RNA Secondary Structure Prediction with Simple Pseudoknots *

Jitender S. Deogun[1] **Ruben Donis**[2] **Olga Komina**[1] **Fangrui Ma**[1]

[1]Department of Computer Science and Engineering
[2]Department of Vet & Biomedical Science
University of Nebraska – Lincoln,
Lincoln, NE 68588-0115, USA,
Email: deogun{okomina, fma}@cse.unl.edu, rdonis1@unlnotes.unl.edu

Abstract

Pseudoknots are widely occurring structural motifs in RNA. Pseudoknots have been shown to be functionally important in different RNAs which play regulatory, catalytic, or structural roles in cells. Current biophysical methods to identify the presence of pseudoknots are extremely time consuming and expensive. Therefore, bioinformatics approaches to accurately predict such structures are highly desirable.

Most methods for RNA folding with pseudoknots adopt different heuristics such as quasi-Monte Carlo search, genetic algorithms, stochastic context-free grammars, and the Hopfield networks, and techniques like dynamic programming (DP). These approaches, however, have limitations. The DP algorithm has worst case time and space complexities of $O(n^{6.8})$ and $O(n^4)$, respectively. The algorithm is not practical for sequences longer than 100 nucleotides.

In this paper, we present a dynamic programming algorithm for prediction of simple pseudoknots in optimal secondary structure of a single RNA sequence using standard thermodynamic parameters for RNA folding. Our approach is based on a pseudoknot technique for maximizing the number of base pairs proposed by Akutsu (Akutsu 2000). The algorithm has worst case time and space complexities of $O(n^4)$ and $O(n^3)$, respectively.

We validate the accuracy of our algorithm by experimental results on the entire set of simple pseudoknot collection in the PseudoBase database. Our program folds 163 pseudoknots out of 169 total in the Pseudobase database predicting the structure of 131 pseudoknots correctly or almost correctly. The algorithm is quite efficient. For example, a sequence of 75 nucleotides takes 55 seconds (compared to 20 minutes with the existing software) and a sequence of 114 nucleotides takes 8 minutes (4 hours 30 min). To our knowledge, this is most accurate and efficient algorithm for predicting simple pseudoknots in optimal secondary structure of a single RNA sequence.

Keywords: RNA, pseudoknot, structure prediction.

*This research was supported in part by NSF EPSCOR Grant No. EPS-0091900 and NSF Digital Government Grant No. EIA-0091530.

1 Introduction

A ribonucleic acid (RNA) is one of the two types of nucleic acids (desoxyribonucleic acid—DNA and ribonucleic acid—RNA) found in living organisms. An *RNA* molecule represents a long chain of monomers called nucleotides. RNAs contain four different nucleotides, adenine (A), guanine (G), cytosine (C), and uracil (U). The sequence of nucleotides of an RNA molecule constitues its primary structure, and the pattern of pairing between nucleotides determines the secondary structure of an RNA.

Different RNAs can play different roles in cells. For example, an RNA can function in the transmission and storage of genetic information, and they may also have structural and catalytic roles. RNA serves as the genetic material of many viruses. In all cellular organisms, information stored in DNA is used to govern cellular activities through the formation of RNA messages. Thus, similar to DNA, messenger RNAs (mRNA) can carry information. Ribosomal RNAs (rRNA) serve as structural scaffolds on which the proteins of the ribosome can be attached and as elements that recognize and bind various soluble components required for protein synthesis. Another example of structural RNAs is that of small nuclear RNAs (snRNA). These form a vital part of sliceosomes that process mRNAs in eukaryotes. Some RNAs play an important role in many chemical reactions. One of the ribosomal RNAs of the large subunit acts as the catalyst for the reaction by which amino acids are covalently joined during protein synthesis. Other RNAs are able to catalyze RNA processing. RNAs that have a catalytic role are called ribozymes.

RNA molecules are continuous single strands that when fold back on themselves they form secondary structures consisting of extensive double stranded segments of variable length and complex tertiary structures. The double-stranded regions are held together by hydrogen bonds between the bases. This is the same principle that is responsible for forming the double helix of DNA molecules. Watson-Crick (WC) (A=U and G≡C), wobble (G=U) and other, non-canonical pairings can occur when the RNA is folded. Base triples are possible but rare. For the purpose of this project we consider WC pairings between complementary bases (A=U, G≡C) and wobble pairing (G=U) only.

Many RNAs fold into structures that have been shown to be important for regulatory, catalytic, or structural roles in a cell. For example, secondary structure explains in part translational control in mRNA (De Smit & van Duin 1990) and replication control in single-

stranded RNA viruses (Mills, Priano, Merz, & Binderow 1990). Some RNAs do not code for proteins or structural RNAs (Brannan, Dees, Ingram, & Tilghman 1990, Brown, Ballabio, Rupert, Lafreniere, Grompe, Tonlorenzi & Willard 1991), and it is likely that the secondary structure of such transcripts defines their regulatory function in the cell. Thus, the knowledge of secondary and tertiary structures of an RNA molecule is highly desirable when investigating its role in a cell.

In recent years, biophysical techniques have been developed to determine the tertiary structures of small RNAs, and it has become evident that RNA molecules can achieve a level of structural complexity approaching that of proteins. Current biophysical methods to determine RNA structures are extremely time consuming and expensive. Therefore, bioinformatics approaches for accurate prediction are highly desirable.

