<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>An Efficient Alignment Algorithm for Searching Simple Pseudoknots over Long Genomic Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Ma, CCC; Wong, KF; Lam, TW; Hon, W; Sadakane, K; Yiu, SM</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, PP, Issue: 99, p. 1</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2012</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/165841">http://hdl.handle.net/10722/165841</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.</td>
</tr>
</tbody>
</table>
An Efficient Alignment Algorithm for Searching Simple Pseudoknots over Long Genomic Sequence

Christopher Ma, Thomas K.F. Wong, T.W. Lam, W.K. Hon, K. Sadakane, and S.M. Yiu

Abstract—Structural alignment has been shown to be an effective computational method to identify structural noncoding RNA (ncRNA) candidates as ncRNAs are known to be conserved in secondary structures. However, the complexity of the structural alignment algorithms becomes higher when the structure has pseudoknots. Even for the simplest type of pseudoknots (simple pseudoknots), the fastest algorithm runs in $O(mn^2)$ time, where $m$, $n$ are the length of the query ncRNA (with known structure) and the length of the target sequence (with unknown structure), respectively. In practice, we are usually given a long DNA sequence and we try to locate regions in the sequence for possible candidates of a particular ncRNA. Thus, we need to run the structural alignment algorithm on every possible region in the long sequence. For example, finding candidates for a known ncRNA of length 100 on a sequence of length 50,000, it takes more than one day. In this paper, we provide an efficient algorithm to solve the problem for simple pseudoknots and it is shown to be 10 times faster. The speedup stems from an effective pruning strategy consisting of the computation of a lower bound score for the optimal alignment and an estimation of the maximum score that a candidate can achieve to decide whether to prune the current candidate or not.

Index Terms—Noncoding RNAs, pseudoknot, structural alignment

1 INTRODUCTION

A NON-CODING RNA (ncRNA) is a functional RNA molecule which is not translated into a protein. There are many different types of ncRNAs such as tRNAs, rRNAs, snoRNAs, microRNAs, and siRNAs. These RNA molecules have been found to be involved in many biological processes. The number of ncRNAs within the genome was underestimated before, but recently some databases reveal over 1,973 ncRNA families [1]. Data accumulated on ncRNAs and these families show that ncRNAs may be as diverse as protein molecules.

Identifying ncRNAs is an important problem. It is known that the structure of an ncRNA molecule usually plays an important role in its biological functions. Some researchers attempted to identify ncRNAs by considering the stability of secondary structures formed by substrings of a given genome [2]. Although, on average, the functional RNAs have lower folding energy than the random sequences of the same length and dinucleotide frequency [3], this method is not quite effective because a random sequence with high GC composition also allows an energetically favorable secondary structure [4].

Another promising method is the comparative approach. The idea is to make use of a known ncRNA and try to identify ncRNA candidates along the genome whose sequence and structure are similar to that of the ncRNA. The resulting regions are the potential ncRNAs candidates of the same family. The key of this approach is to compute the structural alignment between the folded ncRNA (query) and the unfolded region (target). The unfolded sequence will be aligned to the folded ncRNA. The alignment score represents their sequence and structural similarity. RSEARCH [5], FASTR [6], and INFERNAL [7] for Rfam were developed based on this approach.

However, these tools [5], [6], [7] do not support pseudoknots. Given two base-pairs at positions $(i,j)$ and $(i',j')$, where $i < j$ and $i' < j'$, pseudoknots are base-pairs crossing each other, i.e., $i < i' < j < j'$ or $i' < i < j' < j$. In some studies, secondary structures including pseudoknots are found involved in some functions such as telomerase [8], catalytic functions [9], and self-splicing introns [10]. For example, the pseudoknot structure in the human telomerase RNA is conserved in all vertebrates and is essential for telomerase activity [11]. Structural alignment of ncRNAs containing pseudoknots is believed to be NP-complete and the problem is computationally hard.

Matsui et al. developed a method of computing the structural alignment to support some complicated pseudoknot structures which they refer as 2-crossing [12]. Their algorithm runs in $O(mn^5)$ with space complexity of $O(mn^4)$ where $m$ is the length of the query sequence and $n$ is the length of the target sequence. However, their method can only consider the structural similarity but not sequence similarity between the query and the target sequence.

