Report Number: WUCSE-2003-43

2003-03-16

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

Jianhua Ruan and Weixiong Zhang

Motivation: Pseudoknots have generally been excluded from the prediction of RNA secondary structures due to the difficulty in modeling and complexity in computing. 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 (Tabaska et. al, Bioinformatics, 14:691-9, 1998), an algorithm for pseudoknot prediction with comparative analysis, suffers from low prediction accuracy in many cases. Here we present an algorithm, iterative loop matching, for... **Read complete abstract on page 2.**

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

Jianhua Ruan and Weixiong Zhang

Complete Abstract:

Motivation: Pseudoknots have generally been excluded from the prediction of RNA secondary structures due to the difficulty in modeling and complexity in computing. 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 (Tabaska et. al, Bioinformatics, 14:691-9, 1998), an algorithm for pseudoknot prediction with comparative analysis, suffers from low prediction accuracy in many cases. Here we present an algorithm, iterative loop matching, for predict-ing RNA secondary structures including pseudoknots reliably and efficiently. The method can utilize either thermodynamic or comparative information or both, thus is able to predict for both aligned sequences and individual sequences. Results: We have tested the algorithm on a number of RNA families, including both structures with and without pseudoknots. 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. Furthermore, it 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 to the PKNOTS algorithm (Rivas & Eddy, J Mol Biol, 285:2053-68, 1999), 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/.

Department of Computer Science & Engineering

Washington University in St. Louis
School of Engineering & Applied Science

2003-43

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

Authors: Ruan, Jianhua; Zhang, Weixiong

May 20, 2003

Abstract: Pseudoknots have generally been excluded from the prediction of RNA secondary structures due to the difficulty in modeling and complexity in computing. 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 (Tabaska et. al, Bioinformatics, 14:691-9, 1998), an algorithm for pseudoknot prediction with comparative analysis, suffers from low prediction accuracy in many cases. Here we present an algorithm, iterative loop matching, for predicting RNA secondary structures including pseudoknots reliably and efficiently. The method can utilize either thermodynamic or comparative information or both, thus is able to predict for both aligned sequences and individual sequences. We have tested the algorithm on a number of RNA families, including both structures with and without pseudoknots. 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. Furthermore, it 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 to the PKNOTS algorithm (Rivas & Eddy, J Mol Biol, 285:2053-68, 1999), while using much less computational resources.

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

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

March 16, 2003

Abstract

Motivation: Pseudoknots have generally been excluded from the prediction of RNA secondary structures due to the difficulty in modeling and complexity in computing. 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 (Tabaska et. al, Bioinformatics, 14:691-9, 1998), an algorithm for pseudoknot prediction with comparative analysis, suffers from low prediction accuracy in many cases. Here we present an algorithm, iterative loop matching, for predicting RNA secondary structures including pseudoknots reliably and efficiently. The method can utilize either thermodynamic or comparative information or both, thus is able to predict for both aligned sequences and individual sequences.

Results: We have tested the algorithm on a number of RNA families, including both structures with and without pseudoknots. 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. Furthermore, it 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 to the PKNOTS algorithm (Rivas & Eddy, J Mol Biol, 285:2053-68, 1999), 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, zhang}@cse.wustl.edu

* To whom correspondence should be addressed

1 Introduction

In addition to carrying genetic information from DNA to protein, RNA molecules play many important regulatory, catalytic and structural roles in the cell. A complete understanding of the functions of RNA molecules requires knowledge of their three-dimensional structures. Since it is often difficult to crystallize or obtain nuclear magnetic resonance (NMR) 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 without pseudoknots is a list of base-pairs that are compatible with each other. Base-pair (i, j) and (k, l) are said to be *compatible* if they are either juxtaposed (e.g., $i < j < k < l$, Figure 1A) or nested (e.g., $i < k < l < j$, Figure 1B). Otherwise they are called *incompatible* (e.g, $i < k < j < l$, Figure 1C). Such an incompatible structure is known as a *pseudoknot*. More complex pseudoknots may occur if three or more base-pairs cross each other (Figure 1D). Whenever two or more *nested* base-pairs are connected without any interruption, they are said to form a *helix*. Also, consecutively unpaired bases are called a *loop*. Most computational methods for the prediction of RNA secondary structures can be classified into three families: thermodynamic, comparative and hybrid approaches.

