Dynamic matching algorithm for viral structure prediction

Hengwu Li[1]*, Daming Zhu[2], Caiming Zhang[3], Zhengdong Liu[3], Huijian Han[1] and Zhenzhong Xu[1]

[1]School of Computer Science and Technology and Shandong Provincial Key Laboratory of Digital Media Technology, Shandong University of Finance and Economics, Jinan, China
[2]School of Computer Science and Technology and Shandong Provincial Key Laboratory of Software Engineering, Shandong University, Jinan, China
[3]School of Computer Science and Technology, Shandong Jianzhu University, Jinan, China

Abstract: Most viruses have RNA genomes, their biological functions are expressed more by folded architecture than by sequence. Among the various RNA structures, pseudoknots are the most typical. In general, RNA secondary structures prediction doesn't contain pseudoknots because of its difficulty in modeling. Here we present an algorithm of dynamic matching to predict RNA secondary structures with pseudoknots by combining the merits of comparative and thermodynamic approaches. We have tested and verified our algorithm on some viral RNA. Comparisons show that our algorithm and loop matching method has similar accuracy and time complexity, and are more sensitive than the maximum weighted matching method and Rivas algorithm. Among the four methods, our algorithm has the best prediction specificity. The results show that our algorithm is more reliable and efficient than the other methods.

Keywords: RNA structure prediction; dynamic matching; pseudoknot; algorithm, algorithm, dynamic matching, pseudoknot, RNA structure prediction.

INTRODUCTION

Most viruses have RNA genomes. Among the biological molecules, RNAs functional versatility is unique containing encoding or performing catalysis and transferring genetic information. RNA functions greatly depend on RNA structure and folding.

Predicting RNA structure by computer has become an important issue in bioinformatics, because experimental test of RNA structure is expensive and takes lots of time (Staple *et al*., 2005).

RNA viruses generally have very high mutation rates. The disruption of RNA-silencing pathways in the host by virus-encoded counter defense molecules is probably the most common mode of disease induction in virus-infected plants. So RNA fold and pathways is important to disease controlling.

RNA secondary structure can be predicted satisfactorily in polynomial time. Among the various RNA structures, pseudoknots are the most typical. Pseudoknots play a variety of diverse roles in biology (Staple *et al*., 2005). It has been confirmed that pseudokonts exist in many RNAs (Mathews *et al*., 2006).

In general, RNA secondary structures prediction doesn't contain pseudoknots because of its difficulty in modeling. Finding the best secondary structure containing arbitrary pseudoknots has been proved to be NP-hard (Ieong *et al*., 2003).

Most method of predicting pseudoknots adopt heuristic searches, such as genetic algorithm and quasi-Monte Carlo search. They inherently sacrifice optimality (Ren *et al*., 2005).

Dynamic programming is another approach to predict pseudoknot. The tractable subclass of pseudokonts can be predicted in $O(n^4)$-$O(n^6)$ time with it (Rivas *et al*., 1999; Reeder *et al*., 2004; Hengwu *et al*., 2006). It is impractical for long sequences.

The third method of predicting pseudoknts is comparative analysis (Barrette *et al*., 2001). It is more reliable than above approaches, but doesn't fit single sequence.

RNA is folded as the process of transcribed into RNA from DNA. Current RNA structure prediction by calculating the global optimal structure, does not reflect the dynamic folding mechanism of RNA (an Batenburg *et al*., 2000).

Dynamic programming can predict optimal structure with minimum energy, but the native fold often has a different structure of suboptimal energy. A case may be made that the natural folding process of RNA and the simulated folding of RNA using an evolutionary algorithm, which includes intermediate folds, have much in common (Kay *et al*., 2006).

In this paper, an adapted dynamic iterated matching algorithm is presented to predict RNA secondary

**Corresponding authors*: e-mail: hengwu@sdufe.edu.cn

structures including pseudoknots. We tested and verified the algorithm on some viral RNA families. The results show that our algorithm has good sensitivity and specificity.

MATERIALS AND METHODS

Let a single-stranded RNA sequence be $s=s_1s_{2...}s_n$, where each base $s_i \in \{A, U, C, G\}$, $1 \le i \le n$. The segment of s can be expressed as subsequence $s_{i,j} = s_i\, s_{i+1...}s_j$, $1 \le i \le j \le n$.