2 Pseudoknots: structure and functions

Pseudoknots are widely occurring structural motifs in RNA. First described in the early eighties as part of tRNA-like structures in plant viral RNAs, pseudoknots were recognized as a general principle of RNA folding (Pleij, Rietveld, & Bosch 1985). Since then, pseudoknots have been found in virtually all kinds of RNAs, including coding and non-coding regions of cellular mRNAs, viral RNAs, ribosomal RNAs, and snRNAs. Viral RNAs, especially, have been proven to be a rich source for pseudoknot structures. Pseudoknots have been shown to be functionally important in different RNAs which play regulatory, catalytic, or structural roles in cells, e.g. ribosomal frameshifting, regulation of translation and splicing. Pseudoknots are also essential elements of the topology of many structural RNAs such as ribosomal RNAs. Therefore, elucidation of structural features of pseudoknots and reliable prediction of pseudoknotting using sequence data are important for understanding structure-function relationships in many RNA molecules. It is important to mention that a database, PseudoBase, containing a collection of pseudoknotted RNA structures is freely accessible at http://wwwbio.LeidenUniv.nl/~Batenburg/PKB.html.

The simplest pseudoknot, the classical or so-called H-(hairpin) pseudoknot contains two stems (S_1 and S_2) and two loops (L_1 and L_2) (Fig.1.C). Such pseudoknot is formed by the pairing of a hairpin loop, closed by S_1, with the downstream nucleotides forming S_2, or, alternatively, by pairing of a hairpin loop, closed by S_2, with the upstream nucleotides forming S_1. Generally, pseudoknots can have recursive structure (Fig.1.D) where any loop region can be replaced by another secondary structure with or without pseudoknots. Akutsu (Akutsu 2000) proved that RNA secondary structure prediction with generalized pseudoknots is NP-hard, when an arbitrary energy function depending on adjacent base pairs is used.

3 Definitions of RNA secondary structure

A secondary structure, S, on an RNA sequence, $A = a_1, a_2, a_3, \ldots, a_n$, is a set of base pairs. A base pair between nucleotides a_i and a_j $(i < j)$ is denoted by $(a_i - a_j)$ or simply by $(i-j)$. Base triples are prohibited. Sharp turns are prohibited. A turn, called a hairpin loop, must contain at least 3 bases. A set of base pairs

$$M = \{\, (i-j) \mid 1 \le i < j \le n, \; (a_i - a_j) \text{ is a base pair and every } i \text{ and } j \text{ appears at most once} \,\}$$

is called an RNA secondary structure without pseudoknots if no distinct pairs $(a_i - a_j), (a_h - a_k) \in M$ satisfy $i \le h \le j \le k$ (Fig.1.A).

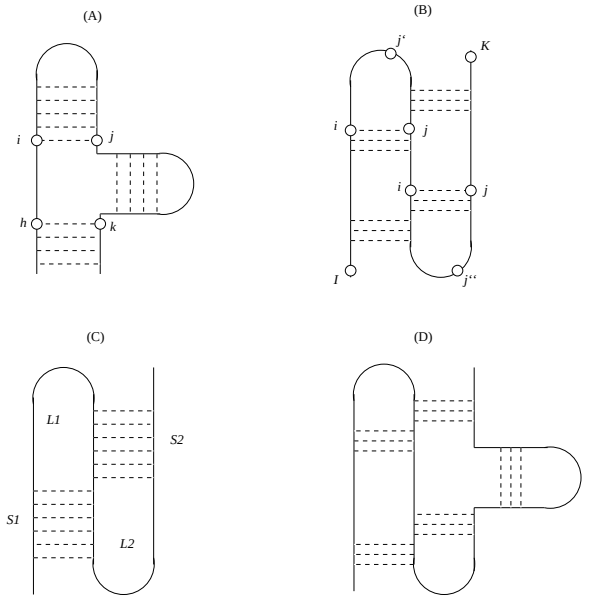

Figure 1: Examples of RNA secondary structure: (A) simple (non-pseudoknotted) structure; (B) pseudoknot; (C) H-pseudoknot (simple pseudoknot); (D) recursive pseudoknot.

A set of base pairs $M_{I,K}$ is a pseudoknot (Fig.1.B) if there exist positions j' and j'', where $(I < j' < j'' < K)$, such that each pair $(i-j) \in M_{I,K}$ satisfies either $I \le i < j' \le j < j''$ or $j' \le i < j'' \le j \le K$, and if pairs $(i-j)$ and $(i_1 - j_1) \in M_{I,K}$ satisfy either $i < i_1 < j'$ or $j' \le i < i_1$, then $j > j_1$ holds.

4 Prediction of RNA folding with pseudoknots

Several studies have been conducted both from a practical or theoretical viewpoints to predict RNA secondary structure with pseudoknots. Most methods, which are capable of predicting pseudoknots adopt different heuristic search procedures, none of which is guaranteed to find an optimal structure. Moreover, current approaches fail to determine how far a given prediction is from optimality given a thermodynamic model. Such approaches include quasi-Monte Carlo search developed by Abrahams et al. (Abraham, Berg, Van Batenburg & Pleij 1990), genetic algorithms by Gultyaev et al. (Gultyaev, van Batenburg, & Pleij 1995), van Batenburg et al. (van Batenburg, Gyltyaev, & Pleij 1995), and by Shapiro and Wu (Shapiro & Wu 1997), methods based on an extension of the stochastic contest-free grammars developed by Brown and Wilson (Brown & Wilson 1996) and on the Hopfield network developed by Akiyama and Kanehisa (Akiyama & Kanehisa 1992).

