From RNA Secondary Structure to Coding Theory: A Combinatorial Approach

Christine E. Heitsch, Anne E. Condon, and Holger H. Hoos*

Department of Computer Science
University of British Columbia
201-2366 Main Mall
Vancouver, B. C. V6T 1Z4
{heitsch, condon, hoos}@cs.ubc.ca

Abstract. We use combinatorial analysis to transform a special case of the computational problem of designing RNA base sequences with a given minimal free energy secondary structure into a coding theory question. The function of RNA molecules is largely determined by their molecular form, which in turn is significantly related to the base pairings of the secondary structure. Hence, this is crucial initial work in the design of RNA molecules with desired three-dimensional structures and specific functional properties. The biological importance of RNA only continues to grow with the discoveries of many different RNA molecules having vital functions other than mediating the production of proteins from DNA. Furthermore, RNA has the same potential as DNA in terms of nanotechnology and biomolecular computing.

1 Introduction

Beyond their essential roles in living organisms, the synthetic importance of nucleotide molecules with particular functions continues to expand. For example [1, 8], biomolecular computations utilize DNA and RNA molecules with specially designed structural properties. Other potential applications of RNA design include "nanorobotics and the rational synthesis of periodic matter," as has been the goal for DNA of Seeman's laboratory [9]. As such, the analysis and design of nucleotide structures is an important problem at the rapidly developing intersection of the biological, mathematical, and computational sciences. Although RNA molecules have been the focus of this work, the same principles apply to the design of DNA base sequences with desired structural arrangements.

A significant initial step in the engineering of RNA molecules with desired functional properties would be solving the **RNA secondary structure design problem** of finding, when possible, a base sequence which folds to a given target RNA secondary structure. Previous work on RNA structure algorithms

* This material is based upon work supported by the U.S. National Science Foundation under Grant No. 0130108, by the National Sciences and Engineering Research Council of Canada, and by GenTel Corp.

M. Hagiya and A. Ohuchi (Eds.): DNA8, LNCS 2568, pp. 215–228, 2003.

has mainly focused on the reverse problem of predicting base pairings from a primary nucleotide sequence, under certain structural and thermodynamic assumptions. Although efficient algorithms have been developed for this prediction question, there is no known efficient deterministic procedure for RNA secondary structure design. The Vienna RNA Package of Schuster *et al.* [4] provides a simple stochastic local search algorithm which works well for the design of small secondary structures. Seeman [5] used a heuristic approach based on sequence symmetry minimization for a design problem closely related to the one studied in this paper. Here we focus on a significantly restricted version of the RNA secondary structure design question, with the ultimate goal of an efficient algorithmic solution for well-characterized subcases of the general problem.

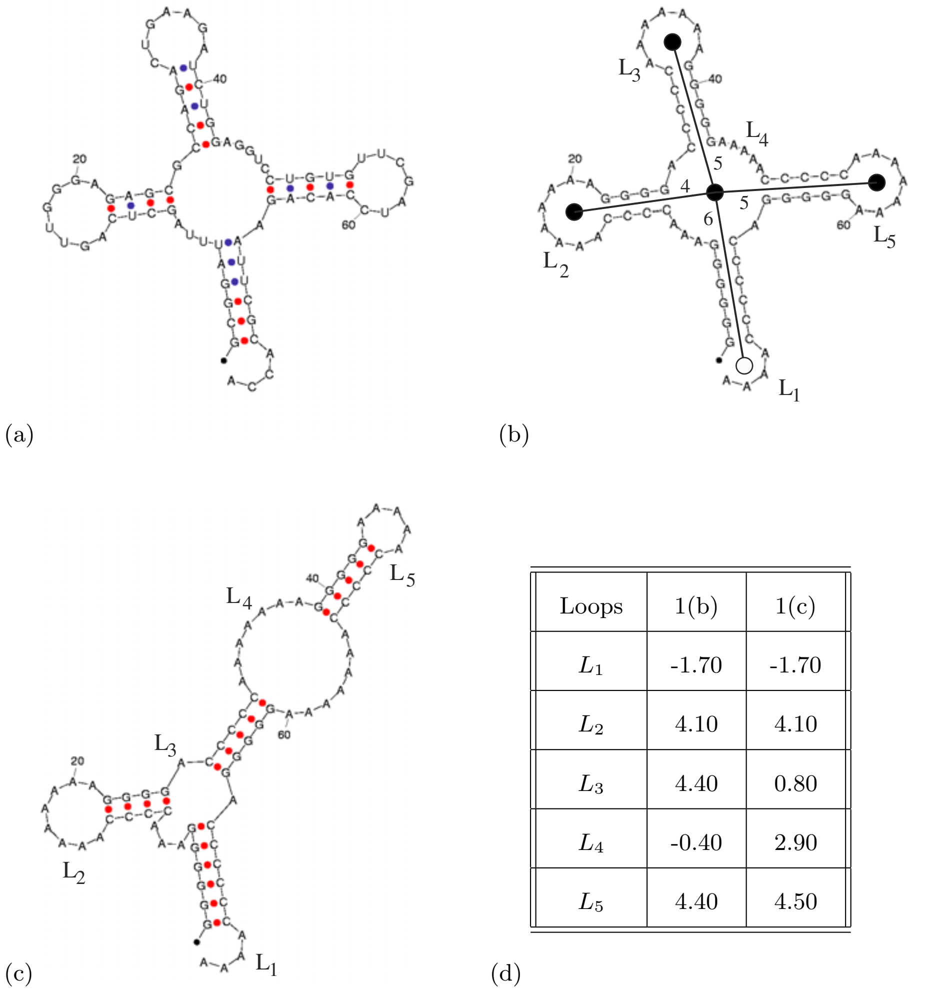
The special case considered in this paper is already nontrivial to resolve, and retains enough characteristics of the full RNA secondary structure design problem to be a very useful first step. A precise problem statement is provided in Section 4, and an outline of our current methodology can be found in Section 5. Section 2 gives an overview of the standard RNA thermodynamic model, while Section 3 illustrates its abstract mathematical interpretation. The example of Figure 1, however, captures both our choice of restrictions and our algorithmic approach, as well as the essential difficulty confronting any solution strategy.