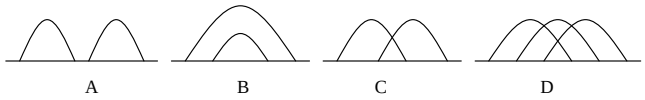

Figure 1: Diagrammatic representation of different types of relationships between base-pairs. An arch represents a base-pair between the two end-points. A. Two base-pairs are juxtaposed. B. Two base-pairs are nested. C. Two base-pairs cross each other, forming a pseudoknot. D. Three base-pairs cross each another, forming three pseudoknots.

Thermodynamic approaches (Zuker & Stiegler, 1981; Ho-

facker *et al.*, 1994) use dynamic programming to compute the optimal secondary structure with globally minimal free energy for a single RNA sequence based on a set of experimentally determined energy parameters (Freier *et al.*, 1986; Mathews *et al.*, 1999). Such methods have been successful in predicting the secondary structures of relatively short RNA sequences. However, several factors limit their accuracy. First, thermodynamic approaches assume RNA structures are in thermodynamic equilibria, independent of environmental conditions. However, many RNAs *in vivo* may be bound to protein factors that may alter their free energies. RNAs may also be kinetically trapped in non-equilibrium states. Second, energy models and parameters used by these approaches are only approximations and may not be able to capture all the details. Therefore, any method designed to find optimal structures using such energy models is also an approximation. Even with suboptimal structures taken into account (Zuker, 1989; McCaskill, 1990), the ability of these methods is still limited.

When a number of homologous sequences are available, comparative approaches are more reliable for determining the secondary structure 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 Chi-square statistics (Chiu & Kolodziejczak, 1991) and mutual information (Gutell *et al.*, 1992). Gulko & 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. However, these methods also have drawbacks. First, they typically require a very large collection of aligned homologous sequences, which may not be always available. Second, if some RNAs are conserved in both sequences, there will be very weak or no covariation, causing the failure of comparative approaches. Third, these methods assume that sequence alignments were performed according to structure conservation rather than sequence conservation, while structure alignment can hardly be done correctly without knowing the structure.

The third family of methods, which have emerged recently, combines the advantages of the first two (e.g. Luck *et al.*, 1999; Juan & Wilson, 1999; Hofacker *et al.*, 2002). These methods take both thermodynamic stability and sequence covariance into consideration and are able to produce positive results on as few as three 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 & Turner, 2002). They were only successful on short sequences due to their high time and space complexity. Eddy & Durbin (1994) and Sakakibara *et al.* (1994) introduced another family of methods, using stochastic context free grammars, to iteratively align homologous sequences and find a consensus structure for them.

An even 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). The rising number of known pseudoknots has triggered the development of a specific pseudoknot database (van Batenburg *et al.*, 2001). 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 & Pedersen (2000*b*), it is NP-complete (Garey & 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 dynamic programming algorithms have been developed recently, which run in polynomial time and space (Rivas & Eddy, 1999; Uemura *et al.*, 1999; Lyngso & Pedersen, 2000*a*; Akutsu, 2000). However, these methods still have very high time and space complexity, 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 a few hundred bases long sequences. Another dilemma for pseudoknot prediction algorithms based on energy models is that there is little experimentally determined thermodynamic data for pseudoknots.

The comparative approaches mentioned above, such as those based on Chi-Square and mutual information, can also be applied to the prediction of pseudoknots and are more practical and reliable than thermodynamic approaches. For example, comparative analysis has revealed the existence of pseudoknots in prion protein mRNA (Barrette *et al.*, 2001), eukaryotic small subunit ribosomal RNA (Wuyts *et al.*, 2000), tmRNA (Zwieb *et al.*, 1999) and vertebrate telomerase RNA (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 algorithm (MWM) (Cary & 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, 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.