A different approach to pseudoknot prediction was introduced by Cary and Stormo (Cary & Stormo 1995) and improved by Tabaska et al. (Tabaska, Cary, Gabow, & Stormo 1998). This method is based on maximum weighted matching (MWM) algorithm (Edmonds 1965, Gabow 1976). The method is capable of finding an optimal structure in the presence of complicated pseudoknots in $O(n^3)$ time and $O(n^2)$ space. Unfortunately, the MWM algorithm is suitable only for sequences for which a multiple alignment exists, so that scores may be assigned to possible base-pairs by comparative analysis. It is not clear whether the MWM algorithm can be applied to a single sequence folding using the complicated Turner thermodynamic model.

Rivas and Eddy (Rivas & Eddy 1999) developed a dynamic programming algorithm for predicting optimal (minimum energy) RNA secondary structure, including pseudoknots. The algorithm has the worst case time and space complexities of $O(n^6)$ and $O(n^4)$ respectively. The implementation of the algorithm uses standard RNA folding thermodynamic parameters augmented by a few parameters describing the thermodynamic stability of pseudoknots and by coaxial stacking energies (Walter, Turner, Kim, Lyttle, Muller, Mathews, & Zuker 1994). The description of the algorithm is complex for both nested and non-nested configurations. The key point of their pseudoknot algorithm is the use of one-hole or gap matrices as a generalization of the matrices required for nested configuration.

Uemura et al. (Uemura, Hasegawa, Kobayashi, & Yokomori 1995) proposed an algorithm based on tree adjoining grammar. The time complexities of their algorithm depends on types of pseudoknots: it is $O(n^4)$ for simple pseudoknots and $O(n^5)$ or more for the other pseudoknots. Although, the algorithm can always find optimal structures, tree adjoining grammars are complicated and impractical for longer RNA sequences. Akutsu (Akutsu 2000) analyzed their method and found that tree adjoining grammar was not crucial but the parsing procedure was crucial. Since the parsing procedure is intrinsically a dynamic programming procedure, Akutsu (Akutsu 2000) re-formulated their method as a dynamic programming procedure without tree adjoining grammar.

In this paper, we present, implement and analyze the performance of a dynamic programming algorithm following Akutsu's for predicting simple pseudoknots in optimal secondary structure of a single RNA sequence using standard thermodynamic parameters for RNA folding. The algorithm has the worst case time and space complexities $O(n^4)$ and $O(n^3)$, respectively. We describe the algorithm using diagrams to illustrate its recurrence relations and initialization. We develop implementation for our algorithm to find minimal energy RNA structures using the current RNA structure thermodynamic model (Zuker, Mathews, & Turner 1999). We validate the accuracy of our algorithm on the entire simple pseudoknot collection in the PseudoBase.

5 Nearest-neighbor thermodynamic model

The nearest neighbor energy rules are widely used in RNA structure prediction. The problem of RNA secondary structure prediction using thermodynamic parameters, is defined as a problem of computing RNA secondary structure with minimum free energy. Zuker et al summarized the details of free energy computation (Zuker et al 1999). In the nearest neighbor model, free energies are assigned to loops rather than to base pairs. These rules are also known as loop dependent energy rules.

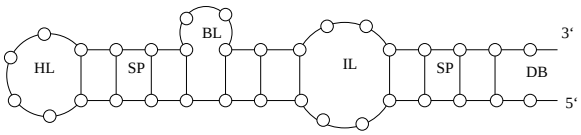

Figure 2: Schematic representation of the most relevant simple RNA secondary structures. The nucleotides of the sequence are represented by circles. A hairpin loop (HL) is a sequence of unpaired bases bounded by one base pair. A stacking pair (SP), a bulge loop (BL) and an internal loop (IL) are all bounded by two base pairs. In a stack, the two base pairs are contiguous at both ends. In a bulge, the two base pairs are contiguous at one end. In an internal loop, the two base pairs are not contiguous. A dangling base (DB) is a single stranded base adjacent to $5'$ or $3'$ end of a base pair.

Any secondary structure, S, uniquely decomposes an RNA molecule into loops. Thus, the total free energy is the sum of free energies of all loops in the structure. Loops may contain zero, one or more base pairs and zero, one or more single-stranded nucleotides. For example, a 1-loop which is called a hairpin loop contains one base pair and at least three single stranded nucleotides, while loop in a stacking region contains two base pairs and no single-stranded nucleotides (Fig.2). Available free energy parameters include values for hairpin loops, bulges, internal loops, terminal mismatched pairs, terminal mismatched pairs for hairpin loops, stacked pairs, multibranch loops, coaxial stacking of adjacent helices and more. Details of energy computation for different kinds of loops are summarized and can be found in the literature (Zuker et al 1999).

