i) - E(\mathcal{T}_0)$, where the max is taken over i in the range $1 \leq i \leq r$. Note that there is always a folding pathway from $\mathcal{I}$ to $\mathcal{F}$, in which first all arcs of $\mathcal{I}$ are removed and then all arcs of $\mathcal{F}$ are added. All secondary structures on such a pathway are pseudoknot free since they are either subsets of $\mathcal{I}$ or of $\mathcal{F}$, both of which are pseudoknot free. However, the energy barrier of this pathway is $|\mathcal{I}|$. The question is whether there is another pathway that avoids such a high energy barrier, by adding arcs of $\mathcal{F}$ before all arcs of $\mathcal{I}$ are removed. The *energy barrier problem* is to determine whether there is a folding pathway from a given initial structure $\mathcal{I}$ to a given final structure $\mathcal{F}$, whose energy barrier is at most k, for some given k.

The results presented here are a first step towards solving the energy barrier problem. Our results pertain to restricted types of folding pathways, namely *direct* folding pathways. Such pathways were introduced by

Morgan and Higgs [18]. A folding pathway from secondary structure $\mathcal{I}$ to $\mathcal{F}$ is *direct* if the only arcs which are added are those from $\mathcal{F} - \mathcal{I}$ and the only arcs which are removed are those from $\mathcal{I} - \mathcal{F}$. Thus, there are exactly $|\mathcal{I} - \mathcal{F}| + |\mathcal{F} - \mathcal{I}|$ steps along a direct folding pathway. All of the designed nucleic acid folding pathway systems with which we are familiar are such that the desired folding pathway is direct [2–4, 7]. The *energy barrier problem for direct pseudoknot free folding pathways (*DPKF-EB PROBLEM*)* is: given initial and final pseudoknot free secondary structures $\mathcal{I}$ and $\mathcal{F}$ and an integer k, is there a direct folding pathway from $\mathcal{I}$ to $\mathcal{F}$ which has energy barrier at most k? In Section 3, we show that the DPKF-EB problem is NP-complete.

The rest of the paper is organized as follows. We provide definitions of pathways, energy barrier, and other useful notation in Section 2. We prove our result in Section 3: Theorem 2 shows that the energy barrier problem for direct folding pathways of pseudoknot free structures is NP-complete. We conclude with a brief discussion of our result and open problems in Section 4.

2 Definitions

Fix initial and final pseudoknot free secondary structures $\mathcal{I}$ and $\mathcal{F}$. A *direct pseudoknot free folding pathway* from $\mathcal{I}$ to $\mathcal{F}$ is a sequence of pseudoknot free secondary structures $\mathcal{I} = \mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_r = \mathcal{F}$, where each $\mathcal{T}_i$ is obtained from $\mathcal{T}_{i-1}$ by either the addition of one arc from $\mathcal{F} - \mathcal{I}$ or the removal of one arc from $\mathcal{I} - \mathcal{F}$. We call each such addition or removal an *arc operation* and we let $+x$ and $-x$ denote the addition and removal of the arc x, respectively. The $\mathcal{T}_i$'s are called *intermediate structures*. A folding pathway can thus be specified by its corresponding sequence of arc operations; we call this a *transformation sequence*. A *direct pseudoknot-free transformation sequence* specifies a folding pathway which is both direct and pseudoknot free.

The *energy barrier* of a folding pathway $\mathcal{I} = \mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_r = \mathcal{F}$ is the maximum of $E(\mathcal{T}_i) - E(\mathcal{I})$, where the max is taken over all integers i in the range $1 \leq i \leq r$. The *energy difference* of each intermediate configuration $\mathcal{T}_i$ is defined as $\mathrm{E}(\mathcal{T}_i) - \mathrm{E}(\mathcal{I})$. If Π is the transformation sequence for this pathway, then the *energy barrier* of transformation sequence Π, denoted as $\Delta\,\mathrm{E}(\mathcal{I}, \mathcal{F}, \Pi)$, is defined to be the energy barrier of the corresponding folding pathway.

In our result, it is convenient to work with weighted arcs. To motivate why, note that the union $\mathcal{I} \cup \mathcal{F}$ of two pseudoknot free secondary structures may be pseudoknotted, i.e., may have crossing arcs, even when both $\mathcal{I}$ and $\mathcal{F}$ are pseudoknot free. In a pseudoknotted structure, we use the term *band* to refer to a set of nested arcs, each of which crosses the same set of arcs. In a folding pathway from $\mathcal{I}$ to $\mathcal{F}$ which minimizes the energy barrier, we can assume without loss of generality that when one arc in a band of $\mathcal{I} \cup \mathcal{F}$ is added, then all arcs in the band are added consecutively. Similarly, we can assume without loss of generality that when one arc in a band is removed, then all arcs in the band are removed consecutively. Thus, it is natural to represent the set of arcs in a band as one arc with a weight equal to the number of arcs in the band.

Hence we generalize the notion of secondary structure as follows. A *weighted arc* $I = (I^b, I^e)^{I^w}$ is specified by start and end indices $I^b < I^e$ and a weight I^w. We say that two weighted arcs I and J are *crossing* if either $I^b \leq J^b$ and $I^e \leq J^e$, or $J^b \leq I^e$ and $J^e \leq I^e$. A *configuration* is a set of weighted arcs. Configuration $\{I_i\}_{i=1}^n$ is *pseudoknot-free* if for all $1 \leq i < j \leq n$, I_i and I_j are not crossing. The energy of configuration $\mathcal{I} = \{I_i\}_{i=1}^n$ is $\mathrm{E}(\mathcal{I}) = -\sum_{i=1}^n I^w$. The previous definitions can easily be generalized to weighted arcs.

