BIOINFORMATICS

Vol. 20 no. 1 2004, pages 58–66
DOI: 10.1093/bioinformatics/btg373

An Iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots

Jianhua Ruan[1,], Gary D. Stormo[1,2] and Weixiong Zhang[1,2,*]*

[1]Department of Computer Science and [2]Department of Genetics, Washington University in St. Louis, St. Louis, MO 63130, USA

Received on March 16, 2003; revised on June 20, 2003; accepted on July 9, 2003

ABSTRACT

Motivation: Pseudoknots have generally been excluded from the prediction of RNA secondary structures due to its difficulty in modeling. Although, several dynamic programming algorithms exist for the prediction of pseudoknots using thermodynamic approaches, they are neither reliable nor efficient. On the other hand, comparative methods are more reliable, but are often done in an *ad hoc* manner and require expert intervention. Maximum weighted matching, an algorithm for pseudoknot prediction with comparative analysis, suffers from low-prediction accuracy in many cases.

Results: Here we present an algorithm, iterated loop matching, for reliably and efficiently predicting RNA secondary structures including pseudoknots. The method can utilize either thermodynamic or comparative information or both, thus is able to predict pseudoknots for both aligned and individual sequences. We have tested the algorithm on a number of RNA families. 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 and produces very few spurious base-pairs for sequences without pseudoknots. Comparisons show that our algorithm is both more sensitive and more specific than the maximum weighted matching method. In addition, our algorithm has high-prediction accuracy on individual sequences, comparable with the PKNOTS algorithm, while using much less computational resources.

Availability: The program has been implemented in ANSI C and is freely available for academic use at http://www.cse.wustl.edu/~zhang/projects/rna/ilm/

Contact: jruan@cse.wustl.edu; zhang@cse.wustl.edu

Supplementary information: http://www.cse.wustl.edu/~zhang/projects/rna/ilm/

INTRODUCTION

RNA molecules play many important regulatory, catalytic and structural roles in the cell and a complete understanding of the functions of RNA molecules requires knowledge of their three-dimensional structures. Since it is often difficult to obtain spectrum data for large RNA molecules to inspect their structures, reliable prediction of RNA structures from their primary sequences is highly desirable.

Much work has been done on automated RNA secondary structure predictions without pseudoknots. A secondary structure is a list of base-pairs. Base-pair (i, j) and (k, l) are said to be *compatible* if they are either juxtaposed (e.g. $i < j < k < l$, Fig. 1A) or nested (e.g. $i < k < l < j$, Fig. 1B). Otherwise they are called *incompatible* (e.g. $i < k < j < l$, Fig. 1C). Such an incompatible structure is known as a *pseudoknot*. More complex pseudoknots may occur if three or more base-pairs cross each other (Fig. 1D).

Most computational methods for the prediction of RNA secondary structures can be classified into three families: thermodynamic, comparative and hybrid approaches. Thermodynamic approaches (Zuker and Stiegler, 1981; Hofacker *et al*., 1994) use dynamic programming to compute the optimal secondary structure for a single RNA sequence with globally minimal free energy, based on a set of experimentally determined energy parameters (Freier *et al*., 1986; Mathews *et al*., 1999). Such methods have been successful for relatively short RNAs. When a number of homologous sequences are available, comparative approaches are more reliable than thermodynamic approaches, and have been used to establish the structures of most known RNA families. These approaches compute a consensus structure on a set of aligned RNA sequences by looking for covariance evidence between each pair of bases. Quantitative measures of covariance have been implemented in χ^2 statistics (Chiu and Kolodziejczak, 1991) and mutual information (Gutell *et al*., 1992). Gulko and Haussler (1996) and Akmaev *et al*. (1999) also extended the approach to take into account explicitly the phylogeny of the sequences and showed some positive results. The third family of methods, which have emerged recently, combines the advantages of the first two (e.g. Luck *et al*., 1999; Juan and Wilson, 1999; Hofacker *et al*., 2002). These methods take both thermodynamic stability

*To whom correspondence should be addressed.

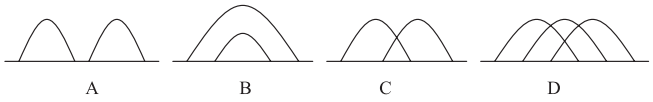

Fig. 1. Diagrammatic representation of different types of relationships between base-pairs. The straight lines represent primary sequences. An arc represents a base-pair between the two end-points. **(A)** two base-pairs are juxtaposed. **(B)** two base-pairs are nested. **(C)** two base-pairs cross each other, forming a pseudoknot. **(D)** three base-pairs cross each another, forming three pseudoknots.

and sequence covariance into consideration and are able to produce positive results on as few as three homologous sequences. There are also methods that cannot be classified into any of these three families. Among them, there are a few methods which attempt to align and fold homologous sequences simultaneously (Sankoff, 1985; Gorodkin *et al*., 1997; Mathews and Turner, 2002). They were only successful on short sequences due to their high time and space complexities. Eddy and Durbin (1994) and Sakakibara *et al*. (1994) introduced stochastic context-free grammars to align homologous sequences iteratively and find a consensus structure for them.

