

An Iterative Loop Matching Approach to the Prediction of RNA Secondary Structures with Pseudoknots

Jianhua Ruan¹, Gary D. Stormo^{2,1} and Weixiong Zhang^{1,2}
Department of Computer Science¹ and Department of Genetics²
Washington University in St. Louis, St. Louis, MO 63130, USA
jruan@cse.wustl.edu, stormo@ural.wustl.edu, zhang@cse.wustl.edu

Abstract

In this paper we present a heuristic algorithm, iterative loop matching, for predicting RNA pseudoknots. The method can utilize either thermodynamic or comparative information or both, thus is able to predict for both aligned and individual sequences. Using 8–12 homologous sequences, the algorithm correctly identifies more than 90% of base-pairs for short sequences and 80% overall. It correctly predicts nearly all pseudoknots, while having very few false predictions. Comparisons show that our algorithm is more sensitive and more specific than existing methods. In addition, our algorithm is very efficient and can be applied to sequences up to several thousands of bases long.

1 Introduction

RNA secondary structures without pseudoknots obey a “nested” constraint: for any two base-pairs (i, j) and (k, l) , where $i < j$ and $k < l$, it must satisfy either $i < j < k < l$ or $i < k < l < j$. Due to this property, prediction of RNA secondary structures can be solved by a dynamic programming procedure, the weighted loop matching algorithm (LM) [6]. Given an RNA sequence $S[1..N]$ and a score matrix B , where $B(i, j)$ is the score for the i th base to form a base-pair with the j th base, the LM algorithm computes the score of the best structure for each subsequence of S , starting from the shortest one. Let these scores be stored in a matrix Z . At the end of the procedure, $Z(1, N)$ stores the score of an optimal structure for sequence $S[1..N]$, which can be obtained by tracing back matrix Z . The computation and trace-back can be done in $O(n^3)$ time and $O(n^2)$ space [6].

Recently several dynamic programming algorithms have been proposed for the prediction of pseudoknots using thermodynamic approaches [7, 10, 1], but unfortunately they are very expensive (e.g., $O(n^6)$ in time and $O(n^4)$ in space). Yet, due to the lack of proper energy parameters, their ac-

curacies are not satisfactory. On the other hand, comparative methods can produce more reliable predictions, but are often done in an ad hoc manner and require expert intervention. Maximum weighted matching (MWM) [9], an algorithm for pseudoknot prediction using covariance information, suffers from low prediction accuracy in many cases since it allows many types of unrealistic interactions to happen.

2 Algorithm

We can extend the basic LM algorithm to accommodate pseudoknots. A pseudoknot can be considered as an interaction between two loop regions of a secondary structure. Therefore we can identify a pseudoknot in two steps. First we predict a secondary structure using the LM algorithm. Then we remove all the paired bases and predict another secondary structure on the remaining sequence. Combining these two secondary structures produces a structure that may contain pseudoknots. However, this idea often fails in practice. The bases of some pseudoknots may be predicted to form base-pairs with wrong partners in the first step, which prevents them from being correctly identified later. To solve this problem, we can run the LM procedure multiple times, each time we only accept the group of base-pairs that appear to be the most reliable, e.g., with the highest scores. The sketch of the new algorithm, called iterative loop matching (ILM), is as follows:

1. Prepare a base-pairing score matrix $B[1..N][1..N]$ from a sequence or a sequence alignment.
2. Run the LM algorithm using matrix B to produce matrix Z and trace back Z to get a base-pair list L . Identify all helices in L and combine helices separated by small internal loops or bulges. If no helix is identified, go to step 5.
3. Pick helix H that has the highest score, merge H into the base-pair list S to be reported. Remove bases of H

from the initial sequence. Update the score matrix B accordingly.

4. Repeat step 2–3 until no remaining bases.
5. Report base-pair list S and terminate.

Score matrix B can be prepared by a variety of ways. When there is more than one homologous sequence, a combination of thermodynamic and phylogenetic scores can be used, while for a single sequence only thermodynamic scores are available. Detailed description of the score matrix B in step 3 simply means to remove rows and columns corresponding to bases that have been paired. Thus in each iteration, only a part of Z needs to be re-computed. Suppose that a previous iteration has selected a base-pair (p, q) . Then the subsequent iteration needs to re-compute $Z(i, j)$ only if i and j are separated by either p or q , i.e., $i < p < j$ or $i < q < j$, since the value of $Z(i, j)$ only depends on the bases between i and j .

