Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Biology and Chemistry

ELSEVIER

Computational Biology and Chemistry 30 (2006) 72–76

www.elsevier.com/locate/compbiolchem

Software note

An RNA folding algorithm including pseudoknots based on dynamic weighted matching

Haijun Liu[a], Dong Xu[a], Jianlin Shao[b], Yifei Wang[a,*]

[a] *Department of Mathematics, Shanghai University, Shanghai 200444,China*
[b] *College of Life Sciences, China Jiliang University, Hangzhou 310018, China*

Received 3 June 2005; received in revised form 3 June 2005; accepted 2 October 2005

Abstract

On the basis of maximum weighted matching (MWM) algorithm, we introduced a dynamic weight related with stem length and used a recursive algorithm to predict RNA secondary structures by searching the stem structure with maximum weight summation step-by-step. This algorithm not only avoids the complicated free energy calculation, but also it could attain higher prediction accuracy. Moreover, our algorithm can predict most types of potential pseudoknots in the RNA structure.

Keywords: RNA secondary structure; Pseudoknots; Dynamic weighted matching

1. Introduction

RNA molecules play an important role in life activities of cells. Besides being transferring intermediaries or carriers of genetic information that is widely known, RNA also has many other important functions such as participating in gene expression and regulation by virtue of its special structural motifs, catalyzing activity of certain reactions or being components of some organelles and so on. In order to entirely understand these functions of RNAs, it needs to further know their three-dimensional structures. However, since RNA is prone to be degraded and hard to be crystallized, it is difficult to determine RNA spatial structure by means of X-ray diffraction or NMR. Therefore it is very necessary to reliably predict RNA structure directly from sequence by computer. RNA secondary structure prediction is a preliminary step for tertiary structure prediction because the base matching relationship on the level of secondary structure is also the main body of tertiary structure. On the other hand, the secondary structure of RNA only involves the arrangement of bases on the two-dimensional plane such that the prediction model is simplified.

According to the number of sequences applied in a method, most of the current methods to predict RNA secondary structure can be classified into two classes: one is ab initio prediction method which only needs one sequence, and the other is comparative sequence analysis method which needs a group of homology sequences. The ab initio prediction method includes the dynamic programming algorithm represented by free energy minimization method, and some heuristic algorithms with stem combinatory. The complexity of free energy minimization method is relatively higher. Its running time is $O(n^4)$ and running space is $O(n^2)$. Moreover, the free energy minimization method cannot predict pseudoknots (Zuker and Stiegler, 1981). Some optimization methods based on stems such as genetic algorithms, simulated annealing cannot guarantee global optimization and accuracy of results (van Batenburg et al., 1995; Schmitz and Steger, 1996). And their implementations are complicated. Generally, the predicted results are most reliable with the comparative sequence analysis. However, because it needs a few homology sequences to make alignment and requires those homology sequences to be enough similar with each other, it does not work when there is only one sequence or few sequences or when the homological degree of given sequences is not enough. Multiple sequence alignment at the beginning of comparative sequence analysis is also a complicated computational process that even needs manual operation (Eddy and Durbin, 1994). There are other algorithms between these two types of methods, among which maximum weighted matching algorithm is typical. At first a matching weight matrix is created using aligned homology sequences, and then the structure

[*] Corresponding author. Tel.: +86 21 66134331; fax: +86 21 66133292.
E-mail address: yifei_wang@staff.shu.edu.cn (Y. Wang).

1476-9271/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.10.001

of target sequence can be predicted based on the matrix (Cary and Stormo, 1995; Tabaska et al., 1998). This algorithm is able to predict secondary structures, pseudoknots and more complex relationship of base pairs in polynomial time. But limited by the accuracy of weight matrix, its behavior in predicting accuracy is not very steady yet. Especially, if alignment sequences are few or their homological relationship is not strong, the weight matrix deduced from them will result in increase of false positive of prediction results (Ruan et al., 2004).

