Approximation Scheme for RNA Structure Prediction Based on Base Pair Stacking

HENGWU LI[1,2], DAMING ZHU[1], ZHENZHONG XU[2], HUIJIAN HAN[2]

1. School of Computer Science and Technology
Shandong University
NO.73 Jingshi Road, Jinan 250061
CHINA
2. Department of Computer Science and Technology
Shandong Economic University
NO.7366 Erhuandong Road, Jinan 250014
CHINA

Abstract: - Pseudoknotted RNA secondary structure prediction is an important problem in computational biology. Existing polynomial time algorithms have no performance guarantee or can handle only limited types of pseudoknots. In this paper for the general problem of pseudoknotted RNA secondary structure prediction, a polynomial time approximation scheme is presented to predict pseudoknotted RNA secondary structure by dynamic programming and branch-bound based on base pair stacking. Compared with existing polynomial time algorithm, it has exact approximation performance and can predict arbitrary pseudoknots.

Key-Words: - RNA; Secondary Structure; Pseudoknot; Algorithm; Approximation Scheme; Dynamic Programming

1 Introduction

RNA secondary structures prediction plays an important role in functional analysis of RNA molecules. Among the most prevalent RNA structures is a motif known as the pseudoknot. Pseudoknots play a variety of diverse roles in biology. These roles include forming the catalytic core of various ribozymes, self-splicing introns, and telomerase. Additionally, pseudoknots play critical roles in altering gene expression by inducing ribosomal frameshifting in many viruses[1]. Plausible pseudoknotted structures have been proposed (Pleij et al.) in 1985 and confirmed (Kolk et al.) in 1998 for the 3' end of several plant viral RNAs, where pseudoknots are apparently used to mimic tRNA structure[2]. Recently, pseudoknots were confirmed in some RNAs of humans and many other species[3][4].

Currently pseudoknot is not included in the majority of the study for RNA secondary structure prediction. The best Zuker algorithm predicts RNA secondary structure without pseudoknots with $O(n^3)$ time and $O(n^2)$ space for a sequence of length n and is implemented by MFOLD and ViennaRNA programs. Finding the best secondary structure including arbitrary pseudoknots has been proved to be NP-hard[5].

Most methods for RNA folding which are capable of folding pseudoknots adopt heuristic search procedures and sacrifice optimality. Examples of these approaches include quasi-Monte Carlo searches and genetic algorithms. These approaches are inherently unable to guarantee that they have found the best structure, and consequently unable to say how far a given prediction is from optimality[6][7].

A different approach to pseudoknot prediction is the maximum weighted matching algorithm, considering only the base paired action and no stacking action. The maximum weighted matching algorithm folds an optimal pseudoknotted structure in $O(n^3)$ time with low accuracy and seems best suited to folding sequences for which a previous multiple alignment exists[8]. Another approach adopts dynamic programming to predict the tractable subclass of pseudokonts based on complex thermodynamic model in $O(n^4)$-$O(n^6)$ time[9][10][11].

The major driving force of structure formation for RNA molecules is Watson–Crick base pair and wobble G, U base pair formation, and in particular stacking of adjacent base pairs[5]. RNA secondary structure is a set of base pair. Base pair and internal unpaired bases construct loops. Stack doesn't contain unpaired bases, and any other kinds of loops contain one or more unpaired bases. Since unpaired bases are destabilizing, stack is the only type of loops that stabilize the secondary structure [5]. So we study stack problem to find the key of RNA structure prediction.

In this paper for the general problem of pseudoknotted RNA secondary structure prediction, considering only stacking energy and neglecting other secondary role, an approximation scheme with $O((n/2dk)^{dk+1})$ time is presented to predict pseudoknotted RNA structure. Compared with existing polynomial time algorithms, which can handle only limited types of pseudoknots or have no performance guarantee, it has exact approximation performance and can predict arbitrary pseudoknots.

In section 2 we give the energy model and PTAS for RNA secondary structure prediction. In section 3 we briefly conclude the paper.