The basic assumption underlying current understanding of RNA secondary structure is that base sequences fold to minimize free energy. Under this hypothesis, the fundamental problem with RNA secondary structure design is ensuring the desired *minimal* free energy configuration of the constructed sequence. Intuitively, a sequence will fold to a configuration which minimizes loop costs while maximizing the beneficial stacked pairs. Thus, to preclude alternate configurations we must ensure that improvements in loop energies are offset by the penalty of lost base pairs. Our simple design strategy isolates loop stretches from helical segments, enabling the clear understanding of helix "mismatches" found in Section 6 and an efficient analysis of the energy trade-offs among various possible loop arrangements as given in Section 7.

Essentially, each base sequence and corresponding complement assigned to a stem must be as different as necessary from all others. Theorem 2 of Section 8 gives a constructive bound on the helix "quality" guaranteeing a unique minimal secondary structure among a subset of all possibilities for the corresponding sequence. Thus, our major advancement in this work is a value on the quantization of "as different as necessary" for a particular subset of secondary structures and a certain class of primary base sequences. A secondary contribution is the theoretical framework surrounding our main result, which may facilitate further insights in the investigation of the nucleotide structure design problem.

2 RNA Secondary Structure and the Free Energy Model

Like DNA, the primary RNA structure is an oriented linear sequence of four nucleotides, denoted A, U, C, and G, with chemically distinct ends referred to as $5'$ and $3'$. These nucleotides may form the usual Watson-Crick, or so-called



Loops	1(b)	1(c)
L_1	-1.70	-1.70
L_2	4.10	4.10
L_3	4.40	0.80
L_4	-0.40	2.90
L_5	4.40	4.50

Fig. 1. (a) The simple secondary structure of S. cerevisiae Phe-tRNA at 37°. The destabilizing effects of single-stranded regions, or "loops," are counterbalanced by the beneficial negative free energies of successive stacked base pairs, or "stems." (b) A basic design method attempts to replicate the structure by wrapping a simple strand around its geometric interpretation, assigning A's as loop segments and the unrelated Watson-Crick base pairs, $C-G$ and $G-C$, to helical stretches. (c) Without careful construction, the sequence exploits symmetries to reduce loop costs and folds to an alternate minimal energy configuration with fewer "expensive" hairpin loops. (d) This table lists the different energies for the corresponding loops L_i for the structures in 1(b) and 1(c) respectively. All foldings courtesy of `mfold version 3.1` by Zuker and Turner [10, 2], available online via `http://www.bioinfo.rpi.edu`.

canonical, base pairings (namely A − U, U − A, G − C, and C − G) as well as other non-canonical matches. Self-bonding forces the single RNA strand into a potentially complicated arrangement of stabilizing helical regions, or "stems," connected by single-stranded regions, or "loops." The RNA secondary structure is characterized as the set of base pairs, including the "wobble" pairing of G and U, inducing these geometric structural arrangements.

Definition 1. *Let $R = b_1 b_2 \dots b_n$ be an RNA sequence of length n. For $1 \leq i < j \leq n$, let $i \cdot j$ denote the pairing of base b_i with b_j. A* **secondary structure** *of R is a set P of base pairs such that, for all $i \cdot j, i' \cdot j' \in P$, $i = i'$ if and only if $j = j'$.*

A basic premise is that RNA molecules will assume foldings which minimize the overall free energy. There currently exist well-regarded and widely-used algorithms such as Zuker's `mfold` [10, 2] which implement an efficient recursive calculation of the minimum free energy configuration under this model. Experimentally derived parameters are used in these computations, such as the RNA thermodynamic values provided by the Turner group [6] and the corresponding ones for DNA calculated by the SantaLucia group [3].

However, the computational efficiency of the current recursive methods is obtained at the expense of a class of foldings, called pseudoknots, which can be conceptualized as "switchbacks" in the RNA secondary structure. Without pseudoknots, all base pairings can be considered as "inside" the planar representation of an RNA secondary structure. Under this exclusion, the free energy can be efficiently calculated by decomposition into an independent sum over the loops and stacked pairs of the interior.