6 Pseudoknot Thermodynamics

While thermodynamic parameters for non-pseudoknotted secondary structure are estimated with reasonable accuracy (Zuker et al 1999), there is no systematic study available of pseudoknot thermodynamics. The free energy of an H-pseudoknot structure is mainly the sum of the free energies of stacking in both stems (stabilizing negative values) and the positive destabilizing loop values. The stacking energies can be calculated using the known nearest-neighbor model parameters of helix propagation (Zuker et al 1999). An attempt to estimate the free energy parameters for pseudoknot loops, based on the general theory of polymer loop thermodynamics, is presented in (Gultyaev, van Batenburg, & Pleij 1999). The authors have restricted themselves to simplest H-pseudoknots that contain two loops and not more than one nucleotide at the junction between stems. Without a claim to be very precise, they proposed a set of free energy values for pseudoknot loops. The values are different for L_1 and L_2 and dependent on the length of the loop and the length of the stem that the loop crosses. For minimal loop size at optimal stem length, a value of 3.5 kcal/mol is suggested (a loop of one nucleotide bridging 6 or 7 base pairs in L_1 or two nucleotides bridging three base pairs in L_2). This set of parameters seems to be the best possible in absence of systematic experimental results. The

folding algorithm presented here does not yet take into account these pseudoknot loop free energies. It is not a trivial matter to decide which combination of loops and stems will form the pseudoknot until the pseudoknot is constructed by the traceback algorithm described below. Thus, the pseudoknot loop energy values can be used to re-evaluate folding energies obtained by our algorithm.

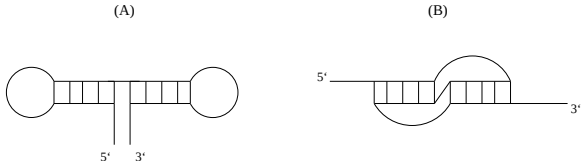

Figure 3: Coaxial stacking of adjacent or almost adjacent helices. (A) simple structure and (B) pseudoknot. Two stems S_1 and S_2 are adjacent helices. There is a stacking interaction between the adjacent closing base pairs of S_1 and S_2. Two helices are almost adjacent when the addition of a single base pair (often non-canonical) results in an adjacent pair.

It has been evident for some time that taking into account the stacking interactions between adjacent helices (Fig.3) improves folding predictions. For helix-helix interfaces, the coaxial stacking is known to have a strong stabilizing effect, about 1 kcal/mol more than the corresponding nearest-neighbor energy in a regular helix (Zuker et al 1999). A similar increase of stacking contribution has been suggested for junctions of pseudoknot stems on the basis of mutational analysis of the pseudoknot in the viral tRNA-like structure (Mans et al 1992). Thus, the bonus of 1 kcal/mol for helixes without intervening nucleotides and the mismatch values for one nucleotide at the junction are assumed to contribute to the free energies of a pseudoknot. *Mfold* does not take into account coaxial stacking of adjacent or almost adjacent helices (Zuker et al 1999). The pseudoknot algorithm presented here takes into account the coaxial stacking of adjacent helices and employ thermodynamic data for adjacent helices reported in (Zuker et al 1999).

7 Dynamic programming algorithm to maximize the number of base pairs in a pseudoknot

A basic version of RNA secondary structure prediction with simple pseudoknots is defined as a problem of finding an RNA secondary structure with simple pseudoknots that has the maximum number of base pairs. An RNA secondary structure prediction with simple pseudoknots which has the maximum number of base pairs can be computed by the dynamic programming algorithm in $O(n^4)$ time and $O(n^3)$ space. Our approach is based on combining a dynamic programming formulation for the basic version of RNA secondary structure prediction with simple pseudoknots presented in (Akutsu 2000) with RNA thermodynamic model (Zuker et al 1999) and dynamic programming formulation for RNA secondary structure prediction without pseudoknots implemented in *mfold* software (Zuker & Stiegler 1981). In order to describe the pseudoknot algorithm we need to introduce three three-dimensional ($N \times N \times N$) matrices, to be called SL, SM, and SR and one triangular $N \times N$ matrix to be called PS. For finding a simple pseudoknot substructure whose endpoints are Ith and Kth residues, the algorithm considers three types of triplets $SL(i,j,k), SM(i,j,k)$, and $SR(i,j,k)$ for each i, j, and k such that $(I \leq i < j < k \leq K)$. These matrices are defined in the following way: $SL(i,j,k)$ is the score of the best folding between positions I and i, and j and k, provided that ith and jth residues make a base pair; $SR(i,j,k)$ is the score of the best folding between positions I and i, and j and k, provided that jth and kth residues make a base pair; $SM(i,j,k)$ is the score of the best folding between positions I and i, and j and k, provided that neither i pairs with j, no j pairs with k. $PS(i,j)$ is the score of the best pseudoknot fold with ending points i and j. Recurrence relations for computing these triplet can be found in (Akutsu 2000). For each pair (I,K) such that $I+6 < K$ we compute the above three matrices and obtain the score of a pseudoknot $PS(I,K)$ with endpoints I and K by

$$PS(I,K) = \max_{I<i<K} \begin{cases} SL(i,i+1,K) \\ SM(i,i+1,K) \\ SR(i,i+1,K) \end{cases}$$

Finally, the optimal score for each pair (i,j) can be computed by the following recurrence:

$$S(i,j) = \max \begin{cases} PS(i,j) \\ \mu(i,j) + S(i+1,j-1) \\ \max_{i \leq m < j} \{ S(i,m) + S(m+1,j) \} \end{cases}$$

8 Dynamic programming algorithm to minimize the free energy of a pseudoknot

We extended the basic version of RNA secondary structure prediction with simple pseudoknots presented above in order to include thermodynamic parameters. Such modification for a secondary structure without pseudoknots is well known and implemented in *mfold* (Zuker et al 1999). *Mfold* is a software package for RNA and DNA secondary structure prediction using nearest neighbor thermodynamic rules (Zuker et al 1999). The updated C++ version of the program was reported in (Mathews, Andre, Kim, Turner, & Zuker 1998). *Mfold* uses four different procedures to compute the energies of specific substructures. Our implementation of pseudoknot free energy computation uses the same set of thermodynamic parameters and employs three *mfold* routines with little or no modifications to computes energies of different kinds of loops composing pseudoknot structures.

