Pacific Symposium on Biocomputing 15:98-107(2010)

DYNAMIC PROGRAMMING ALGORITHMS FOR RNA STRUCTURE PREDICTION WITH BINDING SITES

UNYANEE POOLSAP*, YUKI KATO*†, TATSUYA AKUTSU

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
E-mail: {unyanee,ykato,takutsu}@kuicr.kyoto-u.ac.jp

Noncoding antisense RNAs have recently occupied considerable attention and several computational studies have been made on RNA-RNA interaction prediction. In this paper, we present novel dynamic programming algorithms for predicting the minimum energy secondary structure when binding sites of one of the two interacting RNAs are known. Experimental results on several known RNA-RNA interaction data show that our proposed method achieves good performance in accuracy and time.

Keywords: RNA secondary structure; RNA-RNA interaction; dynamic programming

1. Introduction

In recent years, analysis of noncoding RNAs has attained great importance. They play a crucial role in some biological processes including post-transcriptional regulation of gene expression. Some noncoding RNAs, called *antisense RNAs*, aim at inhibiting their target RNA function through base complementary binding. Some antisense RNAs use full complementarity to their target for binding, whereas a number of antisense RNAs use partial complementarity,[1] and several *kissing hairpin* structures (Fig. 1) caused by loop-loop interaction have been reported.[2]

To predict joint secondary structures of interacting RNAs, several dynamic programming (DP) algorithms have been proposed so far. Andronescu *et al.*[3] developed the PairFold algorithm for secondary structure prediction of two interacting RNAs of minimum free energy. Since this algorithm is based on the Zuker's algorithm[4] for predicting pseudoknot-free structure of a single RNA, its time complexity is $O((n+m)^3)$ where n and m are respective lengths of two input sequences. The PairFold algorithm, however, cannot deal with any kissing hairpins, which are essentially equivalent to pseudoknotted structures when concatenating two interacting sequences. On the other hand, DP algorithms presented by Pervouchine,[5] Alkan *et al.*[6] and Kato *et al.*[7] can predict joint secondary structures including kissing hairpins in $O(n^3m^3)$ time. However, the time complexity of these algorithms is prohibitive in case $n \simeq m$ (i.e., $O(n^6)$ time), which is the same complexity of a prediction algorithm for pseudoknots.[8]

Viewing RNA-RNA interaction prediction from a different angle inspires us to consider the situation where we aim at predicting the secondary structure with binding sites of one of the two interacting RNAs (e.g., target RNA) on condition that interacting sites of the other RNA (e.g., antisense RNA) are known. In fact, we assume that a "profile" of intermolecular binding is given in advance, which can be obtained from the known secondary structure of the antisense RNA. This assumption could be reasonable since we can reduce computational complexity of a kind of interaction prediction and discover new target RNAs for antisense RNAs with known profiles. In this paper, we propose novel DP algorithms for predicting RNA secondary structures with binding site locations. Note that our formulation of the prediction problem requires that the order in which binding sites appear in an antisense RNA should be the same as the order in its target RNA (see Fig. 1). To deal with binding sites as well as base-paired structures, we design an extension of the classical Nussinov's algorithm,[9] which essentially minimizes the sum of base pair energies. In addition, we develop another DP algorithm that can incorporate stacking energy, which is based on the Zuker's algorithm.[4] Both of our proposed algorithms can run in $O(N^3n^3)$ time where N is the number of binding sites and n is a

*These authors contributed equally to this work.
†To whom correspondence should be addressed.

Pacific Symposium on Biocomputing 15:98-107(2010)

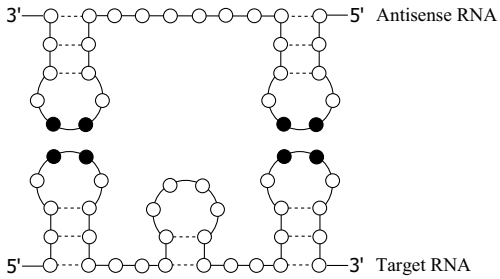

Fig. 1. An example of RNA-RNA interaction containing kissing hairpins. A black circle indicates one base of a binding site.

Table 1. An energy function e cited from Ref. 10.

Base pair	Energy value
{G, C}	−5
{A, U}	−4
{G, U}	−1

sequence length. Since N can be regarded as a constant in most cases, the time complexity of our algorithms can be evaluated as $O(n^3)$. We demonstrate the performance of our approach using the proposed algorithms on some data sets.