Christopher Ma, Thomas K.F. Wong, T.W. Lam, W.K. Hon, K. Sadakane, and S.M. Yiu
In this paper, we consider the structural alignment problem as follows: given a query sequence with pseudo-knot structure and a target sequence with unknown structure, our aim is to output the alignment with maximum score between the whole query sequence and the whole target sequence. The score, which will be formally defined in Section 2.1, is computed according to both of their sequence and structural similarity. Higher score represents higher similarity between the two sequences according to their sequences and structures.

PAL, which was developed by Han et al. [13], can solve this problem in $O(mn^3)$ time with space complexity of $O(mn^3)$ for simple pseudoknot structure. For the definition of these pseudoknot structures, please refer to Section 2. However, their algorithm is still not feasible for long RNA sequences (say the length is over 50k) due to the extensive time required. For example, as shown in Table 1 for the pseudoknotted ncRNA family of sequence length around 100 to 150, when searching a long genome of length 50,000 to find regions of which the sequences and the structures are similar to that of the selected query of the family, it takes 1.3 to 3.3 days to finish. The figure shows that the tool seems impractical for searching ncRNAs along a long genome.

We proposed an efficient algorithm for solving the same structural alignment problem for simple pseudoknot. While maintaining the same sensitivity and specificity, the speed up achieved by our method ranges from 8.5 to 11 as compared to PAL when the query is longer than 65. For an ncRNA family with simple pseudoknot structure of length around 100 to 150, our method requires only a few hours for searching along the same long genome. Although the method presented in this paper is designed for simple pseudoknot, a similar technique can also be applied to other structural alignment algorithms that handle other pseudoknot types.

Please note that there exists another approach like [14] to compute the alignment between the query structure including pseudoknots and the target sequence. Their method requires $O(ktN^2)$ time where $N$ is the length of the target sequence, $t$ is the width of the tree according to the query structure and a parameter $k$ (for details, please refer to [14]). Their approach is a heuristic approach. The choice of $k$ would affect the accuracy of the result and their method do not consider the loop regions (i.e., the regions which do not have any base pair) of the query sequence when performing the alignment. If loop regions are large, the accuracy of the result would be decreased.

### 2 Preliminaries

Let $A = a_1a_2 \ldots a_n$ be an RNA sequence and $M$ be the secondary structure of $A$. $M$ is represented as a set of base pairs $(a_i, a_j), 1 \leq i < j \leq m$. Let $M_{x,y} \subseteq M$ be the set of base pairs in the subsequence $a_xa_{x+1} \ldots a_y, 1 \leq x < y \leq m$. $M_{x,y}$ is a regular structure if there does not exist two pairs $(i,j),(k,l) \in M_{x,y}$ such that $i < k < j < l$ or $k < i < l < j$. Note that an empty set is considered as a regular structure. $M_{x,y}$ is a simple pseudoknot if $\exists \, x < x_1, x_2 < y \, (x_1,x_2$ are referred as pivot points) such that (see Fig. 1a for an example):

1. each $(i,j) \in M_{x,y}$ satisfies either $x \leq i < x_1 \leq j < x_2$ or $x_1 \leq i < x_2 \leq j \leq y$; and
2. $M_L$ and $M_R$ are both regular where $M_L = \{(i,j) \in M_{x,y} | x \leq i < x_1 \leq j < x_2\}$ and $M_R = \{(i,j) \in M_{x,y} | x_1 \leq i < x_2 \leq j \leq y\}$.

In the literature, regular structure is also called nested structure.