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. Unlike many other algorithms, it does not attempt to compute theoretically optimal structures, but rather gives a practical and reliable approximate solution for this hard problem. We test the algorithm on a number of RNA families with sequence lengths ranging from 35nt to 1542nt, including both structures with and without pseudoknots. The results show that our algorithm correctly identifies more than 90% of base-pairs for short sequences (< 300nt) and approximately 80% on average for all sequences tested. Furthermore, the algorithm correctly predicts all pseudoknots except a 3bp pseudoknot in the longest sequence. It produces a very small number of false positive base-pairs on sequences without pseudoknot. The comparison with the maximum weighted matching algorithm (MWM) shows that our algorithm is both more specific and sensitive. In addition, we also apply the algorithm to individual sequences without using covariance information and compare its accuracy with an algorithm based on free energy minimization, the PKNOTS algorithm (Rivas & Eddy, 1999). Our algorithm exhibits an accuracy comparable to that of the PKNOTS algorithm, while having much lower time and space complexity.

2 Algorithms

Our algorithm is based on the loop matching algorithm (Nussinov *et al.*, 1978), which we will briefly describe first. We then introduce a new algorithm, called the iterative loop matching algorithm, to compute a secondary structure including pseudoknots. We will also discuss the score matrix used in our experiments and how the reliability of score matrix affects the prediction accuracy.

2.1 Loop Matching (LM)

Given a matrix B, where $B(i,j)$ is the score for the ith residue forming a base-pair with the jth residue, the loop matching algorithm finds a best-score secondary structure *without pseudoknots*. To reiterate, a secondary structure without pseudoknots is a "nested" structure as shown in Figure 1A and 1B, i.e., for any base-pair (i,j), if another base-pair (k,l) has one end (k) falling between i and j, then the other end (l) must also be in this range. Thanks to this constraint, the secondary structure of a long RNA sequence can be subdivided into shorter pieces. This observation is the core of all dynamic programming algorithms for RNA secondary structure prediction. The algorithm starts from calculating optimal structures of short subsequences, and then uses the structures of these short subsequences to construct optimal structures of longer subsequences, until the whole sequence is included. 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; (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.

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

Let $l = (j-i+1)$ be the length of subsequence $S[i,j]$. Initially $Z(i,i) = Z(i,i+1) = \cdots = Z(i,i+LOOP\_LENGTH+1) = 0$ for all i, where $LOOP\_LENGTH$ is a parameter that describes the minimum distance required between two paired bases (by default $LOOP\_LENGTH = 3$ in our program). The algorithm iteratively computes the values of $Z(i,i+l-1)$ for all i with increasing l values. At the end of the algorithm, $Z(1,N)$ is the score of the optimal structure for sequence $S[1..N]$. The optimal structure can be obtained by tracing back the Z matrix. The computation and trace-back can be done in $O(n^3)$ time and $O(n^2)$ space (Nussinov *et al.*, 1978).

In the simplest case, $B(i,j) = 1$ if the ith residue and the jth 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.

2.2 Iterative Loop Matching (ILM)

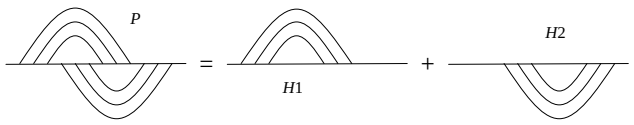

Figure 2: A pseudoknot (P) can be treated as two separate helices ($H1$ and $H2$) and can be identified by a two-iteration loop matching. Assume $H1$ is identified by the basic loop matching, then running the loop matching algorithm on the remaining single-stranded bases identifies the second helix,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

We now extend the basic loop matching algorithm to support pseudoknots. The loop matching algorithm cannot handle pseudoknots because it only allows compatible base-pairs. Notice that a pseudoknot can be thought of interactions between two loop regions of a secondary structure, as illustrated in Figure 2. Therefore we could run the loop matching algorithm twice. First we run the loop matching algorithm to predict a secondary structure as usual. We then apply the loop matching algorithm on the remaining single-stranded regions