8.1 Computation of optimal secondary RNA structure with simple pseudoknots

Our new approach for secondary RNA structure prediction with simple pseudoknots computes the minimum free energy of both simple (non-pseudoknot) substructure and pseudoknot substructure for each RNA segment beginning at position i and ending at position j $(i < j)$. In order to describe the algorithm we need first to introduce two triangular $N \times N$ matrices, to be called $V(i,j)$ and $W(i,j)$. These matrices are defined in the following way: $V(i,j)$ is the score of the best non-pseudoknot folding between positions i and j, provided that i and j can form a base pair; whereas $W(i,j)$ is the score of the best folding between positions i and j regardless of whether i and j pair to each other or not. To recall,

$PS(i,j)$ is the score of best pseudoknot configuration between positions i and j. The energy of simple substructure $V(i,j)$ is computed according to well-known recurrence relations and implemented by *mfold* (Zuker et al 1981). (Fig.4. A, B and C):

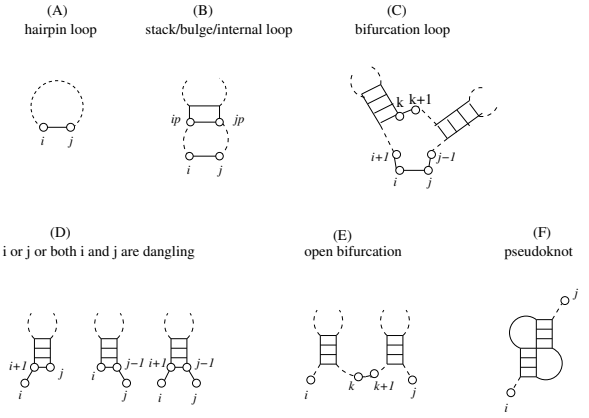

Figure 4: (A), (B) and (C) are three different cases in computation of $V(i,j)$ which is the minimum free energy of a simple structure starting at position i and ending at position j, given that i and j form a base pair. (D), (E), (F) and $V(i,j)$ are four different cases in computation of $W(i,j)$ which is the minimum free energy of a structure possibly containing pseudoknots starting at position i and ending at position j.

$$V(i,j) = \min \begin{cases} hairpin(i,j) \\ \min_{i<ip<jp<j}\{\, V(ip,jp) \\ \quad +loop(i,j,ip,jp)\,\} \\ \min_{i+1<m<j-2}\{\, W(i+1,m) \\ \quad +W(m+1,j-1)\,\} \end{cases}$$

$V(i,j) = \infty$ if i and j can not form a base pair.

The best substructure $W(i,j)$ having minimum free energy is assigned for each (i,j) according the following recurrence relations (Fig.4.):

$$W(i,j) = \min \begin{cases} V(i,j) \\ W(i+1,j) + dangle(i+1,j,i,2) \\ W(i,j-1) + dangle(i,j-1,j,1) \\ W(i+1,j-1) \\ \qquad +dangle(i+1,j-1,i,2) \\ \qquad +dangle(i+1,j-1,j,1) \\ \min_{i<m\leq j}\{\, W(i,m-1) + W(m,j)\,\} \\ PS(i,j) \end{cases}$$

Intuitively, this recursive algorithm works by adding one nucleotide at a time to a sequence, and observing what the best structure is at each step. The last number to be computed, $W(1,n)$, represents the minimum free energy for the whole sequence. It is the minimum energy of an admissible structure on a sequence S. All that remains is the development of the molecule structure. This is achieved by a trace back through the matrices W and V, and matrices SL, SM, and SR when pseudoknot configuration was selected. The traceback procedure for matrices W and V is implemented by *mfold* (Mathews et al 1998).

8.2 Computation of the optimal free energy of a pseudoknot

We extend the basic dynamic programming algorithm presented in (Akutsu 2000) for maximizing the number of base pairs in a pseudoknot and described above to include thermodynamic parameters. Our algorithm computes the minimum free energy of a pseudoknot using nearest-neighbor thermodynamic model (Zuker et al 1999). In order to describe the pseudoknot algorithm we need to introduce three dimensional ($N \times N \times N$) matrices, to be called SL, SM, and SR and two triangular $N \times N$ matrices to be called $stem1$ and $stem2$. For finding a simple pseudoknot substructure with endpoints Ith and Kth residues, the algorithm considers three types of triplets $SL(i,j,k), SM(i,j,k)$, and $SR(i,j,k)$ for each i, j, and k $(I \leq i < j < k \leq K)$. These matrices are defined in the following way: $SL(i,j,k)$ is the energy of the best folding between positions I and i, and j and $k-1$ including energy of the loop between i and j, provided that ith and jth residues make a base pair; $SR(i,j,k)$ is the energy of the best folding between positions I and i, and j and $k-1$ including energy of the loop between i and $j+1$, provided that jth and kth residues make a base pair; $SM(i,j,k)$ is the energy of the best folding between positions I and i, and j and k including energy of the loop between i and $j+1$, provided that neither i pairs with j, no j pairs with k; $stem1(i,j)$ is the free energy of a folding between positions I and i, and j and $k-1$, including energy of the loop between i and j; and $stem2(j,k)$ is the free energy of a folding between positions j and k. The values in $stem1$ and $stem2$ compose the energy of a pseudoknotted structure (i,j,k) such that