2 Problem Formulation

Let s=s_1,s_2, ...,s_n be an RNA sequence, base $s_i \in \{A,U, C,\mathrm{G}\}$, $1 \le i \le n$. The subsequence $\mathrm{s}_{i,j}$ = si, si+1,...,sj is a segment of s, $1 \le \mathrm{i} \le \mathrm{j} \le n$. If s_i&$s_j \in \{A\&U, C\&\mathrm{G}, U\&G\}$, then s_i and s_j may constitute base pair (i, j). Each base can at most take part in one base pair. RNA secondary structure S is a set of base pairs for s. Base pair and internal unpaired bases construct loops.

If (i, j) and (i+1, j-1)$\in S$, base pairs (i, j) and (i+1, j-1) constitute stack (i, i+1: j-1, j), and m($\ge$1) consecutive stacks form the helix (i, i+m: j-m, j) with the length of m+1. The energy of helix (p, p+m-1: i-m+1,i) is denoted as $E(p, p+m-1:i-m+1,i)$.

If base pairs (i, j) and (k, l) are parallel ($i<j<k<l$ or $k<l<i<j$) or nested ($i<k<l<j$ or $k<i<j<l$), then base pairs (i, j) and (k, l) are compatible, otherwise base pairs (i, j) and (k, l) constitute pseudoknots ($i<k<j<l$ or $k<i<l<j$) as Fig.1.

Stack is the only type of loops that stabilize the secondary structure. Therefore for pseudoknotted RNA structure prediction, we give the general energy model considering only stacking energy and neglecting other secondary role.

Definition 1 (stacking energy model of pseudoknotted RNA structure prediction, SEM): For RNA sequence s, $s \in \{A, U, C, G\}^*$, a secondary structure S is a set of base pairs such that if $(i, j) \in S$ then

1) $\forall (i', j') \in S$, if $(i, j) \cap (i', j') \ne \varnothing$, then $(i, j) = (i', j')$.

2) $(i, j) \in \{(A,U), (C,G), (U,G)\}$.

3) if $(i+1, j-1) \in S$, then (i, j) and $(i+1, j-1)$ form stack with the energy of $E(i, i+1:j-1, j)$.

4) if $(i+1, j-1),(i', j'),(i'+1, j'-1) \in S$, $s_i = s_{i'}$, $s_j = s_{j'}$ and $s_{i+1} = s_{i'+1}$, $s_{j-1} = s_{j'-1}$, then $E(i,i+1:j-1,j) = E(i', i'+1: j'-1,j')$. That is, the size of stacking force is determined by base pair itself and adjacent bases pair.

5) if $(i+1, j-1) \in S$, then the energy of S is $E(S) = \sum_{1 \le i<j \le n} E(i, i+1:j-1, j)$.

So the problem of pseudoknotted RNA structure prediction is to find a secondary structure S with maximal energy for given RNA sequence s under SEM model.

We divide sequence into single base, adjacent double bases, ..., and adjacent K ($K \in$ integer and $K \ge 2$) bases in all possible ways. Then assigned the stacking energy of complementary adjacent i bases as weight of matching, we compute the maximum weight matching for each partition, and choose the maximum weight matching of all the partitions as the result.

As each base belongs to adjacent i bases or single base, the number of partitions is K^n, $2 \le i \le K$. For each partition, $O(n^3)$ time is required to compute the maximum weighted matching, so the time complexity is $O(n^3K^n)$ to compute maximum weight matching of all the partitions.

But we need only consider the type and energy of paired adjacent i bases, not paired adjacent i bases themselves, $2 \le i \le K$. So we represent the energy of paired adjacent i bases as weight, and save the number of unpaired adjacent i bases for each type of adjacent i bases in order to pair with back complementary ones. For each type of unpaired adjacent i bases, if two partitions all have the same the number of this type of unpaired adjacent i bases, and they have the same paired weight, then they have the same results. Moreover for each type of unpaired adjacent i bases, if the partitions all have the same the number of this type of unpaired adjacent i bases, we need only choose the one with maximal weight from these partitions according to the theory of optimization.