Definition 2. *An RNA secondary structure P includes a* **pseudoknot** *if there exist two base pairs $i \cdot j, i' \cdot j' \in P$ with $i < i' < j < j'$.*

Definition 3. *A base b_k or base pair $k \cdot l \in P$ is* **accessible** *from a base pair $i \cdot j \in P$ if $i < k$ $(< l) < j$ and if there is no other base pair $i' \cdot j' \in P$ such that $i < i' < k$ $(< l) < j' < j$.*

Definition 4. *The* **loop** *closed by a base pair $i \cdot j \in P$, denoted $L(i \cdot j)$, is the collection of bases and base pairs accessible from $i \cdot j$.*

Note that $L(i \cdot j)$ does not include the closing base pair $i \cdot j$. According to the above terminology, a **stacked pair** is formed by a closing base pair $i \cdot j$ whose loop $L(i \cdot j)$ contains exactly the base pair $(i + 1) \cdot (j - 1)$. In succession, stacked pairs form a helical segment, or **stem**, which stabilizes the secondary structure. For the purposes of this work, we will generally reserve the term "loop" for destabilizing components containing unpaired bases. Loops are distinguished according to whether they contain 0, 1, or more base pairs. Let the term **k-loop** refer to a loop having $k - 1$ accessible base pairs, totaling k base pairs including the closing one. Intuitively, k different stems radiate out from a k-loop; the central loop, labeled L_4, in Figure 1(b) is a 4-loop. A 1-loop, such as L_2 in both Figure 1(b) and 1(c), is also commonly known as a **hairpin**. In general, a 2-loop is called an **internal loop**, as in the case of L_4 in Figure 1(c), except for **bulges** which have all unpaired bases occurring on only one side. For $k \geq 3$, a k-loop is simply called a **multiloop**.

Typically, the energy of a 1-loop or 2-loop is the sum of several terms. In our model, the relevant values are the entropic term, which depends only on the number of unpaired bases in the loop, and the beneficial stacking interaction between a base pair and the adjacent single-stranded nucleotide. In general, these single-stranded stacking energies, also known as the terminal mismatch energies, depend on the orientation of the closing base pair so the values for C − G and G − C are not necessarily symmetric. The standard affine linear energy function for the entropic term of k-loop energies when $k > 2$ is chosen primarily for computational convenience since so little is known experimentally about the stability of multiloops.

The **external loop** of an RNA secondary structure is the set of bases and base pairs without a closing base pair. The loop L_1 in Figures 1(b) and 1(c) is an example of an external loop. For arbitrary RNA secondary structures, it will be denoted L_{e}. The current model assumes that the external loop has no conformal constraints, and hence no associated entropic costs. Thus, it must be treated distinctly from all other loops.

3 Plane Trees and RNA Foldings

Much of the essential arrangement of loops and stems in an RNA secondary structure is captured by a special type of graphical object known as a plane tree. Specifically, as observed in the previous section, the exclusion of pseudoknots induces an interior/exterior orientation to an RNA secondary structure. We utilize this fact to abstract a given set of base pairs to their geometric “skeleton.” Helical segments are associated with edges, and loops to vertices. We preserve information about the length of the stems as a weight on the corresponding edge. Hence, the basic arrangement of an RNA secondary structure may be described by a weighted plane tree, such as in Figure 1(b).

Definition 5. *[7] A* **plane tree** *is a rooted tree whose subtrees at any vertex are linearly ordered.*

The ordering is sufficient to distinguish any vertex of a plane tree, and so labels are unnecessary. The unique root vertex of a plane tree corresponds to the distinct external loop of an RNA secondary structure. According to common graph terminology, a vertex is the “child” of the connected vertex one edge closer to the root. Vertices with no children are called leaves and, for our purposes, correspond to the hairpin loops of an RNA secondary structure.

Definition 6. *[7] A plane tree vertex with n children has* **degree** *n.*

In our association of plane trees and RNA secondary structures, the degree of a vertex corresponds to the number of base pairs in the loop, excluding the closing pair. Thus a k-loop has $k-1$ accessible base pairs and corresponds to a vertex of degree $k-1$ in the associated plane tree. We say that plane tree corresponds to an RNA secondary structure if the arrangement of vertices and edges is the same as that of loops and stems. With the additional information of the loop segments lengths, a weighted plane tree completely specifies the desired configuration of base pairs.

4 Restricted RNA Secondary Structure Design

The question of designing RNA base sequences with desired secondary foldings has important computational, as well as biological, implications. The general problem may be precisely stated as: *Given the specification of an RNA secondary structure, return a primary nucleotide sequence whose minimal energy configuration has the desired base pairings under the current free energy model.* We consider here a special case capturing many essential aspects of this difficult problem. Specifically, we impose restrictions on the input structure, output sequences, and potential reconfigurations.

To begin, in keeping with most RNA prediction algorithms, we exclude pseudoknots. This permits an input configuration abstractly described by a plane tree T. We will also allow specification of the loop segment lengths and edge weights / stem lengths. These input parameters are subject to the requirement that, for any output strand, among the secondary structures corresponding to the desired configuration T, the lowest free energy must have the given loop structures. We call a set of base pairs satisfying these input constraints a **restricted structure**.