$$stem1(i,j) + stem2(j,k) = \min \begin{cases} SL(i,j,k), \\ SM(i,j,k), \\ SR(i,j,k) \end{cases}$$

Both matrices $stem1$ and $stem2$ are used in the computation of matrices SL, SR, and SM. The values of $stem1$ and $stem2$ are assigned according to the minimum energies chosen for $SL(i,j,k)$ and $SR(i,j,k)$, respectively. All five matrices are computed recursively. Details of these computations are omitted for brevity. In the following section we show a few recurrence relations.

8.2.1 Computation of matrix SL)

When ith and jth residues form a base pair, the value of $SL(i,j,k)$ can be obtained in a number of ways. First, the pair $(i-j)$ can close a hairpin loop, then it can stack on the pair $((i-1)-(j+1))$ or with some other pair $(ip-jp)$ can close a bulge or an internal loop. So we can write $SL(i,j,k) = \min\{\, E_1, E_2\,\}$, where E_1 corresponds to a hairpin case in S_1, and E_2 corresponds to the best energy among all possible stack, bulge or internal loop configurations in S_1 (Fig.5).

$$\begin{aligned} E_1 &= hairpin(i,j) + stem2(j+1,k) \\ E_2 &= \min_{I\leq i,\, i+4\leq j<jp<k} \{\, hairpin(i,j) \\ &\quad - hairpin(ip,jp) + loop(ip,jp,i,j) \\ &\quad + SL(ip,jp,k)\,\} \end{aligned}$$

If the hairpin conformation appears to be more favorable $(E1 < E2)$, then $stem1(i,j)$ will store the energy of the hairpin structure, i.e.

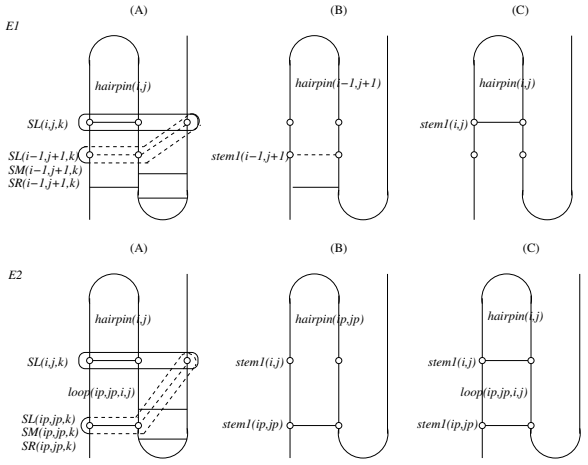

Figure 5: Illustration of the recurrence of the new algorithm for SL computation given that i and j form a base pair. $SL(i,j,k) = \min\{ E_1, E_2 \}$, where E_1 is the minimum energy of a hairpin configuration in S_1, and E_2 is the minimum energy of a stack/bulge/internal loop configuration in S_1. (A) illustrates the recurrence for $SL(i,j,k)$, (B) illustrates the recurrence for $stem1(i,j)$

$$stem1(i,j) = hairpin(i,j)$$

Otherwise, it will store the best energy of the stack/bulge/internal loop structure:

$$\begin{aligned} stem1(i,j) \quad &= hairpin(i,j) - hairpin(ip,jp) \\ &+ loop(ip,jp,i,j) + stem1(ip,jp). \end{aligned}$$

There is a special case possible in SL computation given that i and j form a base pair. When the value of $stem2(j+1,k) = \infty$, E_1 should be computed as follows:

$$E_1 = hairpin(i,j) + penalty$$

$penalty$ is a positive constant assigned to any non-pseudoknot substructure containing only one hairpin loop. Section "Results" includes a short overview on the meaning and application of $penalty$ constant.

When ith and jth residues can not form a base pair the values of SL and $stem1$ can be computed as follows:

$$SL(i,j,k) = \min \begin{cases} SL(i-1,j+1,k) \\ SR(i-1,j+1,k) \\ SM(i-1,j+1,k) \end{cases}$$

$$stem1(i,j) = \min \begin{cases} stem1(i-1,j) \\ stem1(i,j+1) \end{cases}$$

8.3 Complexity of pseudoknot computation

For each pair (I,K) we must compute scores for $O(n^3)$ triplets. Therefore, scores for $O(n^5)$ triplets must be computed. Since K does not appear in the recurrence relations, the score for any triplet depends on I only but not K. Also, scores for (i,j,k) do not depend on scores for (i',j',k'), where $k < k'$. Therefore, we need to compute only $O(n^3)$ scores for each I. Sinceeach score can be computed in constant time, $O(n^4)$ time is sufficient in total. $O(n^3)$ space is required to compute all $PS(I,K)$ for a fixed I. The same space can be reused to compute all $PS(I,K)$ for all Is. Since $O(n^2)$ space is sufficient for storing all values of $PS(I,K)$, the total space complexity is $O(n^3)$.