2. Methods

In this section, we will present dynamic programming (DP) algorithms for predicting RNA secondary structures with binding sites. Before going through the details of the algorithms, let us begin with definitions of RNA secondary structure and the prediction problem considering binding sites.

2.1. *Preliminaries*

Definition 2.1 (RNA secondary structure). *For an RNA sequence $s = s_1 s_2 \cdots s_n$ where $s_i \in \Sigma = \{A, C, G, U\}$ $(1 \leq i \leq n)$, a secondary structure of s is defined as a set R of position pairs (i, j) that satisfies the following conditions:*

- $1 \leq i < i + 1 < j \leq n$;
- $\forall (i, j), (i', j') \in R$; $i = i' \iff j = j'$.

Next, let us formally define the binding site profile.

Definition 2.2 (Binding site profile). *Let N be the number of binding sites and $\bar{b}_p = \bar{s}_{p,1}\bar{s}_{p,2} \cdots \bar{s}_{p,\ell_p} \in \Sigma^*$ $(1 \leq p \leq N)$ denote a binding site (subsequence) of an antisense RNA sequence. Let $s_i s_{i+1} \cdots s_j \in \Sigma^*$ be a subsequence of a target RNA sequence. Then, for each p $(1 \leq p \leq N)$, a binding site profile $I_p(i, j)$ of $s_i s_{i+1} \cdots s_j$ is defined as follows:*

$$I_p(i,j) = \begin{cases} \gamma \sum_{k=1}^{\ell_p} e(s_{i+k-1}, \bar{s}_{p,k}) & (j = i + \ell_p - 1, \text{ and } \forall k;\ s_{i+k-1} \text{ is complementary to } \bar{s}_{p,k}), \\ \infty & (\text{otherwise}), \end{cases} \tag{1}$$

where γ is a positive weight parameter, and e is an energy function that maps from a valid base pair to the corresponding energy value (see Table 1).

It should be noted that we do not know the actual binding sites of the target RNA in advance even though the actual binding sites of the antisense RNA are given. Instead of using the binding site profile, estimates

Pacific Symposium on Biocomputing 15:98-107(2010)

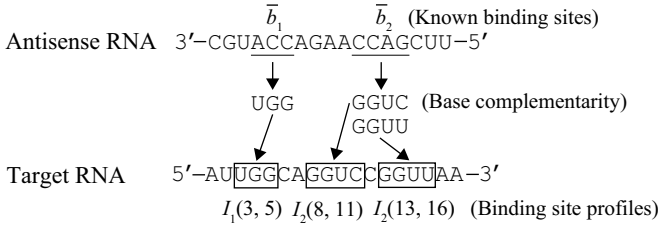

Fig. 2. An example of binding site profile.

of binding energies could be used in our algorithms if the binding sites of the antisense RNA are known and the binding energies depend on the stacking energies of the base pairs involved in the binding sites.

Example 2.1. Consider an antisense RNA sequence shown in Fig. 2 with known binding sites $\bar{b}_1 = \text{ACC}$ and $\bar{b}_2 = \text{CCAG}$. We first compute their complementary subsequences UGG for $\bar{b}_1$ and GGUC, GGUU for $\bar{b}_2$. Notice that we also take the wobble pair {G, U} into account. Then, we search for those complementary subsequences through the target sequence. Since UGG matches at the location from 3 to 5 in the target, we have $I_1(3,5) = -4-5-5 = -14$ where $\gamma = 1$ in Eq. (1) and the energy function shown in Table 1 are used. The rest of the elements of I_1 are equal to ∞ because no other sites of length three in the target match UGG. In a similar way, we obtain $I_2(8,11) = -19$ and $I_2(13,16) = -15$, and the rest of the elements of I_2 are set at ∞.

With these definitions, we define the prediction problem of RNA secondary structure with binding sites.

Definition 2.3 (RNA secondary structure prediction with binding sites).