If s_i & $s_j \in$ (*A*&*U*, *C*&G, *U*&*G*, *U*&*A*, *G*& *C*, *G*&*U*), then s_i and s_j may constitute a base pair (i, j). RNA secondary structure S is a set of non-crossing base pairs, that is, $S=\{(i, j)\}$, where $i, j \in \{1,2,\ldots,n\}$ and $i<j$-4. S is a matching, because no base joins more than one base pair. If (i, j) and (k, l) $\in S$, then they are juxtaposed (e.g. $i<j<k<l$) or nested (e.g. $i<k<l<j$).

Base pair (i, j) and internal unpaired subsequence $s_{i+1, j-1}$ are called loops. If (i, j) and (i+1, j-1)$\in S$, then (i, j) and (i+1, j-1) form stack (i, i+1: j-1, j). The helix (i, i+m: j-m, j) consists of m($\ge$1) consecutive stacks. The length of helix (i, i+m: j-m, j) is m+1.

If (i, j) and (k, l) are base pairs of s, and $i<k<j<l$, then they forms a pseudoknot. A pseudoknot consists of two crossing helixes and three unpaired subsequences.

Our algorithm is based on basic secondary structure prediction algorithm, such as MFOLD algorithm (Zuker *et al*., 1981) or loop matching algorithm. We then introduce dynamic iterated matching (DIM) algorithm, to compute a secondary structure including pseudoknots.

DIM runs MFOLD algorithm multiple times, and each time only selects the helixes with the highest score.

DIM algorithm

It is in the transcription process experiment that RNA is folded with synthesis in the process of transcribing into RNA from DNA. Current RNA structure prediction by calculating the global optimal structure prediction, not reflect the dynamic folding mechanism of RNA.

So we fold RNA sequence with synthesis in the process of transcribed into RNA from DNA, as the process of the RNA folding. By gradually increasing the length of RNA sequence, and each time only select the helix with the highest score, e.g. the most reliable one.

DIM algorithm is as follows:

(1) Input sequence s
(2) Select a score matrix or compute one from a sequence alignment to calculate the score of one helix.
(3) For (i=startLen, i<sequenceLen, i=i+stepLen)
  3.1 Run the basic secondary structure prediction algorithm on $s_{1,i}$. Add the helix to the helix list L.
  3.2 Combine the helixes adjoined by small loops with several bases.
  3.3 Compute the score to each helix in L. Select the helix H with the highest score and add H into the helix list S.
  3.4 Replace the bases of H with x, and clear L.
End for
(4) Run the basic secondary structure prediction algorithm on the rest of s, until there is no helix with the length more than 2. Add the helixes into the helix list S.
(5) Output S.

Initially startLen=stepLen=sequenceLen/k, k is a constant. First, the algorithm computes the secondary structure of $s_{1,\text{startLen}}$. Then, the algorithm computes the secondary structure of $s_{1,i}$ with increasing sequence length. At the end of the algorithm, the algorithm computes the secondary structure of the rest of s. S is the optimal structure under former selection.

Energy parameter

For the nested structures, hairpin loops, bulges, stacks, internal loops and multi-loops, we have used the same set of energies as used in MFOLD. For coaxial stacking, we have adopted the same energy as used in PKNOTS.

Computation of helix score

The scores of helixes in the list L were calculated as hairpin loops. When we combine two helixes adjoined by small loop with several bases, the combination can form internal loop, bulge or coaxing stacking, which depends on the position of the small loop. So the score of the combination was computed by accordingly type.

Computation of pseudokonts

DIM runs MFOLD algorithm multiple times on different sequence, and each time only selects the helixes with the highest score. Once the helix was selected, its bases were replaced with x in the back sequences. The selected helixes in the list S have no limitation to nested or juxtaposed relations, so they can form pseudoknots.

Complexity of DIM

The worst case complexity of the algorithm can be easily determined. The basic secondary structure algorithm, which takes $O(n^3)$ in time and $O(n^2)$ in space, is repeated k times in the third step and p times in the fourth step, where k<10 and $p \le n/6$. But in general p is small, so the worst case time complexity keeps $O(n^3)$. The space complexity remains $O(n^2)$. DIM algorithm sacrifices the optimality to prefer long helixes over arbitrarily crossed helixes.