Additionally, our constructed sequences must satisfy the **loop-protecting property**: all intended loop segments should remain unpaired in any alternate configuration. This requirement is enforced by restricting to a three letter alphabet $\{\mathrm{A}, \mathrm{C}, \mathrm{G}\}$ and assigning A exactly to the unpaired segments. The current thermodynamic model predicts no base pair energetic interactions between A and C or G. Hence, under this restriction, the number of base pairs in any alternate configuration cannot exceed the original count.

As the final restriction, our design must be sufficiently good to preclude any alternate minimal energy configurations from a particular subclass of structures. In our loop-protecting RNA model, there can be no interaction between the intended A loop segments and the $\mathrm{C}-\mathrm{G}$, $G-\mathrm{C}$ base pairs forming the helical stretches. However, various $\{\mathrm{C}, \mathrm{G}\}$ segments may align in the minimal energy configuration even though they are not exact or intended complements. Hence, a helix from the target structure is said to be **partially conserved** in another pseudoknot-free configuration when a $\{\mathrm{C}, \mathrm{G}\}$ nucleotide segment forms base pairs with exactly one other segment of the strand; a helix is fully conserved when it pairs with its intended complement. The set of alternate configurations with partially or fully conserved helices will be called **helix-preserving** for a given strand and target structure.

Thus, given as input a restricted RNA secondary structure, we investigate efficient algorithmic methods for generating a loop-protecting sequence which will not to fold into any helix-preserving alternate configuration.

5 A Constraint Satisfaction Solution Strategy

Subject to our current restrictions, our soluction must encode the given minimal free energy secondary structure configuration into a primary nucleotide sequence. Accordingly, for an RNA molecule with n stems, we need to produce n strings over the alphabet $\{\mathrm{C}, \mathrm{G}\}$, $s_1, \ldots, s_n$. They and their Watson-Crick complements, $\bar{s}_1, \ldots, \bar{s}_n$, would then be appropriately arranged into a single linear strand interspersed by A stretches of the desired length. Thus, our output will have the form $R = (l_0, h_1, l_1, h_2, \ldots, l_{2n-1}, h_{2n}, l_{2n})$ where $h_i \in \{s_1, \ldots, s_n, \bar{s}_1, \ldots, \bar{s}_n\} \subset \{\mathrm{C}, \mathrm{G}\}^+$ and $l_j \in \{\mathrm{A}\}^*$.

We let $\mathbf{R_H} = (h_1, h_2, \ldots h_{2n})$ be the intended helical segments of R, while the loop regions are denoted $\mathbf{R_L} = (l_0, l_1, \ldots, l_{2n-1}, l_{2n})$. We accept as input a plane tree T with edges e_j and weights w_j for $1 \leq j \leq n$ as well as the specification of the loop segments $R_L = (l_0, l_1, \ldots, l_{2n-1}, l_{2n})$ where $l_i \in \{\mathrm{A}\}^*$. In keeping with known thermodynamic constraints on RNA secondary structure, we have two restrictions on the possible lengths of the loop segments. If l_i is the single-stranded segment intended to form a 1-loop, then $|l_i| \geq 3$. Likewise, there cannot exist i and j such that l_i and l_j form a 2-loop and $|l_i| = 0 = |l_j|$.

According to the free energy model, for a weighted plane tree T representing the given restricted RNA secondary structure with loop segments R_L, we can calculate the lowest free energy of strand R in that abstract configuration, $\mathbf{E}(\mathbf{R}, \mathbf{T})$. We will divide this energy value into two components – one involving all the loops segments R_L, denoted $\mathbf{E_L}(\mathbf{R}, \mathbf{T})$, and the other, $\mathbf{E_H}(\mathbf{R}, \mathbf{T})$, for the energies associated with R_H. By our input assumption, $E_L(R, T)$ depends only on the lengths of the l_i and the single-stranded stacking interactions with the base pairs in the loops since the lowest value of $E(R, T)$ corresponds to the foldings with the given loop segments R_L.

It may be, however, that the energy $E(R, T)$ is not minimal over all other helix-preserving configurations T'. If not, there would exist at least one T' such that the lowest free energy of strand R in a configuration corresponding to T', $\mathbf{E}(\mathbf{R}, \mathbf{T'})$, is lower than $E(R, T)$. To preclude such an occurrence, we must ensure that any improvement in the energies of the structures involving the predetermined loop segments R_L is offset by the loss of beneficial stacked pairs in the remaining components. More specifically, we bound from below the $E(R, T')$ value by the sum of two components determined by our loop-protecting strand. Thus, we have $\mathbf{E'_L}(\mathbf{R}, \mathbf{T'})$ which represents the lowest free energy associated with the loop structures / vertices of configuration T' which include all the bases from R_L (and possibly some unpaired bases from the ends of R_H). Likewise, $\mathbf{E'_H}(\mathbf{R}, \mathbf{T'})$ is then defined to be the lowest free energies for the bases associated with the edges of T', which includes most of R_H, and corresponds to the helices which are conserved, partially or fully, when strand R is in configuration T'. Hence, we will refer to $E'_H(R,T')$ as the "helix energy" component of $E(R,T')$.