Input: *A target RNA sequence $s = s_1 s_2 \cdots s_n \in \Sigma^*$ and N binding site profiles $I_1, I_2, \ldots, I_N$ of s.*

Output: *The optimum secondary structure of s whose subsequences match the binding sites in the order from I_1 to I_N.*

2.2. *DP Algorithms*

We develop two prediction models based on DP. The first DP-based model is an extension of the Nussinov's algorithm[9] using a simple base pair energy function. For the second model, we extend the first model to utilize the stacking energy and loop energy functions, which is based on the Zuker's algorithm.[4]

2.2.1. *Base pair energy model*

In the beginning, we define DP tables to design the algorithm. Let $s = s_1 s_2 \cdots s_n$ be an RNA sequence. As in the conventional case, we let $W(i,j)$ denote the minimum free energy of secondary structure formed from a subsequence $s_i s_{i+1} \cdots s_j$ of s. In addition, let $W_{pq}(i,j)$ be the minimum free energy of secondary structure for $s_i s_{i+1} \cdots s_j$ that contains binding sites corresponding to $I_p, I_{p+1}, \ldots, I_q$ $(1 \le p \le q \le N)$. Note that $W_{pq}(i,j)$ is a four dimensional table. We use this notation to facilitate comparison with the other DP table $W(i,j)$.

These DP tables are initialized as follows:

$$W(i,i) = 0,\ W_{pq}(i,i) = \infty \quad (1 \le \forall i \le n;\ 1 \le \forall p \le \forall q \le N).$$

The recursions are classified into three cases as shown below. In the first case, we use the simple Nussinov's algorithm for predicting secondary structure without binding sites. The second case is used for dealing with the structure with just one binding site. The third case is used for predicting the structure with two or more binding sites.

Pacific Symposium on Biocomputing 15:98-107(2010)

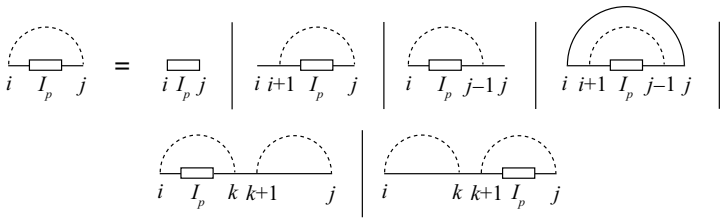

Fig. 3. Recursion for $W_{pp}(i,j)$. A dashed curve indicates that we do not know whether or not two bases connected by the curve form a base pair, and a solid curve shows that two bases connected by it definitely form a base pair.

Case 1 (the Nussinov's algorithm):

$$W(i,j) = \min \begin{cases} W(i+1,j), \\ W(i,j-1), \\ W(i+1,j-1) + e(i,j), \\ \min\limits_{i \le k < j} \{W(i,k) + W(k+1,j)\}, \end{cases} \tag{2}$$

where $e(i,j)$ is the simple energy function for a base pair (s_i, s_j). In the above DP recursion, the first and the second cases of minimization represent the cases where s_i and s_j do not form a base pair. The third case says that s_i and s_j form a base pair, and the resulting energy $e(i,j)$ is added to the present value of W. The forth formula represents the bifurcation structure. Note that k is the position at which the structure bifurcates in such a way that the sum of energies of two substructures is minimized.

Case 2 ($p = q$):

$$W_{pp}(i,j) = \min \begin{cases} I_p(i,j), \\ W_{pp}(i+1,j), \\ W_{pp}(i,j-1), \\ W_{pp}(i+1,j-1) + e(i,j), \\ \min\limits_{i \le k < j} \{W_{pp}(i,k) + W(k+1,j)\}, \\ \min\limits_{i \le k < j} \{W(i,k) + W_{pp}(k+1,j)\}. \end{cases} \tag{3}$$

The first case means that $s_i s_{i+1} \cdots s_j$ is a possible binding site and we adopt the corresponding score $I_p(i,j)$ computed in Eq. (1). The formulas from the second through the fourth are similar to the ones from the first through the third in Eq. (2). The fifth case represents the bifurcation structure where the binding site is contained in the former part of the bifurcation. Since the latter part of the bifurcation does not contain any binding sites, we use W computed in Eq. (2). The last case is a counterpart of the fifth case. Following a diagrammatic representation in Ref. 8, we provide a schematic representation of the recursion for $W_{pp}(i,j)$ in Fig. 3.

