On low energy barrier folding pathways for nucleic acid sequences

Leigh-Anne Mathieson and Anne Condon

U. British Columbia, Department of Computer Science, Vancouver, BC, Canada

Abstract. Secondary structure folding pathways correspond to the execution of DNA programs such as DNA strand displacement systems. It is helpful to understand the full diversity of features that such pathways can have, when designing novel folding pathways. In this work, we show that properties of folding pathways over a 2-base strand (a strand with either A and T, or C and G, but not all four bases) may be quite different than those over a 4-base alphabet. Our main result is that, for a simple energy model in which each base pair contributes -1, 2-base sequences of length n always have a folding pathway of length $O(n^3)$ with energy barrier at most 2. We provide an efficient algorithm for constructing such a pathway. In contrast, it is unknown whether minimum energy barrier pathways for 4-base sequences can be found efficiently, and such pathways can have barrier $\Theta(n)$. We also present several results that show how folding pathways with temporary and/or repeated base pairs can have lower energy barrier than pathways without such base pairs.

1 Introduction

Nucleic acid folding pathways—sequences of structures visited by DNA and RNA molecules as they fold—are interesting because they influence the shape and thus function of key agents of cellular processes [4]. Folding pathways are also very interesting to DNA nanotechnologists and molecular programmers because they are the realization of DNA programs for the creation of nano-materials, robots, logic circuits, artificial neural networks and much more [10, 13, 14, 19, 20].

Kinetics constrain nucleic acids to fold along pathways that tend to have low energy barriers. The energy barrier of a pathway, or simply the barrier, is the largest difference in free energy between any structure on the pathway and a subsequent structure. Specifically, if we are interested in folding pathways for a sequence s from an initial structure $\mathcal{I}$ to a final structure $\mathcal{F}$, where $\mathcal{I}$ has minimum free energy (MFE), then the energy barrier is the largest difference in free energy between $\mathcal{I}$ and any other structure along the pathway. We refer to a folding pathway from $\mathcal{I}$ to $\mathcal{F}$ with minimum barrier (taken over all possible folding pathways) as a min-barrier pathway. Several methods for computationally predicting nucleic acid folding pathways rely on energy barrier estimation [3, 15]. Moreover, designed nucleic acid systems such as DNA strand displacement systems ensure that the desired folding pathways have low energy barriers, while undesired alternatives have high barriers.

Thus there has been substantial work on methods for finding folding pathways between two given structures of a DNA or RNA strand s and in particular, finding min-barrier pathways (or approximations to these) [3, 6]. These methods for computational prediction of folding pathways and energy barriers use reliable RNA or DNA thermodynamic and kinetic parameters [7], and mostly focus on pseudoknot free structures.

However, it can be helpful to work with simpler energy models, e.g., when the goal is to understand the computational complexity of folding pathway or energy barrier estimation, or to gain coarse-grained qualitative information on the shape of RNA folding landscapes [1, 5, 12]. Morgan and Higgs [8] studied how the energy barriers of min-barrier pathways of pseudoknot-free structures scale with strand length, assuming a simple energy model in which each base pair contributes -1 to the free energy of a structure. Their work considered so-called *direct* folding pathways in which the only base pairs that can be added along the folding pathway from structure $\mathcal{I}$ to structure $\mathcal{F}$ are those in $\mathcal{F} - \mathcal{I}$ and the only base pairs that can be removed are those in $\mathcal{I} - \mathcal{F}$. Thachuk et al. [17] showed that the *direct energy barrier problem (Direct-EBP)*, namely to determine whether there is a direct folding pathway from $\mathcal{I}$ to $\mathcal{F}$ with barrier of at most k, is NP-complete. Because of an earlier result of Thachuk et al. [16], the NP-completeness result holds whether or not the pathway can repeatedly remove and add back base pairs of $\mathcal{I}$ or $\mathcal{F}$ along the pathway.

The computational complexity of the more general energy barrier problem (EBP) remains open even for the simple energy model, where the EBP is to determine whether there is a possibly *indirect* pseudoknot-free folding pathway from $\mathcal{I}$ to $\mathcal{F}$ with barrier at most k. A pathway is indirect if so-called temporary and/or repeated base pairs can arise along the pathway, where a *temporary base pair* is one that is not in $\mathcal{I}$ or $\mathcal{F}$ but is in some other structure of the pathway, and a *repeated base pair* is one that is in some structure on the pathway (possibly the initial structure), then is removed and later added back again.

The main result of this paper is that there is indeed an efficient algorithm for the general energy barrier problem *for sequences over a 2-base alphabet.* For concreteness we state our result for sequences over the alphabet $\{\mathsf{A}, \mathsf{U}\}$, which we call AU-sequences. Our result shows that, for the simple energy model, not only is it possible to efficiently find a min-barrier pathway of length $O(|s|^3)$ from any initial MFE structure to any final MFE structure for any AU-sequence s, but that this pathway will have barrier 2 if the number of U's equals the number of A's and will have barrier 1 if the number of U's is not equal to the number of A's. In contrast, the minimum energy barrier of a sequence over a 4-base alphabet may be proportional to the length of the sequence.