RESULTS

We now give some prediction results from DIM algorithm. We compared DIM algorithm with MWM, PKNOTS (Rivas *et al*., 1999), and ILM. We selected these algorithms because they are well-developed algorithms in their respective categories. MWM is the classic algorithm to predict optimal pseudoknotted structures with comparative approach. PKNOTS is the classic dynamic programming algorithm for standard RNA with thermodynamic models, and the prediction accuracy of pseudoknot is high on short sequences with the length less than 150.

Then we tested and verified all four algorithms, MWM, PKNOTS, ILM and DIM, on some viral sequences with default parameters.

Individual sequences from HIV-1-RT virus, HDV ribozyme RNA, TMV RNA, TYMV RNA, and anti-genomic HDV ribozyme RNA were selected. All these sequences has at least one pseudoknot.

We computed the prediction accuracy with sensitivity and specificity. Let *EP* be the number of base pairs in a verified reference structure, *TP* the number of correctly predicted base pairs (true prediction) and *FP* the number of predicted base pairs that do not exist in the verified reference structure (false prediction). We defined sensitivity as *TP/EP*, and specificity as *TP/(TP+FP)* according to Baldi *et al*. (2000).

The sequences are listed in table 1. The results are showed in table 2. We computed the results of PKNOTS with the software PKNOTS5, and do that of ILM from the web site.

EHLX: the expected number of helixes; EK: the expected number of pseudoknots. Only helixes with the length more than 2 are counted.

Table 1: Sequences used in the experiments

RNA	*L* (nt)	EP	EHLX	EK
HIV-1-RT	35	11	2	1
TYMV	86	23	5	1
TMV-3 -up	84	25	6	3
TMV-3 -down	105	34	7	2
HDV	87	28	5	1
Anti-HDV	91	25	5	1

DIM and ILM, MWM and PKNOTS show similar prediction accuracies, and the former group is better than the later group. In specificity side, DIM exhibits the best results, and MWM does the worst results among the four methods. In accuracy side, ILM has the best results, and MWM does the worst results among the four methods.

For short sequence, four methods all have high accuracy. Among the whole 9 pseudoknots, DIM shows the best results and predicts correctly 7 pseudoknots, and MWM does the worst results and folds only 4 pseudoknots. DIM missed a pseudoknot each in HDV and TMV-3-end upstream. PKNOTS and ILM also show good accuracy of pseudoknot. PKNOTS lost all three pseudoknots in TMV-3-up and a short helix in HDV, otherwise it was almost perfect. ILM lost a pseudoknot each in TMV-3-up and TMV-3-down sequences; but predicted two pseudoknots each in HDV and Anti-HDV, which do not exist in the verified reference structure. Also, MWM gave the worst sensitivity and specificity among all four methods.

By using the score matrix from a sequence alignment to calculate the score of one helix, the accuracy of DIM can further improved. It will combine the advantage of thermodynamic methods with comparative approaches.

DISCUSSION

In this paper, DIM algorithm for RNA secondary structure prediction including pseudoknots is presented.

Thermodynamic method can find optimal structures. However, it usually has great time and memory complexity, and is impractical for long sequences of more than two hundred bases. Moreover due to the lack of proper energy parameters and models, its result is often not satisfactory. Comparative method is more reliable on predicting pseudoknot structures, but we don't know how the distance between the predicted structure and the optimal one. By combining the advantages of thermodynamic methods with comparative approaches, DIM can predict RNA secondary structures including pseudoknots efficiently and reliably, using only single or a few sequences. DIM method does not give a theoretically optimal structure, it forms the stable helixes. It turns out that the DIM method significantly improves the specificity, especially with only single or a few sequences.

We have tested the algorithm on some viral RNA families. Comparisons show that our algorithm and loop matching method has similar accuracy and time complexity, and are more sensitive than MWM method and Rivas algorithm. Among the four methods, our algorithm has the best prediction specificity. The results show that our algorithm is more reliable and efficient than the other methods.

