Pseudoknots in RNA Secondary Structures

Rune B. Lyngsø* Christian N. S. Pedersen*

Abstract

RNA molecules are sequences of nucleotides that serve as more than mere intermediaries between DNA and proteins, e.g. as catalytic molecules. Computational prediction of RNA secondary structure is among the few structure prediction problems that can be solved satisfactory in polynomial time. Most work has been done to predict structures that do not contain pseudoknots. Allowing pseudoknots introduce modelling and computational problems. In this paper we consider the problem of predicting RNA secondary structure when certain types of pseudoknots are allowed. We first present an algorithm that in time $O(n^5)$ and space $O(n^3)$ predicts the secondary structure of an RNA sequence of length n in a model that allows certain kinds of pseudoknots. We then prove that the general problem of predicting RNA secondary structure containing pseudoknots is **NP**-complete for a large class of reasonable models of pseudoknots.

1 Introduction

An RNA molecule is a sequence of nucleotides that often is just an intermediary between DNA and proteins. Some RNA molecules do however have vital importance, e.g. in translation of mRNA to proteins. The three dimensional structure of an RNA molecule is to a large extent determined by interactions between pairs of nucleotides, called base pairings. The secondary structure of an RNA molecule is the set of base pairings in the three dimensional structure of the molecule. The secondary structure can thus be used in its own right to look for information, e.g. active sites, or as a stepping stone towards prediction of higher structural levels.

If the three dimensional, or tertiary, structure of an RNA molecule is available it is of course easy to determine the secondary structure. But determining the tertiary structure is a complicated and time consuming task. When the tertiary structure of an RNA molecule is not available, the authoritative way of determining the secondary structure of an RNA molecule is by comparative modelling. Given a number of related RNA sequences the common secondary structure is inferred by identifying compensatory mutations, that is, by identifying pairs of positions where mutations of the base in one of the positions is accompanied by a mutation of the base in the other position to retain their base pairing capability. The drawback of this technique is that it requires several related RNA sequences to be available. Moreover, since expert intervention is often necessary to identify the compensatory mutations, it is difficult to fully automate comparative modelling.

Thus computational methods for predicting the secondary structure of an RNA sequence are in demand. To construct such methods it is necessary to model the biological reality that governs structure formation. Inspired by the laws of thermodynamics this is often done in terms of energy minimisation. Using a model that describes how to assign free energies to legal secondary structures, the secondary structure of an RNA sequence is predicted as the structure of least free energy. The biological relevance of the predicted structure and the computational resources such as time and space that are needed to compute it, depend entirely on the choice of legal structures and free energies. Most work has been devoted to construct algorithms for RNA secondary structure prediction when the legal structures are limited to secondary structures that do not contain pseudoknots, that is, do not contain overlapping base pairs. Nussinov *et al.* in [7] present an algorithm using a simple free energy function that is minimised when the secondary structure contains the maximum number of complementary base pairs. The algorithm takes time

*Basic Research In Computer Science (BRICS), Center of the Danish National Research Foundation, Department of Computer Science, University of Aarhus, Ny Munkegade, 8000 Århus C, Denmark. E-mail: {rlyngsoe,cstorm}@brics.dk.

$O(n^3)$ for predicting the secondary structure of an RNA sequence of length n. A more complex model for the free energy of secondary structures is proposed by Tinoco *et al.* in [15]. This model states that the free energy of a secondary structure is the sum of independent energies for each loop in the structure. Based on this model of free energy, Zuker and Stiegler in [19], and Nussinov and Jacobsen in [6], present algorithms that take time $O(n^3)$ for predicting the secondary structure of an RNA sequence of length n. Since the ideas of these algorithms form the basis of the widely used `mfold` server for RNA secondary structure prediction, they are commonly referred to as `mfold` algorithms, or algorithms of the `mfold` type.

The reason that legal structures are often required not to contain pseudoknots is not that pseudoknots do not occur in real world structures, but rather because of modelling and computational considerations. It is still an open question how to construct a reasonable model of free energy for structures containing pseudoknots that also makes it possible to construct efficient structure prediction algorithms. Rivas and Eddy in [10] present an algorithm that in time $O(n^6)$ and space $O(n^4)$ predicts the secondary structure of an RNA sequence of length n in a model that allows certain kinds of pseudoknots. In this paper we study the problem of predicting RNA secondary structure containing pseudoknots further. In section 2 we briefly review the ideas of the `mfold` algorithms. Extending on these ideas, we in section 3 present an algorithm for predicting RNA secondary structure when certain types of pseudoknots are allowed. We compare the presented algorithm with the algorithm presented by Rivas and Eddy in [10]. In section 4 we show that predicting RNA secondary structures containing pseudoknots of arbitrary types is **NP**-complete for a large class of reasonable free energy functions. Finally, in section 5 we discuss the implications of the **NP**-completeness result.