Our algorithm relies heavily on the assumption of the simple energy model, but variants of the techniques involved, which are relatively straightforward and intuitive, may be useful also for more realistic energy models. The proof of our main result builds on the fact that the minimum free energy pseudoknot free structure of any AU-sequence s has energy $-q$, where q is the lesser of the number of A's in s and the number of U's in s.

Our main result raises two further questions that we address in this paper. First, our algorithm yields indirect barrier-1 or barrier-2 pathways, specifically, pathways with temporary base pairs. Dotu et al. [3] observed that there exist strands over $\{A, C, G, U\}$ whose min-barrier pathways are necessarily indirect. We strengthen Dotu et al.'s observation for the simple energy model, by showing that min-barrier pathways may also necessarily be indirect even for AU-sequences. Specifically we show that for any k, there is a length-$6k$ AU-sequence s, and minimum energy initial and final structures for s, such that any direct pathway from initial to final structure must have barrier at least $k + 1$, while there is a barrier-1 indirect pathway.

As noted above, it is not known whether there is an efficient (polynomial-time) algorithm for the EBP, for strands over {A,C,G,U}. It's conceivable that, because of the possibility that a min-barrier pathway must contain repeated base pairs, there exist infinitely many strands for which any min-barrier, indirect pathway from a given initial to a given target structure must have length that grows exponentially with the strand length. If this is the case, the EBP problem may be complete for PSPACE, a complexity class that is believed to include problems that are even harder than those in NP. Here we present the first example of a sequence s, initial structure $\mathcal{I}$ and final structure $\mathcal{F}$ such that the min-barrier pathway of pseudoknot-free structures has the property that base pairs which are in both $\mathcal{I}$ and $\mathcal{F}$ must be removed along the pathway, and then added back in again. This result for indirect pathways stands in contrast with the result of Thachuk et al. [16] that, for direct pathways, repeated base pairs are not necessary in min-barrier folding pathways.

The rest of this paper is organized as follows. Section 2 introduces notation and a preliminary result. We present our main result, namely our efficient algorithm for finding min-barrier folding pathways for AU-sequences, in Section 3. Our examples that illustrate why indirect pathways can have lower min-barrier than direct pathways for AU-sequences, and why pathways with repeats can have lower min-barrier than pathways without repeats, are in Section 4. Most of the proofs are omitted, because of space limitations. We present conclusions and directions for further work in Section 5.

2 Notation

Here we first introduce notation to describe nucleic acid secondary structure and folding pathways, and present a useful result on the free energy of minimum free energy structures. For an RNA sequence $s = s_1, s_2, \ldots, s_n$ (i.e, string over $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$), a *base pair* is an unordered pair $\{i, j\}$ where indices i and j are in the range $[1, \ldots, n]$, $i \neq j$, and the set of bases $\{s_i, s_j\}$ is either $\{A, U\}$ or $\{C, G\}$. (DNA is similar with T instead of U). A secondary structure $\mathcal{S}$ for s is a set of base pairs of s, such that no two intersect. Secondary structure is often represented as an arc diagram such as that in Fig. 1 (a), in which each base pair is represented as an arc that connects two bases of sequence s. For this reason, we often refer to a base pair as an arc, and refer to its indices as endpoints. We only consider pseudoknot-free structures: these are structures in which no arcs cross in the arc diagram representation. Equivalently, if a structure is pseudoknot free, then for all $\{i, j\}$ and $\{i', j'\}$ in the structure with $i < j$ and $i' < j'$, it is not the case that $i < i' < j < j'$ or $i' < i < j' < j$. Given a set $\mathcal{S}$ of arcs, a *narrowest* arc of $\mathcal{S}$ is an arc $\{i, j\}$ of $\mathcal{S}$ for which $|i - j|$ is minimal.

We use a simple energy model where each bond in a structure contributes -1 to the structure's free energy, and we denote the free energy of a structure $\mathcal{P}$ by $E(\mathcal{P})$. A *folding pathway*, π, from structure $\mathcal{I}$ to structure $\mathcal{F}$ is a sequence of pseudoknot-free secondary structures $\pi = \mathcal{P}_0, \mathcal{P}_1, ..., \mathcal{P}_m$ where $\mathcal{I} = \mathcal{P}_0$ and $\mathcal{F} = \mathcal{P}_m$. Each structure in the sequence differs from the structure directly before it by the addition or removal of exactly one base pair. When the first structure $\mathcal{I}$ on a folding pathway π is a MFE structure (as is always the case in this paper), the *energy barrier* of π is $\max_{1 \leq i \leq m} E(\mathcal{P}_0) - E(\mathcal{P}_i)$. Sometimes, rather than listing a given folding pathway, we list instead its *transformation sequence*, which is the sequence of arcs that are added or removed to obtain successive structures of the folding pathway. When listing the arcs of a transformation sequence, we use the prefices "+" and "−" to indicate whether the arc is added or removed. For example, if $\mathcal{I}$ is the structure $\{a_1, a_2, a_3\}$ with three arcs, then the transformation sequence $-a_1, +a_4, -a_2, +a_1$ corresponds to the folding pathway

$$\{a_1, a_2, a_3\}, \{a_2, a_3\}, \{a_2, a_3, a_4\}, \{a_3, a_4\}, \{a_1, a_3, a_4\}.$$