A more challenging task of RNA folding is the prediction of pseudoknots. Pseudoknots are important structures that occur in RNA and often have important functional roles (Dam *et al*., 1992). However, relatively little effort has been devoted to automated pseudoknot prediction, partially due to the difficulty in modeling and the complexity in computing. Despite the observation of certain types of pseudoknots, there exists no definitive evidence of what types of pseudoknots are legitimate. As proven by Lyngso and Pedersen (2000b), it is NP-complete (Garey and Johnson, 1979) to predict RNA secondary structures with pseudoknots by free energy minimization in general. By restricting the types of pseudoknots that may occur, several polynomial time and space dynamic programming algorithms have been developed recently (Rivas and Eddy, 1999; Uemura *et al*., 1999; Lyngso and Pedersen, 2000a; Akutsu, 2000). However, these methods still have very high time and space complexities, typically $O(n^5)$ to $O(n^6)$ in time and $O(n^3)$ to $O(n^4)$ in space, making them impractical even for sequences of a few hundred bases long. More practical methods thus must adopt heuristic procedures, such as Monte-Carlo simulation (Abrahams *et al*., 1990) and genetic algorithms (Gultyaev *et al*., 1995; van Batenburg *et al*., 1995). These methods, however, are not guaranteed to find the optimal solution and are unable to say how far a prediction is from the optimal solution. Another dilemma for pseudoknot prediction algorithms based on energy models is that there is little experimentally determined thermodynamic data for pseudoknots.

Comparative approaches can also be applied to the prediction of pseudoknots and are more reliable than thermodynamic approaches. For example, comparative analysis has revealed the existence of pseudoknots in several RNAs (Barrette *et al*., 2001; Wuyts *et al*., 2000; Zwieb *et al*., 1999; Chen *et al*., 2000). However, comparative analysis has typically been done in an *ad hoc* manner from an algorithmic point of view. The only published algorithm we have found that automates pseudoknot prediction by comparative analysis is the maximum weighted matching (MWM) algorithm (Cary and Stormo, 1995; Tabaska *et al*., 1998). The MWM algorithm takes as input a matrix of base-pairing scores, typically covariance scores, and computes an optimal structure allowing all possible base-pairs. However, the MWM algorithm is able to produce meaningful predictions only if the number of homologous sequences is large enough and the alignment is accurate so that covariance signals from their alignment are sufficiently strong. It is vulnerable to noisy data and often results in many spurious base-pairs.

In this paper, we present an adapted dynamic programming algorithm that is capable of predicting RNA secondary structures including pseudoknots. Our algorithm uses combined thermodynamic and covariance information and does not depend on any pseudoknot models, thus is able to detect any type of pseudoknots. We test the algorithm on a number of RNA families, including structures with and without pseudoknots. With 8–12 homologous sequences, our algorithm correctly identifies more than 90% of base-pairs for short sequences ($<$300 nt) and approximately 80% on average. Furthermore, the algorithm correctly predicts all pseudoknots except a 3 bp pseudoknot in the longest sequence and produces very few false positive base-pairs on sequences without pseudoknots. The comparison with the MWM algorithm shows that our algorithm is more specific and sensitive. In addition, we also apply the algorithm to individual sequences and compare its accuracy with an algorithm based on free energy minimization, the PKNOTS algorithm (Rivas and Eddy, 1999). Our algorithm exhibits an accuracy comparable with that of the PKNOTS algorithm, while having much lower time and space complexity.

ALGORITHMS

Our algorithm is based on the loop matching (LM) algorithm (Nussinov *et al*., 1978), which we will describe briefly first. We then introduce a new algorithm, called the iterated loop matching (ILM) algorithm, to compute a secondary structure including pseudoknots. We will also discuss the score matrix used in our experiments.

Loop matching

Given a matrix B, where $B(i, j)$ is the score for the i-th residue forming a base-pair with the j-th residue, the LM algorithm finds a best-score secondary structure *without pseudoknots*. To reiterate, a secondary structure without pseudoknots is a 'compatible' structure as shown in Figure 1A and B. Thanks to this constraint, the secondary structure of a long RNA sequence can be subdivided

into shorter pieces. Formally, for any subsequence $S[i..j]$, with $i+1 < j$, there are only three possibilities: (i) i is single-stranded; (ii) i is paired with j; and (iii) i is paired with some k, where $i < k < j$. Thus, the score of an optimal structure for subsequence $S[i..j]$ can be calculated by Equation (1).

$$Z(i,j) = \max \left\{ \begin{array}{l} Z(i+1,j); \\ Z(i+1,j-1) + B(i,j); \\ \max_k \{Z(i+1,k-1) + Z(k+1,j) \\ \quad +B(i,k)\}, \quad \forall k,\ i < k < j. \end{array} \right\} \tag{1}$$

Initially $Z(i,i) = Z(i,i+1) = \cdots = Z(i,i+LOOP\_LENGTH) = 0$ for all i, where *LOOP_LENGTH* is a parameter that describes the minimum distance required between two paired bases (default $LOOP\_LENGTH = 3$). The algorithm uses a dynamic programming strategy to compute the values of $Z(i,j)$ for all i and j with increasing sequence length. At the end of the algorithm, $Z(1,N)$ is the score of the optimal structure for sequence $S[1..N]$, and the optimal structure can be obtained by tracing back the Z matrix. The computation and trace-back can be done in $O(n^3)$ in time and $O(n^2)$ in space.

In the simplest case, $B(i,j) = 1$ if the i-th residue and the j-th residue can form a Waston–Crick or G–U base-pair, and 0 otherwise. The algorithm finds a secondary structure with the maximal number of base-pairs in this case. We can also assign a different score to each potential base-pair in a more sophisticated way, e.g. by comparative analysis.