RNA is folded as the process of transcribed into RNA from DNA. So we fold RNA sequence by gradually increasing the length of RNA sequence, and each time only select the helix with the most reliable one. It fits the natural folding, so DIM algorithm gets the best prediction specificity. In fact, the third step of DIM algorithm forms

Table 2: Summary of prediction results on individual RNA sequences

RNA	DIM			MWM			PKNOTS			ILM		
	T P (*SS*)	*SP*	*K*	*T P (SS)*	*SP*	*K*	*T P* (*SS*)	*SP*	*K*	*T P (SS)*	*SP*	*K*
HIV-1-RT	11(100)	100	1/1	11(100)	84.6	1/1	11(100)	100	1/1	11(100)	100	1/1
TYMV	24(100)	96.0	1/1	24(100)	63.2	1/1	23(100)	96.0	1/1	24(100)	82.8	1/1
TMV-3-up	21(84.0)	91.3.	2/3	17 (68.0)	41.5	1/3	13(52.0)	68.4.	0/3	20(80.0)	80.0	2/3
TMV-3-down	32(94.1)	94.1.	2/2	25(73.5)	49.0	0/2	32(94.1.)	94.1.	2/2	26(76.5)	68.4	1/2
HDV	18(64.3)	62.1	0/1	19(67.8)	45.2	0/1	24(85.7)	75.0	1/1	25(89.3)	62.5	3/1
Anti-HDV	21(84.0)	63.6	1/1	17(70.8)	38.6	1/1	11(44.0)	34.3	1/1	22(88.0)	44.0	3/1
Average	87.7	84.5	7/9	80.0	53.7	4/9	79.3	78.0	6/9	89.0	73.0	11/9

the frame of RNA structure, and it prevents the formation of unnatural helixes.

In general, DIM method and algorithm can used as a tool for the prediction of RNA secondary structures including pseudoknots because of the high accuracy and low need on computational resources.

ACKNOWLEDGMENT

This work was supported in part by NSFC under grant NO.61070019, 61272431, NSFC of Shan Dong under grant NO.ZR2011FL029, ZR2013FM016, the Open Project Program of the Shandong Provincial Key Lab of Software Engineering under grant No.2011SE004, and Program for Scientific Research Innovation Team in Colleges and Universities of Shandong Province.

We thank the anonymous reviewers for their detailed and very useful comments.

REFERENCES

Akmaev V, Kelley S and Stormo G (1999). A phylogenetic approach to RNA structure prediction. *International Conference on Intelligent Systems for Molecular Biology*, **7**: 10-17.

Barrette I, Poisson G, Gendron P and Major F (2001). Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching. *Nucleic Acids Res.*, **29**: 753-778.

Cary RB and Stormo GD (1995) Graph-theoretic Approach to RNA Modeling Using Comparative Data. *The Third International Conference on Intelligent Systems for Molecular Biology,* **3**: 75-80.

Chuanming W, Min P and Kui C (2004). RNA folding. *Nature,* **26**: 249-255.

Eddy SR and Durbin R (1994). RNA sequence analysis using covariance models. *Nucleic Acids Res.,* **22**: 2079-2088.

Hengwu L, Daming Z, Zhendong L and Hong L (2006). Prediction for RNA planar pseudoknots. *Prog. Nat. Sci.,* **17**: 717-724.

Hofacker IL (2003). Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**: 3429-3431.

Ieong S, Kao MY and Lam TW (2003). Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs. *J. Compt. Biol.*, **6**: 981-995.

Kay C, Wiese, Andrew (2006). Hendriks: Comparison of P-RNA Predict and mfold - algorithms for RNA secondary structure prediction. *Bioinformatics,* **22**: 934-942.

Mathews DH, Sabina J, Zuker M and Turner DH (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol*., **288**: 911-940.

Mathews DH and Turner DH (2006). Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, **16**: 270-278.

Reeder J and Giegerich R (2004). Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinf.*, **5**: 104.

Ren J, Rastegari B, Condon A and Hoos HH (2005). Hot Knots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA,* **11**: 1494-1504.

Rivas E and Eddy SR (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol*., **285**: 2053-2068.

Shijie C, Zhijie T, Song C and Wenbing Z (2006). The Statistical Mechanics of RNA Folding. *Physics*, **3**: 218-229.

Staple DW and Butcher SE (2005). Pseudoknots: RNA structures with diverse functions. *PLoS Biol.*, **3**: e213.

van Batenburg F.H., Gultyaev A.P., Pleij C.W., Ng J., Oliehoek J. (2000). Pseudobase: a database with RNA pseudoknots. *Nucleic Acids Res.*, **28**: 201-204.

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