Finally, we define the main problem studied in this paper, namely the DPKF-EB problem, and also the 3-PARTITION problem which is used to show NP-completeness of DPKF-EB.

DPKF-EB PROBLEM (Energy barrier problem for direct folding pathways without pseudoknots). Given two pseudoknot-free configurations $\mathcal{I} = \{I_i\}_{i=1}^n$ (initial) and $\mathcal{F} = \{F_i\}_{i=1}^m$ (final), and integer k, is there a direct pseudoknot-free transformation sequence S such that the energy barrier of S is at most k, i.e., $\Delta\,\mathrm{E}(\mathcal{I}, \mathcal{F}, S) \leq k$.

3-PARTITION PROBLEM. Given $3n$ integers $a_1, \ldots, a_{3n}$ such that $a_1 + \cdots + a_{3n} = nA$ and $A/4 < a_i < A/2$ for each i, the 3-PARTITION PROBLEM asks: is there a partition of the integers $\{1, \ldots, 3n\}$ into disjoint triples $\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n$ such that the sum of a_j, where j belongs to G_i is equal to A, i.e., $\mathrm{c}(\mathrm{G}_i) = \sum_{j \in \mathrm{G}_i} a_j = A$ for each $i = 1, \ldots, n$.

Theorem 1 (Garey, Johnson (1979) [19]). *The* 3-PARTITION *problem is NP-complete even if A is polynomial in n.*

The 3-PARTITION problem is in P if A is a constant.

3 Result

Theorem 2. *The* DPKF-EB *problem, namely the energy barrier problem for direct folding pathways without pseudoknots, is NP-complete.*

Proof. It is straightforward to show that the DPKF-EB problem is in NP. Given an instance $(\mathcal{I}, \mathcal{F}, k)$, it is sufficient to non-deterministically guess a direct folding pathway from $\mathcal{I}$ to $\mathcal{F}$, and to verify that the energy barrier of this path is at most k. Note that the length of any such pathway is at most $|\mathcal{I}| + |\mathcal{F}|$.

To show that the DPKF-EB problem is NP-hard, we provide a reduction from the 3-PARTITION problem. We first provide a formal description of the reduction, then provide some intuition as to why the reduction is correct, and then prove correctness in detail.

Consider an instance of the 3-PARTITION problem $A/2 > a_1 \geq \cdots \geq a_{3n} > A/4$ such that $\sum_{j=1}^{3n} a_j = nA$ and A is polynomial in n. We define an instance $(\mathcal{I}, \mathcal{F}, k)$ of the DPKF-EB problem as follows. The initial configuration $\mathcal{I}$ contains weighted arcs $\{\bar{A}_{j,i};\ j = 1, \ldots, 3n,\ i = 1, \ldots, n\} \cup \{\tilde{A}_{j,i};\ j = 1, \ldots, 3n,\ i = 1, \ldots, n\} \cup \{\tilde{T}_i;\ i = 1, \ldots, n\}$. The final configuration $\mathcal{F}$ is $\{A_{j,i};\ j = 1, \ldots, 3n,\ i = 1, \ldots, n\} \cup \{T_i;\ i = 1, \ldots, n\}$. The arcs are organized as in Figure 3.

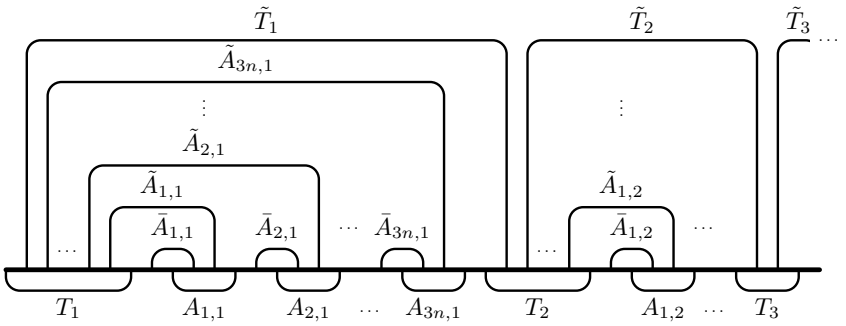


Fig. 3: Organization of weighted arcs in the initial (top) and the final (bottom) configurations.

Formally, the arcs are organized as follows:

$$
\begin{aligned}
T_1^b < \tilde{T}_1^b < \tilde{A}_{3n,1}^b < \cdots < \tilde{A}_{1,1}^b < T_1^e < \bar{A}_{1,1}^b, \\
T_i^b < \tilde{T}_{i-1}^e < \tilde{T}_i^b < \tilde{A}_{3n,i}^b < \cdots < \tilde{A}_{1,i}^b < T_i^e < \bar{A}_{1,i}^b,\ \forall i = 2, \ldots, n, \\
\bar{A}_{j,i}^b < A_{j,i}^b < \bar{A}_{j,i}^e < \tilde{A}_{j,i}^e < A_{j,i}^e,\ \forall i = 1, \ldots, n,\ j = 1, \ldots, 3n, \\
A_{j,i}^e < \bar{A}_{j+1,i}^b,\ \forall i = 1, \ldots, n,\ j = 1, \ldots, 3n-1, \\
A_{3n,i}^e < T_{i+1}^b,\ \forall i = 1, \ldots, n-1, \\
A_{3n,n}^e < \tilde{T}_n^e .
\end{aligned}
$$

The weights of arcs are set up as follows. For all $i = 1, \ldots, n$ and $j = 1, \ldots, 3n$:

$$
\begin{aligned}
\tilde{A}_{j,i}^w &= 4ia_j, \\
\bar{A}_{j,i}^w &= k - (j-1)A - 4ia_j, \\
A_{j,i}^w &= k - jA.
\end{aligned}
$$

Also,

$$
\begin{aligned}
\tilde{T}_1^w &= k - (7n-4)A, \\
\tilde{T}_i^w &= k - (6n+8)nA - 4(n-1)iA,\ \forall i = 2, \ldots, n, \\
T_i^w &= k - (6n+8)nA,\ \forall i = 1, \ldots, n-1, \\
T_n^w &= k\,,
\end{aligned}
$$

where $k > 4(5n^2 + n + 1)A$ is the energy barrier.