A U-index of s is a number u in the range $[1, \ldots, |s|]$ such that the base at position u of sequence s is U. An A-index of s is defined similarly, with A replacing U. If p is an arc then A-index(p) and U-index(p) denote the endpoints of p that are an A-index and a U-index, respectively. We say that an index i is *covered* by an arc p if i is in the range $[\text{A-index}(p) + 1, \text{U-index}(p) - 1]$ if A-index(p) $<$ U-index(p) or the range $[\text{U-index}(p) + 1, \text{A-index}(p) - 1]$ if U-index(p) $<$ A-index(p). Similarly, we say that an arc p is covered by arc p' if both endpoints of p are covered by p'.

An arc p' *separates* an index u from arc p if $p' \neq p$ and p' either covers u or covers p but does not cover both. Arc p' separates u from a set P of arcs if $p' \notin P$ and p' either covers u or all arcs in P, but not both.

For the simple energy model, the number of base pairs that could form in a secondary structure of an AU-sequence s is bounded by the minimum of the number of A's and the number of U's. Without loss of generality, suppose that s has at least as many U's as A's and let q be the number of A's. A simple stack-based algorithm can find a pseudoknot free structure with q base pairs in linear time:

Claim 1 *Let s be an AU-sequence with at least as many U's as A's, and let q be the number of A bases. There is a pseudoknot-free secondary structure $\mathcal{S}$ for s with q base pairs, and $\mathcal{S}$ can be generated in time $O(|s|)$.*

3 Low-barrier Pathways for AU-Sequences

In this section we show how to find a folding pathway with barrier at most 2 from an initial MFE structure $\mathcal{I}$ to a final MFE structure $\mathcal{F}$ of an AU-sequence s. We consider two cases in the following two subsections: first, where the number of U's of s equals the number of A's and second, where there are more U's than A's. The case where there are more A's than U's can be handled in a manner symmetric to the case where there are more U's than A's and we do not discuss it further here.

3.1 AU-Sequences with an Equal Number of A's and U's

In the first case, a simple algorithm works to find a pathway with barrier 2, namely our FindBarrier2Pathway, Algorithm 1. This and later algorithms maintain a current structure S_{curr} which is initially set to $\mathcal{I}$; the algorithm repeatedly removes and adds arcs to S_{curr} until the structure $\mathcal{F}$ is reached, and the resulting sequence of structures forms the folding pathway. In this case, the algorithm adds the arcs of $\mathcal{F}$ to S_{curr} in narrowest-first order. Before adding arc f_{nar}, the two arcs of S_{curr} that share an endpoint with f_{nar} must first be removed; then f_{nar} and one additional arc are added in order to avoid a barrier of more than 2. At the start of each iteration, $\mathcal{F}_{frozen}$ is the set of arcs of $\mathcal{F}$ that have already been added to S_{curr} (these arcs are "frozen" in the sense that they will not be subsequently removed from S_{curr}). Claim 2 asserts that this can be done without introducing pseudoknots. We also note that if the number of U's is not equal to the number of A's, Algorithm 1 is not correct.

Claim 2 *The pathway π produced by FindBarrier2Pathway (Algorithm 1) on input s, $\mathcal{I}$, $\mathcal{F}$ is a valid barrier-2 pathway from $\mathcal{I}$ to $\mathcal{F}$ where no structure in the pathway contains pseudoknots. The pathway produced has length at most 4 times the number of arcs in an MFE structure of s.*

3.2 AU-Sequences with More U's Than A's

If sequence s has more U's than A's, there is a barrier-1 pathway from MFE structure $\mathcal{I}$ to MFE structure $\mathcal{F}$. Here we present our FindPathway algorithm, Algorithm 2, which constructs this pathway.

Starting with a current structure $\mathcal{S}_{curr}$ that is set to the initial structure $\mathcal{I}$, FindPathway repeatedly selects an arc f of $\mathcal{F}$ that is not in the current structure. For each f, it calls the ResolveConflicts algorithm, Algorithm 3, which updates the current structure via a barrier-1 pathway that removes any arcs that conflict with, i.e., form a pseudoknot with, f, while also ensuring that arcs of $\mathcal{F}$ that were added in earlier iterations—so-called *frozen* arcs—are not removed. Once ResolveConflicts is done, the FindPathways algorithm adds f to $\mathcal{S}_{curr}$ and arc f is also frozen. As we show later, the order in which the arcs of $\mathcal{F}$ are added by FindPathway ensures that ResolveConflicts can proceed within barrier 1.

