Volume 54
Numbers 4-5
August–October 2005
ISSN 0233-1934
Optimization
A JOURNAL OF MATHEMATICAL PROGRAMMING AND OPERATIONS RESEARCH
Special Issue:
In Celebration of Prof. Dr. V. F. Demyanov's 65th birthday
Guest Editors:
Franco Giannessi, Diethard Pallaschke and Alexander M. Rubinov
Taylor & Francis

Optimization
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713646500

A new method for prediction of RNA secondary structure with pseudoknots, based on helix removal and refinement

Bayarbaatar Amgalan[a]; Julian Lee[a]
[a] Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743, Korea

To cite this Article Amgalan, Bayarbaatar and Lee, Julian(2009) 'A new method for prediction of RNA secondary structure with pseudoknots, based on helix removal and refinement', Optimization, 58: 7, 861 — 869

To link to this Article: DOI: 10.1080/02331930902944986

URL: http://dx.doi.org/10.1080/02331930902944986

PLEASE SCROLL DOWN FOR ARTICLE

Optimization
Vol. 58, No. 7, October 2009, 861–869

A new method for prediction of RNA secondary structure with pseudoknots, based on helix removal and refinement

Bayarbaatar Amgalan and Julian Lee*

Department of Bioinformatics and Life Science, Soongsil University, 1-1 Sangdo5-dong, Dongjak-gu, Seoul 156-743, Korea

(*Received 20 October 2007; final version received 21 May 2008*)

RNA secondary structure is predicted by computing the structure with the minimum free energy. Although RNA structure without pseudoknots can be found using dynamic programming algorithms, finding general structures with pseudoknots is a nondeterministic polynomial-time (NP) hard problem. Several methods, such as recursive simple pseudoknots, have been developed in the past for obtaining a conformation with globally minimal energy among the restricted class of pseudoknots. In this work, we develop a new method for approximating a conformation with low energy, posing no restrictions on type of pseudoknots contained in the RNA secondary structure. In our method, the low-energy RNA secondary structure is obtained by repeatedly removing helices and performing dynamic programming to obtain the structure with energy lower than that obtained in the previous iteration. This method can be considered as a local minimization, and can be combined with any global optimization method that takes advantage of local minimization. We tested performance and convergency of the method by predicting a secondary structure of several RNA sequences, which is *a priori* known to contain pseudoknots.

Keywords: RNA; secondary structure; pseudoknot

AMS Subject Classifications: 92B05; 90C10; 90C39

1. Introduction

1.1. *The central dogma of molecular biology*

The flow of (genetic) information in the cell is described by the central dogma of molecular biology. The genome (DNA) encodes the sequence information for all the proteins synthesized by the cell. A segment of DNA holding construction information of a protein is called a protein coding gene. However, DNA is not the direct template for protein synthesis, since it is transcribed by an enzyme, due to RNA polymerase, into a messenger RNA (mRNA) carrying the same information as the transcribed gene.

*Corresponding author. Email: jul@ssu.ac.kr

ISSN 0233–1934 print/ISSN 1029–4945 online
© 2009 Taylor & Francis
DOI: 10.1080/02331930902944986
http://www.informaworld.com

The mRNA is then translated into protein by the ribosome. A gene can also be transcribed into an RNA that is never translated into a protein. Such genes are called non-coding genes and the RNAs are called functional RNAs, as they are not translated into protein, but they have functions themselves.