Before getting into the details of the proof, we next describe intuitively the key properties of the construction. The weights are chosen to ensure that the folding pathway with minimum energy barrier has the following properties. Here we list only the arcs that are added and assume without loss of generality that all arc removals happen only when needed.

1. Initially a (possibly empty) sequence of $A_{j,i}$'s are added to the folding pathway. The added $A_{j,i}$'s define a potential solution $\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n$ to the 3-PARTITION problem in a natural way: G_i contains j if $A_{j,i}$ is in this initial sequence. As we will prove later, the weights ensure that the addition of each $A_{j,i}$ raises the energy difference. After $3n$ such additions, the energy difference is so high that no other $A_{j,i}$'s can be added. As a result, the weights impose certain desirable constraints on the G_i's which will help ensure that they (or a slight perturbation of the G_i's) form a valid solution.
2. Following the initially-added sequence of $A_{j,i}$'s, the T_i's must be added in increasing order of i (with no interspersed $A_{j,i}$'s). This is in part because of the placement of the $\tilde{T}_i$'s: adding T_1 requires only the removal of $\tilde{T}_1$, whereas adding T_i, for $i > 1$, requires the costlier removal of both $\tilde{T}_{i-1}$ and $\tilde{T}_i$. Thus, it becomes feasible to add T_i without exceeding the energy barrier only after T_{i-1} is added because, at that point, $\tilde{T}_{i-1}$ has already been removed. In addition after adding T_1, the energy difference increases to the level that none of the $A_{i,j}$'s can be added (and stays there until addition of T_n).
3. Moreover, the T_i's can be added without exceeding the energy barrier only if the G_i's defined by the initial sequence of $A_{j,i}$'s actually is a valid solution. That is, if the G_i's are valid then for each i, at least three of the $A_{j,i}$'s are in the initial sequence and so at least three of the $\tilde{A}_{j,i}$'s (whose weights sum to at least $4iA$) were removed in the initial part of the pathway described in 1 above. This means that at most $n-3$ of the $\tilde{A}_{j,i}$'s remain to be removed before T_i can be added. The total weight of the remaining $\tilde{A}_{j,i}$'s is just low enough to ensure that they can be added without exceeding the energy barrier. In contrast, if the G_i's are not valid then for some i the weight of the $\tilde{A}_{j,i}$'s which must be removed in order to add T_i causes the energy barrier k to be exceeded.

We now prove that the DPKF-EB instance has a solution with energy barrier at most k if and only if the 3-PARTITION instance $a_1, \ldots, a_{3n}$ has a solution.

First, assume that the 3-PARTITION problem has a solution $\mathrm{G}_1, \ldots, \mathrm{G}_n$, where $\mathrm{G}_i = \{j_{i,1}, j_{i,2}, j_{i,3}\}$. Let $f(j) = i$ if $j \in \mathrm{G}_i$, for every $j = 1, \ldots, 3n$. We will show that the transformation sequence

$$-\bar{A}_{1,f(1)}, -\tilde{A}_{1,f(1)}, +A_{1,f(1)}, \ldots, -\bar{A}_{3n,f(3n)}, -\tilde{A}_{3n,f(3n)}, +A_{3n,f(3n)}, \tag{1}$$

$$\underbrace{-\tilde{A}_{1,1}, \ldots, -\tilde{A}_{3n,1}}_{\text{without } -\tilde{A}_{j_{1,1},1}, -\tilde{A}_{j_{1,2},1}, -\tilde{A}_{j_{1,3},1}} , -\tilde{T}_1, +T_1, \ldots,$$

$$\underbrace{-\tilde{A}_{1,n}, \ldots, -\tilde{A}_{3n,n}}_{\text{without } -\tilde{A}_{j_{n,1},n}, -\tilde{A}_{j_{n,2},n}, -\tilde{A}_{j_{n,3},n}} , -\tilde{T}_n, +T_n, \tag{2}$$

$$\underbrace{-\bar{A}_{1,1}, +A_{1,1}, -\bar{A}_{1,2}, +A_{1,2}, \ldots, -\bar{A}_{3n,n}, +A_{3n,n}}_{\text{without indexes } 1, f(1);\ 2, f(2);\ \ldots;\ 3n, f(3n)} \tag{3}$$

is pseudoknot-free with energy barrier exactly k. For clarity, the $-$ sign marks the arcs from the initial configuration which are being removed and the $+$ sign marks the arcs from the final configuration which are being added. It is easy to see that the sequence is pseudoknot-free, since

- each $A_{j,i}$ is crossing only $\tilde{A}_{j,i}$ and $\bar{A}_{j,i}$ and is added only when these two arcs are already removed; and
- each T_i is crossing the following arcs in the initial configuration: $\tilde{T}_{i-1}$ (if $i > 2$), $\tilde{T}_i$ and $\tilde{A}_{1,i}, \ldots, \tilde{A}_{3n,i}$ and they are all removed before T_i is added.