2 Terminology

For an RNA sequence s, $|s| = n$, a secondary structure is a set S of base pairs $i \cdot j$ with $1 \leq i < j \leq n$, such that $\forall i \cdot j, i' \cdot j' \in S : i = i' \Leftrightarrow j = j'$. Each base can thus take part in at most one base pair. The base pairs of a secondary structure describe the base pairing interactions formed by hydrogen bonding in a corresponding tertiary structure. It is usually assumed that RNA secondary structures do not contain pseudoknots. Two base pairs form a pseudoknot if they are overlapping, i.e. two base pairs $i \cdot j, i' \cdot j' \in S$ form a pseudoknot if $i < i' < j < j'$. The term pseudoknot is also used as a shorthand for other overlapping structures than base pairs, e.g. two helices of stacking base pairs, when the base pairs of these structures form pseudoknots.

There are of course good reasons for introducing this restriction, prominent among which is a simplification of legal structures. The simplification of not allowing pseudoknots ensures that two base pairs $i \cdot j, i' \cdot j' \in S$ are either nested, i.e. $i < i' < j' < j$, or disjoint, i.e. $i < j < i' < j'$. In many situations this allows us to first handle one base pair and then the other (if they are nested), or handle them independently (if they are disjoint). The pseudoknot restriction is thus crucial in algorithms for e.g. structure prediction [1,3,6,11,19], partition function calculations [5], comparing secondary structures [18], and simultaneous alignment and structure prediction of RNA sequences [2,12]. In the following we will exemplify this by giving a brief summary of an algorithm of the `mfold` type for secondary structure prediction. The summary is also aimed at introducing the terminology we will use in section 3. A more detailed summary can be found in e.g. Turner *et al.* [16].

An `mfold` algorithm predicts secondary structures by computing minimum (or close to minimum) energy structures in the model proposed by Tinoco *et al.* [14] extended with simplifying assumptions about the nature of the energy function for multibranched loops. Three arrays, $V(i,j)$ holding the minimum energy of a secondary structure on $s[i \mathinner{..} j]$ with bases i and j forming a base pair, $WM(i,j)$ holding the minimum energy of a structure on $s[i \mathinner{..} j]$ that is part of a multibranched loop, and $W(i)$ holding the minimum energy of a structure on $s[1 \mathinner{..} i]$, are computed based on the recursions

$$V(i,j) = \min \Big\{ eH(i,j), \; eS(i,j,i+1,j-1) + V(i+1,j-1), \; \min_{\substack{i<i'<j'<j \\ i'-i+j-j'>2}} \{eL(i,j,i',j') + V(i',j')\}, \; \min_{i+1<k<j} \{WM(i+1,k-1) + WM(k,j-1) + a\} \Big\}, \tag{1}$$

$$WM(i,j) = \min \Big\{ V(i,j) + b, \; WM(i,j-1) + c, \; WM(i+1,j) + c, \; \min_{i<k\leq j} \{WM(i,k-1) + WM(k,j)\} \Big\}, \tag{2}$$

$$W(i) = \min \Big\{ W(i-1), \; \min_{0\leq k<i} \{W(k) + V(k+1,i)\} \Big\}. \tag{3}$$

These recursions employ energy functions for hairpin loops (eH), stacking base pairs (eS), internal loops and bulges (eL), and multibranched loops ($eM(k,k') = a + bk' + ck$, where k' is the number of unpaired bases and k the number of helices in the multibranched loop). With the currently used parameters for the energy functions

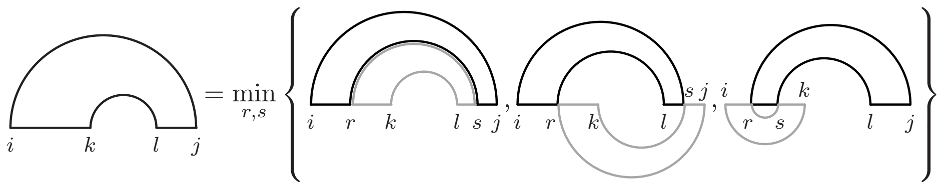

Figure 1: General recursion scheme for the Rivas and Eddy RNA secondary structure prediction algorithm.

these recursions allow for an $O(|s|^3)$ time algorithm, cf. [4, 16], for computing secondary structures of minimum energy for an RNA sequence s.

3 Algorithmic Results

The Tinoco model, cf. [14] describes how to assign energies to secondary structures not containing pseudoknots, but does not address how to handle secondary structures containing pseudoknots. To develop algorithms for predicting secondary structures containing pseudoknots, an important step is to decide on a model, i.e. to give a description of the types of legal secondary structures, and how to assign energies to these structures. As developing an algorithm and deciding on a model are closely connected processes, the description of the model is often only in part given explicitly. Often the types of legal secondary structures are only defined implicitly through the constructed algorithm.

An example of this is the pseudoknot model used by Rivas and Eddy in [10]. This is, to our knowledge, the only rigorous, energy based, polynomial time algorithm for RNA secondary structure prediction including a class of pseudoknots. In figure 1 we briefly sketch the idea of the Rivas and Eddy algorithm. Arrays holding energies of optimal structures for the subsequence from i through j are maintained similar to equations 1 to 3, but with the further restriction that the bases from k through l are yet unpaired (to allow for future pseudoknot interactions). The general recursion scheme for an entry in one of these matrices is to minimise over all possible ways of splitting the subsequence with an unpaired region into two new subsequences with unpaired regions. This defines the legal structures of the model. The energy parameters, cf. [10, table 3], used were partly fine tuned by hand and partly obtained by multiplying similar parameters for unknotted structures by a weighting parameter.