Iterated loop matching

We now extend the basic LM algorithm to accommodate pseudoknots. A pseudoknot can be thought as an interaction between two loop regions of a secondary structure, as illustrated in Figure 2; therefore, we could run the LM algorithm twice to identify it. First, we run the LM algorithm to predict a secondary structure as usual. Then the predicted base-pairs are treated as if they were removed from the original sequence, allowing the next iteration of LM to start. By combining base-pairs obtained from the two iterations, we may be able to predict pseudoknotted base-pairs. Similarly, more complicated pseudoknots such as the one in Figure 1D can be identified with more iterations.

However, this idea often fails in practice. The bases that are supposed to form pseudoknots may be involved in some false positive base-pairs during the previous iteration of the LM, which invalidates our efforts of further searching, as shown and explained in Figure 3. To avoid this problem, we use a least-commitment strategy. We run the LM algorithm multiple times, and each time we only accept the base-pairs that appear to be the most reliable, e.g. with the highest score. This modification attempts to avoid possible false predictions from being included, as illustrated in Figure 3.

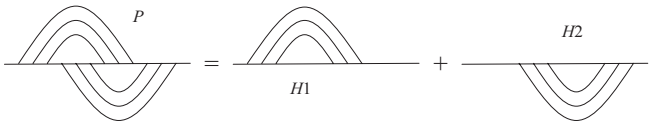

Fig. 2. A pseudoknot (P) can be treated as two separate helices ($H1$ and $H2$) and can be identified by a two-iteration LM. Assume $H1$ is identified by the basic LM, then running the LM algorithm on the remaining single-stranded bases identifies the second helix, $H2$.

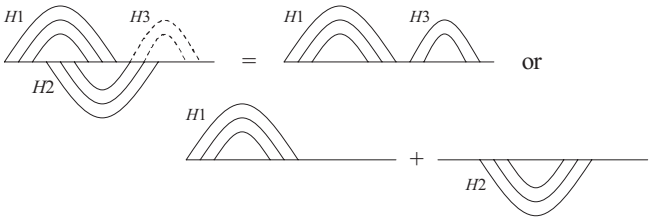

Fig. 3. Pseudoknots that can be correctly identified by the iterated LM algorithm. $H1$ and $H2$ are two true helices forming a pseudoknot. $H3$ is a false helix overlapping $H2$. Scores (R) of the helices satisfy $R(H1) + R(H3) > R(H2)$, $R(H3) < R(H1)$ and $R(H3) < R(H2)$. ILM will correctly predict $H1$ in the first iteration and predict $H2$ in the second iteration. In contrast, basic LM would pick $H1$ and $H3$ together since it gives a higher total score than $H2$ alone. Then even if we run LM again on the remaining single strand, $H2$ cannot be identified correctly since it conflicts with $H3$.

The sketch of the algorithm is as follows:

(1) Prepare a base-pairing score matrix $B[1..n][1..n]$ from a sequence or a sequence alignment, where $B[i][j]$ is the score for the i-th base to pair with the j-th base.

(2) Run the basic LM algorithm using matrix B to produce matrix Z and trace-back Z to get a base-pair list L.

(3) Identify all helices in L and combine helices separated by small internal loops or bulges. If no helix is identified, go to step 7.

(4) Assign a score to each helix by summing up the scores of its constitutive base-pairs. Pick the helix H that has the highest score and merge H into the base-pair list S to be reported.

(5) 'Remove' positions of H from the initial sequence. Update the score matrix B accordingly.

(6) Repeat steps 2–5 until no bases remain.

(7) Report base-pair list S and terminate.

The method to prepare score matrices will be discussed later. Note that in step 5, updating score matrix B simply means removing rows and columns corresponding to bases that have been paired. Alternatively, we use an array M to keep track of the indices of remaining single-stranded bases and run the basic LM algorithm to compute the scores only

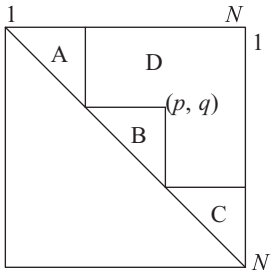

Fig. 4. Three triangle areas of the matrix do not need to be re-computed in each iteration. Let i and j be the row and column index of a cell. (p, q) is the base-pair selected in the previous iteration. A, $i < j < p$; B, $p < i < j < q$; C, $q < i < j$.

for the positions remaining in M. Furthermore, notice that not all elements of Z need to be re-computed in every iteration. 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 $i < q < j$. The optimal score of a subsequence $S[i, j]$, with $1 \leq i < j < q$, does not depend on bases whose indices are greater than q, so it will not change in the next iteration. Thus, three triangle areas of the matrix do not need to be re-computed in each iteration except the first one, as illustrated in Figure 4.

Another issue worth mentioning is that after removing a sequence segment, two previously separated bases may be brought together. Thus the initialization step needs to be modified accordingly. We define the virtual distance of two bases to be the distance between their indices in M. An additional parameter, *VLOOP_LENGTH*, describes the minimum virtual distance required between two paired bases after the first iteration. Two bases with virtual distance less than *VLOOP_LENGTH* are not allowed to pair. The default value of *VLOOP_LENGTH* is set to 3.