In this paper, an RNA folding algorithm based on dynamic weighted matching was presented, which determined RNA secondary structures through recursively searching the stems with maximum weight sum and added a secondary recursion to predict potential pseudoknots. We introduced a dynamic weight correlated with stem length and combined it with experiential constant weight of base pairs as an optimization criterion to avoid the complicated free energy calculation. Optimum RNA secondary structures and pseudoknots were directly found through double recursion and there was no need of traceback stage occurred in dynamic programming algorithms. The space complexity of our algorithm is only O(n^2) and the time complexity is less than O($n^3 \log n$). A group of tRNAs and noncoding RNAs were used to test our algorithm. The prediction accuracy reaches 95.08% for tRNAs, 64.63% for noncoding RNAs. This result maybe has fluctuation with different randomly selected test sets. In addition, because we have not real RNA secondary structures of those test sequences and regard the prediction results of free energy algorithm as reference, some deviation will exist during the comparison of results. However, the results to a certain extent show the validity of our algorithm.

2. RNA secondary structure and pseudoknots

2.1. RNA secondary structure

Like proteins, RNA structures also have such forms as primary structure, secondary structure, tertiary structure and even quaternary structure. RNA secondary structures refer to unpaired single strand and complementary base-paired double strand including various structural motifs such as stems, hairpin loops, interior loops, bugle loops and multi-branch loops. RNA base pairs contain Watson–Crick pair A–U, G–C and wobble pair G–U. There is GC > AU > GU as far as the stability of hydrogen bond is concerned. We give a mathematical representation of RNA secondary structure here.

Given a RNA sequence $R = r_1, r_2 \cdots r_n$ of length n, its secondary structure is defined as a base pairs set $S = \{(r_i, r_j)\}$, where (r_i, r_j) satisfies the following conditions:

(1) $(r_i, r_j)\{(A, U), (U, A), (G, C), (C, G), (G, U), (U, G)\}$, $1 \le i < j \le n$ and $j - i >$ CONST1;
(2) If $(r_i, r_j) \in S$ and $(r_f, r_g) \in S$, then $i = f$ if and only if $j = g$;
(3) If $(r_i, r_j) \in S$, $(r_f, r_g) \in S$ and $i < f$, then these two base pairs only have two types of location relationship namely $i < f < g < j$ or $i < j < f < g$.

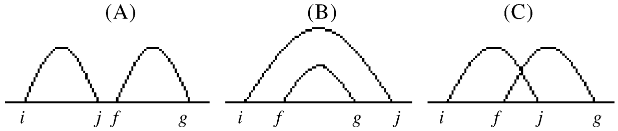

Fig. 1. The position relation between base pairs. (A) Juxtaposed base pairs. (B) Nested base pairs. (C) Pseudoknot.

If there exists a section of successive base pairs $(r_i, r_j),(r_{i+1}, r_{j-1}), \ldots, (r_{i+k-1}, r_{j-k+1})$, where $k \ge$ CONST2, this section is called a stem marked as Stm(i, j, k). Here, i and j represent 5′-end initial site and 3′-end terminal site, respectively, k represents stem length. The single strand between a and b is marked as Str(a, b).

CONST1 limits the minimum span between base r_i and r_j, which is also the minimum loop length of hairpins. CONST2 is the minimum stem length. Condition (2) shows that RNA secondary structures only allow one-to-one base pair and do not take one-to-more-than-one hydrogen bond interaction into consideration. Condition (3) specifies that the position of every two base pairs in RNA secondary structures either is nested fashion or juxtaposed fashion. Otherwise, if a cross occurs, viz. $i < f < j < g$, it is called a pseudoknot shown in Fig. 1.

2.2. Pseudoknots

Pseudoknots, in fact, belong to tertiary structure. Though the pseudoknots base pairs have a very small proportion to total base pairs in secondary structure, they are very important since they are often the active centers or important conservative structure of a certain function. Imagined pseudoknots have 14 types in all (Shapiro and Wu, 1997). Ten types are loop-to-loop interaction and four types are loop to single strand interaction. But in the reality there are some types that appear seldom because of limitation in structural chemistry and thermodynamics. Several common pseudoknots are as Fig. 2 shows.