The requirements of time $O(|s|^6)$ and space $O(|s|^4)$ for this algorithm are observations that follow directly from figure 1. Though polynomial, these time and space requirements are rather steep and in [10] an estimate of 130 – 140 bases is mentioned as a rough upper bound for the size of sequences for which the algorithm is feasible. Though computational power is ever increasing, applying Moore's law (stating that computational power doubles every 18 months) still only allows sequences of 300 bases ten years from now and of 650 bases in twenty years. Nevertheless, the experiments based on this algorithm reported in [10] show the feasibility of energy-based predictions of RNA secondary structures with pseudoknots.

To obtain a faster algorithm, we propose a more restricted model for legal secondary structures. The legal secondary structures of our model are structures where we can split the sequence into four parts (one of which might extend across the ends of the sequence) as illustrated in figure 2. The splitting into four parts divides the sequence into two pairs of opposing subsequences, illustrated in figure 2 as pairs of black and grey parts of the sequence. Each pair of opposing subsequences are allowed to form an unknotted secondary structure and the pseudoknotted secondary structure arises when these two secondary structures are combined.

To further explain the types of secondary structures allowed in this model, consider a pseudoknot of type H as illustrated in figure 3. A pseudoknot of type H consists of two overlapping helices, each closing a hairpin loop, such that some of the bases in the hairpin closed by one of the helices are part of the other he-

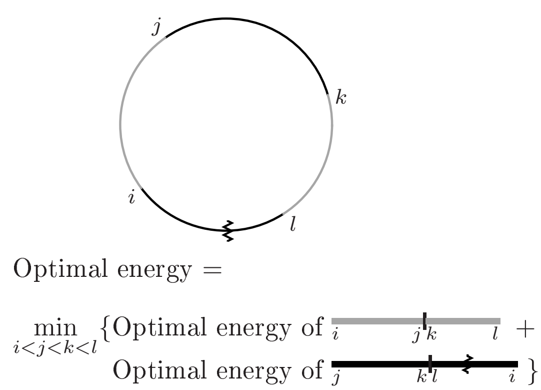

Figure 2: A model for a class of pseudoknots. The sequence has been drawn as a circle to highlight that one of the four parts of the sequence might extend across the sequence ends, here shown with a zigzagged line.

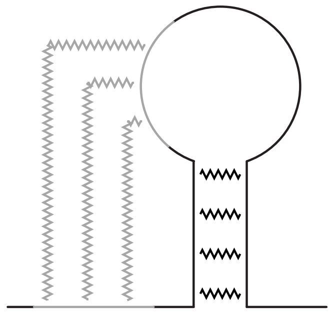

Figure 3: A pseudoknot of type H (cf. [9, figure 1]). Zigzagged lines indicate base pairings.

Algorithm 1 An algorithm for predicting RNA secondary structures containing pseudoknots based on the model illustrated in figure 2.

/ $V_{j,k}(i,l)$ denotes the energy of the optimal structure for $s[i..j]$ concatenated with $s[k..l]$. */*
$E = \infty$
for $\underline{k} = 1$ **to** $|s|$ **do** */* Fix one of the endpoints of the excluded region */*
  Allocate memory for storing and calculating $V_{j,\underline{k}}(i,l)$ and $V_{\underline{k}-1,l}(j,i)$ for $i < j < \underline{k} < l$
  / Compute tables with k (or $k-1$) as right (or left) endpoint of excluded region. */*
  for $j = 1$ **to** $\underline{k} - 1$ **do**
    Compute table $V_{j,\underline{k}}$
  end for
  for $l = \underline{k}$ **to** $|s|$ **do**
    Compute table $V_{\underline{k}-1,l}$
  end for
  / Combine tables. */*
  for $1 \leq i < j < \underline{k} < l \leq |s|$ **do**
    $E = \min\{E, V_{j,\underline{k}}(i,l) + V_{\underline{k}-1,l+1}(j+1, i-1)\}$
  end for
  Free allocated memory
end for

lix. As indicated in figure 3, we can split a pseudoknot of type H into four parts such that only bases in non-neighbouring, or opposing, parts form base pairs. The model in figure 2 can be seen as a generalisation of pseudoknots of type H where

- the overlapping structures can be arbitrary, complex secondary structures not containing pseudoknots.
- the loop regions closed by the overlapping structures do not need to be hairpin loops. They can be part of any type of loop as long as they are consecutive stretches of bases.

The model in figure 2 thus encompasses secondary structures with one pseudoknot of type H (or of type B or type I, cf. [9, figure 3]) among others.

As just explained, our model allows only one (albeit very complex) pseudoknot, so in that respect our model is a step backward compared to the model used by Rivas and Eddy. But if we can develop more efficient algorithms for secondary structure prediction in this model, it finds its justification in cases where using the Rivas and Eddy algorithm is infeasible and we only expect, or are content, to find only one pseudoknot interaction. In the rest of this section we will focus on developing an efficient algorithm for secondary structure prediction in our model.