The recursion for re-computing Z is given in Equation (2),

$$
Z'(M[i], M[j])
$$

$$
= \begin{cases} Z(M[i], M[j]), \\ \quad \text{if } M[j] < p \ \text{ or } \ M[i] > q \ \text{ or } \ p < M[i] < M[j] < q; \\ 0, \quad \text{if } j - i + 1 < VLOOP\_LENGTH; \\ \max \begin{Bmatrix} Z'(M[i+1], M[j]); \\ Z'(M[i+1], M[j-1]) + B(M[i], M[j]); \\ \max_k \{Z'(M[i+1], M[k-1]) + Z'(M[k+1], M[j]) \\ \quad + B(M[i], M[k])\}, \quad \forall k, \ i < k < j. \end{Bmatrix} \\ \quad \text{otherwise.} \end{cases} \tag{2}
$$

where $M[i]$ is the i-th remaining unpaired base, and p and q, with $p < q$, are two end-points of the helix selected in the previous iteration. In the first iteration of ILM, where $M[i] = i$ and p and q are not defined, the recursion is reduced to be equivalent to Equation (1).

The worst case complexity of the algorithm can be easily determined. The basic LM algorithm, which takes $O(n^3)$ in time and $O(n^2)$ in space, is repeated m times, where m is the total number of helices predicted by the algorithm. Since $m \leq n/2k$, with k being the minimal helix length required, the worst case time complexity is $O(n^4)$. However, m is typically small and sequence length n will be reduced after each iteration. Furthermore, generally the Z matrix needs to be only partially re-computed in each iteration, making the average case complexity close to $O(n^3)$. The space complexity remains $O(n^2)$.

Since the total score of a structure can be considered as a measure of its probability among all possible structures, we usually prefer an algorithm to compute a structure with the highest score. The LM algorithm computes such a structure with the constraint that base-pairs must be compatible with each other. If we loosen this constraint, in the extreme case we have the MWM algorithm (Cary and Stormo, 1995; Tabaska *et al*., 1998) that allows all types of base-pairs. A problem of MWM is that it allows a much larger degree of freedom than real structures and does not take into consideration that helices are the most frequent elements of RNA structures; as a result, MWM often introduces many spurious base-pairs. Between LM and MWM are algorithms that compute optimal structures with restricted pseudoknot models (e.g. Rivas and Eddy, 1999; Uemura *et al*., 1999; Lyngso and Pedersen, 2000a; Akutsu, 2000). However, none of these models have been generally accepted. In contrast, without assuming any pseudoknot model, the ILM algorithm sacrifices the optimality to prefer long helices over arbitrarily crossed lone base-pairs. Although ILM does not guarantee optimality, it guarantees that the score of a predicted structure is no less than that of a structure predicted by the basic LM algorithm. We now give a proof of this claim. Let S_{ILM} denote the score of the structure computed by ILM, and let S_{LM} denote the score of the structure computed by the LM algorithm.

PROPOSITION 1. $S_{ILM} \geq S_{LM}$.

PROOF. We prove it by induction. S_{ILM} is computed by multiple iterations of LM. Let $R(H)$ be the score of helix H, which is the sum of the scores of its constitutive base-pairs. Let h_j^i be the j-th helix predicted in the i-th iteration. Helices are ranked in a non-increasing order of their scores. Note that the algorithm selects the helix with the highest score, i.e. h_1^i, for the i-th iteration. Let $L(i)$ be the total score of selected base-pairs after i iterations. Let $N(i)$ be the total score of all base-pairs predicted in the i-th iteration. Assume that ILM will terminate after m iterations when no helix is identified. By definition,

$$
\begin{aligned} L(i) &= R(h_1^1) + R(h_1^2) + \cdots + R(h_1^i) \\ &= L(i-1) + R(h_1^i), \quad \text{and} \\ N(i) &= R(h_1^i) + R(h_2^i) + \cdots + R(h_j^i). \end{aligned}
$$

Note that $L(m-1) = L(m) = S_{ILM}, N(1) = S_{LM}$ and $N(m) = 0$. Let $S(i) = L(i-1) + N(i)$. Then

$$S(1) = L(0) + N(1) = S_{LM}, \quad \text{and}$$
$$S(m) = L(m-1) + N(m) = L(m-1) = S_{ILM}.$$

Hence, to prove $S_{ILM} \geq S_{LM}$, we only need to prove $S(i+1) \geq S(i), \forall i, 1 \leq i < m-1$.

$$\begin{aligned} S(i+1) - S(i) &= N(i+1) - N(i) + R(h_1^i) \\ &= N(i+1) - (N(i) - R(h_1^i)) \end{aligned}$$

Since $N(i)$ and $N(i+1)$ are computed on the same sequence, except that the subsequence corresponding to h_1^i has been removed before computing the latter, it must satisfy $N(i+1) \geq N(i) - R(h_1^i)$. Hence $S(i+1) \geq S(i), \forall i, 1 \leq i < m-1$, which concludes that $S_{ILM} \geq S_{LM}$.

Several observations of the algorithm help to extend the ILM algorithm while retaining the lower-bound property. First, h_1^i can be any helix predicted in the i-th iteration, not necessarily the one with the highest score. We prefer to choose the helix with the highest reliability to reduce the risk of predicting false base-pairs in the early stages. Although in most cases a higher score indeed indicates higher reliability, this may not be always true. Second, if the algorithm is terminated early after i iterations ($i < m$) and all base-pairs predicted in the last iteration are accepted, the total score of the predicted structure is $S(i)$. $S(i) \geq S_{LM}$ since $S(i)$ is monotonically increasing. By doing so, some spurious pseudoknots may be filtrated since they tend to have low scores. Finally, more than one helix may be selected in each iteration. The number of helices selected in each iteration controls the granularity of the algorithm. The smaller the number, the less is the chance to miss pseudoknots, but the more spurious base-pairs the algorithm may introduce.