Case 3 ($q \ge p+1$):

$$W_{pq}(i,j) = \min \begin{cases} \min\limits_{i \le k < j} \min\limits_{p \le r < q} \{W_{pr}(i,k) + W_{r+1,q}(k+1,j)\}, \\ W_{pq}(i+1,j), \\ W_{pq}(i,j-1), \\ W_{pq}(i+1,j-1) + e(i,j), \\ \min\limits_{i \le k < j} \{W_{pq}(i,k) + W(k+1,j)\}, \\ \min\limits_{i \le k < j} \{W(i,k) + W_{pq}(k+1,j)\}. \end{cases} \tag{4}$$

Pacific Symposium on Biocomputing 15:98-107(2010)

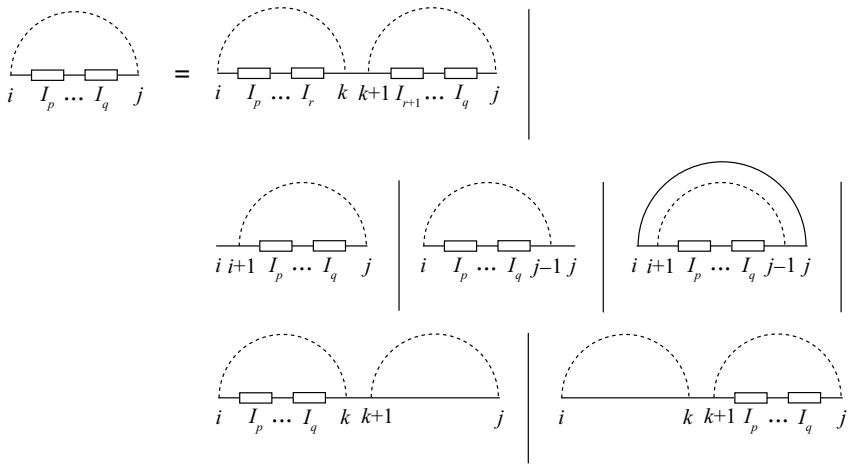

Fig. 4. Recursion for $W_{pq}(i, j)$.

The first case is designed for computing the bifurcation of secondary substructures, each of which contains the binding sites. It should be noted that the former part of the bifurcation contains the binding sites corresponding to $I_p, \ldots, I_r$, whereas the latter part corresponds to the substructure with binding sites for $I_{r+1}, \ldots, I_q$. The other cases can be interpreted as in Case 2. Figure 4 illustrates the above DP recursion.

We now evaluate the complexity of the above algorithm. Computing Eq. (2) takes $O(n^3)$ time. Equations (3) and (4) can be computed in $O(Nn^3)$ and $O(N^3n^3)$ time, respectively. Therefore, the overall time complexity is evaluated as $O(N^3n^3)$. By similar evaluation, we can see that the space complexity is $O(N^2n^2)$.

The minimum energy of the secondary structure of the input sequence is equivalent to $W_{1,N}(1, n)$, and the optimum secondary structure can be retrieved by tracing back the DP tables from $W_{1,N}(1, n)$.

2.2.2. *Stacking energy model*

Since the energy function used in the above DP algorithm is very simple, there is room for further improvement of our DP model. It is widely accepted that calculating contributions for stacking energy rather than individual contributions for each base pair yields better prediction. Hence, we extend the above DP algorithm based on this idea. In order to incorporate stacking energy into our previous DP model, we introduce additional DP tables. Let $V(i, j)$ be the minimum free energy of secondary structure formed from a subsequence $s_i s_{i+1} \cdots s_j$ such that s_i and s_j form a base pair. Let $V_{pq}(i, j)$ be the minimum free energy of secondary structure for $s_i s_{i+1} \cdots s_j$ that contains binding sites corresponding to $I_p, I_{p+1}, \ldots, I_q$ such that s_i and s_j form a base pair. Note that $W(i, j)$ and $W_{pq}(i, j)$ are defined in the same way as in the base pair energy model. Although energies of multi-branched and exterior loops could be incorporated into the recursions of W and W_{pq}, we exclude such energy rules for simplicity.