Second, let us verify that the energy difference of each intermediate configuration is at most k. First, in line (1), by induction, for each $j = 1, \ldots, 3n$, before removing $-\bar{A}_{j,f(j)}, -\tilde{A}_{j,f(j)}$ the energy difference is $(j-1)A$ and after removal it is k. Then after adding $+A_{j,f(j)}$ it decreases to jA. At the end of line (1), the energy difference is $3nA$, cf. Figure 4. Next, we need to check that the sum of weights of arcs

$$\underbrace{-\tilde{A}_{1,1}, \ldots, -\tilde{A}_{3n,1}}_{\text{without } -\tilde{A}_{j_{1,1},1}, -\tilde{A}_{j_{1,2},1}, -\tilde{A}_{j_{1,3},1}} , -\tilde{T}_1$$

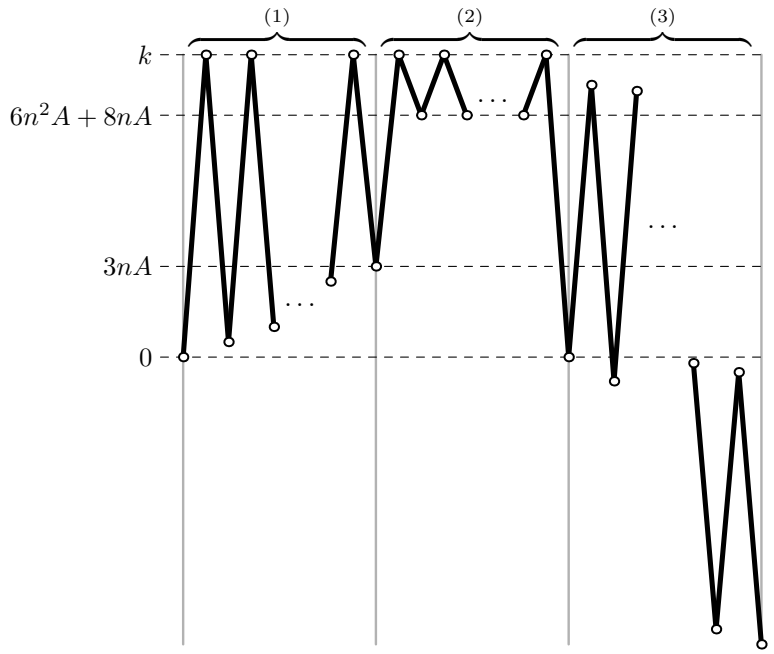


Fig. 4: Illustration of energy difference changes during the process of transitioning from the initial to final configuration of the construction.

is at most $k - 3nA$. The sum of weights of these arcs is exactly

$$\sum_{j=1}^{3n} \tilde{A}^w_{j,1} - \sum_{\ell=1}^{3} \tilde{A}^w_{j_{1,\ell},1} + \tilde{T}^w_1 = \sum_{j=1}^{3n} 4a_j - \sum_{\ell=1}^{3} 4a_{j_{1,\ell}} + k - 7nA + 4A$$
$$= 4nA - 4A + k - 7nA + 4A = k - 3nA\,.$$

Thus, just before adding $+T_1$, the energy difference is again exactly k. And after adding $+T_1$, it is $6n^2A+8nA$. Similar calculations show that the energy difference will alternate between k (after each removal subsequence) and $6n^2A + 8nA$ (after each addition of $+T_i$) in line (2) with exception of the last addition, when the energy difference is 0. In line (3), all remaining arcs from the initial configuration ($-\bar{A}_{j,i}$) are removed and all remaining arcs from the final configuration ($+A_{j,i}$) are added. Note that each removal is possible since $\bar{A}^w_{j,i} < k$ and after processing each pair $-\bar{A}_{j,i}, +A_{j,i}$, energy difference only decreases since $\bar{A}^w_{j,i} - A^w_{j,i} = A - 4ia_j < 0$.

Now, assume that there is a pseudoknot-free transformation sequence S with the energy barrier at most k. From S, we will construct a solution for the original 3-PARTITION instance and show that it is a valid solution. We organize our proof into three parts, in line with the three properties described in the intuition at the start of the proof.

Consider the subsequence of S containing only additions, i.e., arcs from the final configuration. Let S^+ denote this subsequence. We assume without loss of generality that all removals in S happen only when needed, i.e., the next addition would not be possible without those removals. Hence, the subsequence S^+ determines the whole sequence S. By *processing* an arc $+I$ in S^+ we mean removal of all arcs $-J$ in S immediately preceding $+I$ (that is, not preceding any other $+I'$ appearing in S^+ before $+I$) and adding $+I$.

The first part of our proof considers the prefix of S^+ just before the first T_ℓ is added. Let this prefix be:

$$+A_{j_1,i_1}, +A_{j_2,i_2}, \ldots, +A_{j_M,i_M} \tag{4}$$

where M is the number of $+A_{j,i}$'s added before $+T_\ell$. We use this prefix to define a potential solution to the 3-PARTITION problem:

$$\mathrm{G}_i = \{j_\ell;\ i_\ell = i\},$$

for every $i = 1, \ldots, n$.