Let $dk = \sum_{2 \le i \le K} 4^i$, matrices $S_{[x_1][x_2]\ldots[x_{dk}]}$, $SA_{[x_1][x_2]\ldots[x_{dk}]}$ and $SB_{[x_1][x_2]\ldots[x_{dk}]}$ represent respectively the maximal energy of sequences $s_{1,i}$, $s_{1,i-1}$, $s_{1,i-2}$ with x_i unpaired adjacent y_i bases in the ith type ($1 \le i \le dk$, $0 \le x_i \le n_i$). Because each partition has at most n/2 stack, then we can reduce computation by branch-bound method. Base on above principle, we give an approximation scheme for pseudoknotted RNA secondary structure prediction.

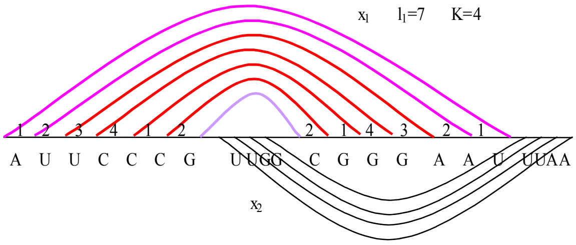

Fig. 1. The partition of pseudoknot

The arcs represent base pair, and level line represents sequence.

```
//Let $s=s_1s_{2\ldots}s_n$ be the input sequence, $K\in$integer and $E(S)$ is the output energy of the algorithm.
//Initially, $E(S)=\varnothing$, matrices $S=0$, $SA=0$, and $SB=0$.
SAA(s)
1. for $m$=2 to $K$ do
        Divide sequence $s$ into $n$-$m$+1 adjacent $m$ bases in all possible ways.
        Compute the number of each types of adjacent $m$ bases.
   end for
2. Sort all type of adjacent bases such that $n_1\le n_2\le\ldots\le n_{dk}$, $dk=\sum_{2\le i\le K}4^i$. $q_i=n_i+1$.
3. for $i$=2 to $n$ do
      for $m$=2 to $K$ do
         Assuming the type of adjacent $m$ bases $s_{i\text{-}m+1\ldots}\ s_{i\text{-}1}s_i$ is the $k$th and that of adjacent $m$ bases $s_p\ s_{p+1\ldots}s_{p+m\text{-}1}$ paired with $s_{i\text{-}m+1\ldots}\ s_{i\text{-}1}s_i$ is the $l$th.
         1) $S_{[x_1]\ldots[x_k+1]\ldots[x_{dk}]}=SB_{[x_1]\ldots[x_k]\ldots[x_{dk}]}$, if $S_{[x_1]\ldots[x_k+1]\ldots[x_{dk}]}<SB_{[x_1]\ldots[x_k]\ldots[x_{dk}]}$ and $x_1y_1+x_2y_2+\ldots+x_{dk}y_{dk}\le i\text{-}m$. That is, $s_{i\text{-}m+1\ldots}s_{i\text{-}1}\,s_i$ is adjacent $m$ bases waiting for pair.
         2) $S_{[x_1]\ldots[x_l\text{-}1]\ldots[x_{dk}]}=SB_{[x_1]\ldots[x_l]\ldots[x_{dk}]}+E(i\text{-}m+1,i\text{: } j,\ j+m\ \text{-}1)$, if $S_{[x_1]\ldots[x_l\text{-}1]\ldots[x_{dk}]}<SB_{[x_1]\ldots[x_l]\ldots[x_{dk}]}+E(p,\ p+m\ \text{-}1\text{:}i\text{-}m+1,i)$ and $x_1y_1+x_2y_2+\ldots+x_{dk}y_{dk}\le i\text{-}m$. That is, $s_{i\text{-}m+1\ldots}\ s_{i\text{-}1}s_i$ forms helix with adjacent $m$ bases waiting for pair.
      end for
      $SB\leftarrow SA$, $SA\leftarrow S$, if $x_1y_1+x_2y_2+\ldots+x_{dk}y_{dk}\le i$.
   end for
4. $E(S)=\max(S_{[x_1][x_2]\ldots[x_{dk}]})$, if $x_1y_1+x_2y_2+\ldots+x_{dk}y_{dk}\le i$.
```