9 Implementation

The algorithm was implemented in C++ using principles of object-oriented programming. The program runs in a Unix environment on Sun Solaris platform. The input to the program consists of a single RNA sequence and thermodynamic data. The output consists of a list of base pairs. In addition, each pseudoknot configuration has an output in the following format :(((:::[[[[)):::]]]]:. One of our future plans is to provide a graphical output of secondary structure for an entire RNA molecule.

10 Experimental Evaluation and Results

Our algorithm finds optimal RNA structures for single sequences that might contain simple pseudoknots by dynamic programming. We implemented the algorithm for energy minimization, extending *mfold* to pseudoknotted structures. It is well known that RNA structure prediction for single sequences without pseudoknots already involves some inaccuracy due to imperfection of termodynamic parameters and limited knowledge on other factors affecting RNA folding. For example, the most popular software for RNA structure prediction without pseudoknots, *mfold* , includes a variety of parameters that may be adjusted to obtain better results, i.e. more accurate predictions (Zuker et al 1999). Our approach obviously inherits this inaccuracy and may possibly introduce still more, since it allows a much wider space of all possible configurations of RNA molecules. Our algorithm for folding of pseudoknots is also sensitive to the accuracy of the existing thermodynamic parameters. We validate the accuracy of our aproach in three consecutive steps. First, we want to show that our program can accurately fold the structures of known pseudoknots, second, that the program predicts those pseudoknots as structures with minimal free energies, third, that the program does not introduce spurious pseudoknots and fourth, that it outputs results similar to *mfold*.

To evaluate our approach we use the set of sequence data collected in the PseudoBase. The accuracy of most structures in the PseudoBase is supported by sequence comparison only which may introduce some variation with the predicted structures. Some pseudoknot structures are supported by structure probing and/or mutagenesis. We examined the entire collection and downloaded those sequences which contain simple pseudoknots. The resulting set consists of 169 sequences with pseudoknots of variable size from 19 to 114 nucleotides. We tested our program on the entire set and found that the program folds 163 pseudoknots and 6 simple structures. Among those 163 pseudoknots the structure of 131 pseudoknots was predicted correctly or almost correctly. Examples of the accuracy of the predictions are shown in Table 1.

For all these sequences predicted by our program to adopt a pseudoknotted cofiguration, the predicted pseudoknots have lower free energy than simple structure configurations. Thus, our program predicts pseudoknots with correct or almost correct structure for 78% sequences, which we believe is much better than existing results. Moreover, futher improvement in accuracy is possible and we are currently working in this direction.

(A)	squenceID	BSBV1
	Sequence	CCCCUUUACUUGAGGGAAAUCAAGC
	PseudoBase	:(((::::[[[[[)))::::]]]]]:
	Prediction	:(((::::[[[[[)))::::]]]]]:
(B)	squenceID	SBWMV2
	Sequence	CCCCCAUCCGGAGGGUUAUCCGGC
	PseudoBase	:(((:::[[[[[)))::::]]]]]:
	Prediction	::(((::[[[[[)))::::]]]]]:
(C)	sequenceID	HiPV_IRES-PK1
	Sequence	CAGCCUUGUAGUUUUAGUGGACUUUAGGCUAAAGAAUUUCACUAG
	PseudoBase	:(((((:::((((:[[[[[)))))::)))))::::::::::]]]]]:
	Prediction	:(((((::::::::[[[[[[[[:[[[)))))):::]]]:]]]]]]]]
(D)	sequenceID	CcTMV_UPD-PK2
	Sequence	UAGGGGCUUACCGAAAUAAGCC
	PseudoBase	:((((:[[[[[)))::::]]]]]:
	Prediction	:::::[[[[[::::::]]]]]:

Table 1: Some typical outputs of the predictions. Under each sequence is the correct pseudoknotted structure given in the database followed by the predicted structure. Examples of correct (A), almost correct (B), different pseudoknotted (C), and simple (D) structures are shown. For almost correct structure up to 4 positions are allowed to differ.

For three out of six sequences for which simple structure was predicted the accuracy of prediction can be improved by adjusting the value of a parameter which we call $penalty$. It is a positive constant assigned to any substructure containing only one hairpin loop. It was introduced in order to build a bias in the algorithm in favor of a structure with two hairpin loops rather than one hairpin loop. The $penalty$ value never contributes to the final minimum energy of the pseudoknot computed. We ascertain the $penalty$ value empirically based on our experiments.

In order to evaluate out approach we compared the effeciency and accuracy of our software to that of the currently best known software developed by Rivas (Rivas et al 1999). Rivas algorithm has worst case time ans space complexities of $O(n^{6.8})$ and $O(n^4)$, respectively. When run on the same set of 169 pseudoknotted sequences the Rivas program was able to predict only 50% of pseudoknots, while our software predicts 95 % of pseudoknots, and 78 % of pseudoknots are predicted withcorrect or almost correct structures.