Ultimately we will show that the G_i's (or a slight perturbation of the G_i's) form a solution to the 3-PARTITION problem. Towards this goal, our first two lemmas below prove some useful properties of the G_i's that can be inferred from the weights of the arcs in the folding pathway prefix (4) and the corresponding removed arcs. Let $|\,\mathrm{G}_i\,|_j$ denote the number of elements in G_i with value at most j. In order for the G_i's to be a valid solution, $|\,\mathrm{G}_i\,|_j$ should be exactly j for all $j, 1 \leq j \leq 3n$. Moreover it should be the case that $\sum_{i=1}^{n} \mathrm{c}(\mathrm{G}_i) = nA$ where $\mathrm{c}(\mathrm{G}_i)$ denotes the sum of a_j for $j \in \mathrm{G}_i$ (see the definition of 3-PARTITION). The statements of the two lemmas below assert somewhat weaker properties of the G_i's.

Lemma 1. *For every $j = 1, \ldots, 3n$, $\sum_{i=1}^{n} |\,\mathrm{G}_i\,|_j \leq j$. Consequently, $M \leq 3n$.*

Proof. Let $+T_\ell$ be the first $+T_i$ in S^+. Consider an $+A_{j,i}$ appearing before $+T_\ell$. Recall that before adding $+A_{j,i}$, we need to remove both $-\tilde{A}_{j,i}$ and $-\bar{A}_{j,i}$. Since, $\tilde{A}^w_{j,i} + \bar{A}^w_{j,i} = k - (j-1)A$, the energy difference has to be at most $(j-1)A$ for $+A_{j,i}$ to be added. Note that processing of each $+A_{j,i}$ appearing in S^+ before $+T_\ell$ will increase the energy difference by A, as it requires both $-\tilde{A}_{j,i}$ and $-\bar{A}_{j,i}$ to be removed first and $\bar{A}^w_{j,i} + \tilde{A}^w_{j,i} - A^w_{j,i} = k - (j-1)A - 4ia_j + 4ia_j - (k - jA) = A$. For instance, an $+A_{1,i}$ can only appear at the first position of the part of the subsequence S^+ before $+T_\ell$, since it requires the energy difference at least 0 and after any $+A_{j,i}$ is added, the energy difference increases to A. Thus, starting from the second position, no $+A_{1,i'}$ can be added before $+T_\ell$. Similarly, $+A_{j,i}$ can appear only in the first j positions of the subsequence of S^+ before $+T_\ell$. The lemma easily follows.

In the next lemma we we use double brackets to denote multisets: for example $\{\{1, 2, 2\}\}$ is the multiset with elements 1, 2, and 2 and $\{\{1, 1, 2\}\} \neq \{\{1, 2, 2\}\}$.

Lemma 2. $\sum_{i=1}^{n} \mathrm{c}(\mathrm{G}_i) \leq nA - (3n - M)A/4$, *where the equality happens only if $M = 3n$ and $\{\{a_{j_1}, \ldots, a_{j_M}\}\} = \{\{a_1, \ldots, a_{3n}\}\}$.*

Proof. Let $b_1 \geq b_2 \geq \cdots \geq b_M$ be the sorted elements of the multiset $\{\{a_{j_1}, \ldots, a_{j_M}\}\}$. Note that $\sum_{i=1}^{n} \mathrm{c}(\mathrm{G}_i) = \sum_{j=1}^{M} b_j$. We will show that $b_j \leq a_j$ for every $j = 1, \ldots, M$. Suppose to the contrary that $b_j > a_j$ for some j. Hence, elements $b_1, \ldots, b_j$ belong to $\{\{a_1, \ldots, a_{j-1}\}\}$, i.e., $|\{\{a_{j_1}, \ldots, a_{j_M}\}\}|_{j-1} \geq j$, a contradiction with Lemma 1. Hence, we have

$$\sum_{i=1}^{n} \mathrm{c}(\mathrm{G}_i) = \sum_{j=1}^{M} b_j \leq \sum_{j=1}^{M} a_j = nA - \sum_{j=M+1}^{3n} a_j \leq nA - (3n - M)A/4\,.$$

The equality happens only if $M = 3n$ (since $a_j > A/4$) and $b_j = a_j$, for every $j = 1, \ldots, 3n$.

We now turn to the second part of our proof: we show that, following the initially-added sequence of $+A_{i,j}$'s, the T_i's must be added in increasing order of i. That is, the arcs $+T_1, \ldots, +T_n$ appear in the subsequence S^+ consecutively (with no $+A_{j,i}$ in between) and in this order. The next lemma shows that the first $+T_i$ in the sequence S^+ must be $+T_1$ and the following lemma reasons about the rest of the sequence of $+T_i$'s.

Lemma 3. *The first $+T_i$ in S^+ is $+T_1$.*

Proof. Let $+T_\ell$ be the first $+T_i$ in S^+. As argued in the proof of Lemma 1, after each $+A_{j,i}$, the energy difference increases by A. Hence, before adding $+T_\ell$, the energy difference is non-negative. Second, if $\ell > 1$ then to add $+T_\ell$, both $-\tilde{T}_{\ell-1}$ and $-\tilde{T}_\ell$ has to be removed. After their removal the energy difference would be at least $2k - 2(6n + 8)nA - 4(n-1)(2\ell - 1)A > k$, a contradiction. The last inequality follows by $k > 4(5n^2 + n + 1)A = 2(6n + 8)nA - 4(n-1)(2n-1)A$.

Hence, by the above lemma, the subsequence S^+ has the following form

$$+A_{j_1,i_1}, +A_{j_2,i_2}, \ldots, +A_{j_M,i_M}, +T_1$$

followed by the all remaining $+A_{j,i}$'s and $+T_i$'s. The following lemma gives more detailed insight into order of arcs in S^+.

