

On Using Tabu Search for RNA Secondary Structure Prediction

^{*1,2} Yongguo Liu and ¹Jianrui Hao

^{*1, Corresponding Author} School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu 611731, P. R. China

²Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education
Xiangtan University, Xiangtan 411105, P. R. China
doi:10.4156/jdcta.vol4.issue9.30

Abstract

In this article, a novel approach is proposed to predict RNA secondary structure called RNA secondary structure prediction based on Tabu Search (RNATS). In the RNATS algorithm, two search models, intensification search and diversification search, are developed to exploit the local regions around the current solution and explore the unvisited solution space, respectively. Simulation experiments are conducted on eight RNA sequences to show that the proposed method is feasible and effective.

Keywords: RNA Secondary Structure Prediction, Tabu Search, Minimum Free Energy

1. Introduction

Ribonucleic acid (RNA), a kind of biological molecules, plays a key role in the synthesis of protein. In many cases, RNA structure is essential for its biological function. There are three structural levels in RNA: primary, secondary, and tertiary structure [1]. The primary structure is defined as the sequence of bases. The secondary structure is defined as the set of stems formed by base pairs, loops among stems, and single strands of ribonucleotides. The tertiary structure which plays an important role in the functionality of RNA molecule is formed by packing secondary structure elements and turning into compact globular units. Due to difficulty in determining RNA 3D structure (tertiary structure) by experimental techniques, many attempts have so far been made at predicting secondary structure given an RNA sequence (primary structure) [2]. RNA secondary structure prediction becomes one of the most important fundamental problems in biological sequence information analysis.

RNA secondary structure can be predicted by laboratory tools and computer simulations [3,4]. Laboratory tools such as X-ray and nuclear magnetic resonance can predict RNA secondary structure accurately, but they are time consuming and require high cost [4]. Recently, researchers focused on computing techniques to predict RNA secondary structure, such as comparative sequences analysis [5], minimum free energy [6,7], and genetic algorithm [8,9]. The comparative sequences analysis method estimates the similarity between the RNA sequence to be predicted and the known RNA sequences according to the rule of covariant alignment so as to output the common RNA secondary structure model [5]. The minimum free energy method mimics the laws of thermodynamics, computes the free energy of RNA secondary structure, and views the structure of the minimum free energy as the best result [6,7]. Genetic algorithm encodes RNA secondary structure as a chromosome, establishes a population of individuals, and adopts genetic operators to evolve RNA secondary structure [8,9].

In this paper, we introduce tabu search (TS) to predict RNA secondary structure and develop a novel prediction approach called RNA secondary structure prediction based on Tabu Search (RNATS). To the best of our knowledge, this is the first reported study that reflects upon the usage of tabu search in the prediction of RNA secondary structure. Our aim is to explore the applicability of tabu search to RNA secondary structure prediction and to predict the proper RNA secondary structure based on tabu search according to the minimum free energy technique. Two search procedures, intensification search and diversification search, are developed in the RNATS algorithm to exploit the local area around the current solution and explore the unvisited solution space. In the RNATS method, Tabu List (TL) is employed to prevent solution cycling and Visited Region List (VRL) is introduced to encourage the

search to explore the unvisited space. Computer simulations show that the RNATS method is effective for predicting RNA secondary structure.

The remaining part of this article is organized as follows. In Section 2, RNA secondary structure is briefly introduced. The RNATS method and its components are described in Section 3. Section 4 shows the results of computer simulations. Finally, some conclusions are drawn in Section 5.

2. RNA secondary structure

RNA molecule consists of a chain of ribonucleotides linked by covalent chemical bonds. Each ribonucleotide represents one of four bases, adenine (A), cytosine (C), guanine (G), and uracil (U). Two bases in close proximity form a chemical bond called base pair if they are complementary: A with U, G with C, and G with U [6]. In the following, RNA secondary structure is described.

Definition 1 Given RNA sequence $R = r_1 \cdots r_i \cdots r_{L_S}$, where r_i denotes the i^{th} base, $1 \leq i \leq L_S$, L_S denotes the length of sequence R , r_1 terminus is called the 5' terminus, r_{L_S} terminus is called the 3' terminus, and $P = \{(r_i, r_j)\}$ is defined as the set of base pairs and should satisfy the following conditions.

- (1) $(r_i, r_j) \in \{AU, CG, GU\}$, where $1 \leq i < j \leq L_S$ and $j - i > 3$;
- (2) if $(r_i, r_j) \in P$, $(r_i, r_k) \in P$, then $j = k$;
- (3) if $(r_i, r_j) \in P$, $(r_k, r_l) \in P$, $i < k < j$, then $i < k < l < j$.

There are some substructures in RNA secondary structure, stem, hairpin loop, internal loop, bulge loop, multibranch loop, and pseudoknot. They are described as follows.