Lemma 1: *Let OPT(I) be the maximal energy that can be formed by any secondary structure of sequence I. Let SAA[I] be the output by algorithm SAA. Then, OPT(I) /SAA[I]≤1+1/(K-1), K∈integer and K≥2.*

Proof: Let the helices in $OPT(I)$ are $x_1,x_2,\ldots,x_m$ with the length of $l_1,l_2,\ldots,l_m$ and the energy of Ex_1, $Ex_2,\ldots, Ex_m$, $m\ge1$.

$\forall x_q\in OPT(I), 1\le q\le m$, if $l_q\le K$, then we choose that $E_q=Ex_q$; otherwise we divide x_q into helices with the length of 2, and group these helices into K set $X_{q1},X_{q2},\ldots,X_{qK}$.

X_{q1}={ (i,i+1: j-1,j), (i+K+1,i+K+2: j-K-2,j-K-1),...}

X_{q2}={ (i+1,i+2: j-2,j-1), (i+K+2,i+K+3: j-K-3, j-K-2), ... }

....

X_{qk}={ (i+K,i+K+1:j-K-1,j-K), (i+2K+1,i+2K+2: j-2K-2, j-2K-1) ,.... }

Let the energy of $X_{q1},X_{q2},\ldots,X_{qK}$ is EX_{q1}, $EX_{q2},\ldots,EX_{qK}$ respectively, then $Ex_q=EX_{q1}+EX_{q2}+\ldots+X_{qK}$.

After that, we sort $EX_{q1},EX_{q2},\ldots,EX_{qK}$ such that $EX_{qa1}\ge EX_{qa2}\ge\ldots\ge EX_{qaK}$ and delete the energy EX_{qaK} in order to just divide x_q into helices whose length is not more than K. For example, for $x_1,x_2\in OPT(I)$ in Fig.1, when K=4, we divide x_1 into four groups of 1-4, then delete the energy of the second group so that x_1 is divided into two helices with the length of 2 and 4.

Let the sum of left energy is E_q, then

$E_q\ge(EX_{q1}+EX_{q2}+\ldots+EX_{qK})(K\text{-}1)/K=(K\text{-}1)\,Ex_q/K$.

After above handle, all helices in $OPT(I)$ become the structures formed by the helices whose length is not more than K, then $\sum_{1\le q\le m}E_q\ge\sum_{1\le q\le m}(K\text{-}1)Ex_q/K=(K\text{-}1)OPT(I)/K$.

Also the length of sequence $s_{1,i}$ is i, so each partition of s meets the condition $x_1y_1+x_2y_2+\ldots+x_{dk}y_{dk}\le i$. Obviously $SAA[I]$ is the optimal structure formed by helices whose length is not more than K.

Therefore, $SAA[I]\ge\sum_{1\le q\le m}E_q\ge(K\text{-}1)OPT(I)/K$.

$OPT(I)/SAA[I]\le K/(K\text{-}1)=1+1/(K\text{-}1)$.

Lemma 2: *Given an RNA sequence s of length n, algorithm SAA computes the maximal energy that can be formed by s in $O((n/2dk)^{dk+1})$time and $O((n/dk)^{dk})$space.*

Proof: The time complexity of Step1 is $O(Kn)$.

The time complexity of Step2 is $O(Kn\log Kn)$.

The time complexity of Step3 is $O(K\sum_{2\le i\le n}(x_1+1)(x_2+1)\ldots(x_{dk}+1))$.

The time complexity of Step4 is $O((x_1+1)(x_2+1)\ldots(x_{dk}+1))$.