In the remaining lemmas we adopt notation which was introduced by Graham, Knuth and Patashnik [20]:

$$[i > j] = \begin{cases} 1, & \text{if } i > j; \\ 0, & \text{otherwise,} \end{cases} \qquad \text{and} \qquad [i = j] = \begin{cases} 1, & \text{if } i = j; \\ 0, & \text{otherwise.} \end{cases}$$

Lemma 4. *All T_i's appear in S^+ in one sequence and in increasing order.*

Proof. Assume to the contrary that subsequence $+T_1, +T_2, \ldots, +T_p$ is followed by an arc $+I$ different from $+T_{p+1}$ in S^+, where $p < n$. This arc could be either $+A_{j,i}$ or $+T_\ell$, where $\ell > p+1$. We will show that both cases leads to contradiction by lower bounding the energy difference of the intermediate configuration after adding $+T_p$.

As argued in the proof of Lemma 1, processing of each $+A_{j_m,i_m}$ will contribute A to the energy difference. Hence, before adding $+T_1$, the energy difference is non-negative. We will lower bound contributions of processing $+T_1, \ldots, +T_p$ to the energy difference. For every $i = 1, \ldots, p$, to process $+T_i$, we need to remove $-\tilde{T}_i$ and all $-\tilde{A}_{j,i}$ which were not yet removed. This will add to the energy difference

$$\begin{aligned} \tilde{T}_i^w + \sum_{j \notin T_i} \tilde{A}_{j,i}^w &\geq k - 3nA - [i > 1](6n+5)nA - 4(n-1)iA + 4i \sum_{j=1}^{3n} a_j - 4|\,\mathrm{G}_i\,|iA/2 \\ &> k - 3nA + [i > 1](6n+5)nA - 2|\,\mathrm{G}_i\,|nA\,, \end{aligned}$$

since only $|\,\mathrm{G}_i\,|$ arcs $-\tilde{A}_{j,i}$ have been removed before processing $+T_i$ and each $\tilde{A}_{j,i}^w = 4ia_j < 2nA$. Hence, the contribution of processing $+T_1$ is at least $k - 3nA - 2|\,\mathrm{G}_1\,|nA - T_1^w = (6n+5)nA - 2|\,\mathrm{G}_1\,|nA$, and the contribution of processing $+T_i$, for $i = 2, \ldots, p$, is at least $-2|\,\mathrm{G}_i\,|nA$. Since $\sum_{i=1}^{p} |\,\mathrm{G}_i\,| \leq M$ and by Lemma 1, $M \leq 3n$, the total contribution of adding $T_1, \ldots, T_p$ is at least $6n^2A + 5nA - 6n \cdot nA = 5nA$. Hence, the energy difference of the intermediate configuration before processing $+I$ is at least $5nA$.

Now, let us consider two cases depending on type of arc $+I$. First, assume that $+I$ is a $+T_\ell$, for some $\ell > p+1$. Since $+T_{\ell-1}$ appears in S^+ after $+T_\ell$, to add $+T_\ell$, we need to remove both $-\tilde{T}_{\ell-1}$ and $-\tilde{T}_\ell$. Since the energy difference before removing $-\tilde{T}_{\ell-1}$ and $-\tilde{T}_\ell$ is positive (at least $5nA$), the lemma follows by the argument used in the proof of Lemma 3.

Second, assume that $+I$ is an $+A_{j,i}$. Before adding $+A_{j,i}$, the arc $-\bar{A}_{j,i}$ needs to be removed. Since $\bar{A}_{j,i}^w = k - (j-1)A - 4ia_j > k - (3n-1)A - 2nA > k - 5nA$, the energy difference after removing $-\bar{A}_{j,i}$ would be greater than $5nA + k - 5nA = k$, a contradiction.

Hence, by the above lemmas, the subsequence S^+ has the following form

$$+A_{j_1,i_1}, +A_{j_2,i_2}, \ldots, +A_{j_M,i_M}, +T_1, +T_2, \ldots, +T_n$$

followed by the all remaining $+A_{j,i}$'s.

Moving on to the last part of the proof: we show that the G_i's defined by the initial sequence of $+A_{i,j}$'s form a valid solution (or can be perturbed slightly to form a valid solution) by arguing that only in this case can all of the T_ℓ's be added without exceeding the energy barrier. Specifically, we will show that $M = 3n$ and $\{\{a_{j_1}, \ldots, a_{j_{3n}}\}\} = \{\{a_1, \ldots, a_{3n}\}\}$. For this purpose, the next two lemmas prove lower bounds on sums of the $\mathrm{c}(\mathrm{G}_i)$'s.

Lemma 5. *For every $\ell = 1, \ldots, n$, $\sum_{i=1}^{\ell} i(\mathrm{c}(\mathrm{G}_i) - A) \geq (M - 3n)A/4$.*

Proof. To process $+T_\ell$, $-\tilde{T}_\ell$ and all remaining $-\tilde{A}_{1,\ell}, \ldots, -\tilde{A}_{3n,\ell}$ need to be removed, that is those $-\tilde{A}_{j,\ell}$'s for which $j \notin \mathrm{G}_\ell$. Hence, the total weight of arcs which need to be removed is

$$\begin{aligned} \tilde{T}_\ell^w + \textstyle\sum_{j \notin T_\ell} \tilde{A}_{j,\ell}^w &= k - 3nA - [\ell > 1](6n+5)nA - 4(n-1)\ell A + 4\ell(nA - \mathrm{c}(\mathrm{G}_\ell)) \\ &= k - 3nA - [\ell > 1](6n+5)nA + 4\ell(A - \mathrm{c}(\mathrm{G}_\ell))\,. \end{aligned}$$

After removing these arcs, the energy difference will increase by this amount and then decrease by $T_\ell^w = k - (6n+8)nA$. Hence, the total change of the energy difference for adding $+T_\ell$ is $[\ell = 1](6n+5)nA + 4\ell(A - \mathrm{c}(\mathrm{G}_\ell))$.