Definition 2 Given two subsequences $R_1 = r_i \cdots r_{i+k-1}$ and $R_2 = r_j \cdots r_{j-k+1}$, if base pair (r_{i+t}, r_{j-t}) can be matched, $t = 0, \dots, k-1$, then R_1 and R_2 form stem $s(i, i+k-1, j, j-k+1, k)$. Here, i and j denote the start position of stem s on sequences R_1 and R_2 , respectively, $(i+k-1)$ and $(j-k+1)$ denote the end position of stem s on sequences R_1 and R_2 , respectively, k denotes the length of stem s , and $k \geq 3$ [5].

Definition 3 Stems $s_1 = (i_1, i_1 + k_1 - 1, j_1, j_1 - k_1 + 1, k_1)$ and $s_2 = (i_2, i_2 + k_2 - 1, j_2, j_2 - k_2 + 1, k_2)$ are compatible, if and only if the following condition holds:
 $\{(i_1, i_1 + k_1 - 1) \cup (j_1 - k_1 + 1, j_1)\} \cap \{(i_2, i_2 + k_2 - 1) \cup (j_2 - k_2 + 1, j_2)\} = \Phi$.

Suppose $S = \{s_1, \dots, s_Q\}$ as the stem pool and S' as the set of compatible stems, then $S' \subseteq S$, where Q is the number of stems. Five other substructures are described as follows [10].

Definition 4 Given stem $s_m \in S'$, if no other stem $s_l \in S'$ satisfies $(i_m + (i_m + k_m - 1))/2 < (i_l + (i_l + k_l - 1))/2 < (j_m + (j_m - k_m + 1))/2$, then stem set S' determines a hairpin loop of length $(j_m - k_m + 1) - (i_m + k_m - 1) - 1$.

Definition 5 Given stems $s_m, s_l \in S'$, where $(i_m + (i_m + k_m - 1))/2 < (i_l + (i_l + k_l - 1))/2$ and $(j_l + (j_l - k_l + 1))/2 < (j_m + (j_m - k_m + 1))/2$, if no other stem $s_t \in S'$ satisfies $(i_m + (i_m + k_m - 1))/2 < (i_t + (i_t + k_t - 1))/2 < (i_l + (i_l + k_l - 1))/2$ and $(j_l + (j_l - k_l + 1))/2 < (j_t + (j_t - k_t + 1))/2 < (j_m + (j_m - k_m + 1))/2$, then stem set S' determines an internal loop. Let $h_1 = i_l - (i_m + k_m - 1) - 1$ and $h_2 = (j_m - k_m + 1) - j_l - 1$, if both $h_1 \neq 0$ and $h_2 \neq 0$, the length of the internal loop is $h_1 + h_2$, otherwise S' determines a bulge loop of length h_1 when $h_2 = 0$ or length h_2 when $h_1 = 0$.

Definition 6 If stems $s_1, \dots, s_q \in S'$, where q is the number of stems and $q \geq 3$, $(i_1 + (i_1 + k_1 - 1))/2 < (i_2 + (i_2 + k_2 - 1))/2$ and no other $(i_1 + (i_1 + k_1 - 1))/2 < (i_c + (i_c + k_c - 1))/2 < (i_2 + (i_2 + k_2 - 1))/2$, $s_c \in S'$, for each b , $2 \leq b \leq q-1$, $(j_b + (j_b - k_b + 1))/2 < (i_{b+1} + (i_{b+1} + k_{b+1} - 1))/2$, and no other $(j_b + (j_b - k_b + 1))/2 < (i_g + (i_g + k_g - 1))/2 < (i_{b+1} + (i_{b+1} + k_{b+1} - 1))/2$, $s_g \in S'$, and $(j_q + (j_q - k_q + 1))/2 < (j_1 + (j_1 - k_1 + 1))/2$ and no other $(j_q + (j_q - k_q + 1))/2 < (i_h + (i_h + k_h - 1))/2 < (j_1 + (j_1 - k_1 + 1))/2$

$1)/2, s_h \in S'$, then S' determines a multibranch loop, of which the length is $(i_2 - (i_1 + k_1 - 1) - 1) + ((j_1 - k_1 + 1) - j_q - 1) + \sum_{b=2}^{q-1} (i_{b+1} - j_b - 1)$.

Definition 7 Given stems $s_m, s_l \in S'$, where $i_m < i_l < j_m < j_l$, then stem set S' determines a pseudoknot.

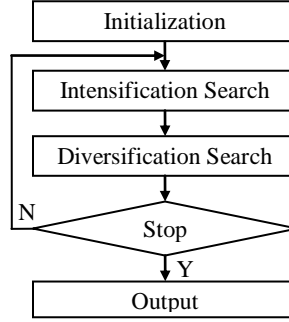


Figure 1. General description of the RNATS method

3. The RNATS algorithm

Figure 1 gives the general description of the RNATS method. Its main procedures observe the architecture of tabu search. The following subsections consider the design approaches in detail.