1.2. *RNA*

An RNA is one of the two type of nucleic acids (deoxyribo nucleic acid (DNA) and ribonucleic acid (RNA)) found in living organisms. RNA plays many roles within cells. An RNA molecule is a sequence of nucleotides, or bases of four possible types, denoted by the letters A, C, G and U (for Adenine, Cytosine, Guanine and Uracil) connected by a backbone. The function of an RNA within the cell is determined in large part by the three-dimensional structure of the RNA molecule, when it folds. In turn, the three-dimensional structure is partly determined by the secondary structure of the molecule. The secondary structure is simply a list of the bonds that are formed between the individual bases within the molecule. Determining the secondary structure of an RNA molecule is an integral part of understanding the function of the RNA molecules. Several secondary structure prediction methods are available today, that predict secondary structure from a single RNA sequence or from a set of homologous RNA sequence. The methods that predict structure from a single sequence usually maximize the number of base pairs [1] or minimize the free energy [2–8]. RNA structure prediction is a typical global optimization problem [9]. Although RNA structure without pseudoknots can be found using dynamic programming algorithms, finding general structures with pseudoknots is a nondeterministic polynomial-time (NP) hard problem. In this work, we develop a new method for approximating a conformation with low energy, posing no restrictions on the type of pseudoknots contained in the RNA secondary structure.

2. RNA secondary structure

2.1. *RNA secondary structure without pseudoknots*

A single-stranded RNA folds into a functional shape by forming intramolecular base pairs among some of its bases. The set of these base pairs is known as the secondary structure of RNA. For a molecule with n nucleotides, we index the nucleotides from 1 to n, starting at the $5'$ end. We write $i.j$ if the nucleotide with index i is paired with the nucleotide with index j, and $i < j$. Then a secondary structure R is a set of base pairs such that if $i.j$ and $i'.j'$ are distinct base pairs in R, then i, j, i', and j' are distinct.

Definition 2.1 (RNA sequence) An RNA sequence is a string $s \in \sum^*$, $s = r_1 r_2 \dots r_n$, $r_i \in \sum$, where alphabet is $\sum = \{A, C, G, U\}$ Each of the r_i is called a base with position i in the sequence. The length of s, noted as $|s|$, is n.

Definition 2.2 (Base pair) A base pair in an RNA sequence, typed as (i,j), $i \neq j$, is a pair of bases with positions i and j in the sequence.

In RNA the bases corresponding to the positions i and j are usually the Watson Crick base pairs (G, C) and (A, U) and the non-canonical base pair (G, U). Informally, a secondary structure is a collection of matching base pairs. Formally, we can define it as follows.

Definition 2.3 (RNA secondary structure) A secondary structure of an RNA sequence S is a set, R, of ordered base pairs (i,j), where $\forall(i,j) \in R : 1 \leq i < j \leq n$, and n is the length of the sequence S. Furthermore, there is the following restrictions on the set R of base pairs:

(1) $j - i > 3$, for all base pairs (i,j) in R.
(2) A base i can be part of at most one pairing (i,j),

i.e. if (i,j) and (i,j') are two base pairs in R, then $j = j'$.

A pseudoknot is special substructure in the RNA secondary structure, which consists of base pairs which cross over each other in the sequence. Some RNA sequences have natural occurrences of this substructure, and therefore it is also of practical interest to model them.

Definition 2.4 (RNA secondary structure without pseudoknots) A secondary structure R (without pseudoknots) of an RNA sequence S is the same as Definition 2.3, with an extra restriction on base pairs: For all base pairs (i,j) and (i',j'), assume $i < i'$, in $R : \neg(i < i' < j < j')$,

i.e. the positions j and i' must not cross each other.

When RNA is folded, some bases remain unpaired, forming loops in the molecule. A hairpin loop contains exactly one base pair. An internal loop contains exactly two base pairs. A bulge is an internal loop with one base from each of its two base pairs adjacent on the backbone. A stacked pair is a loop formed by two adjacent base pair $i.j$ and $(i+1).(j-1)$, thus they have both ends adjacent on the backbone. A multi-branched loop is a loop that contains more than two base pairs. An external base is a base not contained in any loop (Figure 1).

Another representation of a secondary structure is an arc diagram. The arc diagram of the structure of Figure 1 is given in Figure 2. In an arc diagram, points on a line represent nucleotides in order from the 5′ end, and arcs represent the base pairs.