Pseudoknots prediction is a difficult point in RNA folding simulation. On one hand no precise definitions or facts prove at present which kind of pseudoknots is rational, therefore it is very difficult in its modeling. On the other hand, there is no suitable optimization index for pseudoknots because of lack of special free energy parameters of pseudoknots so far. Lyngso and Pedersen (2000) have proved that it is a NP-complete problem to predict RNA secondary structures and pseudoknots using general free energy minimization method. Among the published RNA secondary prediction methods with pseudoknots, some of

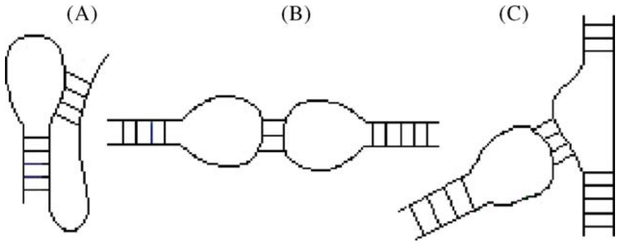

Fig. 2. Three types of common pseudoknots.

them are not practical due to high complexity (some time complexity reaches O(n^6) while space complexity reaches O(n^4)), for example, the extended dynamic programming algorithm put forward by Rivas and Eddy (1999). And some of them cause relatively greater prediction errors for whole structure because of the consideration of pseudoknots, for instance, genetic algorithm and maximum weighted matching algorithm. Our algorithm is able to find out most of the potential pseudoknots in less than O($n^3 \log n$) time. It not only can reach a higher predicting accuracy, but also can detect potential pseudoknots especially for RNA sequences within 100nt length. Because we do not have the real secondary structures of test sequences, we can only say the predicted pseudoknots are potential and only for references.

3. Methods

Our algorithm belongs to ab initio algorithm so that RNA structure prediction is regarded as an optimization problem. According to the principle of more steady structure with smaller energy, free energy minimization method searches the structure with minimum energy as real RNA secondary structure using free energy as optimization criterion. Similarly, we were inspired from maximum weighted matching algorithm and introduced a compound weight as the optimization criterion in our algorithm, which looks for the structure with maximum whole weight value as predicted RNA secondary structure. The so-called compound weight is constant weight plus dynamic weight. Given a section of stem Stm(i, j, k), its compound weight is calculated by the following formula:

$$W_{(i,j,k)} = \sum_{l=0}^{k-1} w_{i+l,j-l} + \frac{1}{3}(w_{\mathrm{GC}} + w_{\mathrm{AU}} + w_{\mathrm{GU}})\sqrt{k}$$

where the first term is the constant weight sum of base pairs in stem and the constant weight value of GC, AU, GU is marked as w_{GC}, w_{AU} and w_{GU}, respectively. The second term is a bonus weight for successive base pairs in stem that is a product of average weight and square root of stem length. Since its dynamical change with stem length it is called a dynamic weight.

Based on above weight, we presented a double recursive algorithm to search the stems with maximum weight sum and potential pseudoknots. The main process of our algorithm is like this: Firstly find out a section of stem Stm(i, j, k) with maximum weight sum in the whole sequence $R = r_1, r_2 \cdots r_n$, which divides the remaining strand of the sequence into three parts as Fig. 3 shows. Sometimes the length of strand Str(1, i) or Str(j, n) may be 0; Secondly continue to search the maximum weight sum stem in strand Str(1, i), Str($i+k, j-k$) and Str(j, n), respectively until there is no base pair or until base pair cannot meet the conditions limited by CONST1, CONST2 in the foregoing definition; Finally jointly search the two sections of Str(1, i) and Str(j, n) for successive base pairs matched between unmatched bases of strand Str(1, i) and those of Str(j, n). If a maximum weight sum stem is found out then the searching for remaining sections will be continued. Because the last step for joint search is executed in each layer of recursive procedure, the whole algorithm is a double recursive algorithm.