3.1. Solution representation

Given RNA sequence $R = r_1 \cdots r_{L_S}$, build matrix $\mathbf{C} = [c_{ij}]_{L_S \times L_S}$ as

$$c_{ij} = \begin{cases} 1 & \text{if } (r_i, r_j) \in \{AU, CG, GU\} \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

where $1 \leq i, j \leq L_S$, then stem pool $S = \{s_1, \dots, s_Q\}$ can be set up, where Q denotes the size of stem pool S . As a result, matrix $\mathbf{D} = [d_{lm}]_{Q \times Q}$ is built to represent the similarity between stems s_l and s_m , $1 \leq l, m \leq Q$. The value of stem s is defined as $sv = sl + st + sp$, where sl denotes the number of base pairs belonging to stem s , st denotes the type value of stem s , and sp denotes the position of stem s from the $5'$ terminus. The type value st is defined as the product of type position and type factor. Here, type position p_k^{st} denotes the position of the k^{th} base pair belonging to stem s , that is, $p_k^{st} = k$. In addition, type factor f^{st} is defined as

$$\begin{cases} f^{st} = 1 & (r_i, r_j) = GU \\ f^{st} = 2 & (r_i, r_j) = AU \\ f^{st} = 3 & (r_i, r_j) = CG \end{cases} \quad (2)$$

The following example may be helpful in understanding how to calculate the type value st of stem s . Given RNA sequence $R = AUGCAGAACUUGGC$ and stem $s = (4, 6, 12, 10, 3)$, then stem s includes base pairs CG , AU , and GU . Here, the number of base pairs $sl = 3$ and the position value $sp = 4$. The type value st of stem s is calculated as $st = \sum_{k=1}^{sl} (p_k^{st} \times f_k^{st}) = 1 \times 3 + 2 \times 2 + 3 \times 1 = 10$. As a result, the value sv of stem s is calculated as $sv = sl + st + sp = 3 + 10 + 4 = 17$. According to the definition

of the stem value, the similarity between stems s_l and s_m is defined as $d_{lm} = |sv_l - sv_m|$, where sv_l and sv_m denote the values of stems s_l and s_m , respectively. If stems s_l and s_m are not compatible, then let $d_{lm} = -1$.

In this paper, the solution is defined as $X = x_1 x_2 \cdots x_Q$, where Q denotes the number of stems. If stem s_i is selected, then stem s_i is activated and $x_i = 1$, otherwise stem s_i is disabled and $x_i = 0$, $1 \leq i \leq Q$. Here, all active stems belonging to solution X should be compatible with each other. The distance between solutions X_p and X_q is defined as

$$\begin{cases} D(X_p, X_q) = \max(d(s_l, s_m)) \\ x_l \in X_p, x_l = 1 \\ x_m \in X_q, x_m = 1 \end{cases} \quad (3)$$

3.2. Objective function computation

The free energy of RNA secondary structure E is defined as the objective function in this paper as follows [6].

$$E = E_{stem} + E_{hairpin} + E_{internal} + E_{bulge} + E_{multi}, \quad (4)$$

where E_{stem} is the energy of the stem, $E_{hairpin}$ is the energy of the hairpin loop, $E_{internal}$ is the energy of the internal loop, E_{bulge} is the energy of the bulge loop, and E_{multi} is the energy of the multibranch loop. Pseudoknot is not considered in this paper. The parameter settings of thermodynamic model can be found in [11].

3.3. Intensification search

At the intensification search stage, we exploit the local area of the current solution X_c in order to find the solution of low free energy. In addition, tabu list is employed to avoid revisiting recently visited solutions. The intensification search procedure is implemented as follows:

- Step 1: Given the current iteration number i_c , if $i_c = 1$, then create initial solution X_0 with the longest active stem from the stem pool, let the current solution $X_c = X_0$, add solution X_c into the tabu list, and set the counter of the tabu list $t = t + 1$. Otherwise proceed to Step 2.
- Step 2: Given the number of neighboring solutions N_l , generate neighboring solution X'_i of the current solution X_c , $i = 1, \dots, N_l$, as follows:
- If $N_c^l = 1$, that is, there is only one active stem s_c in the current solution X_c , then neighboring solution X'_i is established by keeping stem s_c active and activating another stem different from and compatible with stem s_c . When no compatible stem can be found, neighboring solution X'_i is created by randomly activating a stem different from stem s_c , that is, $N_i^l = 1$.
 - If $N_c^l > 1$, we first keep activated stems belonging to the current solution X_c active, then randomly activate stem s_d different from and compatible with activated stems, and finally disable one activated stem different from stem s_d so as to build neighboring solution X'_i . When no compatible stem can be found, neighboring solution X'_i is set up by randomly disabling one activated stem belonging to the current solution X_c , that is, $N_i^l = N_c^l - 1$.

After all neighboring solutions of the current solution X_c are created, we order them in an ascending order according to their free energy values, then the ordered neighboring solutions and their free energy values are denoted as X'_1, \dots, X'_{N_j} and $E(X'_1), \dots, E(X'_{N_j})$, respectively. Set $j=1$.