A straightforward algorithm to solve this problem would be to run through all the $O(|s|^4)$ choices of splits and compute the energy of the optimal structures of the two pairs of subsequences. This would require time $O(|s|^7)$ and space $O(|s|^2)$. One can however observe, that when we compute the energy of the optimal structure of the subsequence from base i to base l with the subsequence from base j to base k removed, we also compute the energy of the optimal structure of the subsequence from base i' to base l' with the subsequence from base j to base k removed for all $i \leq i' \leq j$ and $k \leq l' \leq l$. Hence, by using these intermediate results from the dynamic programming algorithm, we can reduce the time requirement to $O(|s|^5)$ by just running through all the $O(|s|^2)$ choices of the removed subsequence. Unfortunately, we then have to store some intermediate results until other results become available. This increases the space requirement to $O(|s|^4)$. However, a more thorough investigation shows that the intermediate results computed with $k-1$ as the right endpoint of the removed subsequence are only combined with intermediate results computed with k as the left endpoint of the removed subsequence. This allows us to split the computation into n independent phases, each requiring only space $O(|s|^3)$, thus reducing the overall space requirement to $O(|s|^3)$ while maintaining the $O(|s|^5)$ time requirement.

The formal specification of the sketched algorithm for predicting RNA secondary structures containing pseudoknots is given in algorithm 1. The specification is rather abstract. It is more an algorithm schema than a ready-to-implement algorithm. More specifically, an implementation would require several different arrays, storing energies under various assumptions of base pairings of flanking bases. In algorithm 1 we only show have to maintain one type of array (V). But the same technique can be used for maintaining several types of interdependent arrays used in an actual implementation of the algorithm.

The $O(|s|^5)$ running time of algorithm 1 should make it feasible for longer RNA sequences than the Rivas and Eddy algorithm. For example, if we assume that the constants hidden by the O notation are similar for the two algorithms, the 130 – 140 bases upper bound for the Rivas and Eddy algorithm implies an upper bound of 350 – 375 bases for our algorithm. This increase might justify the restricted model of allowing only one pseudoknot. If this restriction is to severe, we could extend our model by allowing the sequence to be split into segments for each of which the optimal secondary structure is calculated using the model of figure 2. Such an extended model is more comparable to the model used by Rivas and Eddy in terms of legal structures (though still more restricted). It is also comparable to the model used by Rivas and Eddy in allowing secondary structure prediction in time $\mathrm{O}(|s|^6)$. The space requirement can still be limited to $\mathrm{O}(|s|^3)$ though.

We could keep playing this game of modifying models and algorithms to obtain the best possible combination of a fast algorithm and broad class of legal secondary structures. But for any class of secondary structures with pseudoknots we should probably not expect to do better than the requirements of time $\mathrm{O}(|s|^3)$ and space $\mathrm{O}(|s|^2)$ of the classic `mfold` algorithm. Furthermore, in the following section we provide evidence that we should not set hopes to high for developing efficient algorithms handling secondary structures with general types of pseudoknots.

4 Complexity Results

In this section we prove that RNA secondary structure prediction with pseudoknots is **NP**-complete in a simple nearest neighbour model, cf. definition 1. This model might seem too simple, and probably would be if we wanted to base a secondary structure prediction algorithm on it. But when proving complexity results, we want to do so in a model that is as simple as possible. If the problem in the simple model is **NP**-complete, it will remain so in any more complex and realistic model if fixing some of the parameters in the complex model turns it into the simple model.

Definition 1 (The Nearest Neighbour Pseudoknot Model) *Let $\mathcal{S}$ be a secondary structure on a sequence $s \in \{A, C, G, U\}^*$, with $|s| = n$, that is, $\mathcal{S}$ is a set of base pairs $i \cdot j$ where $1 \le i < j \le n$ and $\forall i \cdot j, i' \cdot j' \in \mathcal{S} : i = i' \Leftrightarrow j = j'$. The energy of $\mathcal{S}$ is an independent sum of energies of each of the base pairs in $\mathcal{S}$,*

$$E(\mathcal{S}) = \sum_{i \cdot j \in \mathcal{S}} E(i \cdot j, i+1, j-1),$$

where the energy of a base pair $i \cdot j$ only depends on

- *the base pair itself, that is, the types of bases forming the pair.*
- *the two neighbouring bases $i+1$ and $j-1$, that is, the types of these two bases. Furthermore, if $i+1 \cdot j' \in \mathcal{S}$ (or $i' \cdot j-1 \in \mathcal{S}$) the energy can depend on j' (or on i').*

Note that the Nearest Neighbour Pseudoknot Model allows arbitrarily complex pseudoknots as there is no restriction that base pairs are not allowed to overlap. The energy of a base pair in the Nearest Neighbour Pseudoknot Model is allowed to depend on non-neighbouring bases, but only through a base pairing with a neighbouring base. If we compare this to the Tinoco model, cf. [14], the Tinoco model allows the energy of a base pair to depend, not only on the neighbouring bases and the base pairs they might participate in, but on all bases and base pairs in the loop it closes. If we consider the model assumed by the `mfold` server, this is more restricted than the Tinoco model. Still it allows the energy of a base pair to depend on the type of loop it closes, the size of the loop, and coaxial stacking of base pairs involving neighbouring bases. The Nearest Neighbour Pseudoknot Model can be seen as a further restriction of this where we only allow the energy of a base pair to depend on stacking effects with unpaired neighbouring bases and base pairs involving neighbouring bases. The value of these stacking effects can however depend on whether the involved base pairs form a helix, an ordinary loop (a bulge or multibranched loop), or a pseudoknot.