Algorithm 1 Find a barrier-2 pathway for an AU-sequence with #U's = #A's

```
procedure FindBarrier2Pathway ($s, \mathcal{I}, \mathcal{F}$)
  Input:
    a sequence $s \in \{A,U\}^*$, with an equal number of U's and A's
    an initial MFE structure $\mathcal{I}$ for $s$
    a final MFE structure $\mathcal{F}$ for $s$
  Output:
    a valid pathway $\pi$ from $\mathcal{I}$ to $\mathcal{F}$ with barrier 2

  $\mathcal{S}_{curr} = \mathcal{I}$; $\pi \leftarrow$ empty pathway; $\mathcal{F}_{frozen} \leftarrow \emptyset$
  while $\mathcal{F}_{frozen} \neq \mathcal{F}$ do
    $f_{nar} \leftarrow$ a narrowest arc such that $f_{nar} \in \mathcal{F}$ but $f_{nar} \notin \mathcal{F}_{frozen}$
    if $f_{nar} \notin \mathcal{S}_{curr}$ then
      $f_a \leftarrow$ the arc of $\mathcal{S}_{curr}$ with endpoint A-index($f_{nar}$)
      $f_u \leftarrow$ the arc of $\mathcal{S}_{curr}$ with endpoint U-index($f_{nar}$)
      $p \leftarrow$ the arc with endpoint U-index($f_a$) and A-index($f_u$)
      remove $f_a$ from $\mathcal{S}_{curr}$; $\pi \leftarrow \pi, S_{curr}$
      remove $f_u$ from $\mathcal{S}_{curr}$; $\pi \leftarrow \pi, S_{curr}$
      add $p$ to $\mathcal{S}_{curr}$; $\pi \leftarrow \pi, S_{curr}$
      add $f_{nar}$ to $\mathcal{S}_{curr}$; $\pi \leftarrow \pi, S_{curr}$
    end if
    add $f_{nar}$ to $\mathcal{F}_{frozen}$
  end while
  return $\pi$
```

We next describe the ResolveConflicts algorithm, while also introducing definitions that are used in the algorithm descriptions. These definitions are with respect to the inputs to ResolveConflicts, namely a "current" pseudoknot-free secondary structure $\mathcal{S}_{curr}$ for s, a subset $\mathcal{F}_{frozen}$—the frozen arcs of $\mathcal{S}_{curr}$, and an additional arc f of $\mathcal{F}$ that is not yet in $\mathcal{S}_{curr}$. Let conflict(f) be the set of arcs of $\mathcal{S}_{curr}$ that form a pseudoknot with f, except that the arc of $\mathcal{S}_{curr}$ from the A-index endpoint of f is excluded.

ResolveConflicts repeatedly removes the arcs of conflict(f), keeping the barrier low by "repairing" the A-indices of these conflicting arcs with other available U-indices. To do this, ResolveConflicts first identifies a set $\mathcal{U}$ of currently unpaired U-indices that can *indirectly repair* conflict(f). A U-index u of s can indirectly repair conflict(f) if u is unpaired in $\mathcal{S}_{curr}$ and no arc of $\mathcal{F}_{frozen} \cup \{f\}$ separates an index of $\mathcal{U}$ from conflict(f). (If an arc p separates u from some arc of conflict(f) then p must separate u from all arcs of conflict(f).) It is the case (details omitted) that conflict(f) is indeed *repairable*, that is, there is a set $\mathcal{U}$ of $|\,\text{conflict}(f)|$ U-indices that can indirectly repair conflict(f). However, it may not be possible for ResolveConflicts to simply remove an arc p from conflict(f) and pair its A-index with a U-index of $\mathcal{U}$ without creating a pseudoknot. We say that an unpaired U-index u *can directly repair* an arc p if no arc of $\mathcal{S}_{curr} \cup \{f\} - \{p\}$ separates u from A-index(p). The inner while loop of ResolveConflicts finds a pathway that can "convert" an unpaired base u of $\mathcal{U}$ into an unpaired base that can directly repair an arc p of conflict(f). The outer loop of ResolveConflicts then removes p and adds (A-index(p), u) to $\mathcal{S}_{curr}$, thereby reducing the number of arcs that conflict with f. ResolveConflicts ends once all conflicts are removed.

Claims 3 and 4 assert that the ResolveConflicts and FindPathway algorithms are correct, leading to our main result of this section, Theorem 1.