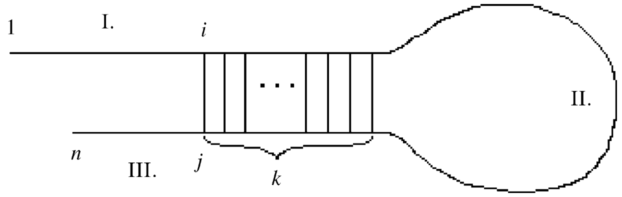

Fig. 3. Sketch map of the recursive algorithm. The stem formed in every step of iteration divides a segment of RNA sequence into three parts.

The detailed steps of the present algorithm is as follows:

(1) Initialize weight matrix wMatrix[n][n]. If (r_i, r_j) is paired, wMatrix[i][j] is set as w_{GC}, w_{AU} or w_{GU} according to the situation, else set as 0. Initialize base-paired state matrix sMatrix[n][n], all its elements are put 0.

(2) Enter the recursive procedure WeightMatch(p, q). At the beginning $p = 1$, $q = n$:

- (2.1) If $p + \mathrm{CONST1} \geq q$, then return, else run the next step.
- (2.2) Search weight matrix for successive base-paired stem with maximum weight sum according to formula (1). If find out, mark it as Stm(i, j, k) and run the next step, else return. Here must be $k \geq \mathrm{CONST2}$.
- (2.3) Record the stem Stm(i, j, k) and set sMatrix[i][j], . . ., sMatrix[$i + k - 1$][$j - k + 1$] as 1.
- (2.4) Let $p = p$, $q = i - 1$, go to (2.1).
- (2.5) Let $p = i + k$, $q = j - k$, go to (2.1).
- (2.6) Let $p = j + 1$, $q = q$, go to (2.1).
- (2.7) Enter the secondary recursive procedure SecMatch(a, b, c, d). At this moment set $a = p$, $b = i - 1$, $c = j + 1$, $d = q$.
  - (2.7.1) If min($b - a + 1$, $d - c + 1$) < CONST2, return, else go to the next step.
  - (2.7.2) Search the unmatched bases in section ab and cd and see whether successive base pairs can be formed between them. If exist, then find out the stem with maximum weight sum and mark it as Stm(i', j', k') and go to next step, else return.
  - (2.7.3) Record the stem Stm(i',j',k') and set the corresponding base-paired states as 1.
  - (2.7.4) Let $a = a$, $b = i' - 1$, c=j' + 1, $d = d$, go to (2.7.1).
  - (2.7.5) Let $a = i' + k'$, $b = b$, $c = c$, $d = j' - k'$, go to (2.7.1).

(3) Traverse the base-paired state matrix sMatrix, output the base pairs (r_i, r_j) with the state value of 1.

Above-mentioned algorithm has followed two main principles. One is the whole weight sum maximization; another is first-near-last-far principle that means juxtaposed stems are considered first and nested stems are second. We assume RNA sequence forms short distance base pairs first in local regions when it is folding, and then carries on long distance base pairing. van Batenburg et al. (1995) have adopted this principle in their genetic algorithm and thought this accorded with the RNA folding way better such that the predicted results were closer to real structures.

Table 1
Test set and the predicted results

RNA	GenBank AC	*L* (nt)	EP/ES	PP/PS	TP	FP	SS (%)	SP (%)	PK
B. subtilis Met-tRNA	K00310	77	20/4	20/4	20	0	100.00	100.00	3/1
B. subtilis Tyr-tRNA	K00269	85	28/5	27/5	27	0	96.43	100.00	0
E. coli Ala-tRNA	M10927	76	23/4	23/4	23	0	100.00	100.00	6/2
E. coli Cys-tRNA	K00179	74	25/5	22/4	20	2	80.00	90.91	0
E. coli Gly-tRNA	M25087	75	19/4	21/4	16	5	84.21	76.19	0
E. coli Tyr-tRNA	M10878	85	24/5	24/5	24	0	100.00	100.00	0
S. cerevisiae Phe-tRNA	K01553	76	21/4	21/4	21	0	100.00	100.00	4/1
S. typhi. Gly-tRNA	K00197	74	22/4	23/4	22	1	100.00	95.65	6/2
Average of the tRNA:							95.08	95.34	
E. coli 5S rRNA	X00414	120	39/9	39/7	27	12	69.23	69.23	0
E. coli RprA RNA	AF326576	106	28/6	27/4	18	9	64.29	66.67	10/3
E. coli UptR RNA	AF272839	96	28/5	26/5	17	9	60.71	65.38	3/1
S. typhi. RprA RNA	NC_003197	107	28/6	27/4	18	9	64.29	66.67	11/3
Average of the noncoding RNA							64.63	66.99	