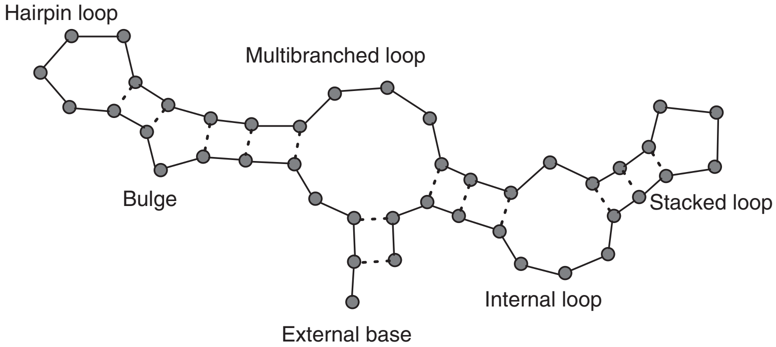

Figure 1. RNA secondary structure without pseudoknots.

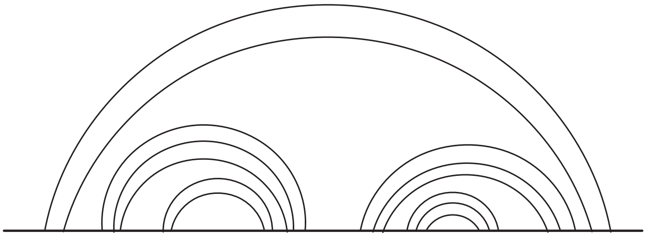

Figure 2. Arc diagram for Figure 1.

Throughout the thesis, we use n to denote the length of a strand. We assume that the bases of the RNA molecule are numbered from 1 to n, starting from the $5'$ end and finishing at the $3'$ end.

The substructures are described further. But first, it should be defined which base pairs are accessible from each other:

Definition 2.5 (Accessibility and loops) A base k is accessible from a base pair (i,j) if $i < k < j$ and there is no base pair between (i,j) and k, i.e. no base pair (m,n) such that $i < m < k < n < j$ and k is accessible from (m,n). The loop closed by (i,j) is the set of all accessible bases from the base pair (i,j). An interior base pair in a loop closed by (i,j) is a base pair (m,n) where both m and n are accessible from (i,j).

Note, that in a loop no bases can be accessible from an interior base pair by definition. The various parts of Figure 1 are described below:

Hairpin loop: A hairpin loop contains one closing base pair and all the bases between the paired bases are unpaired. The hairpin marked in Figure 1 contains four free bases. Formally, the tuple (i,j) defines a hairpin loop in a given secondary structure if i and j are paired, and k is a free base, $\forall k,\ i < k < j$.

Stacked loop: A stacked loop, also called stacked pair, contains two consecutive base pairs. The tuple (i,j) defines a stacked pair if i and j are paired and $i+1$ and $j+1$ are paired. A stem or helix is made of a consecutive number of stacked loops.

Internal loop: An internal loop, sometimes called interior loop, is a loop having two closing base pairs, and all bases between them are free. The 4-tuple (i,j,i',j'), with $i+1 < i' < j' < j-1$, defines an internal loop if i and j are paired, i' and j' are paired, and k is a free base, $\forall k,\ i < k < i'$ and $j' < k < j$.

Bulge loop: A bulge loop, or simply bulge, is a special case of an internal loop, which has no free base on one side, and at least one free base on the other side. Note that, in fact, a stacked loop is also a special case of an internal loop, with no free bases on either side. In this work, we will consider stacked loops and internal loops to be distinct structures, but we include bulges in the internal loop case, unless otherwise specified.

Multi-branched loop: A multi-branched loop, or multi-loop is a loop which has at least three closing base pairs. The 2(m + 1)-tuple $(i,j,i_1,j_1,\ldots,i_m,j_m)$, with $m \geq 2$, $i < i_1 < j_1 < \cdots < i_m < j_m < j$ defines a multi-loop with $m+1$ branches if i pairs with j, i_1 pairs with $j_1,\ldots,i_m$ pairs with j_m and k is a free base, $\forall k,\ i < k < i_1$, $j_1 < k < i_2,\ldots,j_m < k < j$.