Algorithm 2 Find a valid barrier-1 folding pathway from initial structure $\mathcal{I}$ to final structure $\mathcal{F}$, for a sequence s that has more U's than A's.

```
algorithm FindPathway($\mathcal{I}, \mathcal{F}, s$)
  Input:
    a sequence $s \in \{A,U\}^*$, with more U's than A's
    an initial MFE pseudoknot-free structure $\mathcal{I}$ for $s$
    a final MFE pseudoknot-free structure $\mathcal{F}$ for $s$
  Output:
    a valid pseudoknot-free folding pathway $\pi$ from $\mathcal{I}$ to $\mathcal{F}$ with barrier 1

  $\mathcal{S}_{curr} = \mathcal{I}$; $\pi \leftarrow$ empty pathway; $\mathcal{F}_{frozen} \leftarrow \emptyset$
  if in $\mathcal{F}$, some U-index is unpaired and not covered by any arc then
    let U-chosen be any such U-index
  else
    let U-chosen be any U-index that is unpaired in $\mathcal{F}$ and is covered
    by a narrowest arc of $\mathcal{F}$ (among those arcs covering unpaired U-indices)
  end if

  while some arc of $\mathcal{F} - \mathcal{F}_{frozen}$ does not cover U-chosen do
    let $f$ be a narrowest such arc in $\mathcal{F} - \mathcal{F}_{frozen}$
    $(\mathcal{S}', \pi') \leftarrow$ ResolveConflicts($s, \mathcal{S}_{curr}, \mathcal{F}_{frozen}, f$)
    append $\pi'$ to $\pi$; $\mathcal{S}_{curr} \leftarrow \mathcal{S}'$
    remove the arc of $\mathcal{S}_{curr}$ containing A-index($f$) as an endpoint; $\pi \leftarrow \pi, \mathcal{S}_{curr}$
    add $f$ to $\mathcal{S}_{curr}$; $\pi \leftarrow \pi, \mathcal{S}_{curr}$; $\mathcal{F}_{frozen} \leftarrow \mathcal{F}_{frozen} \cup \{f\}$
  end while// all arcs of $\mathcal{F} - \mathcal{F}_{frozen}$ cover U-chosen

  while $\mathcal{S}_{curr} \neq \mathcal{F}$ do
    let $f$ be the widest arc in $\mathcal{F} - \mathcal{F}_{frozen}$
    $(\mathcal{S}', \pi') \leftarrow$ ResolveConflicts($s, \mathcal{S}_{curr}, \mathcal{F}_{frozen}, f$)
    append $\pi'$ to $\pi$; $\mathcal{S}_{curr} \leftarrow \mathcal{S}'$
    remove the arc of $\mathcal{S}_{curr}$ containing A-index($f$) as an endpoint; $\pi \leftarrow \pi, \mathcal{S}_{curr}$
    add $f$ to $\mathcal{S}_{curr}$; $\pi \leftarrow \pi, \mathcal{S}_{curr}$; $\mathcal{F}_{frozen} \leftarrow \mathcal{F}_{frozen} \cup \{f\}$
  end while
  return $\pi$
```

Procedure 3 Find a valid barrier-1 pathway from an input MFE structure $\mathcal{S}_{curr}$ for sequence s to an updated MFE structure $\mathcal{S}_{curr}$ for s, where the updated $\mathcal{S}_{curr}$ contains all arcs in $\mathcal{F}_{frozen}$, a subset of $\mathcal{S}$, and such that conflict(f) is empty.

```
procedure ResolveConflicts ($s$,$\mathcal{S}_{curr}$,$\mathcal{F}_{frozen}$, $f$)
  Input:
   sequence $s \in \{$A,U$\}^*$, with more U's than A's, MFE structure $\mathcal{S}_{curr}$ for $s$,
   $\mathcal{F}_{frozen} \subset \mathcal{S}_{curr}$ and arc $f \notin \mathcal{F}_{frozen}$ such that conflict($f$) is repairable
  Output:
   updated MFE structure $\mathcal{S}_{curr}$ for $s$ such that $\mathcal{F}_{frozen} \subseteq \mathcal{S}_{curr}$ and conflict($f$) is empty
   a valid barrier-1 pathway $\pi'$ from the input $\mathcal{S}_{curr}$ to the output $\mathcal{S}_{curr}$

  $\pi' \leftarrow$ empty pathway
  let $\mathcal{U}$ be a set of $|$conflict($f$)$|$ U-indices that can indirectly repair conflict($f$)
  while $|$conflict($f$)$| > 0$ do
    // create a U-index that can directly repair some arc of conflict($f$)
    select some $u$ in $\mathcal{U}$ and remove $u$ from $\mathcal{U}$
    while $u$ cannot directly repair any arc of conflict($f$) do
      let $p$ be an arc that separates $u$ from conflict($f$), such that $u$ can directly repair
      $p$
      remove $p$ from $\mathcal{S}_{curr}$; $\pi \leftarrow \pi, S_{curr}$
      add $\{$A-index$(p), u\}$ to $S_{curr}$; $\pi \leftarrow \pi, S_{curr}$
      $u \leftarrow$ U-index($p$)
    end while
    choose arc $p \in$ conflict($f$) such that $u$ can directly repair $p$
    remove $p$ from $\mathcal{S}_{curr}$; $\pi \leftarrow \pi, S_{curr}$
    add $\{$A-index$(p), u\}$ to $S_{curr}$; $\pi \leftarrow \pi, S_{curr}$
  end while
  return ($\mathcal{S}_{curr}, \pi'$)
```

Claim 3 *ResolveConflicts is correct, that is, produces an output with the properties specified at the top of the algorithm description, given an input with the properties specified at the top of the algorithm description.*

Claim 4 *FindPathway is correct.*

Theorem 1. *Let $(s, \mathcal{I}, \mathcal{F})$ be an AU-instance of the EBP. A barrier-2 pathway of length $O(|s|)$ can be found in $O(|s|)$ steps for (s, I, F). Moreover, if the number of A's of s does not equal the number of U's of s, a barrier-1 pathway of length $O(|s|^3)$ can be found in $O(|s|^3)$ time.*