Consequently, configuration T will be the minimal energy secondary structure of a loop-protecting strand R among helix-preserving foldings as long as for all alternate T' with the same number of vertices and edges:

$$E(R,T) - E(R,T') \leq [E_H(R,T) - E'_H(R,T')] + [E_L(R,T) - E'_L(R,T')] \leq 0$$

Thus, an alternate minimal energy configuration of a primary base sequence will be prevented by ensuring that any benefit from improving the loop arrangements does not outweigh the cost in free energy terms for the mismatched helical segments. Towards this end, we will introduce a notion of "quality" with respect to a set of nucleotide strings which is a measure of their mutual differences, in a thermodynamic sense. An RNA strand has a **q-quality design** if it is loop-protecting and the quality of its helical segments is at least q.

Understanding the interplay between loop arrangements and the loss of stacked base pairs in helical mismatches is essential to the design of a strand with a specific minimal energy configuration. Given a bound on loop energy improvements, the design problem reduces to finding base sequences for stems which satisfy constraints, hence precluding any beneficial trade-off for an alternate configuration. The value in this is that the latter problem is amenable to solution using RNA word design techniques.

Hence, given a specified input from a subclass of RNA secondary structures, we provide a means of calculating a lower bound on the quality, as a function of the input, which is sufficient to preclude a large number of alternate configurations. Further analysis will be necessary to extend the method beyond the subset of helix-preserving to all possible alternate secondary structures.

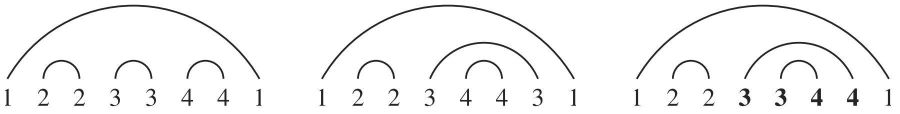
6 Helix Mismatches in Alternate Configurations

Plane trees naturally fall into distinct equivalence classes according to number of edges. Since a tree always has one more vertex than edge, we can also partition plane trees according to the number of vertices. Thus, let $\mathcal{T}_\mathbf{n}$ be the set of plane trees having n edges and $n+1$ vertices. For an RNA strand R designed to have n helices in configuration T, we are concerned about the free energy minimality of other configurations $T' \in \mathcal{T}_n$. This corresponds to restricting our attention to the mismatches in helix-preserving alternate structures.

We will use two other equivalent plane tree representations. The first is a string over the set $\{1,2,\ldots,n\}$ such that each number appears exactly twice and there are no *subsequences* of the form $ijij$. (This restriction corresponds to the pseudoknot exclusion in RNA secondary structures.) The second follows easily from the first by replacing each pair of numbers by the endpoints of an arc, pictured as n nonintersecting semi-circles whose $2n$ endpoints all lie below them on the same line. See Figure 2 for an example. Because each number appears exactly twice, there is no ambiguity in the assignment of arcs to numbered pairs.

Definition 7. *Given $T, T' \in \mathcal{T}_n$, there are m* **mismatches** *between T and T' if there are m arcs from the arc specification of T whose endpoints align with nonequal numbers from the string over $\{1, 2, \ldots, n\}$ for T'.*

Note that two different plane trees with n edges can have at most n mismatches and must have at least two. However, although mismatches are symmetric, they are not necessarilty additive since the mismatches between T, T' and T', T'' may "propagate." Hence, let $\mathcal{T}_{\mathbf{n},\mathbf{i}}(\mathbf{T})$ be the set of trees $T' \in \mathcal{T}_n$ having *up to i* mismatches between T and T'.



(a) Fig. 1(b) helices. (b) Fig. 1(c) helices. (c) The mismatches.

Fig. 2. The first two figures illustrate the arc and string specification for the plane trees representing the corresponding RNA secondary structures from Fig. 1. They have two mismatches as shown in 2(c).

For a strand $R = l_0 h_1 l_1 \ldots l_{2n-1} h_{2n} l_{2n}$, we need a better understanding of the difference in helix energy, $E_H(R, T) - E'_H(R, T')$ between the desired configuration T and a potential reconfiguration T'. We can further refine the helix energy calculations according to the plane tree edges corresponding to the helical segments of R. The helix energy associated with an edge is the minimum free energy of the corresponding two nucleotide segments $5' - h_i - 3'$, $5' - h_j - 3'$, denoted $E_e(R, T)$ or $E'_{e'}(R, T')$, for edges $e \in T$ or $e' \in T'$ respectively. When $h_i = \bar{h}_j$, as is always the case for $E_e(R, T)$, this is just the sum of the stacked pair energies. In the "mismatched" case, when h_i and h_j are not Watson-Crick complements, it is still possible to calculate the free energy as the minimum over all possible partial alignments. (Alternately, the energy could be estimated by using some (generalized) Hamming distance between the strings times the minimum energy of a stacked base pair.)