### Table 1

Comparison on the Running Time between Our Method and Han's

<table>
<thead>
<tr>
<th>Family ID</th>
<th>Family name</th>
<th>Query length</th>
<th>Number of members embedded</th>
<th>Our running time (seconds)</th>
<th>Han’s running time (seconds)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF00165</td>
<td>Corona_pk3</td>
<td>62</td>
<td>30</td>
<td>2220</td>
<td>16213</td>
<td>7.3</td>
</tr>
<tr>
<td>RF00390</td>
<td>UPSK</td>
<td>23</td>
<td>25</td>
<td>1215</td>
<td>4613</td>
<td>3.8</td>
</tr>
<tr>
<td>RF00505</td>
<td>RydC</td>
<td>66</td>
<td>75</td>
<td>2058</td>
<td>17489</td>
<td>8.5</td>
</tr>
<tr>
<td>RF00507</td>
<td>Corona_FSE</td>
<td>85</td>
<td>23</td>
<td>4064</td>
<td>37350</td>
<td>9.2</td>
</tr>
<tr>
<td>RF01072</td>
<td>TMV_UPD-PK3</td>
<td>21</td>
<td>40</td>
<td>1124</td>
<td>4471</td>
<td>4.0</td>
</tr>
<tr>
<td>RF01074</td>
<td>RF_site1</td>
<td>40</td>
<td>12</td>
<td>1731</td>
<td>10025</td>
<td>5.8</td>
</tr>
<tr>
<td>RF01076</td>
<td>RF_site2</td>
<td>72</td>
<td>38</td>
<td>2670</td>
<td>23990</td>
<td>9.0</td>
</tr>
<tr>
<td>RF01080</td>
<td>UPD-Pkb</td>
<td>32</td>
<td>16</td>
<td>1543</td>
<td>7845</td>
<td>5.1</td>
</tr>
<tr>
<td>RF01081</td>
<td>SBWMV1_UPD-Pks</td>
<td>26</td>
<td>14</td>
<td>1341</td>
<td>7418</td>
<td>5.5</td>
</tr>
<tr>
<td>RF01087</td>
<td>PK-repZ</td>
<td>148</td>
<td>37</td>
<td>25725</td>
<td>284186</td>
<td>11.0</td>
</tr>
<tr>
<td>RF01089</td>
<td>PK-repBA</td>
<td>116</td>
<td>25</td>
<td>10481</td>
<td>114237</td>
<td>10.9</td>
</tr>
<tr>
<td>RF01093</td>
<td>RF_site5</td>
<td>60</td>
<td>38</td>
<td>2274</td>
<td>15899</td>
<td>7.0</td>
</tr>
<tr>
<td>RF01095</td>
<td>PK-CuYV_BPYV</td>
<td>56</td>
<td>4</td>
<td>2132</td>
<td>15127</td>
<td>7.1</td>
</tr>
<tr>
<td>RF01097</td>
<td>RF_site6</td>
<td>52</td>
<td>15</td>
<td>2188</td>
<td>14635</td>
<td>6.7</td>
</tr>
<tr>
<td>RF01100</td>
<td>PK-BYV</td>
<td>40</td>
<td>4</td>
<td>1697</td>
<td>9997</td>
<td>5.9</td>
</tr>
<tr>
<td>RF01102</td>
<td>PK1-TEV_CVMV</td>
<td>36</td>
<td>8</td>
<td>1524</td>
<td>9477</td>
<td>6.2</td>
</tr>
<tr>
<td>RF01104</td>
<td>SBWMV2_UPD-PKb</td>
<td>29</td>
<td>8</td>
<td>1488</td>
<td>7334</td>
<td>4.9</td>
</tr>
<tr>
<td>RF01109</td>
<td>SBWMV1_UPD-PKd</td>
<td>22</td>
<td>12</td>
<td>1185</td>
<td>4689</td>
<td>4.0</td>
</tr>
<tr>
<td>RF01118</td>
<td>PK-G12rRNA</td>
<td>124</td>
<td>44</td>
<td>17714</td>
<td>184441</td>
<td>10.4</td>
</tr>
</tbody>
</table>
MA ET AL.: AN EFFICIENT ALIGNMENT ALGORITHM FOR SEARCHING SIMPLE PSEUDOKNOTS OVER LONG GENOMIC SEQUENCE

where $\mathcal{C}_1$ is the character set $\{A,C,G,U\}$. The similarity between two sequences according to their sequences and structures.

$T$ is from $M$ structure. The calculation of structural alignment for sequence similarity and the score for structural similarity, respectively. The calculation of structural alignment score is not restricted to any kind of secondary structure. It works in the same way for pseudoknot structure. The objective is to find an alignment such that the corresponding score is maximized.