by hypothetically removing the paired bases. By doing this, we may be able to identify base-pairs that belong to pseudoknots at this stage. 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 loop matching, which invalidates the efforts of further searching, as shown and explained in Figure 3. To circumvent this problem, we use a least-commitment strategy. We run the loop matching algorithm multiple times, each time we only accept the base-pairs that appear to be the most reliable, e.g. with the highest score. After each run of loop matching, the paired bases were considered as if they were removed from the original sequence, allowing the next iteration of loop matching to identify otherwise pseudoknotted base-pairs. This modification attempts to avoid possible false predictions from being included. Figure 3 illustrates the idea. Suppose that a structure consists of two helices $H1$ and $H2$, forming a pseudoknot, and $H3$ is a false helix that overlaps $H2$. Since loop matching can only predict compatible base-pairs, it will pick either $H1$ and $H3$ together or $H2$ alone. Let $R(H)$ be the score of helix H. Assume that $R(H1) + R(H3) > R(H2)$. Also assume that $R(H3) < R(H1)$ and $R(H3) < R(H2)$. Loop matching would predict $H1$ correctly but would also include $H3$, preventing $H2$ from being recognized, even though $H3$ has a lower score than $H2$. On the contrary, iterative loop matching would only accept $H1$ in the first iteration, allowing $H2$ to be successfully identified in the next iteration.

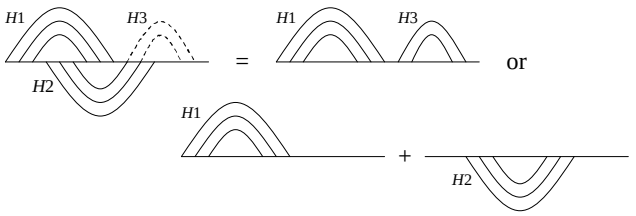

Figure 3: Pseudoknots that can be correctly identified by the iterative loop matching algorithm. $H1$ and $H2$ are two true helices forming a pseudoknot. $H3$ is a false helix overlapping $H2$. Scores (R) of the helices satisfy $R(H1) + R(H3) > R(H2)$, $R(H3) < R(H1)$ and $R(H3) < R(H2)$. Iterative loop matching will correctly predict $H1$ in the first iteration and predict $H2$ in the second iteration. In contrast, basic loop matching would pick $H1$ and $H3$ together since it gives a higher total score than selecting $H2$ alone. Then even if we run loop matching again on the remaining single strand, $H2$ cannot be correctly identified, 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 ith base to pair with the jth base.
2. Run the basic loop matching algorithm using matrix B and trace back 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 scores of its constitutive base-pairs. Pick helix H that has the highest score, 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 step 2–5 until no remaining sequence.
7. Report base-pair list S and terminate.

The method to prepare a score matrix from a single sequence or a sequence alignment will be discussed in section 2.4. Notice that at step 5, updating score matrix B in most cases simply means to remove rows and columns corresponding to bases that have been paired, or, alternatively, set their values to zero. However, this is rather inefficient in terms of running time. A better solution is to use an array M to keep track of the indices of remaining single-stranded bases and run the basic loop matching algorithm to compute the scores only 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 only to re-compute $Z(i, j)$ if i and j are separated by either p or q, i.e., $i < p < j$ or $i < q < j$. This is illustrated by Figure 4.

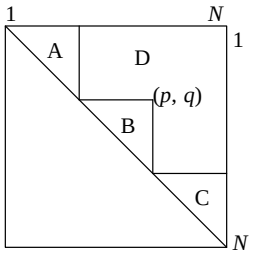

Figure 4: Three triangle areas of the matrix do not need to be re-computed in each iteration. Suppose that a base-pair (p, q), with $q < p$, is selected and deleted in one iteration. The optimal score of a subsequence $S[i, j]$, with $1 \leq i < j < q$, does not depend on bases whose indices are greater than q, so it will not change in the next iteration. Therefore all entries in area A of the score matrix need not to be re-computed. Similarly, entries in triangle areas B and C need not to be re-computed as well.

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. Define virtual distance of two bases to be the distance between them after paired bases

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

in the middle of them have been removed, i.e., 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 distances less than $VLOOP\_LENGTH$ are not allowed to form a base-pair, thus $Z(M[i],M[j]) = 0$ if $j-i+1 < VLOOP\_LENGTH$. The default value of $VLOOP\_LENGTH$ is set to be the same as $LOOP\_LENGTH$, which is equal to 3 in our program.