Finally, we establish that out algorithm is more efficient by several orders for both small and large RNA sequences. For example, a sequence of 75 nucleotides takes 55 seconds (compared to 20 minutes with the existing software), and a sequence of 114 nucleotides takes 8 minutes (compared to 4 hours 30 minutes). To our knowledge, our approach leads to currently most efficient and accurate algorithm for predicting simple pseudoknots in optimal secondary structure of a single RNA sequence. Furhter optimization of the algorithm may provide the means to mine whole genome sequences to explore possible novel roles of RNA pseudoknots in biology.

11 Acknowledgments

The authors would like to acknowledge the contribution of Mr. Zhaohui Sun who introduced the object-oriented design into the implementation of *mfold* algorithm.

References

Abrahams JP, Berg M, van Batenburg E, Pleij CWA, Prediction of RNA secondary structure, including pseudoknots by computer simulation, *Nucleic Acids Res.* **18** 3035-3044.

Akiyama Y, Kanehisa M, NeuroFold: an RNA secondary structure prediction system using a Hopfield neural networks, *Proceedings of the Genome Informatics Workshop III*, Universal Academy Press, Tokyo, 199-202 (in Japanese).

Akutsu T, Dynamic programming algorithm for RNA secondary structure prdiction with pseudoknots, *Discrete Appl. Math.* **104** 45-62.

van Batenburg FDH, Gyltyaev AP, Pleij CWA, An APL-programmed genetic algorithm for the prediction of RNA secondary sreucture, *J Theor Biol* **174** 269-280.

Brannan CI, Dees EC, Ingram RS, and Tilghman SM. The product of the h19 gene may function as an RNA. Mol Cell Biol **10** 28-36.

Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M,Tonlorenzi R, and Willard HF. A gene from the region of the human X inactivation center is expressed exclusively from the inactive X chromosome. Nature **349** 38-44.

Brown M, Wilson C, RNA pseudoknots modeling using intersections of stochastic contex free grammars with applications to database search, in: L.Hunter, T.E.Klein (Eds.), *Pacific Symposium on Biocomputing'96*, World Sientific, Singapore, 109-125.

Cary R, Stormo GD, Graph-theoretic approach to RNA modeling using comparative data. In *ISMB-95* (Rawling, C., *et al.*, eds), AAAI Press, 75-80.

Cech TR. Self-splicing of group I introns. *Ann Rev Biochem* **59** 543-568.

De Smit MH and van Duin J. Control of prokaryotic translation initiation by mRNA secondary structure. *Progress in Nucleic acid research in Molecular Biology* **38** 1-35.

Edmonds J, Maximum matching and polyhedron with 0, 1-vertices, *J Res Nat Bur Stand* **69B** 125-130.

Gabow HN, An efficient implementation of Edmonds' algorithm for maximum matching on graphs, *J Asc Com Mach* **23** 221-234.

Gultyaev AP, van Batenburg FH, Pleij CWA, The computer simulation of RNA folding pathway using a genetic algorithm, *J Mol Biol* **250** 37-51.

Gultyaev AP, van Batenburg FHD, Pleij CWA. An approximation of loop free energy values of RNA H-pseudoknots. *RNA* **5** 609-617.

Mans RMW, van Steeg MH, Verlaan PWG, Pleij CWA, Bosch L. Mutational analysis of the pseudoknot in the tRNA-like structure of turnip yellow mosaic virus RNA: Aminoacylation efficiency and pseudoknot stability. *J Mol Biol* **223** 221-257.

Mathews DH, Andre TC, Kim J, Turner DH, Zuker M. An updated recursive algorithm for RNA secondary structure prediction with improved free energy parameters., chapter 15, pages 246-257. *American Chemical Society Symposium Series* 682. American Chemical Society, Washington, DC.

Mills DR, Priano C, Merz PA, and Binderow BD. Qβ RNA bacteriophage: mapping cis-acting elements within an RNA genome. *J Virol* **64** 3872-3881.

Pleij CWA, Rietveld K and Bosch L. A new principle of RNA folding based on pseudoknotting. *Nucleic Acid Res* **13** 1717-1731.

Rivas E, Eddy SR, A dynamic programming algorithm for RNA structure prediction including pseudoknots, *J Mol Biol* **285** 2053-2068.

Shapiro BA, Wu JC. Predicting RNA H-Type pseudoknots with the massively parallel genetic algorithm. *Comput Appl Biosci* **13** 459-471.

Tabaska JE, Cary RB, Gabow HN, Stormo GD, An RNA folding method capable of identifying pseudoknots and base triples, *Bioinformatics* **8** 691-699.

Uemura Y, Hasegawa A, Kobayashi S, and Yokomori. Grammatically modeling and predicting RNA secondary structures. In: M. Hagiya et al. (eds.), *Proceedings of the Genome Informatics Workshop*, Universal Academy Press, Tokyo, pp.67-76.

Walter A, Turner D, Kim J, Lyttle M, Muller P, Mathews D, Zuker M, Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding, *Proc Natl Acad Sci* **91** 9218-9222.

Zuker M and Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acid Res* **9** 133-148.

Zuker M, Mathews DH, Turner DH. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In RNA Biochemistry and Biotechnology, J. Barciszewski & B.F.C. Clark, eds., *NATO ASI Series*, Kluwer Academic Publishers. Up- to-date reproduction of the article is available at http://www.bioinfo.rpi.edu/˜zukerm/seqanal/.