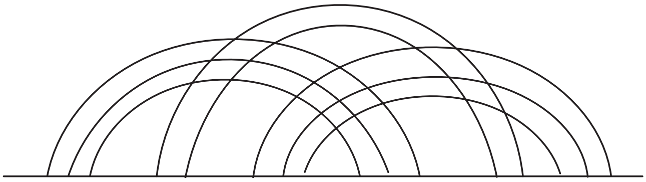

Figure 3. The arc diagram for a structure with pseudoknots.

2.2. *Secondary structure with pseudoknots*

A folding of RNA secondary structure such as that shown in Figure 3 is a pseudoknotted structure. A pseudoknot is an RNA secondary structure containing two-stem loop structures in which the first stem's loop forms part of the second stem.

Definition 2.6 (Hairpin pseudoknot) Two base pairs (i,j) and (k,l), ordered so that $i<k$, are said to form part of a pseudoknot if $i<k<j<l$, i.e. the two base pairs cross over each other in the sequence.

Usually, several base pairs will be involved in a cross matching (a pseudoknot). In an arc diagram of a secondary structure with pseudoknot, at least one arc crosses another arc in the structure. More complicated types of loop appear in pseudoknotted structures.

3. Prediction of RNA secondary structure

3.1. *Prediction of RNA secondary structure without pseudoknots*

The quickest and easiest route to RNA structure prediction is through the use of simple energy rules. One way is to assign an energy $e(a,b)$ to base-pairing of acid types a and b in a secondary structure. The free energy of a secondary structure without pseudoknot for sequence S is then given by:

$$E(S)=\sum_{i,j\in S} e(r_i,r_j).$$

Reasonable values of e are -3, -2 and $-1\,\mathrm{kcal/mole}^{-1}$ for GC, AU and GU base pairs, respectively. For base pair dependent energy rules, the minimum energy can be easily obtained by a simple dynamic programming that recursively computes minimum energy for sequence segment $i\cdots j$ [10]:

$$E(i,j)=\min\begin{cases}0 \quad \text{if} \quad j-i<4 \\ \min\begin{cases}E(i+1,j),E(i,j-1), \\ E(i+1,j-1)+e(i,j), \\ \min_{i\le k\le j}(E(i,k)+E(k+1,j))\end{cases} & \text{otherwise.}\end{cases}$$

The minimum energy of the whole sequence is given by $E(i,n)$, and corresponding secondary structure is obtained by dynamic programming algorithm. Segments of length ≤ 4 have 0 folding energy, since they cannot fold. Bases i or j either do not

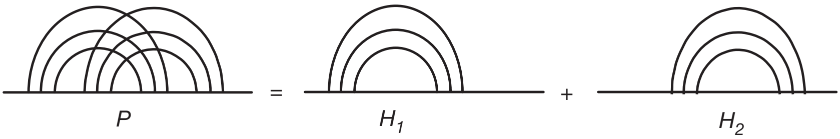

Figure 4. Representation for pseudoknot.

pair, or else they pair with some bases $k_1 < k_2$, respectively, so that the structure splits in 2, or else i and j pair with each other.

3.2. *Prediction of RNA secondary structure with pseudoknots*

We extend the basic dynamic programming algorithm to secondary structure containing pseudoknots. An idea is to consider pseudoknot as an interaction between two separated helices and loops (Figure 4). Starting from a structure without pseudoknot, if we remove a helix and run the dynamic programming again, the helices rearrange themselves to obtain conformation with lower energy, which was originally forbidden due to the removed helix. We put the removed helix back to its place, and we get a secondary structure with pseudoknot that has lower energy than the original conformation. Starting from the optimal secondary structure without pseudoknot, one can repeat the procedure of removing helix and running dynamic programming until there is no more helices left to remove. Then all the removed helices are placed back to their original places.