The recursion for re-computing Z is shown as Equation 2, where $M[i]$ is the ith remaining unpaired base, and p and q, with $p < q$, are two end-points of the helix selected in the previous iteration. Note that if B needs to be completely re-computed at step 5, Z also has to be completely re-computed and thus the above recursion cannot reuse an entry even if it represents a base-pair separated by p or q. 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 loop matching 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 $m \leq \frac{n}{2k}$, where k is the minimal helix length required, the worst case time complexity is $O(n^4)$. However, the average case complexity is close to $O(n^3)$ since 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 algorithm more efficient. The space complexity remains $O(n^2)$.

2.3 Suboptimality

One fundamental difference between ILM and many other pseudoknot prediction algorithms is that ILM does not attempt to find the theoretically optimal structure. Since the total score of a structure can be considered as a measure of its probability among a structure ensemble, 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 maximum weighted matching (MWM) algorithm (Cary & Stormo, 1995; Tabaska *et al.*, 1998) which allows all types of base-pairs. One of the most severe problems of MWM is that it allows a much larger degree of freedom than real 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 & Eddy, 1999; Uemura *et al.*, 1999; Lyngso & Pedersen, 2000*a*; Akutsu, 2000). However, none of these models has been generally agreed. 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 ensures to compute a structure whose score is at least no less than the score of that predicted by the basic loop matching algorithm. We now give a proof to validate 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 jth helix predicted in the ith iteration. Helices are ranked decreasingly by their scores. Note that the algorithm selects the helix with the highest score, i.e., h_1^i, for the ith 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 ith iteration. Assume that ILM will terminate after m iterations when no helix is identified. By definition,

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

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

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

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

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

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 from the proof help to extend the ILM algorithm while retaining the suboptimality property. First, h_1^i can be any helix predicted in the ith 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, as to be discussed in section 2.4. Second, if the algorithm is terminated early after i iterations ($i < m$), and all base-pairs predicted in the ith 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. In the extreme case if all base-pairs predicted in each iteration are accepted, we get the algorithm discussed at the beginning of section 2.2. These extensions are not used in the implementation of the algorithm by default.

2.4 Base-Pairing Score Matrix

The base-pairing score matrix can be obtained from a variety of sources of evidence or their combinations. Generally, for aligned sequences it can be obtained from the combination of energy scores and covariance scores, while for individual sequences only energy scores are available.

A number of score matrices have been previously constructed based on an alignment of multiple homologous sequences (Cary & Stormo, 1995; Luck *et al.*, 1999; Juan & 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 average 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 & Stormo (1995) and Tabaska *et al.* (1998) for more details.

Mutual information score. 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:

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

Helix plot scores. For each sequence in a multiple alignment, a score matrix is formed by assigning a score to each cell of the matrix based on whether the two bases corresponding to the cell can form a Waston-Crick or G-U base-pair. Other types of base-pairs or gaps receive penalty scores. The matrix is then scanned and base-pairs that may form long helices are given bonus scores. Individual score matrices for the sequences in the alignment are finally averaged to yield a single score matrix.

In practice, mutual information scores are usually multiplied by a constant factor to be converted to integers. Mutual information and helix plot scores are summed together to generate the final score matrix to be used by ILM. Different weights can be optionally assigned 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.

For a single sequence, a meaningful score matrix can be obtained by the partition function algorithm (McCaskill, 1990) which computes the base-pairing probability of each possible base-pair. Since pseudoknots are often represented as alternative foldings in this matrix, they may be identified by the ILM algorithm. A drawback of this matrix is that it may only be able to provide information for pseudoknots consisting of two (groups of) helices of similar lengths due to the following reason. The probability of a possible base-pair is inversely proportional to the exponential of the free energy of a structure containing this base-pair. If two incompatible helices have significantly different lengths, including the shorter one may result in significantly higher free energy than including the longer one, causing the probability of each base-pair of the shorter helix to be much lower than that of the longer one. This weak signal may be lost easily due to noises. As a remedy, we compute a new score matrix at the beginning of each iteration, i.e., base-pairing probabilities are re-computed in each iteration after accepted base-pairs are removed from the sequence. This slows down the overall computation by a constant factor since each calculation of the partition function takes $O(n^3)$ time.