- Step 3: If $E(X'_j) < E(X_c)$ or $E(X'_j) > E(X_c)$ but solution X'_j is not taboed, then let $X_c = X'_j$, $E(X_c) = E(X'_j)$, add solution X'_j into the tabu list, set the counter of the tabu list $t = t + 1$, and proceed to Step 6. If $t > T$, then remove the first item of the tabu list, and set $t = t - 1$, where T denotes the size of tabu list.
- Step 4: If $j < N_j$, then $j = j + 1$, go to Step 3, otherwise proceed to Step 5.
- Step 5: Since all neighboring solutions are unavailable for updating the current solution X_c , we let $X_c = X'_1$ and $E(X_c) = E(X'_1)$.
- Step 6: Return the current solution X_c and its free energy $E(X_c)$.

3.4. Diversification search

At the diversification search stage, Visited Regions List (VRL) [12] is adopted to guide RNATS to explore the unvisited space. Let $VRL = \{(X_k, f_k)\}_{k=1}^M$, where solution X_k denotes the center of the visited region in which the distance between solution X_k and the solutions located in this region is less than or equal to the radius of diversification search γ , f_k denotes the frequency of visiting this region, and M denotes the number of all listed visited regions. Here, the radius of diversification search $\gamma = r(\max(\mathbf{D}) - \min(\mathbf{D}))$, where $\max(\mathbf{D}) = \max(d_{lm})$, $\min(\mathbf{D}) = \min(d_{lm})$, $1 \leq l, m \leq Q$, and r denotes the region coefficient used to determine the range in which the radius of diversification search γ varies.

Here, we try to generate new solutions outside the visited regions so as to explore the unvisited space. So, generating the solution near to more frequently visited regions is discouraged. We employ function Φ to distinguish between more and less frequently visited regions. The function Φ is defined as

$$\Phi(f_k) = \eta(1 - e^{-\eta(f_k-1)}), \quad (5)$$

where $\eta \in (0, 1]$. It is seen that the function Φ is strictly increasing and bounded above by parameter η . The diversification search procedure is implemented as follows:

- Step 1: Given the current iteration number i_c , if $i_c = 1$, set $k = 1$, let $X_k = X_c$, add solution X_k into the visited regions list, and update f_k . Otherwise proceed to Step 2.
- Step 2: Given the current solution X_c and the number of diversified solutions N_D , diversified solution \bar{X}_j of the current solution X_c , $j = 1, \dots, N_D$, is generated as follows:
- a) If $N_c^D = 1$, that is, there is only one active stem s_c in the current solution X_c , then diversified solution \bar{X}_j is created by keeping stem s_c active and activating another stem compatible with stem s_c . When no compatible stem can be found, diversified solution \bar{X}_j is created by randomly activating a stem different from stem s_c , that is, $N_j^D = 1$.
 - b) If $N_c^D > 1$, we randomly select an active stem s_c from solution X_c , then diversified solution \bar{X}_j is established by keeping stem s_c active and activating the stem from the stem pool which is different from and compatible with stem s_c . When no compatible stem can be found, we just keep stem s_c active in diversified solution \bar{X}_j .

After building all diversified solutions, we order these diversified solutions in an ascending order by their free energy values, then the ordered diversified solutions and their free energy values are denoted as $\bar{X}_1, \dots, \bar{X}_{N_D}$ and $E(\bar{X}_1), \dots, E(\bar{X}_{N_D})$, respectively. Set $l = 1$.

- Step 3: For the center of the visited region X_k , compute $\sigma_{lk} = D_{lk} / (1 + \Phi_{f_k})$, where D_{lk} denotes the distance between diversified solution \bar{X}_l and the center of the visited region X_k .
- Step 4: If $\min_{1 \leq k \leq M} \sigma_{lk} / \gamma \geq 1$ and solution \bar{X}_l is not in TL and VRL, then let $X_c = \bar{X}_l$, $E(X_c) = E(\bar{X}_l)$, update TL and VRL, and proceed to Step 6. If $\min_{1 \leq k \leq M} \sigma_{lk} / \gamma < 1$ and solution \bar{X}_l is not in TL, but $E(\bar{X}_l) < E(X_c)$, then let $X_c = \bar{X}_l$, $E(X_c) = E(\bar{X}_l)$, update TL and f_k , and proceed to Step 6. Otherwise, if $l < N_D$, then $l = l + 1$ and go to Step 3, otherwise proceed to Step 5.
- Step 5: Since all diversified solutions cannot be used to update the current solution X_c , we let $X_c = \bar{X}_l$ and $E(X_c) = E(\bar{X}_l)$.
- Step 6: Return the current solution X_c and its free energy value $E(X_c)$.