The Algorithm for Pseudoknot: An additional parameter, $H^k_{\min}(p)$, describes the position of the helix that has minimum energy in the k-th iteration,

Input: RNA sequence S, $S_1 = S$, $k = 1$.

Step 1 Build two matrices e^k and E^k using S_k sequence.

Step 2 Run the dynamic programming algorithm to predict a secondary structure as usual.

$$E^k(r_i, r_j) = \min \begin{cases} 0 \quad \text{if } j - i < 4 \\ \min \begin{cases} E^k(r_{i+1}, r_j), E^k(r_i, r_{j-1}), \\ E^k(r_{i+1}, r_{j-1}) + e^k(r_i, r_j), \\ \min_{i \le l \le j}(E^k(r_i, r_l) + E^k(r_{l+1}, r_j)) \end{cases} \quad \text{otherwise} \end{cases}$$

Step 3 Identify all helices in S_k and classify helices separated by internal and bulge loops. If there is no base-pair identified, report list R and terminate the computation.

Step 4 Select the helix $H^k_{\min}$ that has minimum energy

$$E(H^k_{\min}) := \min\{E(H^k_1), E(H^k_2), \ldots, E(H^k_m)\}$$

and put $H^k_{\min}(p)$ into the base-pairs list R to be reported.

Step 5 Remove position of $H^k_{\min}$ from the initial sequence

$$S_{k+1} := S_k \setminus H^k_{\min}(p)$$

$k := k+1$ and go to Step 1.

Output: Predicted structure list R. The fact that energy of the structure is lowered at each iteration is formally proved as follows.

THEOREM 3.1 *Let $E(m)$ be the energy of sequence S after m-th iteration of helix removal and dynamic programming, where we omitted the symbol S for simplicity of the notation. Then $E(m+1) \le E(m)$.*

Proof The position of helix is removed from the initial sequence in the every iteration.

Hence, we can write

$$S = S_1 \supset S_2 \supset \cdots \supset S_m.$$

Let S_k be the sequence left after removing helices $H^1_{\min}, \ldots, H^{k-1}_{\min}$ at k-th iteration and $V(S_i)$ be the minimum energy obtained by dynamic programming performed on the sequence segment S_i.

Since we can write $E(k) = \sum_{i=1}^{k-1} E(H^i_{\min}) + V(S_k)$ such that $E(H^k_{\min}) \le 0$, $\forall k$, $1 < k < m$, we have

$$E(k) - E(k+1) = V(S_k) - V(S_{k+1}) - E(H^k_{\min}) \ge V(S_k) - V(S_{k+1}).$$

Since $S_k \supset S_{k+1}$, we have $V(S_k) - V(S_{k+1}) \ge 0$ and $E(k) \ge E(k+1)$ which completes the proof. ■

Note that our algorithm does not guarantee finding the secondary structure corresponding to the global minimum of the energy function. We guarantee that after each iteration conformation with pseudoknots has been found and folding energy is decreased. Therefore our algorithm can be considered as a local minimization algorithm. Our algorithm is reasonably efficient in terms of space and time complexity, as shown by the following theorem.

THEOREM 3.2 *An RNA secondary structure with pseudoknots which has the minimum energy of base-pairs can be computed by the dynamic programming algorithm with the worst case time complexity is lower than $O(n^4)$ and using $O(n^2)$ space.*

Proof The basic dynamic programming algorithm [10] that takes $O(n^3)$ in time complexity and $O(n^2)$ in space complexity is repeated m times, where m is the total number of helices predicted by the algorithm, and since $m \le n-6$ the proof is immediate from the following inequality

$$mO(n^3) \le (n-6)O(n^3) = nO(n^3) - 6O(n^3) = O(n^4) - 6O(n^3) < O(n^4).$$

■

4. Test results

To evaluate our approach we use the set of sequence data collection in the PseudoBase, consisting of 50 sequences with pseudoknots of variable size from 26 to

98 nucleotides. We tested our program on the entire set and found that the program folds 45 pseudoknots and 5 simple structures. Among those 45 pseudoknots the structure of 39 pseudoknots were predicted correctly or almost correctly. Examples indicating accuracy of the predictions are presented in the following table.