Base-pairing score matrix

A number of score matrices have been previously constructed based on an alignment of multiple homologous sequences (Cary and Stormo, 1995; Luck *et al*., 1999; Juan and Wilson, 1999; Hofacker *et al*., 2002). In our implementation of ILM we used the sum of mutual information and helix plot scores as suggested by Tabaska *et al*. (1998), which is essentially a combination of covariance and thermodynamic scores. Another type of combinatorial score matrix based on averaging thermodynamic scores (Luck *et al*., 1999) was also tested (data not shown). We found that the combination of mutual information and helix plot is faster to compute and has comparable prediction accuracy. Here, we briefly describe the calculation of mutual information and helix plot scores. Readers are referred to Cary and Stormo (1995) and Tabaska *et al*. (1998) for more details.

Mutual information scores Assume that we are given a multiple sequence alignment of N sequences. Let $f_i(X)$ be the frequency of base X at aligned position i and let $f_{ij}(XY)$ be the frequency of finding X at position i and Y at position j. The mutual information score between positions i and j, M_{ij}, is calculated as:

$$M_{ij} = \sum_{X,Y} f_{ij}(XY) \log \frac{f_{ij}(XY)}{f_i(X) f_j(Y)} \tag{3}$$

Helix plot scores For each sequence in a multiple alignment, a score matrix is formed by assigning good-pair scores to cells that represent Waston–Crick or G–U base-pairs, bad-pair scores to other base-pairs and penalty scores to gaps. The matrix is then scanned and base-pairs that can form sufficiently long helices are given bonus scores. Individual score matrices for sequences in the alignment are finally summed together to yield a single-score matrix. Default parameters of the helix plot program (Tabaska *et al*., 1998) are used (good-pair score = 1, bad-pair score = 2, paired gap penalty = 3 and helix bonus = 2 × helix length).

Mutual information and helix plot scores are then added to generate the final score matrix to be used by ILM. Different weights can be assigned optionally to individual matrices to give preferences. One may assign a higher weight to the helix plot score when the number of sequences is small or vice versa, since the mutual information score works the best with a large number of sequences. Let HP_{ij} be the helix plot score of a potential base-pair, N the number of sequences in the alignment, α and β the relative weights of mutual information and helix plot scores. The combined score B_{ij} is calculated as:

$$B_{ij} = \alpha \times 1000 \times M_{ij} + \beta \times 20 \times HP_{ij}/N \tag{4}$$

The coefficients 1000 and 20 are used to convert mutual information and helix plot scores to integers that cover approximately the same range of values. Default values of α and β are both equal to 1.

Extended helix plot scores For individual sequences, mutual information is not available, and helix plot score alone does not provide sufficient information. However, we can extend the ILM algorithm to utilize the standard RNA folding thermodynamic parameters. The principle will remain the same: iteratively predicting a non-pseudoknotted secondary structure, selecting the most reliable helix and removing it from the sequence. By taking both loop destabilizing energies and base-pair stacking energies into account, the algorithm should be able to produce reliable predictions for single sequences, although the actual implementation requires more effort than the current one. Fortunately, RNA-folding thermodynamics can be incorporated to extend helix plot. In the original helix plot score, the score for a potential base-pair consists of two parts: a good-pair score that is the same for Waston–Crick and G–U base-pairs, and a bonus score that is proportional to the length of the helix it belongs to. We extend this to allow

more elaborated energy rules. First, we make a good-pair score depend on the type of a base-pair. Second, we let a helix bonus score be proportional to the total stacking energy of a helix. The extended helix plot score *EXT_HP* is calculated as:

$$EXT\_HP_{ij} = GP_{ij} + BONUS_{ij} \quad (5)$$

where GP_{ij} is a good-pair score, and $BONUS_{ij}$ is the bonus score contributed by the helix that base-pair (i, j) belongs to. Default good-pair scores for G–C, A–U and G–U pairs are 80, 50 and 30, respectively. Bonus scores are calculated as:

$$BONUS_{ij} = 100 \times \frac{Total\ Stacking\ Energy}{\sqrt{Helix\ Length}} \quad (6)$$

Parameters of stacking energies are extracted from the Vienna RNA package 1.4 (http://www.tbi.univie.ac.at/~ivo/RNA/).

RESULTS

We now present some prediction results from our new algorithm. We compared our algorithm with the MWM algorithm (Tabaska *et al*., 1998) and the PKNOTS algorithm (Rivas and Eddy, 1999), which were implemented by their original authors. We chose these two algorithms since they are well-developed algorithms in their respective categories. MWM is the only published algorithm we found for predicting optimal pseudoknotted structures using comparative analysis. PKNOTS is the only dynamic programming algorithm that fully exploits the standard RNA secondary structure thermodynamic models, and has high-pseudoknot-prediction accuracy on short sequences.

We carried out two sets of experiments separately. First, we compared our algorithm and the MWM algorithm on a set of aligned homologous sequences, using combined helix plot and mutual information scores. We then tested all three algorithms, MWM, PKNOTS and ILM, on a set of individual sequences, using the extended helix plot scores. In all cases, all programs were run with default parameters unless otherwise specified (for ILM, minimum loop length = minimum virtual loop length = 3; minimum helix length = 2; number of helices selected per iteration = 1; number of iterations before termination = unlimited).

Five sets of aligned sequences were used, including 16S rRNA, 5S rRNA, srpRNA, tmRNA and telomerase RNA. Individual sequences were taken from HIV-1-RT virus, TYMV RNA, TMV RNA, HDV ribozyme RNA, and anti-genomic HDV ribozyme RNA. Except 5S rRNA, all sequences are known to contain at least one pseudoknot. Table 1 lists some information about the test sequences and their structures. Sequences and their structures were retrieved from academic literatures or publicly accessible databases listed in the Table 1 caption.