We note that when $h_i = \bar{h}_j$ for an edge $e' \in T'$ as well as for $e \in T$, then $E'_{e'}(R, T') = E_e(R, T)$. In this case, the energy component for the helix $h_i, \bar{h}_j$ does not enter into the difference $E_H(R, T) - E'_H(R, T')$. Thus, we need only consider the energies for helices which are only partially conserved in T' in order to calculate the difference in helix energies for an alternate configuration.

7 Bounding Possible Loop Energies

In order to analyze the potential benefit of a configuration other than the one for which an RNA strand was designed, we must be able to bound from below the loop energies for a class of structures. Recall that the minimum free energy calculation is a sum of the independent loop energies. Hence for a given strand $R = l_0 h_1 l_1 \ldots l_{2n-1} h_{2n} l_{2n}$ and an alternate plane tree configuration T' with $n+1$ vertices v', the lower bound on the loop component of the free energy is the sum $E'_L(R,T') = \sum_{v' \in T'} E'_{v'}(R,T')$ where $R_L = (l_0, l_1, \ldots, l_{2n-1}, l_{2n})$ and $E'_{v'}(R,T')$ is a lower bound on the energy of the loop corresponding to vertex v'.

Furthermore, the calculation of lower bounds on loop free energies in our model is a function of the number of single-stranded bases as well as the number and composition of the base pairs, with special cases for 1-loops, 2-loops, and the external loop. Thus, to calculate each $E'_{v'}(R,T')$, except for the root node, we only need to know the number of base pairs and single-stranded nucleotides. But a vertex of degree i represents a loop containing $i+1$ base pairs, including the closing one. Additionally, there are $i+1$ associated single-stranded regions, containing $l_{j_1}, \ldots, l_{j_{i+1}}$, so that a lower bound on the total length is easily calculated. Since the situation is similar, although slightly more complicated for the root node/external loop, we have that $E'_{v'}(R,T')$ can be calculated in time $\mathcal{O}(i)$ for a vertex of degree i.

However, we are interested in the bounds on the possible energies among all helix-preserving configurations. The necessary minimum value may be calculated by adapting the standard dynamic programming method for RNA secondary structure prediction.

Definition 8. *Let* $\mathbf{M}(\mathbf{R},\mathbf{T})$ *be the minimum over all lower bounds* $E'_L(R,T')$ *for* $T' \in \mathcal{T}_n$ *and* $T' \neq T$.

Theorem 1. *There is an efficient algorithm to compute* $M(R,T)$ *under the current energy model.*

8 The Quality of an RNA Encoding

For a strand R designed to be in configuration T, the maximum difference in loop energies over all possible helix-preserving alternate configurations is $E_L(R,T) - M(R,T)$. In order for T to be the minimum free energy configuration in the class $\mathcal{T}_n$, it must be that this improvement is offset by the loss of stacked pairs. Hence, the helical segments must be of a certain "quality."

Definition 9. *For a strand* R *and two plane tree configurations* T *and* T', *let* $Q(R,T,T') = E_H(R,T) - E'_H(R,T')$.

If the value of $Q(R,T,T')$ is negative, then the configuration T is a more beneficial one for strand R, in terms of (partially conserved) helix energies, than the arrangement T'. And vice versa for a positive value of $Q(R,T,T')$ since a more negative free energy is optimal.

In order to produce the necessary helical segments to solve our RNA secondary structure design problem, we must be able to generate sufficient strings and assign them to the edges of the input structure T. Suppose that $S =$

$\{s_1, \ldots, s_k\}$ is a set of distinct strings over the alphabet $\{\mathrm{C}, \mathrm{G}\}$ with $k \leq n$. Let $\bar{S}$ be the set of Watson-Crick complements of S; $s \in S$ if and only if $\bar{s} \in \bar{S}$. Recall that the edges e_j of T have weights w_j for $1 \leq j \leq n$ and that T may be specified by a string over $\{1, \ldots, n\}$, $T = (f_1, \ldots, f_{2n})$. We identify an edge e_j with the two instances of $j = f_i, f_{i'}$ for $1 \leq i, i' \leq 2n$. We say that $\alpha(\mathbf{S}, \mathbf{T}) = (h_1, \ldots, h_{2n}) = R_H$ is a **stem assignment** of S to T if for every edge e_j, for i and i' such that $j = f_i = f_{i'}$, there exists $1 \leq l \leq k$ such that $h_i = s_l$, $h_{i'} = \bar{s}_l$, and $|s_l| = w_i$. Let A be the set of all stem assignments of S to T.

Definition 10. *The* **quality** *of S with respect to T up to i mismatches is*

$$Q_i(S,T) = \min_{\alpha \in A} \max_{T' \in \mathcal{T}_{n,i}(T)} Q(\alpha(S,T), T, T')$$

For the moment, we are concerned with the quality of S over all possible helix-preserving alternate structures $T' \in \mathcal{T}_{n,n}(T) = \mathcal{T}_n$. Because we have restricted to loop-protecting RNA strands over $\{\mathrm{C}, \mathrm{G}, \mathrm{A}\}$, a stem assignment α of S to T will typically have the maximum number of stacked base pairs, and hence the most negative value possible for $E_H(R,T)$. The closer another arrangement T' comes to preserving this value, the less negative the quantity $Q(\alpha(S,T), T, T')$ will be. Thus, by maximizing the value of $Q(\alpha(S,T), T, T')$ over a set of potential configurations, we obtain a measure of the minimum loss in free energy per misaligned helix. We can then optimize this value by chosing the best possible stem assignment – the one which minimizes.