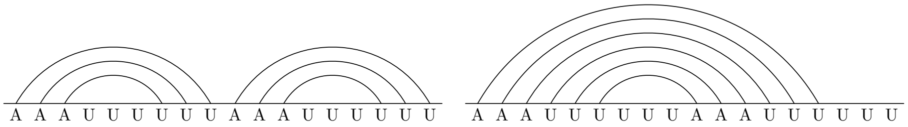
Proof. Claim 2 shows that Algorithm 1, FindBarrier2Pathway, finds a barrier-2, length $O(n)$ pathway for an AU-instance $(s, \mathcal{I}, \mathcal{F})$ of the EBP. The number of steps of the algorithm is $O(|s|)$, since there are $\mathcal{F} \leq |s|$ iterations of the whle loop, each taking $O(1)$ steps.

Claim 4 shows that FindPathway, Algorithm 2 finds a barrier-1 pathway when the AU-instance is such that the number of U's is greater than the number of A's (and by swapping U's and A's in the algorithm works when the number of A's is greater than the number of U's). To bound the length of the pathway, we first need to bound the number of steps in ResolveConflicts, Algorithm 3 (which is called by FindPathway). Each iteration of the inner while loop of ResolveConflicts reduces the number of arcs that separate u from conflict(f) by 1, and thus the number of iterations is $O(|s|)$. Each iteration has $O(1)$ steps and thus the total number of steps per iteration of the inner while loop, and the length of the pathway segment generated, is $O(|s|)$. Each iteration of the outer while loop reduces the size of conflict(f) by 1, using $O(1)$ steps beyond those of the inner while loop. Therefore, the total number of steps of ResolveConflicts is $O(|s|^2)$, and the total length of the pathway segment generated is also $O(|s|^2)$. For each arc of $\mathcal{F}$ that is added to $\mathcal{F}_{frozen}$, FindPathway calls ResolveConflicts once, and takes $O(1)$ additional steps. Thus, the overall length of the pathway generated by FindPathway is $O(|s|^3)$, and this also bounds the total number of steps taken by the algorithm (including calls to ResolveConflicts).

4 On Min-Barrier Pathways That Are Necessarily Indirect Pathways or Contain Repeat Base Pairs

Theorem 2. *For any k, there is a length-$6k$ AU-sequence with minimum energy initial and final structures such that any direct pathway from initial to final structure must have barrier at least $k+1$, while there is a barrier-1 indirect pathway.*

Proof. The length-$6k$ AU-sequence is $\mathsf{A}^k\mathsf{U}^k\mathsf{U}^k\mathsf{A}^k\mathsf{U}^k\mathsf{U}^k$, where here X^k is the letter X repeated k times. The initial and final structures are $\mathcal{I} = (^k.^k)^k(^k.^k)^k$ and $\mathcal{F} = (^k(^k.^k)^k)^k.^k$ respectively. That is, $\mathcal{I}$ has two disjoint hairpin-forming stems that we refer to as the left and right stems, while $\mathcal{F}$ has one stem nested in another; we refer to these as the inner and outer stems. Note also that the set of A-indices of $\mathcal{I}$'s left stem equals the set of A-indices of $\mathcal{F}$'s outer stem, and the set of A-indices of $\mathcal{I}$'s right stem equals the set of A-indices of $\mathcal{F}$'s inner stem. Figure 1 illustrates the sequence and initial and final structures for $k=3$.


A A A U U U U U U A A A U U U U U U
A A A U U U U U U A A A U U U U U U

(a) Initial structure $\mathcal{I}$

(b) Final structure $\mathcal{F}$

Fig. 1. Illustration of the construction of Theorem 2 for $k=3$.

We first show that any direct pathway must have barrier at least $k+1$. Let $P = p_1, p_2, \ldots, p_{|P|}$ be a direct pathway from $\mathcal{I}$ to $\mathcal{F}$. Let a be the first arc of

$\mathcal{F}$ that appears in a structure of pathway P, say structure p_i. By definition of a direct pathway, the only arcs that can be in p_{i-1} are either arcs from $\mathcal{I}$ or $\mathcal{F}$. However, since a is the first arc of $\mathcal{F}$ to appear in a structure of P, with a appearing first in p_i, p_{i-1} contains no arc of $\mathcal{F}$. If a is in the outer stem of $\mathcal{F}$, then p_{i-1} also contains none of the k arcs from the right stem of $\mathcal{I}$; otherwise such arcs would cause a pseudoknot with a in p_i. Furthermore, at least one arc from the left stem of $\mathcal{I}$, namely the arc that shares an endpoint with a, is not in p_{i-1}. Therefore at most $k-1$ arcs of P are in p_{i-1}; since $\mathcal{I}$ and $\mathcal{F}$ have $2k$ arcs, p_{i-1} causes the barrier of the path to be $k+1$. A similar argument shows that if a is in the inner stem of $\mathcal{F}$, then p_{i-1} also contains at most $k-1$ arcs and thus the barrier is $k+1$.