Prediction accuracy is measured by both sensitivity and specificity. Let *EP* be the number of base-pairs in a published 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 reference structure (false prediction). Following Baldi *et al*. (2000), sensitivity is defined as TP/EP, and specificity is defined as $TP/(TP + FP)$.

Table 1. Sequences used in the experiments

RNA	NSEQ	Reference structure				
		Organism	*L* (nt)	EP	EHLX	EK
5S rRNA	12	*Escherichia coli*	120	40	5	0
SRP RNA	12	*Bacillus subtilis*	271	78	14	1
Telomerase RNA	9	*Homo sapiens*	210	50	5	1
tmRNA	8	*Escherichia coli*	362	106	12	4
16S rRNA	10	*Escherichia coli*	1542	478	67	2
HIV-1-RT	1	—	35	11	2	1
TYMV	1	—	86	24	5	1
TMV-3′-up	1	—	84	25	6	3
TMV-3′-down	1	—	105	34	6	2
HDV	1	—	87	28	4	1
Anti-HDV	1	—	91	24	4	1

NSEQ: number of sequences used; L: sequence length; EP: expected number of base-pairs (in a published structure for this molecule); EHLX: expected number of helices; EK: expected number of pseudoknots. Only helices with length >2 are counted. Sequence alignment and structure were obtained from the following sources: 5S rRNA and 16S rRNA, Cannone *et al*. (2002), SRP RNA, Gorodkin *et al*. (2001), Telomerase RNA, Chen *et al*. (2000), tmRNA, Knudsen *et al*. (2001). HIV-1-RT, Tuerk *et al*. (1992), TYMV, Rietveld *et al*. (1982), TMV, van Belkum *et al*. (1985), HDV and anti-genomic HDV, Ferre-D'Amare *et al*. (1998).

Prediction accuracy using aligned sequences

In the first set of experiments, where we compared MWM and ILM, we generated a score matrix from each sequence alignment (5S rRNA, SRP RNA, tmRNA, Telomerase RNA and 16S rRNA) using a combination of the mutual information (MI) and helix plot (HP) scores. Default parameters are used to compute HP scores. The sequences in each family used for alignment are listed online as supplementary materials (http://www.cs.wustl.edu/~zhang/projects/rna/ilm/). MI and HP scores are weighted with a ratio of 1 : 3 for alignments with less than 10 sequences and 1 : 1 in all other cases. Different ratios were chosen simply because MI, being a statistical measure, tends to be less reliable for a small number of sequences. We then run the ILM and the MWM algorithms respectively using the score matrix to produce a consensus structure, which was aligned back to the reference sequence to remove gaps. The predicted structure was compared with the reference structure to measure prediction quality. The results are listed in Table 2.

With 8–12 homologous sequences, our method correctly identified more than 90% of the base-pairs for short sequences (<300 nt), and 80% on average (computed as the number of correctly predicted base-pairs for all sequences divided by the total number of base-pairs in reference structures). In contrast, MWM identified 60–85% bp for short sequences and

Table 2. Summary of prediction results on aligned RNA sequences

RNA	MWM			ILM		
	$TP(SS)$	SP	K	$TP(SS)$	SP	K
5S rRNA	32 (80.0)	58.2	0/0	38 (95.0)	95.0	0/0
SRP RNA	68 (87.2)	59.6	1/1	76 (97.4)	75.2	1/1
Telomerase RNA	29 (58.0)	24.0	1/1	45 (90.0)	60.0	1/1
tmRNA	73 (68.9)	42.7	3/4	93 (87.7)	73.8	4/4
16S rRNA	243 (50.8)	35.5	0/2	351 (73.4)	68.2	1/2

TP = number of correctly predicted base-pairs; $S = 100 \times TP/EP$; $SP = 100 \times TP/(EP+FP)$; K = (number of correctly predicted pseudoknots)/(expected number of pseudoknots); EP = expected number of base-pairs; FP = number of predicted base-pairs that do not exist in the reference structure.

59.2% on average. ILM correctly predicted all pseudoknots for aligned sequences except 16S rRNA, for which a long-range pseudoknot of length 3 bp was missed, while MWM missed a pseudoknot in tmRNA and both pseudoknots in 16S rRNA. The most striking result is perhaps on tmRNA, which contains a total of four pseudoknots. With as few as eight sequences, ILM successfully identified all four pseudoknots and 11 of its 12 helices. ILM is also more specific in predicting only true positive base-pairs and outperforms MWM by a factor of 2 in terms of prediction specificity. The base-pairs predicted by MWM are often discontinuous and thus it is up to the user's discernment to determine whether some scattered base-pairs are indeed a part of a helix. When sequences are relatively long, such as 16S rRNA, our method showed a drastic improvement over MWM. The result on 5S rRNA shows that our algorithm is also superior to the MWM algorithm when no pseudoknot exists in the real structure, where our method produced very few spurious base-pairs, whereas almost half of the base-pairs predicted by the MWM algorithm do not exist in the reference structure.

Prediction accuracy using individual sequences

The second set of experiments was carried out on a set of individual sequences to compare MWM, PKNOTS and ILM. The results are listed in Table 3. The score matrices used by ILM and MWM were calculated using the extended helix plot with default parameters. The prediction results of PKNOTS were obtained from Rivas (personal communication). ILM and PKNOTS exhibit similar prediction accuracies and are both better than MWM. ILM correctly identified all base-pairs except for TMV-3′-end, missed a pseudoknot each in upstream and downstream sequences. PKNOTS missed all three pseudoknots for TMV-3′-end upstream and a short helix for HDV, but was otherwise almost perfect. ILM shows slightly better sensitivity than PKNOTS, while the latter has better specificity. In addition, MWM has the worst sensitivity and specificity among all three methods. We should note that the score matrix was probably biased against MWM. When we varied the parameters, we found that the default parameters used for score matrix generation were not (but close to) optimal for MWM. However, we were unable to tune MWM's parameters to make it better than ILM.