EP: expected base pair number, ES: expected stem number; PP: predicted base pair number, PS: predicted stem number; TP: correctly predicted base pair number, FP: incorrectly predicted base pair number; SS: sensitivity, SP: specificity; PK: predicted pseudoknots base pair number and stem number.

4. Results

We randomly selected 12 RNA sequences from GenBank database as test set, among which there are eight tRNA sequences and four noncoding RNA sequences. Because the recognized RNA sequences whose real secondary structures have been known are few, we do not know the actual structures of the test set. So we mainly compare our results with the ones predicted by free energy minimization algorithm to assess the accuracy of our algorithm. Table 1 lists the test sequences and their predicted results, where EP/ES is the reference results of free energy minimization algorithm and PP/PS is the results of dynamic weighted matching algorithm.

In this paper the widely used indexes of sensitivity and specificity are adopted to measure the result accuracy. If SS denotes sensitivity and SP denotes specificity, according to the definition of Baldi et al. (2000), there are

$$\mathrm{SS}=\frac{\mathrm{TP}}{\mathrm{EP}},\qquad \mathrm{SP}=\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}$$

where (TP + FP) is the total number of predicted base pairs. Here we do not take count of predicted pseudoknots. We count them alone.

In practical testing we let CONST1 = 3, CONST2 = 3 and three types of constant weight value are, respectively set as $w_{\mathrm{GC}}=11$, $w_{\mathrm{AU}}=8$ and $w_{\mathrm{GU}}=3$. The values of these parameters can also be adjusted properly according to actual circumstances, which are the consulting values offered here. In addition, we stipulate that the GU pair cannot occur at the internal end of stems according to general experience. Namely for stem Stm(i, j, k), the base pair (r_{i+k-1}, r_{j-k+1}) in it cannot be GU pair. Whether allowing occurrence of the GU pair at the both end of a stem or not is also an optional parameter in some algorithms. We regard it as an adjustable option too in our program.

5. Discussion

It can be found out from Table 1 that our algorithm is more effective on the prediction of tRNA sequences while its prediction results of noncoding RNA sequences are not ideal. One reason is that with the increase of the sequence length and more and more long distance interactions, our algorithm which pays the utmost attention to the short distance interaction leads to greater deviation of the final results. Another reason is that it may relate with the parameters we set. After all different kinds and different length of RNAs have their own structural characteristics. It is impossible to have the same effect on RNAs of all types with only one kind of models and parameters. In order to improve the prediction accuracy of long noncoding sequences, we need further study the relationship between long distance base pairing and short distance base pairing in the structures.

Compared with free energy minimization method the greatest advantage of our algorithm is that it can predict pseudoknots. The structure of *S. cerevisiae* Phe-tRNA is known very clearly at present. Fig. 4 shows two results predicted by free energy minimization method and dynamic weighted matching algorithm. Compared with the real structure of *S. cerevisiae* Phe-tRNA, both methods accurately predicted the four sections of main stems. Furthermore, our algorithm predicted a section of pseudoknots pairs between D loop and TψC loop such that our predicted structure was closer to real structure.

Additionally, free energy minimization method needs a set of complicated free energy parameter tabulations and five dynamic programming matrices for different types of structural motifs in the course of calculation. After matrices fill stage there is still a complicated traceback stage. However, our algorithm only needs a double recursion using weighted matching that simplifies the prediction process for RNA structure and its complexity.