Thus, if we compare the Nearest Neighbour Pseudoknot Model to the energy model used by Rivas and Eddy, cf. [10], it should be of little surprise that the Nearest Neighbour Pseudoknot Model is a restriction of the model used by Rivas and Eddy. The Nearest Neighbour Pseudoknot Model can be obtained from the energy model used by Rivas and Eddy by fixing some of the parameters. Thus an **NP**-hardness result for secondary structure prediction in the Nearest Neighbour Pseudoknot Model immediately implies that secondary structure prediction in the energy model used by Rivas and Eddy is **NP**-hard.

Proposition 1 *The problem of determining whether the optimal secondary structure in the Nearest Neighbour Pseudoknot Model has energy lower than some energy value E is* **NP***-complete.*

As the problem trivially is in **NP** (guess the optimal secondary structure and verify in polynomial time that it has an energy value lower than E), all we need to do is to prove that it is **NP**-hard. We will do this by a reduction to the special case of 3SAT where each literal occurs at most two times, cf. [8, proposition 9.3]. Throughout the proof of the proposition we will allow only Watson-Crick base pairs, i.e. A pairing with U and C pairing with G. This will become explicit in the final specification of the base pair energy function, and is only a technical limitation to reduce the complexity of the proof. Before proving proposition 1 we need some building blocks.

Definition 2 *The* d *digit* binary representation *of* k*, where* $0 \le k \le 2^d - 1$*, over the alphabet* $\{A, U\}$*, is the string* $b_{\{A,U\}}(k, d)$ *of length* d *that interpreted as a binary number with* A *representing* 0 *and* U *representing* 1 *has the value* k*. Similarly* $b_{\{C,G\}}(k, d)$ *is the* d *digit binary representation of* k *over the alphabet* $\{C, G\}$*.*

The k'th distinct $\{A, U\}$ pattern *using* d *digit binary representations is the string*

$$\underbrace{A \ldots A}_{d+2} U b_{\{A,U\}}(k, d) AUA b_{\{A,U\}}(k, d) U \underbrace{A \ldots A}_{d+2}.$$

The k'th distinct $\{C, G\}$ *pattern using* d *digit binary representations is defined similarly.*

Definition 3 *For a string* s *the* complementary string $\bar{s}$ *is the string constructed by reversing* s *and replacing each* A *with a* U*, each* U *with an* A*, each* C *with a* G*, and each* G *with a* C*.*

The need for these distinct patterns is to circumvent the fact that we only have four letters in the alphabet of RNA sequences. They will be used to construct an RNA sequence corresponding to a boolean formula on restricted 3SAT form, such that the energy of an optimal secondary structure of the constructed RNA sequence implies whether the formula is satisfiable. The constructed RNA sequence will consist of two parts, a part where the literals are grouped according to the clauses and a part where the literals are grouped according to the variables.

If we had an alphabet of arbitrary size we could use two symbols to represent each occurrence of a literal, one symbol in the clauses part and the other symbol in the literals part. A score of minus one could be assigned for each pairing of two such symbols with some extra pairs of symbols being used to form structures nullifying the benefits of pairing more than one symbol in a clause, or pairing a symbol representing a variable as well as pairing a symbol representing this variables negation.

Without an alphabet of arbitrary size we will instead use distinct $\{C, G\}$ patterns and their complementary strings in the clauses and variables parts, respectively, to represent the literals of the formula. A helix formed between a $\{C, G\}$ pattern and its complementary string will indicate that the corresponding literal is true and we will choose energy parameters ensuring that such a helix usually contributes negatively to the total energy. The distinct $\{A, U\}$ patterns and their complementary strings will be used to form structures nullifying benefits of having more than one true literal in each clause, and of having both a literal representing a variable and a literal representing its negation being true at the same time. This is ensured by choosing energy parameters such that helices formed by the distinct $\{A, U\}$ patterns also contribute negatively to the total energy, except if the case they should nullify occurs. In that situation they contribute zero to the total energy. The formal specification of the energy parameters is postponed till the end of this section.

Definition 4 *Let* $C = l_1 \vee l_2 \vee l_3$ *be a boolean disjunction of three literals. The* clause block $\mathcal{C}$ *of* C *using* d *digit binary representations is the string*

$$\underbrace{\qquad}_{S_1}\underbrace{\qquad}_{L_1}\underbrace{\qquad}_{\bar{S}_1}\underbrace{\qquad}_{S_2}\underbrace{\qquad}_{L_2}\underbrace{\qquad}_{S_1}\underbrace{\qquad}_{\bar{S}_2}\underbrace{\qquad}_{L_3}\underbrace{\qquad}_{S_2},$$

where the S_i*'s are distinct* $\{A, U\}$ *patterns using* d *digit binary representations for two different k's, and the* L_i*'s are distinct* $\{C, G\}$ *patterns using* d *digit binary representations for three different k's.*

The rationale behind this construction is that we can form two helices between distinct $\{A, U\}$ patterns and their complementary strings within the clause block. These two helices will span different L_i's, except for the case where the S_1 and S_2 flanking L_2 both form helices with their complementary string. In this case, the innermost base pair of the S_1 helix and the outermost base pair of the S_2 helix (and vice versa) will be neighbouring base pairs forming pseudoknots.