Initialization conditions for W and V are as follows:

$$W(i, i) = \infty,\ V(i, i) = \infty,\ W_{pq}(i, i) = \infty,\ V_{pq}(i, i) = \infty \quad (1 \le \forall i \le n;\ 1 \le \forall p \le \forall q \le N).$$

The revised version of the DP recursions is as follows:

Case 1 (the Zuker's algorithm):

$$W(i, j) = \min \begin{cases} W(i + 1, j), \\ W(i, j - 1), \\ V(i, j), \\ \min\limits_{i \le k < j} \{W(i, k) + W(k + 1, j)\}, \end{cases} \tag{5}$$

Pacific Symposium on Biocomputing 15:98-107(2010)

$$V(i,j) = \min \begin{cases} eh(i,j), \\ V(i+1,j-1) + es(i,i+1,j-1,j), \\ \min\limits_{i<i'<j'<j} \{V(i',j') + ebi(i,i',j',j)\}, \\ \min\limits_{i<k<j-1} \{W(i+1,k) + W(k+1,j-1)\} + b, \end{cases} \tag{6}$$

where $eh(i,j)$ is the destabilizing energy of a hairpin loop closed by a pair of (s_i, s_j), $es(i,i+1,j-1,j)$ is the stacking energy of two pairs (s_i, s_j) and (s_{i+1}, s_{j-1}), $ebi(i,i',j',j)$ is the destabilizing energy of a bulge or an interior loop closed by pairs (s_i, s_j) and $(s_{i'}, s_{j'})$, and b is a penalty for a bifurcation structure. Notice that in Eq. (5), the case where s_i and s_j form a base pair is represented by $V(i,j)$. As can be seen in Eq. (6), $V(i,j)$ is computed by minimizing among the four cases. The first case represents the energy of a hairpin loop closed by (s_i, s_j). The second formula adds the stacking energy of (s_i, s_j) and (s_{i+1}, s_{j-1}) to the present value of V. The third case represents a substructure where a bulge or an interior loop occurs in $s_i \cdots s_{i'}$ and $s_{j'} \cdots s_j$. The fourth formula is used for computing bifurcation.

Case 2 $(p = q)$:

$$W_{pp}(i,j) = \min \begin{cases} I_p(i,j), \\ W_{pp}(i+1,j), \\ W_{pp}(i,j-1), \\ V_{pp}(i,j), \\ \min\limits_{i \le k < j} \{W_{pp}(i,k) + W(k+1,j)\}, \\ \min\limits_{i \le k < j} \{W(i,k) + W_{pp}(k+1,j)\}, \end{cases} \tag{7}$$

$$V_{pp}(i,j) = \min \begin{cases} \min\limits_{i<i'<j'<j} \{W_{pp}(i',j') + er(i,i',j',j)\}, \\ V_{pp}(i+1,j-1) + es(i,i+1,j-1,j), \\ \min\limits_{i<i'<j'<j} \{V_{pp}(i',j') + ebi(i,i',j',j)\}, \\ \min\limits_{i<k<j-1} \{W_{pp}(i+1,k) + W(k+1,j-1)\} + b, \\ \min\limits_{i<k<j-1} \{W(i+1,k) + W_{pp}(k+1,j-1)\} + b, \end{cases} \tag{8}$$

where $er(i,i',j',j)$ is the approximate destabilizing energy of a pair of subsequences $(s_{i+1} \cdots s_{i'-1}, s_{j'+1} \cdots s_{j-1})$, which is obtained by removing $s_{i'} \cdots s_{j'}$ from $s_{i+1} \cdots s_{j-1}$. $V_{pp}(i,j)$ is computed by minimizing among the five choices. The first formula represents the case where the binding site corresponding to I_p is contained in the sequence closed by a base pair (s_i, s_j). The other cases are similar to those of the $V(i,j)$ recursion.

Case 3 $(q \ge p+1)$:

$$W_{pq}(i,j) = \min \begin{cases} \min\limits_{i \le k < j} \min\limits_{p \le r < q} \{W_{pr}(i,k) + W_{r+1,q}(k+1,j)\}, \\ W_{pq}(i+1,j), \\ W_{pq}(i,j-1), \\ V_{pq}(i,j), \\ \min\limits_{i \le k < j} \{W_{pq}(i,k) + W(k+1,j)\}, \\ \min\limits_{i \le k < j} \{W(i,k) + W_{pq}(k+1,j)\}, \end{cases} \tag{9}$$

Pacific Symposium on Biocomputing 15:98-107(2010)

Table 2. Results of the base pair energy model (BPEM), where n is the length of a target sequence and N is the number of binding sites. Note that for the ATP sensitive ribozyme-Substrate, n indicates the length of the antisense sequence. Since the substrate does not fold into secondary structure, only the binding sites can be detected by the algorithms, which is too simple. Therefore, we used the substrate to compute the binding site profile and predicted secondary structure and binding sites of the ATP sensitive ribozyme.