For a secondary structure specification, we need to determine a lower bound on the quality of the set of code strings S which would prevent the corresponding primary sequence from alternate helix-preserving foldings.

Theorem 2. *Let $T \in \mathcal{T}_n$ be the specification of a restricted RNA secondary structure with stem lengths $w_1, \ldots, w_n$ and loop segments $R_L = (l_0, \ldots, l_{2n})$ with $l_i \in \{\mathrm{A}\}^*$. Suppose that S is a set of strings over $\{\mathrm{C}, \mathrm{G}\}$ of quality $Q_n(S,T) = -(E_L(R,T) - M(R,T))$. Then there exists a stem assignment $\alpha(S,T) = R_H$ such that the helix-preserving minimum free energy configuration of R has the plane tree structure T.*

Proof. Let α be a stem assignment which minimizes $Q_n(S,T)$. Now, for another arbitrary helix-preserving configuration $T' \in \mathcal{T}_n$ of R with $R_H = \alpha(S,T)$:

$$\begin{aligned}
E(R,T) - E(R,T') &\leq [E_H(R,T) - E'_H(R,T')] + [E_L(R,T) - E'_L(R,T')] \\
&\leq [E_H(R,T) - E'_H(R,T')] + [E_L(R,T) - M(R,T)] \\
&= Q(R,T,T') + [E_L(R,T) - M(R,T)] \\
&= Q(R,T,T') + (-Q_n(S,T)) \\
&\leq Q_n(S,T) + (-Q_n(S,T)) \\
&= 0
\end{aligned}$$

Hence, T has lower free energy than any other T' which proves the theorem.

9 Methods for Optimizing the Quality Calculation

Theorem 2 guarantees the existence of a strand R which folds to the desired minimal free energy configuration T provided we can generate a set of strings S with a certain quality $Q_n(S,T)$. Determining whether a candidate set of strings has the necessary quality is a nontrivial task, however, since it involves considering all possible helix-preserving alternate configurations for any suitable stem assignment. Since the number of plane trees with n vertices is the nth Catalan number, $C_n = \frac{1}{n+1}\binom{2n}{n}$, simply maximizing over $T' \in \mathcal{T}_n$ would be an exponential calculation.

We can approximate this aspect of the quality calculation, though, restricting the number of alternate configurations which we have to consider. Recall that the set $\mathcal{T}_{n,i}(T)$ gives the trees $T' \in \mathcal{T}_n$ having up to i mismatches with T. Hence, we can restrict to calculating the quality of S with respect to T only up to m mismatches, $Q_m(S,T)$, where m depends on our ability to approximate $Q(\alpha(S,T),T,T')$ for $T' \in \mathcal{T}_n \setminus \mathcal{T}_{n,m}(T)$. Specifically, we consider pairs $(h_i, h_{i'})$ from $R_H = (h_1, \ldots, h_{2n})$ where h_i and $h_{i'}$ do not correspond to two sides of the same edge in T.

As we did with $E'_{e'}(R,T')$ for helices partially or fully conserved in T', we can calculate the energy of these two helical segments h_i and $h_{i'}$, denoted $\mathbf{H(i,i')}$. Then we know that the calculation of $E'_H(R,T')$, for any $T' \in \mathcal{T}_n \setminus \mathcal{T}_{n,m}(T)$ having greater than m mismatches with T, must include the sum $\sum_{j=1}^{m} H(i_j, i'_j)$ where each i_j and i'_j occurs in at most one term since each helix h_{i_j} can pair with exactly one other $h_{i'_j}$. Then we can formulate the following constraint based solution to our restricted secondary structure design problem.

Again, suppose that $S = \{s_1, \ldots, s_k\}$ is a set of distinct strings over the alphabet $\{\mathrm{C}, \mathrm{G}\}$ with $k \leq n$ and $\bar{S}$ the set of Watson-Crick complements of S. Assume that $\alpha(S,T) = (h_1, \ldots, h_{2n}) = R_H$ is a stem assignment of S to T where, for every edge e_j identified with integer $j = f_i = f_{i'}$ from the string representation $T = (f_1, \ldots, f_{2n})$, there exists $1 \leq l \leq k$ such that $h_i = s_l$, $h_{i'} = \bar{s}_l$, and $|s_l| = w_j$.