Furthermore, the L_i's spanned by such a helix will be *screened.* By screened, we mean that at least one of the flanking bases of the L_i pattern cannot form a base pair with a base not spanned by the helix without forming a pseudoknot with the innermost base pair of the helix. The L_i pattern thus cannot form the intended helix with its complementary string in the variable block, that we will describe shortly, without introducing a pseudoknot of neighbouring base pairs. Without introducing neighbouring pseudoknotted base pairs, for a clause block we can thus form helices of two of the distinct patterns straightaway, and a third helix if we can pair one of the L_i patterns with its complementary string in the variables part.

Definition 5 *Let* x *be a variable occurring in a boolean formula where each literal occurs at most twice. The* variable block $\mathcal{V}$ *of* x *using* d *digit binary representations is the string*

$$\underbrace{\qquad}_{S_1}\underbrace{\qquad}_{\bar{P}_1}\underbrace{\qquad}_{\bar{P}_2}\underbrace{\qquad}_{\bar{S}_1}\underbrace{\qquad}_{\bar{N}_1}\underbrace{\qquad}_{\bar{N}_2}\underbrace{\qquad}_{S_1},$$

where S_1 *is a distinct* $\{A, U\}$ *pattern for some* k*, the* $\bar{P}_i$*'s are complementary strings to the distinct* $\{C, G\}$ *patterns used for the at most two positive occurrences of* x *(if* x *occurs positive only once, one of the* $\bar{P}$ *patterns is omitted from* $\mathcal{V}$*) and the* $\bar{N}_i$*'s are complementary strings to the distinct* $\{C, G\}$ *patterns used for the at most two negative occurrences of* x *(if* x *occurs negative only once, one of the* $\bar{N}$ *patterns is omitted from* $\mathcal{V}$*).*

The rationale behind this construction is once again to use a helix formed by one of the occurrences of S_1 and its complementary string to screen the complementary strings corresponding to either the (at most) two positive occurrences of x or the (at most) two negative occurrences of x. If we are to avoid introducing neighbouring base pairs forming a pseudoknot, either none of the distinct S_1 patterns form a helix with the complementary string, the complementary strings corresponding to the positive occurrences of x do not form helices, or the complementary strings corresponding to the negative occurrences of x do not form helices. We are now ready to construct the RNA sequence representing a boolean formula on restricted 3SAT form.

Definition 6 *Let ϕ be a boolean formula on conjunctive normal form where each clause has three literals and each literal occurs at most two times. Assume that ϕ consists of c clauses and uses v variables. The RNA sequence corresponding to ϕ is the sequence*

$$s_\phi = \mathcal{C}_1\mathcal{C}_2\ldots\mathcal{C}_c\mathcal{V}_1\mathcal{V}_2\ldots\mathcal{V}_v,$$

where $\mathcal{C}_i$ is the clause block using $\lceil\log_2(3c+v)\rceil$ digit binary representations corresponding to the i'th clause of ϕ, $\mathcal{V}_i$ is the variable block using $\lceil\log_2(3c+v)\rceil$ digit binary representations corresponding to the i'th variable of ϕ, no distinct pattern is used more than once, and the patterns corresponding to a literal and their complementary strings occur in reverse order.

The choice of number of digits we use in the binary representations ensures that we can choose at least $\max\{3c, 2c+v\}$ different values for distinct patterns. Each clause block uses two distinct $\{A,U\}$ patterns and three distinct $\{C,G\}$ patterns, while each variable block uses one distinct $\{A,U\}$ pattern. Thus we do not run out of patterns. We will use the term *complementary pattern* for the deliberate occurrences of the complementary string to a distinct pattern, that is, the strings indicated by a barred pattern in definitions 4 and 5.

So far we have assumed that helices only form between a distinct pattern and the complementary string designed to form a helix with it. Helices can of course form between parts of distinct patterns not designed to form helices together, but the following lemma limits the length of such helices.

Lemma 1 *Let s_ϕ be an RNA sequence constructed from a boolean formula ϕ according to definition 6. In any structure $\mathcal{S}$ of s_ϕ, any helix of consecutively stacking pairs of length at least $4d+5$, where d is the number of digits used for the binary representations, will have at least $2d+3$ bases at the end of a distinct pattern forming base pairs with the intended bases of the complementary pattern to this distinct pattern.*

Proof. By construction any substring of s_ϕ of length at least $4d+5$ will contain at least $2d+3$ bases from one of the ends of a distinct pattern or its complementary pattern. Consider one of the two substrings forming the helix. This will be of length at least $4d+5$ and thus contain at least $2d+3$ bases from a distinct pattern or its complementary pattern. Assume without loss of generality that it contains the first $2d+3$ bases from the k'th distinct $\{A,U\}$ pattern using d digit representations, that is, it contains the substring $A^{d+2}Ub_{\{A,U\}}(k,d)$. By construction, the only occurrences of $d+2$ consecutive U's preceded by an A in s_ϕ are at the ends of complementary patterns to distinct $\{A,U\}$ patterns, and thus $A^{d+2}Ub_{\{A,U\}}(k,d)$ forms base pairs with $\bar{b}_{\{A,U\}}(k',d)AU^{d+2}$ for some k' (by the assumption that only Watson-Crick base pairs are allowed). As $b_{\{A,U\}}(k,d)$ pairs with $\bar{b}_{\{A,U\}}(k',d)$ it follows that $k=k'$. □

We have now established that any helix of considerable length will contain at least part of a designed pairing. The next lemma establishes that this will be all it contains.