4. Experimental results

In this paper, computer simulations are conducted in Matlab on an Intel Pentium Dual-Core processor running at 2.8 GHz with 2 GB real memory. The impact of the parameters of RNATS is first investigated. Performance comparison between RNATS and another prediction method based on evolutionary algorithm is then conducted on eight RNA sequences. Each experiment includes 20 independent trials.

4.1. Performance analysis

In order to explore the good performance of RNATS, we here discuss the choice of experimental parameters in this subsection. The parameters to be discussed include the region coefficient r , parameter η , the size of tabu list T , the number of neighboring solutions N_l , the number of diversified solutions N_D , and the number of iterations I_T . The well-known RNA sequence CVV-3 [13] is used as a benchmark to evaluate performance.

In this paper, parameters r and η are used to select diversified solutions and explore the unvisited search space. Here, we investigate their contributions to the prediction algorithm as shown in Figure 2.

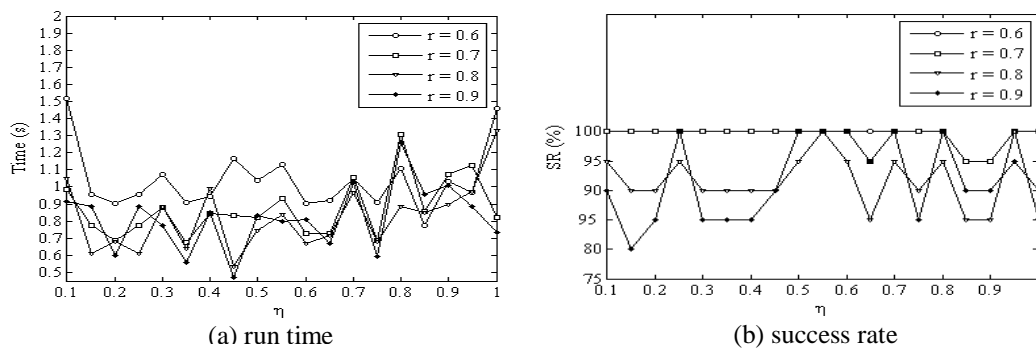


Figure 2. Comparison of different region coefficients r and parameters η

Three indicators are used to evaluate parameters r and η . Firstly, the accuracy rate is defined as the number of bases belonging to correctly predicted base pairs and free bases divided by the total number of bases. We find that RNATS can output the RNA secondary structure with the energy -18.85 kcal/mol lower than the factual one in most cases. At this time, its accuracy rate is equal to 46.15%. Secondly, the average run time values when the lowest energy is firstly provided by different parameters r and η are recorded as shown in Figure 2a. Finally, the success rate (SR) is defined as the number

of trials where the lowest energy is obtained divided by the number of total trials. Figure 2b shows the success rates attained by different parameters r and η . It is seen that as the growth of the region coefficient r , RNATS spends less run time to find the lowest energy but its success rate decreases in most runs. As $r=0.7$ and $\eta=0.35$ can provide high success rate within low run time, we adopt them to choose diversified solutions.

Table 1 shows the prediction results of different tabu lists. Equipped with different tabu lists, RNATS can find the RNA structure with the lowest energy but lead to different standard deviation (SD) and run time values. It is found that when the size of the tabu list is equal to 40, the best performance is attained. So, we choose this parameter to be 40.

Table 1. Comparison of different tabu list T

T	5	10	20	30	40
SD	1.1992	0.4247	1.6538	0.2128	0.0403
Time (s)	1.0148	0.9399	0.9953	0.9775	0.7311

In this paper, we let $N_I = N_D$ and analyze the effect of intensification search and diversification search on the prediction method. The standard deviation and run time results provided by different parameters N_I and N_D are shown as Table 2. In this experiment, the secondary structure with the lowest energy can be found by the prediction methods equipped with different parameters N_I and N_D . As the increase of parameters N_I and N_D , RNATS has more choices to select from and spends more computational effort in finding the proper structure. Therefore, deciding the proper value for parameters N_I and N_D is the process of exploring a balance between quality and cost. In this article, the case that $N_I = N_D = 20$ can reach our goal. In this case, RNATS can obtain the high quality and stable result at the expense of proper computational cost.

Table 2. Comparison of different parameters N_I and N_D

$N_I (N_D)$	10	20	30	40	50
SD	1.6876	0.0403	0.3366	2.1055	1.4113
Time (s)	1.1452	0.7311	0.9328	1.1820	1.4562

Table 3 shows the prediction results of different I_T . When $I_T > 10$, RNATS can find the secondary structure with the lowest energy. It is known that the more the number of iterations I_T , the better the results. However, this is done at the expense of more computational effort. In this article, we set the number of iterations I_T is equal to 90 so as to predict the high quality and stable structure at the cost of low run time.