The worst case time complexity of the algorithm is $O(n^4)$ while in average case it is close to $O(n^3)$. The space complexity remains $O(n^2)$.

Unlike the existing algorithms in [7, 10, 1], the ILM algorithm does not guarantee optimality. However, it does ensure that the score of the predicted structure is at least no worse than that predicted by the basic LM algorithm.

3 Results and Discussion

We compare the ILM algorithm with two existing methods. We first compare ILM and MWM [9] on aligned sequences, using combined thermodynamic and phylogenetic scores. Then we test ILM, MWM and pknots [7] on individual sequences, using thermodynamic scores alone. Partial results are listed in Table 1. In the first experiment, with 8–12 homologous sequences, ILM correctly identifies more than 90% of the base-pairs for short sequences and 80.0% over all sequences, while the corresponding percentages for MWM are 60-85% and 59.2%. ILM correctly predicts all pseudoknots except one in 16S rRNA that involves a long-range 3bp helix. Furthermore, ILM produces much fewer false positive base-pairs than MWM. For sequences without pseudoknots, such as 5S rRNA, ILM produces very few spurious base-pairs. When individual sequences are used (e.g., TMV and HDV), ILM and pknots show similar prediction accuracies and are both better than MWM (data for MWM not shown), but ILM is much faster than pknots. For example, it takes 65 minutes for pknots to fold an RNA of 105 bases long, but only less than 0.1 second for ILM.

The ILM algorithm has been implemented in ANSI C and is freely available at <http://www.cse.wustl.edu/~zhang/projects/rna/ilm/>.

Table 1. Summary of prediction results.

sequence information				MWM/pknots		ILM	
1	2	3	4	5	6	7	8
5S	12	40	0	32 / 55	0	38 / 40	0
SRP	12	78	1	68 / 114	1	76 / 101	1
tmRNA	8	106	4	73 / 171	3	93 / 126	4
16S	10	478	2	243 / 684	0	351 / 515	1
HDV	1	28	1	24 / 32	1	28 / 34	1
TMV	1	25	3	13 / 33	0	20 / 25	2

Column 1 - 4: information about the test sequences (1: sequence name; 2: number of sequences used for alignment; 3: number of base-pairs; 4: number of pseudoknots). Column 5 and 6: prediction results of MWM or pknots (pknots were used for HDV and TMV). Column 7 and 8: prediction results of ILM. Column 5 and 7: number of correctly predicted base-pairs / total predicted base-pairs. Column 6 and 8: number of correctly predicted pseudoknots. Alignment and structure of SRP RNA are from [4], tmRNA from [5], and 5S and 16S rRNA from [2]. Structures of TMV and HDV are from [11] and [3], respectively.

References

- [1] T. Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104(1-3):45–62, 2000.
- [2] J. Cannone et. al. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other rnas. *BMC Bioinformatics*, 3(1):2, 2002.
- [3] A. Ferre-D’Amare, K. Zhou, and J. Doudna. Crystal structure of a hepatitis delta virus ribozyme. *Nature*, 395(6702):567–74, Oct 1998.
- [4] J. Gorodkin et. al. SRPDB (signal recognition particle database). *Nucleic Acids Res*, 29(1):169–70, Jan 2001.
- [5] B. Knudsen et. al. tmRDB (tmRNA database). *Nucleic Acids Res*, 29(1):171–12, Jan 2001.
- [6] R. Nussinov et. al. Algorithms for loop matchings. *SIAM J. Appl. Math.*, 35(1):68–82, 1978.
- [7] E. Rivas and S. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol*, 285(5):2053–2068, Feb 1999.
- [8] J. Ruan, G. Stormo, and W. Zhang. An iterative loop matching approach to the prediction of rna secondary structures with pseudoknots. *Bioinformatics*, accepted.
- [9] J. Tabaska et. al. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14(8):691–69, 1998.
- [10] Y. Uemura et. al. Tree adjoining grammars for RNA structure prediction. *Theoretical Computer Science*, 210(2):277–303, 1999.
- [11] A. van Belkum et. al. Five pseudoknots are present at the 204 nucleotides long 3’ noncoding region of tobacco mosaic virus RNA. *Nucleic Acids Res*, 13(21):7673–786, Nov 1985.