It is worth discussing whether a higher score indeed indicates that the helix is more likely to be a correct one. This is generally true for scores derived from covariance analysis since they have statistical meanings. However, for scores derived from energy data, this measurement is inherently ill-defined due to the uncertainty of energy parameters and approximations made in the energy models of secondary structures. It has been reported that very small changes in the energy parameters sometimes cause drastic changes in the pre-

dicted structures (Zuker *et al.*, 1991). Nevertheless, several energy-based algorithms have successfully predicted pseudoknots for individual sequences. Considering the lack of proper energy parameters and models for pseudoknots, we suspect that the success of these algorithms are probably due to the fact that there are not many competing alternative helices in the test sequences and thus it is relatively easy to correctly identify the true ones. If this is the case, then our algorithm may also be able to predict them correctly, which is one of the main motivations for applying our algorithm to individual sequences.

3 Results

We now present some prediction results from our new algorithm. We compare our algorithm with the maximum weighted matching algorithm (MWM) (Tabaska *et al.*, 1998) and the PKNOTS algorithm (Rivas & Eddy, 1999), both of which were implemented by the original authors. We choose these two algorithms since both 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 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 partition function scores.

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 Table 1 caption.

Prediction accuracy is measured by both sensitivity and specificity. Let EP be the expected number of base-pairs in a 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 $TP/(TP+FP)$. There is a trade-off between these two types of measurement. In RNA structure prediction, we are generally more interested in sensitivity, i.e., how many true base-pairs are identified. However, too many false positive base-pairs are also not desirable as it would be very hard to determine the structure given many spurious base-pairs. To reflect this requirement, we measure the percentage of correctly predicted base-pairs (PCP), defined as $PCP = 100 \times TP/EP$, to indicate the prediction sensitivity. We also define the prediction accuracy $AC = 100 \times TP/(EP + FP)$, to combine both sensitivity and specificity.

3.1 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) score as described in section 2.4. 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 to the reference structure to measure prediction quality. The results are listed in Table 2.

With 8 to 12 homologous sequences, our method correctly identified more than 90% of the base-pairs for short sequences (< 300nt), and 80.0% 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% base-pairs for short sequences and 59.2% on average. ILM correctly predicted all pseudoknots for aligned sequences except 16S rRNA, for which a long-range pseudoknot of length 3bp 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 8 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 accuracy. 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.

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. 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).

Table 2: Summary of prediction results on aligned RNA sequences.

RNA	MWM			ILM		
	$TP(PCP)$	AC	K	$TP(PCP)$	AC	K
5S rRNA	32 (80.0)	50.8	0/0	38 (95.0)	90.5	0/0
SRP RNA	68 (87.2)	40.4	1/1	76 (97.4)	74.5	1/1
Telomerase RNA	29 (58.0)	24.0	1/1	45 (90.0)	60.0	1/1
tmRNA	73 (68.9)	35.8	3/4	93 (87.7)	69.9	4/4
16S rRNA	243 (50.8)	24.9	0/2	351 (73.4)	54.7	1/2

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

3.2 Prediction Accuracy Using Individual Sequences

The second set of experiments was carried on a set of single sequences to compare MWM, PKNOTS and ILM. The partition function algorithm implemented in the Vienna RNA package (Hofacker *et al.*, 1994) was used to prepare the score matrix to be fed to ILM and MWM. ILM can be run with or without the option of computing new score matrices in the algorithm as discussed in section 2.4. The prediction results for PKNOTS were obtained from Rivas (2003). The results are listed in Table 3. With the option of re-computing score matrices, ILM and PKNOTS exhibit similar prediction accuracy and are both better than MWM. ILM correctly identified all base-pairs except for TMV-3'-end, missed a pseudoknot for both upstream and downstream sequence. PKNOTS missed all three pseudoknots for TMV-3'-end upstream and a short helix for HDV, but was otherwise almost perfect. Without the option of re-computing score matrices, ILM showed similar sensitivity to MWM. Both were unable to identify pseudoknots in TMV-3'-end and anti-HDV, although ILM showed better prediction specificity (data for ILM not shown). A careful inspection of the partition function score matrices of TMV-3'-end and anti-HDV revealed that the signals for the existence of the pseudoknots were very weak in these cases, thus any algorithm based on them without the option of re-computing was doomed to fail.