Antisense-Target	n	N	SEN (%)	PPV (%)	Time (s)
Tar-Tar*[11]	16	1	100.00	90.00	0.20
R1inv-R2inv[12]	19	1	100.00	100.00	0.23
DIS-DIS[13]	35	1	82.35	73.68	0.85
CopA-CopT[14]	57	3	100.00	93.94	17.76
ATP sensitive ribozyme-Substrate[15]	59	2	52.17	36.36	9.01
IncRNA_{54}-RepZ[16]	61	2	100.00	94.87	9.72
RyhB-SodB[17]	87	1	37.50	25.00	10.27
OxyS-fhlA[14]	100	2	59.09	52.00	43.25
Average			78.89	70.73	11.41

$$V_{pq}(i,j) = \min \begin{cases} \min\limits_{i<i'<j'<j} \{W_{pq}(i',j') + er(i,i',j',j)\}, \\ V_{pq}(i+1,j-1) + es(i,i+1,j-1,j), \\ \min\limits_{i<i'<j'<j} \{V_{pq}(i',j') + ebi(i,i',j',j)\}, \\ \min\limits_{i<k<j-1} \{W_{pq}(i+1,k) + W(k+1,j-1)\} + b, \\ \min\limits_{i<k<j-1} \{W(i+1,k) + W_{pq}(k+1,j-1)\} + b. \end{cases} \tag{10}$$

$V_{pq}(i,j)$ in Case 3 differs from $V_{pp}(i,j)$ in Case 2 in that the present subsequence $s_i s_{i+1} \cdots s_j$ contains at least two binding sites.

Finally, we evaluate the complexity of this algorithm. Obviously, complexity for computing Eqs. (9) and (10) dominates the overall complexity of the algorithm. Computing the first formula of Eq. (9) takes $O(N^3 n^3)$ time. Exact analysis of the first and third formulas of Eq. (10) reveals time complexity of $O(N^2 n^4)$. In actual case, however, the loop size is bounded by a constant, and thus the complexity can be reduced to $O(N^2 n^2)$. Therefore, the overall time complexity is evaluated as $O(N^3 n^3)$. The space complexity is $O(N^2 n^2)$.

3. Results

Our two DP models were tested on the data set comprising eight antisense-target RNA complexes with known structures, taken from several literatures (see Tables 2–4). In fact, an antisense sequence was used for constructing a binding profile, whereas the corresponding target sequence was used for predicting its structure with binding sites. For the binding site profile computation, we used $\gamma = 2$ in Eq. (1). We employed Table 1 for the simple energy parameter e, and adopted sophisticated energy parameters for folding at 37°C provided by the Turner Group[18] (the recent version is available online at http://www.bioinfo.rpi.edu/zukerm/rna/energy/) for other parameters including eh, es, etc. We limited the size of interior and bulge loops to at most four nucleotides. The penalty for a bifurcation structure b was set at 1. We implemented the algorithms in Java on a machine with Intel Core 2 Duo CPU 1.20GHz and 2.00GB RAM. Prediction accuracy was measured using sensitivity (SEN) and positive predictive value (PPV) defined below:

$$\text{SEN} = \frac{\sharp \text{ of correctly predicted base pairs} + \sharp \text{ of correctly predicted bases of binding sites}}{\sharp \text{ of observed base pairs} + \sharp \text{ of observed bases of binding sites}},$$

$$\text{PPV} = \frac{\sharp \text{ of correctly predicted base pairs} + \sharp \text{ of correctly predicted bases of binding sites}}{\sharp \text{ of predicted base pairs} + \sharp \text{ of predicted bases of binding sites}}.$$

Note that $\sharp$ represents the number.

Tables 2 and 3 show the prediction accuracy of the base pair energy model (BPEM) and that of the stacking energy model (SEM), respectively. Figure 5 depicts predicted structures of the fhlA RNA of the longest sequence in the data set. We can see that SEM outperforms BPEM in terms of accuracy.

Pacific Symposium on Biocomputing 15:98-107(2010)

Table 3. Results of the stacking energy model (SEM).