We can see by the condition $x_1y_1+x_2y_2+\ldots+x_{dk}y_{dk}\le i$ that $x_1x_2\ldots.x_{dk}\le(i/2dk)^{dk}$ when i is big enough. So the time complexity of algorithm SAA is $O(K\sum_{2\le i\le n}(x_1+1)(x_2+1)\ldots.(x_{dk}+1))=O(K\sum_{2\le i\le n}(i/2dk)^{dk})=O((n/2dk)^{dk+1})$.

Similarly by the condition $n_1+n_2+\ldots+n_{dk}\le(K\text{-}1)n$ and $n_1\le n_2\le\ldots.\le n_{dk}$, $n_1\,n_2\ldots.n_{16}\le(n/dk)^{dk}$ when i is big enough. So the space complexity of algorithm SAA is $O(q_1q_2\ldots.q_{dk})=O((n/\ dk)^{dk})$.

Theorem 1: *The Algorithm SAA is a 1+ε approximation algorithm for the problem of* *constructing a secondary structure S with maximal energy for given RNA sequence s under SEM model, $\varepsilon=1/(K-1)$, $K\in$integer and $K\geq 2$.*

Proof: By Lemmas 1 and 2, the result follows.

3 Conclusion

In this paper SEM model is built based on base pair stacking force and neglecting other secondary role, and an approximation scheme with $O((n/2dk)^{dk+1})$ time is presented to predict pseudoknotted RNA structure under the model. Compared with existing polynomial time algorithms, which can handle only limited types of pseudoknots or have no performance guarantee, it has exact approximation performance and can predict arbitrary pseudoknots.

It would be of interest to improve these approximation ratios and time complexity of RNA structure prediction problem.

4 Acknowledgments

The work is supported by NSFC under grant NO.60573024 and 60573181, and the Research Foundation of Shandong Economic University (01610784).

References:

[1] Staple DW, Butcher SE., Pseudoknots: RNA Structures with Diverse Functions, *PLoS Biol,* Vol.3, 2005, pp. e213

[2] Kolk, M. H., van der Graff, M., Wijmenga, S. S., Pleij, C. W. A., Heus, H. A., Hilbers, C. W., NMR structure of a classical pseudoknots: interplay of single- and double-stranded RNA,. *Science,* Vol.280,1998, pp. 434-438

[3] Mathews, D. H., Turner, D. H., Prediction of RNA secondary structure by free energy minimization, *Current Opinion in Structural Biology*, Vol.16 ,2006, pp.270-278

[4] Barette, I., Poisson, G., Gendron, P. et al, Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching, *Nucleic Acids Research*, Vol. 29, 2001, pp. 753-758

[5] Ieong, S., Kao, M.Y., Lam, T.W., Sung, W.K., Yiu, S.M., Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs, *Journal of Computational Biology,* Vol.6, 2003, pp. 981-995

[6] Ren J, Rastegari B, Condon A, Hoos HH, HotKnots: heuristic prediction of RNA secondary structures including pseudoknots, *RNA*, Vol.11, 2005, pp. 1494-1504.

[7] Ruan, J., Stormo, GD. & Zhang, W., An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots, *Bioinformatics*, Vol.20,2004, pp. 58-66

[8] Tabaska JE, Cary RB, Gabow HN, Stormo GD, An RNA folding method capable of identifying pseudoknots and base triples, *Bioinformatics,* Vol.14,1998, pp. 691-699.

[9] Rivas, E., Eddy, SR., A dynamic programming algorithm for RNA structure prediction including pseudoknots, *Journal of Molecular Biology,* Vol. 285,1999, pp. 2053-2068

[10] Reeder J, Giegerich R, Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics, *BMC Bioinformatics*, Vol. 5, 2004, pp.104.

[11] Li Hengwu, Zhu Daming, Liu Zhendong, Li Hong, *Prediction for RNA planar pseudoknots***,** Progress in Nature Science, Vol.17(6),2006, pp. 717-724