Table 3. Comparison of different sizes of iterations I_T

I_T	10	30	50	70	90
SD	0.3398	0.2870	0.9109	0.1530	0.1655
Time (s)	0.9288	0.7475	0.8437	0.7609	0.7376

4.2. Performance comparison

In this section, the RNATS algorithm is applied to eight RNA sequences and compared with the RnaPredict method reported by Wiese et al. [9]. Eight RNA sequences are described as follows, AIMV-3, CiLRV-3, TSV-3, CVV-3, APMV-3, PDV-3 [13], Geobacillus stearothermophilus, and Thermus aquaticus [9]. AIMV-3 is the alfalfa mosaic virus with 39 bases, 2 stems, 2 hairpin loops, and 3 single-strand of ribonucleotides. CiLRV-3 is the citrus leaf rugose virus with 50 bases, 2 stems, 2 hairpin loops, and 3 single-strand of ribonucleotides. TSV-3 is the tobacco streak virus with 49 bases, 2 stems, 2 hairpin loops, and 3 single-strand of ribonucleotides. CVV-3 is the citrus variegation virus with 52 bases, 2 stems, 2 hairpin loops, 1

internal loop, and 3 single-strand of ribonucleotides. APMV-3 is the apple mosaic virus with 41 bases, 2 stems, 2 hairpin loops, and 3 single-strand of ribonucleotides. PDV-3 is the prune dwarf ilarvirus with 41 bases, 2 stems, 2 hairpin loops, and 3 single-strand of ribonucleotides. *Geobacillus stearothermophilus* is with 117 bases, 7 stems, 2 hairpin loops, 3 internal loops and 1 multibranch loop. *Thermus aquaticus* is with 123 bases, 6 stems, 2 hairpin loops, 3 internal loops, and 1 multibranch loop. Experimental RNA secondary structures are shown as Figure 3.

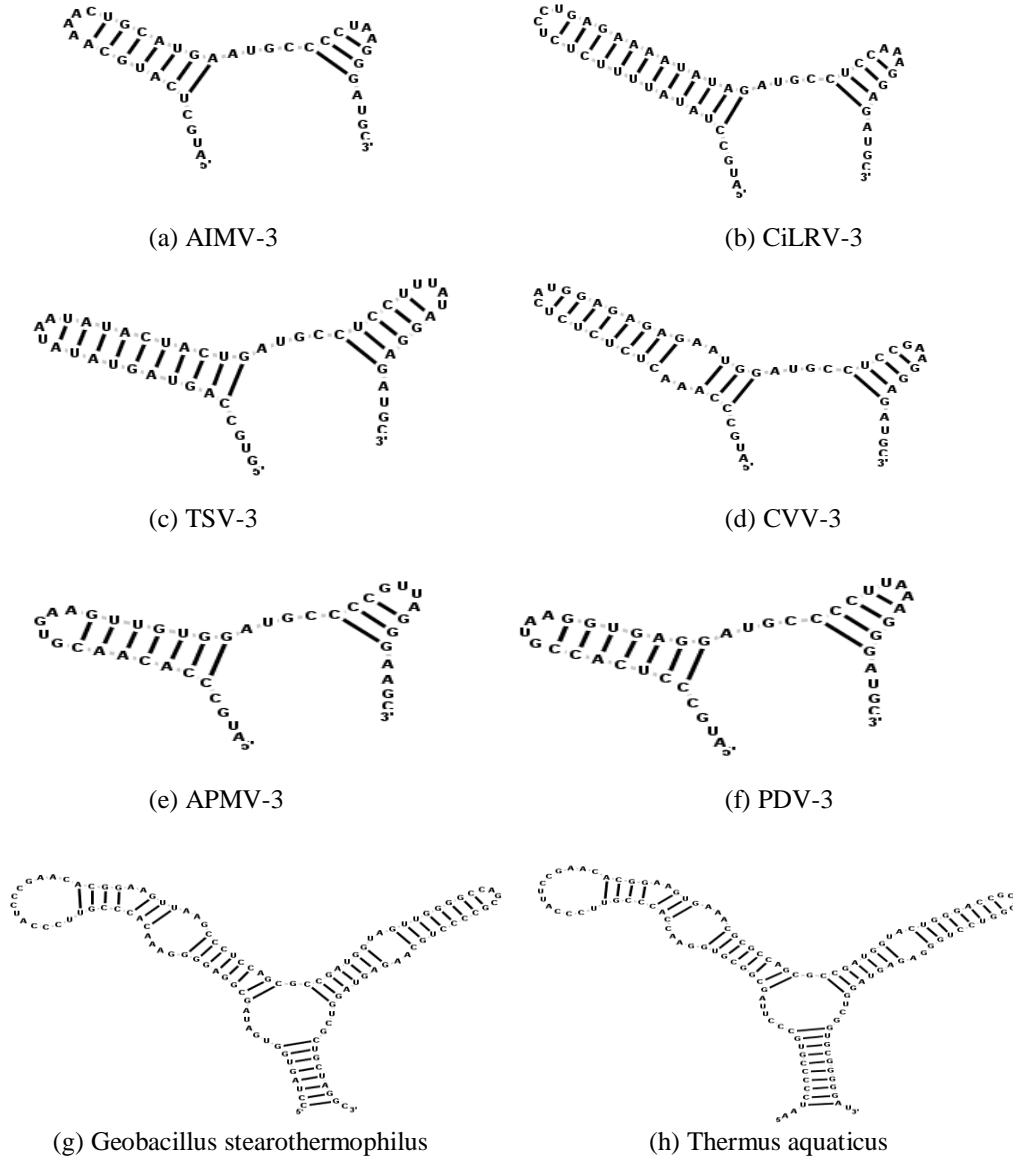


Figure 3. Experimental RNA secondary structures

The parameter settings of RnaPredict are shown as Table 4 and their detail descriptions can be found in [9].