Antisense-Target	SEN (%)	PPV (%)	Time (s)
Tar-Tar*	100.00	100.00	0.39
R1inv-R2inv	92.31	100.00	0.53
DIS-DIS	100.00	100.00	2.92
CopA-CopT	96.77	100.00	50.09
ATP sensitive ribozyme-Substrate	100.00	92.00	29.11
IncRNA$_{54}$-RepZ	100.00	97.37	30.23
RyhB-SodB	83.33	64.52	31.77
OxyS-fhlA	90.91	90.91	115.40
Average	95.42	93.10	32.56

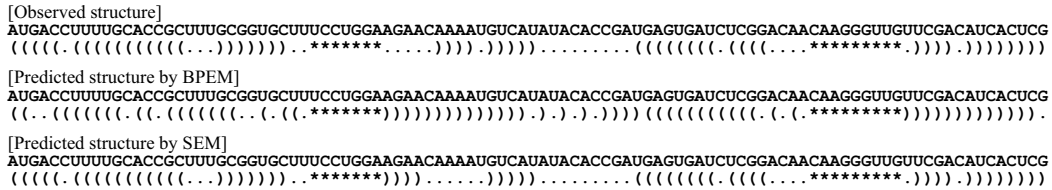

Fig. 5. Prediction results for the fhlA RNA. A pair of parentheses denotes a base pair and a series of asterisks represents a binding site.

Table 4. Comparison of F-measure (%) between our models, the stacked pair model (SPM) and the loop model (LM) presented in Ref. 6. Note that LM returned no base pair when ATP sensitive ribozyme-Substrate was used as an input.

Antisense-Target	**BPEM**	**SEM**	SPM	LM
Tar-Tar*	94.74	**100.00**	90.00	90.00
R1inv-R2inv	**100.00**	96.00	**100.00**	**100.00**
DIS-DIS	77.78	**100.00**	82.35	82.35
CopA-CopT	96.88	**98.36**	83.33	78.79
ATP sensitive ribozyme-Substrate	42.86	**95.83**	55.32	0.00
IncRNA$_{54}$-RepZ	97.37	**98.67**	81.58	81.58
RyhB-SodB	30.00	**72.73**	53.97	52.78
OxyS-fhlA	55.32	**90.91**	80.00	78.72
Average	74.37	**94.06**	78.17	70.68

We then compared the performance of our proposed methods with that of existing DP-based models, called the stacked pair model (SPM) and the loop model (LM) presented in Ref. 6, using the inteRNA web server[19] (see Table 4). In the web interface, we set gap penalty and maximum substructure length at 0 and 61, respectively, for all eight RNA pairs. SPM and LM took much time for prediction due to their high time complexity as stated in Sec. 1. We calculated F-measure F, which is the harmonic mean of SEN and PPV defined by $F = 2 \cdot \mathrm{SEN} \cdot \mathrm{PPV}/(\mathrm{SEN} + \mathrm{PPV})$. As Table 4 shows, the prediction performance of SEM is the best of all four models on average.

To demonstrate the applicability of our profile-based approach to target discovery, we further tried analyzing some RNAs with unknown structures and binding sites, given binding profiles of specific antisense RNAs. Test sequences were selected to be homologous to the known target sequence (query) for the antisense RNA using BLASTN. Figures 6 and 7 illustrate two candidate prediction results using SEM. We must note that in the figures it is not clear whether or not the predicted binding sites actually interact with the antisense RNAs since the number of detected binding sites was fewer than that of the known homologous target RNAs. The prediction result of CP001122 (see Fig. 6), which has 90% sequence identity with the known target CopT, shows that SEM detected two binding sites on CP001122. Note that the second binding site inside the hairpin loop (CUGC in Fig. 6) was ascribed to the profile computed from the third binding site located in the exterior loop of CopT, which leads to more uncertainty of predicted target sites. As for another example, SEM recognized only one binding sites on CU928158 (see Fig. 7), which has 89% sequence

Pacific Symposium on Biocomputing 15:98-107(2010)

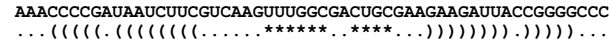

Fig. 6. A prediction result of a sequence CP001122 Salmonella enterica subsp. enterica serovar Kentucky str. CVM29188 plasmid pCVM29188_146, which has 90% sequence identity with the CopT RNA to which the antisense CopA binds. Notice that only two binding sites were predicted as compared with the known target CopT with three binding sites.