To sum up, as RNA secondary structure prediction methods become mature and diversified day by day, it is gradually a

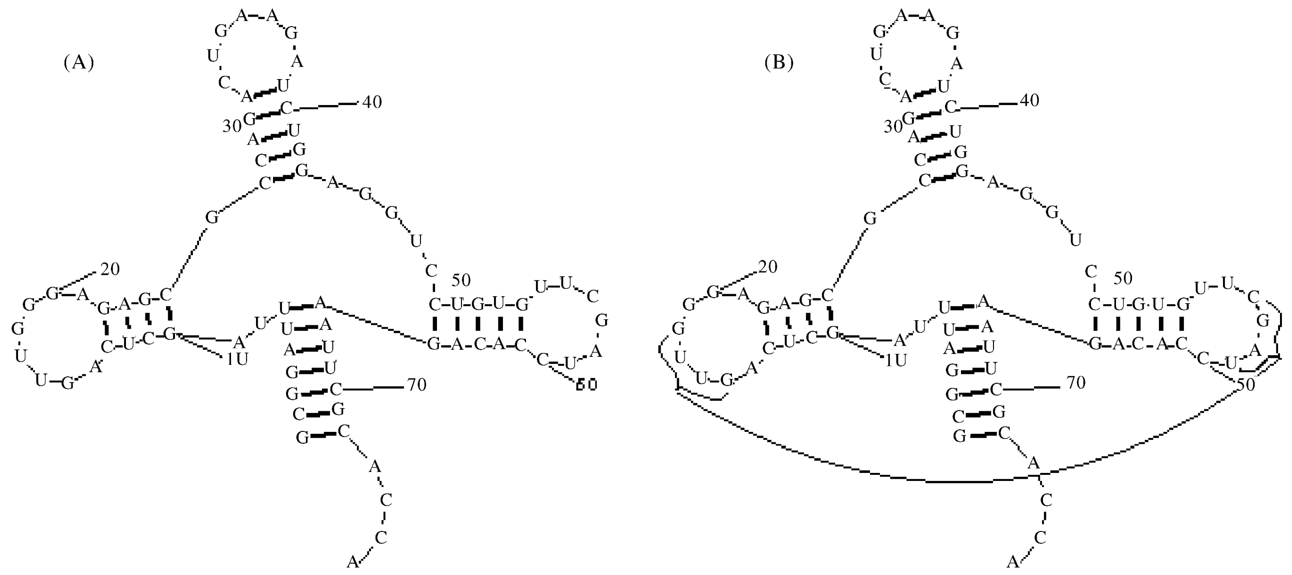

Fig. 4. The two predicted results of *S. cerevisiae* Phe-tRNA. (A) Result of minimum free energy method. (B) Result of dynamic weighted matching algorithm.

developing trend of study to pursue more convenient and more effective methods with pseudoknots and more complicated tertiary structure prediction. The RNA folding algorithm based on dynamic weighted matching that we presented is a very beneficial trial towards this direction.

Acknowledgements

This work was supported by the National High Technology Research and Development Program of China (Grant No. 2002AA234021).

References

Baldi, P., Brunak, S., et al., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424.

Cary, R., Stormo, G., 1995. Graph-theoretic approach to RNA modeling using comparative data. Proc. Int. Conf. Intell. Syst. Mol. Biol. 3, 75–80.

Eddy, S.R., Durbin, R., 1994. RNA sequence analysis using covariance models. Nucleic Acids Res. 22, 2079–2088.

Lyngso, R., Pedersen, C., 2000. RNA pseudokont prediction in energy-based models. J. Comput. Biol. 7, 409–427.

Rivas, E., Eddy, S., 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J. Mol. Biol. 285, 2053–2068.

Ruan, J., Stormo, G.D., Zhang, W., 2004. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. Bioinformatics 20, 58–66.

Schmitz, M., Steger, G., 1996. Description of RNA folding by "simulated annealing". J. Mol. Biol. 255, 254–266.

Shapiro, B.A., Wu, J., 1997. Prediction RNA H-type pseudoknots with the massively parallel genetic algorithm. CABIOS 13, 459–471.

Tabaska, J., Cary, R., et al., 1998. An RNA folding method capable of identifying pseudokonts and base triples. Bioinformatics 14, 691–699.

van Batenburg, F., Gultyaev, A., Pleij, C., 1995. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. J. Theor. Biol. 174, 269–280.

Zuker, M., Stiegler, P., 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 9, 133–148.