It is easy to see, by induction on ℓ, that the energy difference before removing arc for $+T_\ell$ is $MA + [\ell > 1](6n+5)nA + \sum_{i=1}^{\ell-1} 4i(A - \mathrm{c}(\mathrm{G}_i))$, since after processing subsequence $+A_{j_1,i_1}, \ldots, +A_{j_M,i_M}$, the energy difference is MA. Since the energy difference after removing arcs needed for adding $+T_\ell$ must be at most k, we have

$$\begin{aligned} MA + [\ell > 1](6n+5)nA + \sum_{i=1}^{\ell-1} 4i(A - \mathrm{c}(\mathrm{G}_i)) \\ + k - 3nA - [l > 1](6n+5)nA + 4\ell(A - \mathrm{c}(\mathrm{G}_\ell)) \leq k \end{aligned}$$

which simplifies to

$$\sum_{i=1}^{\ell} i(\mathrm{c}(\mathrm{G}_i) - A) \geq (M - 3n)A/4\,.$$

Using the inequalities from Lemma 5, we will lower bound the sum of $\mathrm{c}(\mathrm{G}_i)$'s.

Lemma 6. *We have $\sum_{i=1}^{n} \mathrm{c}(\mathrm{G}_i) \geq nA - (3n - M)A/4$, where the equality happens only if $\mathrm{c}(\mathrm{G}_1) = A - (3n - M)A/4$ and $\mathrm{c}(\mathrm{G}_i) = A$, for every $i = 2, \ldots, n$.*

Proof. We will multiply each inequality of Lemma 5 with the positive constant $1/\ell - [n > l]/\ell + 1$ and sum the inequalities:

$$\sum_{\ell=1}^{n} (1/\ell - [n > l]/(\ell+1)) \sum_{i=1}^{\ell} i(\mathrm{c}(\mathrm{G}_i) - A) \geq \sum_{\ell=1}^{n} (1/\ell - [n > l]/(\ell+1))\,(M - 3n)A/4\,.$$

Changing the order of the sums on the left hand side and using the fact that $\sum_{\ell=i}^{n}(1/\ell - [n > \ell]/(\ell+1)) = 1/i$ we obtain:

$$\sum_{i=1}^{n}(\mathrm{c}(\mathrm{G}_i) - A) = \sum_{i=1}^{n} i(\mathrm{c}(\mathrm{G}_i) - A) \sum_{\ell=i}^{n}(1/\ell - [n > \ell]/(\ell+1)) \geq (M - 3n)A/4,$$

and the lemma easily follows. The equality in the resulting inequality happens only if we have equality in all inequalities used in the summation. This would imply that

$$\sum_{i=1}^{\ell} i(\mathrm{c}(\mathrm{G}_i) - A) = (M - 3n)A/4, \tag{5}$$

for all $\ell = 1, \ldots, n$. For $\ell = 1$, we have $\mathrm{c}(\mathrm{G}_1) - A = (M - 3n)A/4$, i.e., $\mathrm{c}(\mathrm{G}_1) = A - (3n - M)A/4$. Subtracting Equation (5) for ℓ and Equation (5) for $\ell - 1$, we obtain $\ell(\mathrm{c}(\mathrm{G}_\ell) - A) = 0$, i.e., $\mathrm{c}(\mathrm{G}_\ell) = A$.

By Lemmas 2 and 6, we have $\sum_{i=1}^{n} \mathrm{c}(\mathrm{G}_i) = nA - (3n - M)A/4$, i.e., we have equality in both Lemma 2 and Lemma 6. Thus, by Lemma 2, we have that $M = 3n$ and $\{\{a_{j_1}, \ldots, a_{j_{3n}}\}\} = \{\{a_1, \ldots, a_{3n}\}\}$.

Although this does not imply that $\mathrm{G}_1, \ldots, \mathrm{G}_n$ forms a decomposition of set $\{1, 2, \ldots, 3n\}$, for instance, if $a_1 = a_2$, the multiset $\{\{j_1, \ldots, j_{3n}\}\}$ could contain zero 1's and two 2's, the sets $\mathrm{G}_1, \ldots, \mathrm{G}_n$ could be mapped to the decomposition of $\{1, 2, \ldots, 3n\}$ just by a sequence of replacements i's with j's assuming $a_j = a_{j+1} = \cdots = a_i$. Furthermore, by Lemma 6, we have $\mathrm{c}(\mathrm{G}_1) = A - (3n - M)A/4 = A$ and also $\mathrm{c}(\mathrm{G}_i) = A$ for all $i = 2, \ldots, n$. Hence, the sets $\mathrm{G}_1, \ldots, \mathrm{G}_n$ (possibly modified as described above) are the solution to the 3-PARTITION problem.

The reduction is polynomial as the sum of weights of all arcs (which is the total number of arcs in the unweighted instance) is

$$\sum_{i=1}^{n} \left(\tilde{T}_i^w + T_i^w + \sum_{j=1}^{3n} (\tilde{A}_{j,i}^w + \bar{A}_{j,i}^w + A_{j,i}^w) \right) < n \cdot 2k + 3n^2 \cdot 2k = \mathcal{O}(n^2 k) = \mathcal{O}(n^4 A)\,,$$

and A is assumed to be polynomial in n.