Table 4. Parameter settings of the RnaPredict method

Population	200
Generation	100
Crossover operation	CX
Crossover probability	0.7
Mutation probability	0.8
Selection strategy	KBR
Elitism	1
Random seed	30

The average (Avg), standard deviation (SD), accuracy rate (AR), and run time values of the prediction results obtained by RnaPredict and RNATS for experimental sequences are shown as Table 5. The column E_{original} presents the free energy of the factual secondary structures of experimental RNA sequences. As there are base pairs CU, GA, GG, and UU in the factual secondary structures of *Geobacillus stearothermophilus* and *Thermus aquaticus*, which are not considered in thermodynamic model, their free energy values are not reported. In face of APMV-3 and PDV-3, both RnaPredict and RNATS lead to the correct secondary structures in each run. But RNATS finds the optimal results much sooner than RnaPredict. Considering AIMV-3, the energies reported by RnaPredict and RNATS are less than the factual one. In addition, RNATS provides higher accuracy rate than RnaPredict within much less run time. In face of CiLRV-3, RnaPredict obtains the correct secondary structure in each run. In addition, the accuracy rate of RNATS for CiLRV-3 is more than 90% and its the convergence speed is much faster than that of RnaPredict. For TSV-3, RNATS is superior to RnaPredict and reports the maximum accuracy rate within the minimum run time. In face of CVV-3, both RnaPredict and RNATS lead to the same accuracy rate but RNATS requires much less computational cost.

Table 5. Experimental results of RnaPredict and RNATS

RNA	E_{original}	RnaPredict				RNATS			
		Avg	SD	AR(%)	Time(s)	Avg	SD	AR(%)	Time(s)
AIMV-3	-12.60	-15.50	0	38.46	2.145	-15.48	0.01	38.59	0.736
CiLRV-3	-16.50	-16.50	0	100.00	5.816	-16.40	0.10	91.77	1.563
TSV-3	-17.70	-14.23	1.49	76.12	7.198	-19.55	0	87.76	1.153
CVV-3	-17.45	-18.85	0	46.15	4.830	-18.85	0	46.15	0.738
APMV-3	-12.60	-12.60	0	100.00	3.390	-12.60	0	100.00	0.370
PDV-3	-17.40	-17.40	0	100.00	2.798	-17.40	0	100.00	0.252
<i>geobacillus stearothermophilus</i>	-	-45.35	0.07	33.21	49.978	-53.54	7.25	29.66	30.788
<i>thermus aquaticus</i>	-	-63.43	0.37	37.24	57.810	-69.07	9.52	35.93	53.084

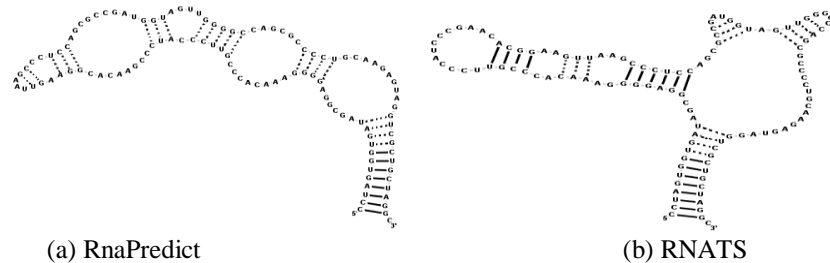


Figure 4. Prediction results for *geobacillus stearothermophilus*

Considering *geobacillus stearothermophilus*, RNATS can provide the minimum free energy sooner than RnaPredict. But the accuracy rate of RnaPredict is higher than that of RNATS. RNA secondary structures of *geobacillus stearothermophilus* provided by RnaPredict and RNATS are shown as Figure 4. In Figure 4, the base pairs correctly predicted are labeled by the real line and the incorrect ones are

labeled by the dashed line. That base pairs CU, GA, GG, and UU are not considered in thermodynamic model leads to the low accuracy rate. It is seen that RNATS can provide more correct base pairs than RnaPredict, and its secondary structure is closer the factual one than the one obtained by RnaPredict.

In face of thermus aquaticus, RNATS can provide lower free energy than RnaPredict within less run time. But the accuracy rate of RNATS is lower than that of RnaPredict. Prediction results for thermus aquaticus reported by RnaPredict and RNATS are shown as Figure 5. As base pairs GA and GG belonging to the factual secondary structure of thermus aquaticus are not considered in thermodynamic model, experimental prediction methods cannot attain high accuracy rate. In this experiment, RNATS still can output more correct base pairs than RnaPredict.