CPU time and memory usage

Table 4 lists the CPU time and memory usage for each algorithm. All experiments were conducted on a machine with an AMD 1600 MHz processor and 2 GB RAM. Running time for the MWM and ILM programs includes time for the preparation of score matrices with extended helix plot. Unlike the PKNOTS which takes 102 h of CPU time and 1.2 GB of memory to fold a 210 nt sequence, ILM and MWM require moderate CPU time and memory. ILM and MWM take less than 10 and 5 MB of memory and less than 5 and 1 min, respectively, to fold a 1542 nt sequence. Although the worst-case time complexity for the ILM algorithm is $O(n^4)$, in practice we observed its average case time complexity close to $O(n^3)$.

DISCUSSION

In this paper, we presented an algorithm for RNA secondary structure prediction with pseudoknots, based on the combination of thermodynamic and comparative approaches. Prior to this work, automated prediction of RNA secondary structure with pseudoknots has not been very successful in practical use. Thermodynamic approaches based on minimum free energy are theoretically important for finding optimal structures. However, they usually have very high time and memory complexity, making them impractical even for sequences of a few hundred bases long. Yet due to the lack of proper models and energy parameters, their results are often not satisfactory even for short sequences. Comparative approaches are more reliable on detecting pseudoknot structures, but are typically done in an *ad hoc* manner. The only published algorithm that we are aware of, the MWM algorithm, is able to produce meaningful predictions only if the number of homologous sequences is large so that covariance signals are sufficiently strong. This algorithm is vulnerable to noisy data such as misalignment, since it allows many types of unrealistic interactions to happen and does not take into consideration that helices are the most frequent structural elements of RNA structures.

By combining the advantages of both thermodynamic and comparative approaches, our method is able to predict RNA secondary structures efficiently and reliably including pseudoknots, using only a few sequences. Although our method does not compute a theoretically optimal structure, it sacrifices some optimality in exchange for forming stable helices. It turns out that this compromise significantly improves the overall prediction accuracy, especially in the cases where data is relatively insufficient for methods such as MWM to produce reliable predictions using unrestricted models.

Table 3. Summary of prediction results on individual RNA sequences

RNA	MWM			PKNOTS			ILM		
	$TP(SS)$	SP	K	$TP(SS)$	SP	K	$TP(SS)$	SP	K
HIV-1-RT	11 (100)	84.6	1/1	11 (100)	100	1/1	11 (100)	100	1/1
TYMV	24 (100)	63.2	1/1	24 (100)	96.0	1/1	24 (100)	82.8	1/1
TMV-3′-up	17 (68.0)	41.5	1/3	13 (52.0)	59.1	0/3	20 (80.0)	80.0	2/3
TMV-3′-down	25 (73.5)	49.0	0/2	33 (97.0)	97.0	2/2	26 (76.5)	68.4	1/2
HDV	19 (67.8)	45.2	0/1	24 (85.7)	75.0	1/1	28 (100)	82.4	1/1
Anti-HDV	17 (70.8)	38.6	1/1	23 (95.8)	69.7	1/1	24 (100)	66.7	1/1

TP, SS, SP and K are defined in Table 2.

Table 4. Comparison of CPU time and memory usage for each algorithm

Sequence length (nt)	MWM CPU time (S)	Memory	PKNOTS CPU time	Memory	ILM CPU time (S)	Memory
86	0.02	448 KB	16.4 min	40 MB	0.02	468 KB
210	0.13	532 KB	102 h	1.2 GB	0.14	620 KB
1542	40	5.0 MB	—	—	306	9.8 MB

Running time of the MWM and the ILM algorithm include the preparation of the score matrix using extended helix plot. Memory usage includes both data and code.

The Monte-Carlo simulation method proposed by Abrahams *et al*. (1990) shares some similarity with ours. Their method first compiles a list of all possible candidate helices, and then predicts a structure by iteratively selecting the highest-scored helix that does not overlap with previous selected ones. Their method is implemented using energy rules for a single sequence. However, there are some difficulties when applying their method to arbitrary score matrices. It is possible for a score matrix to have positive values in all cells, for example when mutual information is used. It is thus difficult to decide the boundary of each helix. In our method, helix boundaries are determined automatically by the LM procedure. Moreover, although both methods do not guarantee optimality, our method finds a solution whose score is at least no worse than that obtained by the basic LM where pseudoknots are forbidden.

Our algorithm can also be applied to individual sequences where no covariance information is available. Using the extended helix plot score, our algorithm has similar prediction accuracy as PKNOTS, and we believe that a more sophisticated implementation using the standard energy model would improve our prediction accuracy significantly. Considering the simplicity of the scoring scheme we used, we would not conclude that our algorithm is able to predict pseudoknotted structures reliably using thermodynamic information alone. What we can conclude is that PKNOTS or similar algorithms, being much more complex and resource demanding than our algorithm, do not necessarily produce more accurate predictions. Despite their theoretical importance for finding optimal thermodynamic structures, such energy-based algorithms are intrinsically limited by the approximations of energy models and the uncertainty in energy parameters.