4 Conclusions

We have shown that the energy barrier problem for direct folding pathways is NP-complete, via a reduction from the 3-PARTITION problem. Thus, unless NP = P, there is no polynomial-time algorithm for calculating the energy barrier of direct folding pathways. This justifies the use of heuristics for estimating energy barriers [17, 14, 18, 16]. An interesting open question is whether there is an algorithm that is guaranteed to return the energy barrier and which works well on practical instances of the problem (while not in the worst case).

Our proof can help shed insight on energy landscapes. Consider an instance $(\mathcal{I}, \mathcal{F}, k)$ of the DPKF-EB problem which is derived from a "yes" instance of 3-PARTITION according to our construction. There are exponentially many possible prefixes (of the type shown in (4)) which could precede the addition of T_1, all of which do not exceed the energy barrier k. Of these, it may be that only one defines a valid solution of G_i's. Thus, if pathways are followed according to a random process, it could take exponential time for the random process to find the pathway with energy barrier k. This is because there are exponentially many initial prefixes which could lead to such a pathway of which only one can be extended to a pathway with barrier k.

We do not resolve the computational complexity of the general energy barrier problem, in which the pathway need not be direct. Two challenges in understanding the complexity of this problem which need to be considered are repeat arcs and temporary arcs. By repeat arcs, we mean arcs which are added or removed multiple times along the pathway. By temporary arcs, we mean arcs that are not in the initial or final structures of the pathway. It is possible that multiple additions and removals of arcs is necessary, in order to minimize the barrier, and that the resulting pathway has length which is exponential in the length of the molecule. Thus, the general energy barrier problem for pseudoknot free structures may be PSPACE-complete. It is also possible that repeat arcs do not help minimize the barrier, even when temporary arcs are allowed in the pathway, in which case the general problem would be in NP. And, it is possible that permitting repeat arcs and/or temporary arcs along the pathway makes the problem tractable, and thus in P. We hope to resolve which of these possibilities is the case in future work.

References

1. Kameda, A., Yamamoto, M., Uejima, H., Hagiya, M., Sakamoto, K., Ohuchi, A.: Hairpin-based state machine and conformational addressing: Design and experiment. Natural Computing **4** (2005) 103–126
2. Yurke, B., Turberfield, A.J., Mills, A.J.J., Simmel, F.C., Neumann, J.L.: A DNA-fuelled molecular machine made of DNA. Nature **406** (2000) 605–608
3. Seelig, G., Soloveichik, D., Zhang, D.Y., Winfree, E.: Enzyme-free nucleic acid logic circuits. Science **314** (2006) 1585–1588
4. Simmel, F.C., Dittmer, W.U.: DNA nanodevices. Small **1** (2005) 284–299
5. Uejima, H., Hagiya, M.: Secondary structure design of multi-state DNA machines based on sequential structure transitions. Lecture Notes in Computer Science,Springer Berlin / Heidelberg **2943** (2004) 74–85
6. Hagiya, M., Yaegashi, S., Takahashi, K.: Computing with hairpins and secondary structures of DNA. Nanotechnology: Science and Computation, Natural Computing Series, Springer Berlin Heidelberg (Editors J. Chen, N. Jonoska, and G. Rozenberg (2006) 293–308
7. Yin, P., Choi, H., Calvert, C., Pierce, N.: Programming biomolecular self-assembly pathways. Nature **451** (2008) 318–322
8. Uejima, H., Hagiya, M.: Analyzing secondary structure transition paths of DNA/RNA molecules. DNA Computing, Lecture Notes in Computer Science, Springer Berlin Heidelberg **2943** (2004) 86–90
9. Chen, S.J., Dill, K.A.: RNA folding energy landscapes. Proc. Nat. Acad. Sci. **97**(2) (January 2000) 646–651
10. Russell, R., Zhuang, X., Babcock, H., Millett, I., Doniach, S., Chu, S., Herschlag, D.: Exploring the folding landscape of a structured RNA. Proc. Nat. Acad. Sci. **99** (2002) 155–160
11. Shcherbakova, I., Mitra, S., Laederach, A., Brenowitz, M.: Energy barriers, pathways, and dynamics during folding of large, multidomain RNAs. Curr. Opin. Chem. Biol. (2008) 655–666
12. Treiber, D.K., Williamson, J.R.: Beyond kinetic traps in RNA folding. Curr. Opin. Struc. Biol. **11** (2001) 309–314
13. Flamm, C., Fontana, W., Hofacker, I.L., Schuster, P.: RNA folding at elementary step resolution. RNA (2000) 325–338
14. Tang, X., Thomas, S., Tapia, L., Giedroc, D.P., Amato, N.M.: Simulating RNA folding kinetics on approximated energy landscapes. J. Mol. Biol. **381** (2008) 1055–1067

15. van Batenburg, F.H.D., Gultyaev, A.P., Pleij, C.W.A., Ng, J., Oliehoek, J.: Pseudobase: a database with RNA pseudoknots. Nucl. Acids Res. **28**(1) (2000) 201–204
16. Wolfinger, M.T.: The energy landscape of RNA folding. Master's thesis, University Vienna (2001)
17. Flamm, C., Hofacker, I.L., Stadler, P.F., Wolfinger, M.T.: Barrier trees of degenerate landscapes. Zeitschrift für Physikalische Chemie **216** (2002) 155–174
18. Morgan, S.R., Higgs, P.G.: Barrier heights between ground states in a model of RNA secondary structure. J. Phys. A: Math. Gen. **31** (1998) 3153–3170
19. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., New York, NY, USA (1979)
20. Graham, R., Knuth, D., Patashnik, O.: Concrete Mathematics: a foundation for computer science. Addison-Wesley, Reading, MA, USA (1989)