Next we show that there is an indirect, barrier-1 pathway from $\mathcal{I}$ to $\mathcal{F}$. The pathway has several stages. First, the right stem of $\mathcal{I}$ is replaced by a narrower stem to obtain the structure $(^k.^k)^k(^k)^k.^k$. This can be done via a barrier-1 pathway in which the arcs of $\mathcal{I}$'s right stem are replaced by narrower arcs, in narrowest-first order. Then, the left stem of $\mathcal{I}$ can be replaced by a stem that spans from the leftmost A's to the rightmost U's of the sequence, via a barrier-1 pathway, thereby reaching current structure $(^k.^k.^k(^k)^k)^k$ Then replace the inner stem of the current structure with the inner stem of $\mathcal{F}$. Finally, replace the wide stem of the current structure with the outer stem of $\mathcal{F}$.

Theorem 3. *There exists an AU-sequence s, with corresponding initial structure $\mathcal{I}$ and final structure $\mathcal{F}$ where there is an indirect pathway with repeats with a lower energy barrier than the energy barrier than that of any direct pathway.*

Proof. Consider the sequence and structures $\mathcal{I}$ and $\mathcal{F}$ of Fig. 2.

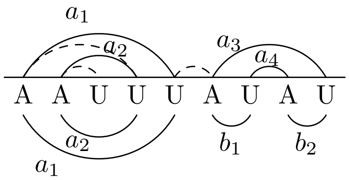


Fig. 2. An initial structure $\mathcal{I} = \{a_1, a_2, a_3, a_4\}$ (top) and a final structure $\mathcal{F} = \{a_1, a_2, b_1, b_2\}$ (bottom) for sequence AAUUUAUAU, such that there is no barrier-1 pathway without repeats from $\mathcal{I}$ to $\mathcal{F}$. Additional dashed arcs are required for a barrier-1 pathway.

We first consider possible barrier-1 pathways without repeats from structure $\mathcal{I}$. Note that since a_1 and a_2 are in $\mathcal{F}$ that in any pathway without repeats they cannot be removed as re-adding either of them would cause a repeat. So we move onto adding b_1 and b_2 without introducing a repeat, and to add either requires first removing both a_3 and a_4, which means that any pathway that does not allow repeats must be barrier-2.

So, we are left to demonstrate that there exists a barrier-1 pathway from $\mathcal{I}$ that contains repeats. We will need to add the dashed arcs in Fig. 2, so of the two nested dashed arcs, let's denote the narrower one by t_1 and the wider one by t_2, and the remaining dashed arc shall be t_3.

The following transformation sequence is barrier-1, and requires a_1 and a_2 to repeat; as an arc is added immediately after every arc that is removed, we have a barrier-1 pathway.

$$\mathcal{T} = -a_2, +t_1, -a_1, +t_2, -a_3, +t_3, -a_4, +b_2, -t_3, +b_1, -t_2, +a_1, -t_1, +a_2$$

5 Conclusions and Future Work

In this paper, for sequences over two bases, we show how to efficiently find min-barrier, pseudoknot-free pathways from initial to final MFE structures, for an energy model that assigns "-1" to each base pair (Theorem 1). In contrast, the computational complexity of finding such min-barrier pathways for sequences over four bases is unknown, and the problem may well be computationally intractable. We also show that min-barrier pathways for sequences over two bases may necessarily be indirect, i.e., involve base pairs that are neither in the initial nor final structures, and that direct pathways for such sequences may have a minimum energy barrier that is proportional to the length of the sequence (Theorem 2). Thirdly, we show that a weak form of arc repetition may be necessary in a min-barrier pathway (Theorem 3).

There are several ways in which our results could be improved. Our algorithm yields a $O(n^3)$ bound on the length of a barrier-1 pathway between two MFE structures of a length-n AU-sequence. We expect that this can be reduced, by carefully choosing the order in which u's are chosen from $\mathcal{U}$ in the while loop of the ResolveConflicts algorithm, the order in which conflicts are repaired, and perhaps also the order in which arcs are added to $\mathcal{F}_{frozen}$. Can the pathway length be reduced to $O(n)$? Another question is whether the problem of finding min-barrier, direct, pseudoknot-free pathways has an efficient algorithm (recall that for 4-base sequences, the problem is NP-hard [17]).

A significant limitation of our results is that the simple energy model ignores critical aspects of real RNA thermodynamics, such as base stacking energies, the energy costs of helix formation and loops, and the fact that hairpin loops have at least three unpaired nucleotides between their innermost paired bases. Another concern is that the model ignores pseudoknots, particularly given that pseudoknots may occur in intermediate structures along a folding pathway to a native structure, even if there is no pseudoknot in the native structure [18]. A first step forward in improving the model would be to have an energy of "-1" per stacked pair. It may be feasible to provide proofs as to whether, for this model, the energy barrier for sequences over two bases is bounded. NP-hardness of the energy barrier problem for the stacked pair model, for either two-base or four-base sequences, would suggest that molecular programs could perhaps be encoded within a DNA or RNA strand; the program could be executed via the strand's folding pathway, with the number of steps being exponential in the strand length. Alternatively, an efficient algorithm might indicate limits to the potential for long computations with a single nucleic acid strand, but could be useful in practice for finding folding pathways.