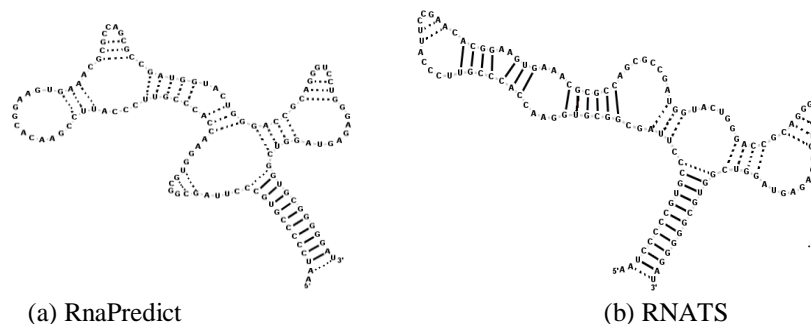


Figure 5. Prediction results for thermus aquaticus

In our study, we find the thermodynamic model employed in this paper cannot deal with different RNA sequences efficiently because of the lack of some constraint conditions. As a result, RNATS can find the minimum free energy values for most RNA sequences but cannot lead to the correct results.

5. Conclusions

As a kind of biological molecules, RNA plays an increasing important role in the research field of bioinformatics. In this article, we introduce tabu search to deal with the problem of RNA secondary structure prediction and propose the RNATS algorithm based on the minimum free energy technique. In the RNATS method, two search procedures, intensification search and diversification search, are developed to exploit the local regions around the current solution and explore the unvisited space, respectively. Computer simulations are conducted on eight RNA sequences to demonstrate the feasibility and effectiveness of the RNATS method for predicting RNA secondary structure. By comparison with the RnaPredict method, the RNATS algorithm can provide the minimum free energy values of experimental secondary structures within much less run time in most cases. In future, further improving the convergence speed of RNATS in complicate cases and designing a new thermodynamic model will be the subject of future research.

6. Acknowledgements

This research was supported in part by the National Natural Science Foundation of China (Grant No. 60903074) and the National High Technology Research and Development Program (863 Program) of China (Grant No. 2008AA01Z119).

7. References

- [1] J. Tinoco, C. Bustamante, "How RNA folds", *Journal of Molecular Biology*, vol. 293, no. 2, pp. 271-281, 1999.
- [2] U. Poolsap, Y. Kato, T. Akutsu, "Prediction of RNA secondary structure with pseudoknots using integer programming", *BMC Bioinformatics*, vol. 10, no. S1, pp. S38, 2009.

- [3] M. Zuker, D. Sankoff, "RNA secondary structure and their prediction", *Bulletin of Mathematical Biology*, vol. 46, no. 4, pp. 591-621, 1984.
- [4] J. Cohen, "Bioinformatics - an introduction for computer scientists", *ACM Computing Surveys*, vol. 36, no. 2, pp. 122-158, 2004.
- [5] J. Parsch, J. M. Braverman, W. Stephan, "Comparative sequence analysis and patterns of covariation in RNA secondary structures", *Genetics*, vol. 154, no. 2, pp. 909-921, 2000.
- [6] M. Zuker, P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information", *Nucleic Acids Research*, vol. 9, no. 1, pp. 133-148, 1981.
- [7] D. H. Mathews, "Predicting RNA secondary structure by free energy minimization", *Theoretical Chemistry Accounts*, vol. 116, no. 1-3, pp. 160-168, 2006.
- [8] S. K. Pal, S. Bandyopadhyay, S. S. Ray, "Evolutionary computation in bioinformatics: a review", *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, vol. 36, no. 5, pp. 601-615, 2006.
- [9] K. C. Wiese, A. A. Deschenes, A. G. Hendriks, "RnaPredict—an evolutionary algorithm for RNA secondary structure prediction", *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, vol. 5, no. 1, pp. 25-41, 2008.
- [10] V. Ferretti, D. Sankoff, "A continuous analog for RNA folding", *Bulletin of Mathematical Biology*, vol. 51, no. 1, pp. 167-171, 1989.
- [11] [Http:// www.bioinfo.rpi.edu/zukerm/cgi-bin/efiles-3.0.cgi](http://www.bioinfo.rpi.edu/zukerm/cgi-bin/efiles-3.0.cgi).
- [12] A. R. Hedar, M. Fukushima, "Tabu search directed by direct search methods for nonlinear global optimization", *European Journal of Operational Research*, vol. 170, no. 2, pp. 329-349, 2006.
- [13] F. Bai, D. Li, T. Wang, "A new mapping rule for RNA secondary structures with its applications", *Journal of Mathematical Chemistry*, vol. 43, no. 3, pp. 932-943, 2008.