Let $\mathit{Constraints}(S,\alpha,T)$ be the following set of constraints on S, with respect to T and α:

There is one constraint in $\mathit{Constraints}(S,\alpha,T)$ for each $T' \in \mathcal{T}_{n,m}$. Let I' be the set of pairs of indices (i,i') such that h_i and $h_{i'}$ correspond to a mismatched edge in T'. We know that $2 \leq |I'| \leq m$. Let I be the corresponding set of matched index pairs. Thus if $(i,i') \in I'$ then there exists in I either (i,j) or (j,i) and either (j',i') or (i',j') where the helices h_i, h_j and $h_{j'}, h_{i'}$ are correctly matched in T. The constraint for T' is then:

$$[\sum_{(k,l)\in I} H(k,l) - \sum_{(i,i')\in I'} H(i,i')] + [E_L(R,T) - E'_L(R,T')] \leq 0$$

Additionally, constraints are needed to handle trees in $T' \in \mathcal{T}_n \setminus \mathcal{T}_{n,m}(T)$. Let J' be a set of m pairs of indices (j,j') where $1 \leq j \leq j' \leq 2n$, each j and

j' occurs in at most one pair in J', and the helical segments h_j and $h_{j'}$ are not matched in T. Let J be the corresponding multiset of matched index pairs, where for $(j, j') \in J'$ there exists in J either (i, j) or (j, i) and either (j', i') or (i', j') where the helices h_i, h_j and $h_{j'}, h_{i'}$ are correctly matched in T. We note that for some correctly matched h_i, h_j, if both i and j each appear in a mismatched pair in J', then the pair i, j occurs twice in J. Then for each possible J' we have the following constraint:

$$[\sum_{(k,l)\in J} \frac{H(k,l)}{2} - \sum_{(j,j')\in J'} H(j,j')] + [E_L(R,T) - M(R,T)] \leq 0$$

Theorem 3. *Let $T \in \mathcal{T}_n$ be the specification of a restricted RNA secondary structure with stem lengths $w_1, \ldots, w_n$ and loops segments $R_L = (l_0, \ldots, l_{2n})$ with $l_i \in \{\mathrm{A}\}^*$. Suppose that S is a set of strings over $\{\mathrm{C}, \mathrm{G}\}$ with a stem assignment $\alpha(S, T) = (h_1, \ldots, h_{2n}) = R_H$. Suppose all constraints in* Constraints(S, α, T) *are satisfied. Then the helix-preserving minimum free energy of R has the plane tree structure T.*

Finally, in terms of optimizing over such stem assignments, we note that one possible approximation strategy would require that $|S| = n$ and to naively assign strings and their Watson-Crick complements solely on the basis of equality between string length and edge weight. Although these methods for increasing the algorithmic efficiency may force a higher value than strictly necessary, they achieve a significant improvement in the running time of an implementation.

10 Conclusions and Future Work

In this paper, we studied the problem of designing RNA sequences with a given secondary structure, under a standard model of free energy minimization. For a restricted case, we derived conditions on the base sequences assigned to the local helical structures (stems) of the desired structure that can be satisfied using using word design strategies. Hence, we have effectively reduced this special case of the RNA secondary structure design problem to a code design question.

In future work, we will relax the restrictions imposed for these initial results on the secondary structure inputs, RNA sequence outputs, and possible refoldings. In particular, we will study cases which allow arbitrary types of bases in any loop and more general stem composition. We will also consider possible alternate structures that are not helix-preserving. We expect that these extensions may require the use of "capping strategies," which additionally stabilize loops by restricting the initial and terminal bases of the helix segments.

We will also investigate efficient algorithmic solutions to the RNA word design questions arising from RNA structure design problems. While known DNA word design methods and results from coding theory provide a good starting point for this endeavor, we anticipate that additional specific techniques will be needed in cases where the target structure is difficult to stabilize (e.g., because of very short helices or highly asymmetric bulges).

Finally, based on this work and its future extensions, we expect to obtain a much better understanding of the class of RNA secondary structures that can be designed easily and efficiently.

References

[1] K. Komiya, K. Sakamoto, H. Gouzu, S. Yokoyama, M. A. A. Nishikawa, and M. Hagiya. Successive state transitions with I/O interface by molecules. In *Preliminary Proc. 6th Intl. Meeting on DNA Based Computers*, pages 21 – 30, June 2000.

[2] D. Mathews, J. Sabina, M. Zuker, and D. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[3] J. SantaLucia Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA*, 95:1460–1465, 1998.

[4] P. Schuster, W. Fontana, P. Stadler, and I. Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. *Proc R Soc Lond B Biol Sci*, 255(1344):279–284, 1994.

[5] N. Seeman. De novo design of sequences for nucleic acid structural engineering. *Journal of Biomolecular Structure and Dynamics*, 8(3):573–581, 1990.

[6] M. Serra, D. Turner, and S. Freier. Predicting thermodynamic properties of RNA. *Meth. Enzymol.*, 259:243–261, 1995.

[7] R. P. Stanley. *Enumerative combinatorics. Vol. 2.* Cambridge University Press, Cambridge, 1999. With a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin.

[8] E. Winfree, F. Liu, L. Wenzler, and N. Seeman. Design and self-assembly of 2D DNA crystals. *Nature*, 394:539–544, August 1998.

[9] H. Yan, X. Zhang, Z. Shen, and N. C. Seeman. A robust DNA mechanical device controlled by hybridization topology. *Nature*, 415:62–65, 2002.

[10] M. Zuker, D. Mathews, and D. Turner. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In J. Barciszewski and B. Clark, editors, *RNA Biochemistry and Biotechnology*, NATO ASI Series, pages 11–43. Kluwer Academic Publishers, 1999.