3.3 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 1600MHz processor and 2 GB RAM. The running time for the MWM and ILM programs include the time for the preparation of score matrices from partition function. Unlike the PKNOTS which takes 102 hours of CPU time and 1.2GB of memory to fold a 210nt 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 10 and 2 minutes, respectively, to fold a 1542nt sequence. Although the worst-case time complexity for the ILM algorithm is $O(n^4)$,

Table 3: Summary of prediction results on individual RNA sequences.

RNA	MWM			PKNOTS			ILM		
	$TP(PCP)$	AC	K	$TP(PCP)$	AC	K	$TP(PCP)$	AC	K
HIV-1-RT	11 (100)	84.6	1/1	11 (100)	100	1/1	11 (100)	100	1/1
TYMV	24 (100)	85.7	1/1	24 (100)	92.3	1/1	24 (100)	92.3	1/1
TMV-3'-up	13 (52.0)	33.3	0/3	13 (52.0)	39.4	0/3	20 (80.0)	58.8	2/3
TMV-3'-down	24 (70.6)	47.0	0/2	33 (97.0)	97.0	2/2	26 (76.5)	60.5	1/2
HDV	26 (92.8)	70.3	1/1	24 (85.7)	66.7	1/1	28 (100)	82.4	1/1
Anti-HDV	17 (70.8)	39.5	0/1	23 (95.8)	67.6	1/1	24 (100)	72.7	1/1

TP, PCP, AC and K are defined in Table 2

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

Sequence	MWM		PKNOTS		ILM	
Length(nt)	CPU time	Memory	CPU time	Memory	CPU time	Memory
86	0.04 sec	448 KB	16.4 min	40 MB	0.05 sec	468 KB
105	0.07 sec	460 KB	65.3 min	86 MB	0.08 sec	484 KB
210	0.4 sec	532 KB	102 hour	1.2 GB	0.5 sec	620 KB
362	1.7 sec	532 KB	–	–	2.9 sec	972 KB
1542	1.9 min	5.0 MB	–	–	8.4 min	9.8 MB

Running time of the MWM and the ILM algorithm include the preparation of the score matrix using partition function algorithm implemented in the Vienna RNA package by Hofacker *et al.* (1994). Memory usage includes both data and code.

in practice we observed average case time complexity close to $O(n^3)$, slightly slower than the MWM algorithm.

4 Discussion

In this paper we presented a reliable and efficient 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 typically have very high time and memory complexity, making them impractical even for a few hundred bases long sequences. 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 maximum weighted matching 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 efficiently and reliably predict RNA secondary structures 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 cases where the evidence is relatively weak for methods such as MWM to produce reliable predictions using unrestricted models.

Our algorithm can also be applied to individual sequences where no covariance information is available. Our algorithm has slightly better prediction accuracy than PKNOTS on the tested sequences. The objective of the test is not to convince readers that our algorithm is able to reliably predict pseudoknotted structures 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.

Finally, 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.

Acknowledgment

This research was supported in part by NSF grants IIS-0196057 and ITR/EIA-0113618. Thanks to Gary Stormo for many inspiring discussions related to RNA folding and for the advice on combining thermodynamic and covariance scores. Thanks to Ivo L. Hofacker and colleagues for the Vienna RNA package and to Elena Rivas and Sean Eddy for the PKNOTS program and results.

References

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

Akutsu, T. (2000) Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, **104** (1-3), 45–62.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. & Nielsen, H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16** (5), 412–24.

Barrette, I., Poisson, G., Gendron, P. & Major, F. (2001) Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching. *Nucleic Acids Res*, **29** (3), 753–78.

Cannone, J., Subramanian, S., Schnare, M., Collett, J., D'Souza, L., Du, Y., Feng, B., Lin, N., Madabusi, L., Muller, K., Pande, N., Shang, Z., Yu, N. & Gutell, R. (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** (1), 2.

Cary, R. & 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. & Greider, C. (2000) Secondary structure of vertebrate telomerase RNA. *Cell*, **100** (5), 503–14.

Chiu, D. & Kolodziejczak, T. (1991) Inferring consensus structure from nucleic acid sequences. *Comput Appl Biosci*, **7** (3), 347–52.

Dam, E., Pleij, K. & Draper, D. (1992) Structural and functional aspects of RNA pseudoknots. *Biochemistry*, **31** (47), 11665–1176.

Eddy, S. & Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*, **22** (11), 2079–88.