Lemma 2 *Let s_ϕ be an RNA sequence constructed from a boolean formula ϕ according to definition 6 using d digit binary representations. In any structure $\mathcal{S}$ of s_ϕ, there are no helices of more than $4d+9$ consecutively stacking base pairs containing only A's and U's or containing only C's and G's. The only helices of length $4d+9$ containing only A's and U's or containing only C's and G's are helices formed by distinct patterns and their complementary pattern.*

Proof. By lemma 1 we know that a helix of length $4d+9$ will contain one of the ends of a distinct pattern paired with its complementary pattern. All we have to show is, that we cannot extend a helix formed by a distinct pattern and its complementary pattern with an extra stacking pair of bases of the same type.

If the distinct pattern is a $\{C,G\}$ pattern this is straightforward, as it will be in a clause block and thus bordered by an A and a U, or by two A's. Similarly, the complementary pattern of a distinct $\{A,U\}$ pattern from a variable block will be bordered by two G's. Finally, the complementary pattern to a distinct $\{A,U\}$ pattern from a clause block will be bordered by an A on one side, cf. definition 4. But taking the $\bar{S}_1$ complementary pattern as example, this A should form an illegal (by the Watson-Crick base pair assumption) base pair with either the leftmost A of the preceding clause block or the rightmost C in the L_2 pattern to extend the helix. □

Proof (of proposition 1). As mentioned above the reduction will be from 3SAT with the restriction that each literal appears at most twice. So let ϕ be a valid formula for this restriction of 3SAT with c clauses and v variables. In polynomial time, we can construct s_ϕ according to definition 6, and the base pair energy function

$$E(X_i \cdot Y_j, V_{i+1}, W_{j-1}) =$$

$$\begin{cases} -1 & \text{if } V_{i+1} \cdot W_{j-1} \in \mathcal{S} \text{ and either} \\ & X \cdot Y, V \cdot W \in \{A \cdot U, U \cdot A\} \\ & \text{or } X \cdot Y, V \cdot W \in \{C \cdot G, G \cdot C\} \\ 4d+7 & \text{if } X \cdot Y \in \{A \cdot U, U \cdot A, C \cdot G, G \cdot C\} \\ & \text{and for } j' \notin \{i+1, \dots j-1\} \text{ we have} \\ & V_{i+1} \cdot Z_{j'}, W_{j-1} \cdot Z_{j'}, Z_{j'} \cdot V_{i+1}, \\ & Z_{j'} \cdot W_{j-1} \notin \mathcal{S} \\ 4d+8 & \text{otherwise} \end{cases}$$

where d is the number of digits used for the binary representations in s_ϕ and $\mathcal{S}$ is the structure for which the energy is calculated. The notation X_i is used as a shorthand to indicate that the i'th base is of type X.

We claim that the optimal secondary structure of s_ϕ with the above energy function has energy $-(3c+v)$ if and only if ϕ is satisfiable. By the energy function, the only helices for which the base pairs combined yields a negative contribution to the energy of the secondary structure are helices of at least $4d+9$ base pairs, base pairs that are either all A's pairing with U's or all C's pairing with G's. By lemma 2, the only such helices that can be formed are between distinct patterns and their complementary patterns; these helices will consist of exactly $4d+9$ base pairs and thus contribute -1 to the total score of a secondary structure, *provided* that the innermost base pair of the helix does not have a neighbouring base pair that forms a pseudoknot. Hence, if a distinct pattern is screened by a helix, it can not form a helix yielding a negative contribution to the total energy.

If there is an assignment of truth values to the variables of ϕ satisfying ϕ, we can construct a secondary structure $\mathcal{S}$ on s_ϕ with energy $-(3c+v)$ based on this assignment by forming the following base pairs.

- For each variable block forming the helix of the distinct $\{A, U\}$ pattern and the complementary pattern screening the complementary patterns of the literals that become **false** by the assignment.
- For each clause block forming the helices between the distinct $\{A, U\}$ patterns that leave the distinct $\{C, G\}$ pattern of a literal that becomes **true** by the assignment unscreened.
- Forming the helices between the unscreened distinct patterns of literals in the clauses part and their complementary patterns (that are unscreened as the assignment satisfies ϕ, and as the reverse order requirement in definition 6 ensures the two complementary patterns corresponding to the same literal not having neighbouring base pairs forming a pseudoknot) in the variables part.