In short, due to the high-prediction accuracy and low requirement on computational resources, we believe that the new algorithm can be used as a desktop tool for the prediction of RNA secondary structures with pseudoknots.

ACKNOWLEDGEMENTS

We thank Elena Rivas and Sean Eddy for providing the PKNOTS program and results. We also thank the anonymous reviewers for their very useful comments. This research was supported in part by NSF grants IIS-0196057 and ITR/EIA-0113618. G.D.S. was supported by NIH grant HG00249.

REFERENCES

Abrahams,J., van den Berg,M., van Batenburg,E. and Pleij,C. (1990) Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Res.*, **18**, 3035–3344.

Akmaev,V., Kelley,S. and Stormo,G. (1999) A phylogenetic approach to RNA structure prediction. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **7**, 10–17. AAAI Press.

Akutsu,T. (2000) Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl.Math.*, **104**, 45–62.

Baldi,P., Brunak,S., Chauvin,Y., Andersen,C. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

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.

Cannone,J., Subramanian,S., Schnare,M., Collett,J., D'Souza,L., Du,Y., Feng,B., Lin,N., Madabusi,L., Muller,K. *et al*. (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other rnas. *BMC Bioinformatics*, **3**, 2.

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

Chen,J., Blasco,M. and Greider,C. (2000) Secondary structure of vertebrate telomerase RNA. *Cell*, **100**, 503–514.

Chiu,D. and Kolodziejczak,T. (1991) Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci.*, **7**, 347–352.

Dam,E., Pleij,K. and Draper,D. (1992) Structural and functional aspects of RNA pseudoknots. *Biochemistry*, **31**, 11665–11176.

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

Ferre-D'Amare,A., Zhou,K. and Doudna,J. (1998) Crystal structure of a hepatitis delta virus ribozyme. *Nature*, **395**, 567–574.

Freier,S., Kierzek,R., Jaeger,J., Sugimoto,N., Caruthers,M., Neilson,T. and Turner,D. (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. USA*, **83**, 9373–9377.

Garey,M. and Johnson,D. (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, San Francisco.

Gorodkin,J., Heyer,L. and Stormo,G. (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.*, **25**, 3724–3732.

Gorodkin,J., Knudsen,B., Zwieb,C. and Samuelsson,T. (2001) SRPDB (signal recognition particle database). *Nucleic Acids Res.*, **29**, 169–170.

Gulko,B. and Haussler,D. (1996) Using multiple alignments and phylogenetic trees to detect RNA secondary structure. *Proc. Pac. Symp. Biocomput.*, **1**, 350–367.

Gultyaev,A., van Batenburg,F.H. and Pleij,C. (1995) The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.*, **250**, 37–51.

Gutell,R., Power,A., Hertz,G., Putz,E. and Stormo,G. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, **20**, 5785–5795.

Hofacker,I., Fekete,M. and Stadler,P. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.

Hofacker,I., Fontana,W., Stadler,P., Bonhoeffer,L., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.

Juan,V. and Wilson,C. (1999) RNA secondary structure prediction based on free energy and phylogenetic analysis. *J. Mol. Biol.*, **289**, 935–947.

Knudsen,B., Wower,J., Zwieb,C. and Gorodkin,J. (2001) tmRDB (tmRNA database). *Nucleic Acids Res.*, **29**, 171–172.

Luck,R., Graf,S. and Steger,G. (1999) ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res.*, **27**, 4208–4217.

Lyngso,R. and Pedersen,C. (2000a) Pseudoknots in RNA secondary structures. *Proceedings of the fourth annual international Conference on Computational Molecular Biology*, pp. 201–209. ACM Press.

Lyngso,R. and Pedersen,C. (2000b) RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.*, **7**, 409–427.

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

Mathews,D. and Turner,D. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.

Nussinov,R., Pieczenik,G., Griggs,J. and Kleitman,D. (1978) Algorithms for loop matchings. *SIAM J. Appl. Math.*, **35**, 68–82.

Rietveld,K., Poelgeest,R.V., Pleij,C., Boom,J.V. and Bosch,L. (1982) The tRNA-like structure at the 3′ terminus of turnip yellow mosaic virus RNA: differences and similarities with canonical tRNA. *Nucleic Acids Res.*, **10**, 1929–1946.

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

Sakakibara,Y., Brown,M., Hughey,R., Mian,I., Sjolander,K., Underwood,R., and Haussler,D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.*, **22**, 5112–5120.

Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.

Tabaska,J., Cary,R., Gabow,H. and Stormo,G. (1998) An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, **14**, 691–699.

Tuerk,C., MacDougal,S. and Gold,L. (1992) RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc. Natl Acad. Sci. USA*, **89**, 6988–6992.

Uemura,Y., Hasegawa,A., Kobayashi,S. and Yokomori,T. (1999) Tree adjoining grammars for RNA structure prediction. *Theor. Comp. Sci.*, **210**, 277–303.

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

van Belkum,A., Abrahams,J., Pleij,C. and Bosch,L. (1985) Five pseudoknots are present at the 204 nucleotides long 3′ noncoding region of tobacco mosaic virus RNA. *Nucleic Acids Res.*, **13**, 7673–7686.

Wuyts,J., Rijk,P.D., de Peer,Y.V., Pison,G., Rousseeuw,P. and Wachter,R.D. (2000) Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA. *Nucleic Acids Res.*, **28**, 4698–4708.

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

Zwieb,C., Wower,I. and Wower,J. (1999) Comparative sequence analysis of tmRNA. *Nucleic Acids Res.*, **27**, 2063–2071.