Given that it will likely be difficult to prove rigorous results for more realistic energy models, empirical computational studies could be very useful in elucidating whether the contrasting properties of two-base and four-base folding pathways described in this paper reflect the properties of two-base versus four-base sequences with respect to realistic energy models. The following questions could fruitfully be investigated empirically. Are there significant differences in min-energy barriers of pathways between low-energy structures of random versus biological sequences? Of two-base and four-base sequences? In particular, is the the min-energy barrier of any two-base sequence bounded by a constant that is independent of the sequence length? Are two-base sequences more likely to quickly fold to their MFE structures, compared with four-base sequences? Or alternatively, is it possible to design a two-base sequence with a kinetic trap that causes the sequence to fold slowly to its MFE state? Insights on questions such as these may be relevant to a hypothesis that in the early history of life, a precursor to RNA contained only two nucleotides [2, 9]. Are there examples of biological molecules that follow indirect folding pathways, or which repeatedly add and remove base pairs or stems (rather than following shorter, possibly higher-barrier pathways)? We plan to study these questions using available software tools for folding pathway and energy barrier prediction.

6 Acknowledgments

We thank the reviewers of the paper for their detailed and thoughtful comments, for raising their significant concerns about the value of our results in light of the underlying simplistic energy model, and for pointing us to the work of Reader and Joyce [11]. Their comments on follow-on work are reflected in Section 5 and, while beyond the scope of what we could address in our revisions, will guide us in our future work.

Bibliography

[1] P. Clote. An efficient algorithm to compute the landscape of locally optimal RNA seconary structures with respect to the nussinov-jacobson enery model. *J. Comput. Biol.*, 12:83–101, 2005.

[2] F.H.C. Crick. The origin of the genetic code. *J. Mol. Biol.*, 38:367–379, 1968.

[3] I. Dotu, W.A. Lorenz, P. Van Hentenryck, and P. Clote. Computing folding pathways between RNA secondary structures. *Nucleic Acids Research*, 38(5):1711–1722, 2010.

[4] C. Flamm, W. Fontana, I.L. Hofacker, and P. Schuster. RNA folding at elementary step resolution. *RNA*, pages 325–338, 2000.

[5] C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger. Barrier trees of degenerate landscapes. *Zeitschrift für Physikalische Chemie*, 216:155–174, 2002.

[6] M. Geis, C. Flamm, M.T. Wolfinger, A. Tanzer, I.L. Hofacker, M. Middendorf, C. Mandl, P.F. Stadler, and C. Thurner. Folding kinetics of large RNAs. *J. Mol. Biol.*, 379(160), 2008.

[7] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911940, 1999.

[8] S. R. Morgan and P. G. Higgs. Barrier heights between ground states in a model of RNA secondary st ructure. *J. Phys. A: Math. Gen.*, 31:3153–3170, 1998.

[9] L.E. Orgel. Evolution of the genetic apparatus. *J. Mol. Biol.*, 38:381–393, 1968.

[10] L. Qian, E. Winfree, and J. Bruck. Neural network computation with dna strand displacement cascades. *Nature*, 475:368372, 2011.

[11] J.S. Reader and G.F. Joyce. A ribozyme composed of only two different nucleotides. *Nature*, 420:841844, 2002.

[12] P. Schuster, W. Fontana, P. Stadler, and I.L. Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. In *Proceedings-Royal Society of London, Biological sciences*, pages 279–284, 1994.

[13] G. Seelig, D. Soloveichik, D.Y. Zhang, and E. Winfree. Enzyme-free nucleic acid logic circuits. *Science*, 314:1585–1588, 2006.

[14] F. C. Simmel and W. U. Dittmer. DNA nanodevices. *Small*, 1:284–299, 2005.

[15] X. Tang, S. Thomas, L. Tapia, D. P. Giedroc, and N. M. Amato. Simulating RNA folding kinetics on approximated energy landscapes. *J. Mol. Biol.*, 381:1055–1067, 2008.

[16] C. Thachuk, J. Manuch, A. Rafiey, L.A. Mathieson, L. Stacho, and A. Condon. An algorithm for the energy barrier problem without pseudoknots and temporary arcs. In *Proc. Pacific Symposium on Biocomputing*, 2010.

[17] C. Thachuk, J. Manuch, L. Stacho, and A. Condon. NP-completeness of the direct energy barrier height problem without pseudoknots. *Natural Computing*, 10(1):391–405, 2011.

[18] N.J.P. Wiebe and I.M. Meyer. Transat a method for detecting the conserved helices of functional rna structures, including transient, pseudo-knotted and alternative structures. *PLoS Computional Biology*, 6(6), 2010.

[19] P. Yin, H.M.T. Choi, C.R. Calvert, and N.A. Pierce. Programming biomolecular self-assembly pathways. *Nature*, 451:318–322, 2008.

[20] B. Yurke, A.J. Turberfield, A.J. Jr. Mills, F.C. Simmel, and J.L. Neumann. A DNA-fuelled molecular machine made of DNA. *Nature*, 406:605–608, 2000.