By the discussion following definition 4, the distinct patterns of a clause block can form at most three helices, each yielding a contribution of -1, and each variable block introduces only one new distinct pattern; hence the energy of $\mathcal{S}$ of $-(3c+v)$ is optimal.

Assume now that s_ϕ has an optimal structure $\mathcal{S}$ of energy $-(3c+v)$. By the above and the discussion following definition 4, we get that each clause block will contain a distinct pattern corresponding to a literal forming a helix with its unscreened complementary pattern in the variables part, and that the complementary patterns corresponding either to a variable or to its negation will be screened. We can thus infer a truth assignment to the variables of ϕ satisfying ϕ from the unscreened complementary patterns of literals in $\mathcal{S}$. $\square$

The energy function specified in the proof of proposition 1 rewards stacking some base pairs, penalises loops by penalising the first base pair in a helix, and further penalises neighbouring base pairs that form a pseudoknot. The only two remarkable oddities are the disallowance of base pairings between G and U, and penalising stacking an A, U base pair with a C, G base pair.

One can observe that we could allow G, U base pairs without changing anything but inserting a C between the two complementary patterns corresponding to the same literal. As for penalising stacking A, U base pairs with C, G base pairs, this was chosen to ease establishing the fact that no energy benefits are obtained by extending a helix formed by a distinct pattern and its complementary pattern by further stacking base pairs. A proof where the energy function rewards stacking of all combinations of A, U base pairs, C, G base pairs and G, U base pairs can be achieved by a more involved construction of the clauses part of s_ϕ. However, to limit the complexity of the proof, we have chosen to present the above version.

5 Discussion

The **NP**-completeness of the RNA secondary structure prediction problem in the Nearest Neighbour Pseudoknot Model tells us, that any algorithm allowing energy functions sufficiently general to be specialised to the energy functions in the Nearest Neighbour Pseudoknot Model, and running in worst case polynomial time, would imply $\mathbf{P} = \mathbf{NP}$. The question whether or not $\mathbf{P}$ is equal to $\mathbf{NP}$ is one of the fundamental open problems in computer science. Based on strong evidence, the large majority of computer scientists believe that $\mathbf{P} \neq \mathbf{NP}$. The **NP**-completeness of the RNA secondary structure prediction problem in the Nearest Neighbour Pseudoknot Model thus hints that there is only little hope for a worst case polynomial time algorithm for RNA secondary structure prediction in the Nearest Neighbour

Pseudoknot Model, or models extending it. Moreover, it hints that any algorithm for predicting RNA secondary structures with general pseudoknots most likely have to exploit specific properties of a fixed energy function to obtain polynomial running time.

One approach to obtain a polynomial time algorithm for RNA secondary structure prediction with pseudoknots is to limit the types of legal pseudoknots. This is the approach taken by Rivas and Eddy in [10] and by us in section 3. Another approach is taken by Tabaska *et al.* in [13], where interactions between neighbouring base pairs are ignored, thus reducing the problem of RNA secondary structure prediction (with pseudoknots) to compute a maximal weighted matching. If we are satisfied to find not necessarily the structures of least free energy, then heuristics can be applied to search for structures of low energy. For example, van Batenburg *et al.* in [17] report on successful experiments with applying genetic algorithms to the problem of finding low energy RNA secondary structures containing pseudoknots.

References

[1] S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22:2079–2088, 1994.

[2] J. Gorodkin, L. J. Heyer, and G. D. Stormo. Finding common sequence and structure motifs in a set of RNA sequences. In *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 120–123, 1997.

[3] B. Knudsen and J. Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15:446–454, 1999.

[4] R. B. Lyngsø, M. Zuker, and C. N. S. Pedersen. Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics*, 15(6):440–445, 1999.

[5] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[6] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Science of the USA*, 77(11):6309–6313, 1980.

[7] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35:68–82, 1978.

[8] C. H. Papadimitriou. *Computational Complexity*. Addison-Wesley Publishing Company, 1994.

[9] C. W. A. Pleij. RNA pseudoknots. In R. F. Gesteland and J. F. Atkins, editors, *The RNA World*. Cold Spring Harbor Laboratory Press, 1993.

[10] E. Rivas and S. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285:2053–2068, 1999.

[11] Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjölander, R. C. Underwood, and D. Haussler. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 22:5112–5120, 1994.

[12] D. Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45:810–825, 1985.

[13] J. E. Tabaska, R. B. Cary, H. N. Gabow, and G. D. Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14(8):691–699, 1998.

[14] I. Tinoco, P. N. Borer, B. Dengler, M. D. Levine, O. C. Uhlenbeck, D. M. Crothers, and J. Gralla. Improved estimation of secondary structure in ribonucleic acids. *Nature New Biology*, 246:40–41, 1973.

[15] I. Tinoco, O. C. Uhlenbeck, and M. D. Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230:362–367, 1971.

[16] D. H. Turner, N. Sugimoto, and S. M. Freier. RNA structure prediction. *Annual Review of Biophysics and Biophysical Chemistry*, 17:167–192, 1988.

[17] F. H. D. van Batenburg, A. P. Gultyaev, and C. W. A. Pleij. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *Journal of Theoretical Biology*, 174(3):269–280, 1995.

[18] K. Zhang and D. Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 18(6):1245–1262, 1989.

[19] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.