Ferre-D'Amare, A., Zhou, K. & Doudna, J. (1998) Crystal structure of a hepatitis delta virus ribozyme. *Nature*, **395** (6702), 567–74.

Freier, S., Kierzek, R., Jaeger, J., Sugimoto, N., Caruthers, M., Neilson, T. & Turner, D. (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci U S A*, **83** (24), 9373–7.

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

Gorodkin, J., Heyer, L. & Stormo, G. (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res*, **25** (18), 3724–332.

Gorodkin, J., Knudsen, B., Zwieb, C. & Samuelsson, T. (2001) SRPDB (signal recognition particle database). *Nucleic Acids Res*, **29** (1), 169–70.

Gulko, B. & Haussler, D. (1996) Using multiple alignments and phylogenetic trees to detect RNA secondary structure. In *Proc Pac Symp Biocomput* pp. 350–67.

Gutell, R., Power, A., Hertz, G., Putz, E. & 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** (21), 5785–595.

Hofacker, I., Fekete, M. & Stadler, P. (2002) Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, **319** (5), 1059–166.

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

Juan, V. & Wilson, C. (1999) RNA secondary structure prediction based on free energy and phylogenetic analysis. *J Mol Biol*, **289** (4), 935–47.

Knudsen, B., Wower, J., Zwieb, C. & Gorodkin, J. (2001) tmRDB (tmRNA database). *Nucleic Acids Res*, **29** (1), 171–12.

Luck, R., Graf, S. & Steger, G. (1999) ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res*, **27** (21), 4208–417.

Lyngso, R. & Pedersen, C. (2000*a*) Pseudoknots in RNA secondary structures. In *Proceedings of the fourth annual international conference on Computational molecular biology* pp. 201–209 ACM Press.

Lyngso, R. & Pedersen, C. (2000*b*) RNA pseudoknot prediction in energy-based models. *J Comput Biol*, **7** (3-4), 409–27.

Mathews, D., Sabina, J., Zuker, M. & Turner, D. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, **288** (5), 911–40.

Mathews, D. & Turner, D. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol*, **317** (2), 191–203.

McCaskill, J. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29** (6-7), 1105–119.

Nussinov, R., Pieczenik, G., Griggs, J. & Kleitman, D. (1978) Algorithms for loop matchings. *SIAM J. Appl. Math.,* **35** (1), 68–82.

Rietveld, K., Poelgeest, R. V., Pleij, C., Boom, J. V. & 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** (6), 1929–146.

Rivas, E. (2003). Personal communication.

Rivas, E. & Eddy, S. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol,* **285** (5), 2053–2068.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I., Sjolander, K., Underwood, R. & Haussler, D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res,* **22** (23), 5112–520.

Sankoff, D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.,* **45** (5), 810–825.

Tabaska, J., Cary, R., Gabow, H. & Stormo, G. (1998) An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics,* **14** (8), 691–69.

Tuerk, C., MacDougal, S. & Gold, L. (1992) RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc Natl Acad Sci U S A,* **89** (15), 6988–692.

Uemura, Y., Hasegawa, A., Kobayashi, S. & Yokomori, T. (1999) Tree adjoining grammars for RNA structure prediction. *Theoretical Computer Science,* **210** (2), 277–303.

van Batenburg, F., Gultyaev, A. & Pleij, C. (2001) Pseudobase: structural information on RNA pseudoknots. *Nucleic Acids Res,* **29** (1), 194–15.

van Belkum, A., Abrahams, J., Pleij, C. & Bosch, L. (1985) Five pseudoknots are present at the 204 nucleotides long 3' noncoding region of tobacco mosaic virus RNA. *Nucleic Acids Res,* **13** (21), 7673–786.

Wuyts, J., Rijk, P. D., de Peer, Y. V., Pison, G., Rousseeuw, P. & 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** (23), 4698–708.

Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science,* **244** (4900), 48–52.

Zuker, M., Jaeger, J. & Turner, D. (1991) A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic Acids Res,* **19** (10), 2707–214.

Zuker, M. & Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res,* **9** (1), 133–48.

Zwieb, C., Wower, I. & Wower, J. (1999) Comparative sequence analysis of tmRNA. *Nucleic Acids Res,* **27** (10), 2063–271.