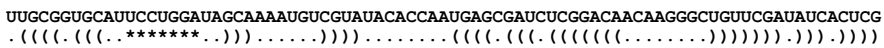

Fig. 7. A prediction result of a sequence CU928158 of the Escherichia fergusonii ATCC 35469 chromosome, which has 89% sequence identity with the fhlA RNA to which the antisense OxyS binds. Notice that only one binding site was predicted as compared with the known target fhlA with two binding sites.

identity with the known target fhlA. On the other hand, exactly two binding sites were found for another test sequence of 99% sequence identity with fhlA (the result is not shown here).

4. Conclusion

We proposed new dynamic programming algorithms for predicting RNA secondary structures with binding sites. The performance of the algorithms for the base pair energy model (BPEM) and the stacking energy model (SEM) was demonstrated for several known RNA-RNA interaction data. Judging from the results, it can be said that the advantages of BPEM are faster running time than that of SEM and the simplicity of DP, which is a natural extension of the classical prediction algorithm for single RNA folding. However, prediction accuracy of BPEM plummets for some inputs. On the other hand, SEM is a more complex model than BPEM but a robust one in a sense that prediction accuracy hardly fluctuates according to the length of an input sequence. We further compared the prediction results with those of other prediction methods, which shows that SEM outperforms those earlier models.

Our approach is a novel method of RNA-RNA interaction prediction from a different point of view (i.e., use of profile of intermolecular interaction), and achieves lower time complexity compared with earlier methods. Moreover, use of our profile-based method can improve prediction performance. Our method will also be useful in discovering new target sites for an antisense RNA with a known binding profile. To ensure this advantage, exhaustive search for candidate targets has to be performed as our future work. If the profile of an interacting protein is available, our method could be applied to RNA-protein interaction prediction, which is also left as a challenging task.

Acknowledgments

The first author thanks the Japanese Government (Monbukagakusho:MEXT) Scholarship from the Ministry of Education, Culture, Sports, Science and Technology. This work was supported in part by Grant-in-Aid for Young Scientists (Start-up) (KAKENHI) #20800023 from Japan Society for the Promotion of Science (JSPS), and by Grant-in-Aid for Scientific Research (KAKENHI) #19200022 from JSPS.

References

1. S. Brantl, *Biochim. Biophys. Acta* **1575**, 1–3 (2002).
2. C. Brunel, R. Marquet, P. Romby and C. Ehresmann, *Biochimie* **84**, 9 (2002).
3. M. Andronescu, Z.C. Zhang and A. Condon, *J. Mol. Biol.* **345**, 5 (2005).
4. M. Zuker and P. Stiegler, *Nucl. Acids Res.* **9**, 1 (1981).
5. D.D. Pervouchine, *Genome Inform.* **15** (2004).
6. C. Alkan, E. Karakoç, J.H. Nadeau, S.C. Şahinalp and K. Zhang, *J. Comput. Biol.* **13**, 2 (2006).
7. Y. Kato, T. Akutsu and H. Seki, *Pattern Recogn.* **42**, 4 (2009).
8. E. Rivas and S.R. Eddy, *J. Mol. Biol.* **285**, 5 (1999).
9. R. Nussinov, G. Pieczenik, J.R. Griggs and D.J. Kleitman, *SIAM J. Appl. Math.* **35**, 1 (1978).

Pacific Symposium on Biocomputing 15:98-107(2010)

10. P. Clote and R. Backofen, *Computational Molecular Biology*, John Wiley & Sons, Ltd (2000).
11. K.-Y. Chang and I. Tinoco Jr, *J. Mol. Biol.* **269**, 1 (1997).
12. M.J. Rist and J.P. Marino, *Nucl. Acids Res.* **29**, 11 (2001).
13. J.-C. Paillart, E. Skripkin, B. Ehresmann, C. Ehresmann and R. Marquet, *Proc. Natl. Acad. Sci. USA* **93**, 11 (1996).
14. E.G.H. Wagner and K. Flardh, *Trends Genet.* **18**, 5 (2002).
15. J. Tang and R.R. Breaker, *Nucl. Acids Res.* **26**, 18 (1998).
16. K. Asano and K. Mizobuchi, *J. Biol. Chem.* **275**, 2 (2000).
17. T.A. Geissmann and D. Touati, *The EMBO J.* **23**, 2 (2004).
18. D.H. Turner, N. Sugimoto and S.M. Freier, *Ann. Rev. Biophys. Biophys. Chem.* **17** (1988).
19. C. Aksay, R. Salari, E. Karakoç, C. Alkan and S.C. Şahinalp, *Nucl. Acids Res.* **35** (2007).