1	PKB	PKB74 (bacteriophage T4)
	Sequence	UGCCAGCUAUGAGGUAAAGUGUCAUAGC
	PseudoBase	((((:[[[[[[[)))):::::]]]]]]]
	Prediction	.((((.([[[[[......)))]]]]]))
2	PKB	PKB49 (E.coli)
	Sequence	CGAGGGGCGGUUGGCCUCGUAAAAAGCCGC
	PseudoBase	(((((:[[[[[[::))))):::::]]]]]]
	Prediction	.[.(((((([[[[.))))).[...]]]]].]
3	PKB	PKB95
	Sequence	CCGUGGCGAGUACGAUAACUCGUA
	PseudoBase	:(((:[[[[[[))):::]]]]]]:
	Prediction	.(((..[[[[[)))...]]]]]..
4	PKB	PKB108 (Tetrahymena thermophila)
	Sequence	UAUAACCUUCACCAAUUAGGUUCAAAUAAGUGGUA
	PseudoBase	:::((((:::[[[[[[[[))))::::]]]]]]]]:
	Prediction	(((...[[[.)((((((.(...).)))]]])))))

5. Discussion

In this work, we developed a new method for finding a RNA secondary structure with low energy without any restriction on types of pseudoknots. The low-energy RNA secondary structure is obtained by repeatedly removing helices and performing dynamic programming to obtain the structure with energy lower than that obtained in the previous iteration. Although we considered only base-pairing energy, the idea can be generalized to the case where there are destabilization energies associated with loops. Also, since this algorithm is a local minimization algorithm, it should be combined with efficient global optimization algorithm that takes advantage of a local minimization, in order to efficiently find the global minimum energy conformation. All these issues are left for the future study.

Acknowledgement

This work was supported by the Soongsil University Research Fund.

References

[1] R. Nussinov, G. Piecznik, R.G. Jerrold, and J.K. Daniel, *Algorithms for loop matching*, SIAM J. Appl. Math. 35 (1978), pp. 68–82.

[2] T. Akutsu, *Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots*, Discrete Appl. Math. 104 (2000), pp. 45–62.

[3] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster, *Fast folding and comparison of RNA secondary structures*, Monatsh Chemie 125(2) (1994), pp. 167–188.

[4] R. Nussinov and A. Jacobson, *Fast algorithm for prediction for predicting the secondary structure of single-stranded RNA*, The Proceedings of the National Academy of Sciences USA 77 (1980), pp. 6309–6313.

[5] E. Rivas and S. Eddy, *A dynamic programming algorithm for RNA structure prediction including pseudoknots*, J. Mol. Biol. 285 (1999), pp. 2053–2068.

[6] E. Rivas and R. Eddy, *The language of RNA: A formal grammar that includes pseudoknots*, Bioinformatics 16 (2000), pp. 334–340.

[7] M. Zuker, *Computer prediction of RNA structure*, Meth. Enzymol. 180 (1989), pp. 262–283.

[8] M. Zuker and P. Stiegler, *Optimal computer folding of large RNA sequence using thermodynamics and auxiliary information*, Nucleic Acid Res. 9 (1981), pp. 133–148.

[9] A. Chinchuluun, P.M. Pardalos, and R. Enkhbat, *Global minimization algorithms for concave quadratic programming problems*, Optimization 54(6) (2005), pp. 627–639.

[10] J. Setubal and J. Meidanis, *Introduction to computational molecular biology*, in *Biological Sequence Analysis*, Chapter 8 and 10, R. Durbin et al., eds.