Definition 1. Optimal structural alignment between $S[1 \ldots m]$ and $T[1 \ldots n]$ is to find $S'[1 \ldots r]$ and $T'[r \ldots r]$, where $r \geq m, n, S'$ is from $S$ and $T'$ is from $T$ with spaces inserted in between the characters to make both sequences of the same length. A space cannot appear in the same position of $S'$ and $T'$. The score of the alignment, which determines the sequence and structure similarity between $S'$ and $T'$, is defined as follows:

$$\text{score} = \sum_{i=1}^{r} \gamma(S'[i], T'[i]) + \sum_{\mathcal{C}_1 \times \mathcal{C}_1 \times \mathcal{C}_1 \times \mathcal{C}_1 \times \mathcal{C}_1} \delta(S'[i], S'[j], T'[i], T'[j])$$

where $\eta(i)$ is the corresponding position in $S$ according to the position $i$ in $S$; $\gamma(t_1, t_2)$ and $\delta(x_1, y_1, x_2, y_2)$, where $t_1, t_2 \in \{A,C,G,U\}$ and $x_1, x_2, y_1, y_2 \in \{A,C,G,U\}$, are the score for sequence similarity and the score for structural similarity, respectively. The calculation of structural alignment score is not restricted to any kind of secondary structure. It works in the same way for pseudoknot structure. The objective is to find an alignment such that the corresponding score is maximized.

Definition 2. The optimal semiglobal structural alignment problem between $S[1 \ldots m]$ and $T[1 \ldots n]$ is to find a substring $T[p \ldots q]$ where $1 \leq p, q \leq n$ (i.e., $T[p \ldots q]$ is an empty string when $p > q$) such that the score of the optimal structural alignment between $S[1 \ldots m]$ and the substring $T[p \ldots q]$ is maximum.

In the following sections, for the sake of simplicity, we regard each subregion $T[k \ldots m + (k + 1)s]$ as $T$ during the illustration of the algorithms.

3 ALGORITHMS

The optimal semiglobal structural alignment can be computed by using dynamic programming. The algorithm was developed by Han et al. [13] (although their method is designed to compute the optimal structural alignment, a minor modification on their method will be able to compute the optimal semiglobal structural alignment). The key idea behind their algorithm is the concept of a substructure which allows the optimal alignment to be computed recursively. The algorithm takes $O(mn^2)$ time where $m$ is the length of the query sequence with known simple pseudoknot structure and $n$ is the length of the target sequence with unknown structure.

3.1 Han’s Algorithm

Let $S[i_0 \ldots i_0]$ be the query sequence with simple pseudoknot structure $M_{i_0}$. As defined in Section 2, the pivot points $x_1, x_2$ for $S[i_0 \ldots i_0]$ are known. Let $v = (i, j, k)$ be a triple with $i_0 \leq i < x_1 \leq j < x_2 \leq k \leq i_0$. Define the subregion $R(S, v) = [i_0 \ldots i] \cup [j \ldots k]$ (as shown in Fig. 3). Let $\text{Struct}(R) = \{(i, j) \in M[i, j \in R]\}$ where $R$ is a subregion. The subregion $R$ defines a valid substructure (Struct$(R)$) of researchers usually use the sliding-window technique to perform the searching. If it is assumed that the maximum length of a candidate is $m$ and there does not exist two candidates inside a region of length $m + s$, by setting the window size as $m + s$ and a sliding distance as $s$ (see Fig. 2), for each subregion $T[1 \ldots m + s], T[s + 1 \ldots m + 2s], \ldots, T[k + 1]$, where $1 \leq k \leq (n - m)/s - 1$, one could compute the optimal semiglobal structural alignment (which will be formally defined later) between the query and each of the subregions. This method is commonly used because one would only need to compute the structural alignment for every $s$ position, and so there would lead to a speedup. However, according to our experiment, we found that even using this approach, it still takes a long time for a long genome sequence of length, say 50k.
The situation, where neither \((i, j)\) nor \((j, k)\) is a base pair, is also similar. The only modifications on Han’s algorithm for semiglobal alignment are the additional cases for \(i = i_0\) and \(k = k_0\) when inserting a space on \(S\). To fill the dynamic programming table, not all entries for all possible subranges of \(S\) need to be filled. For any given subregion \(v = (i, j, k)\), a function \(\zeta(v)\) can be defined as follows to determine for which subregions in \(S\) we need to fill the corresponding \(B\) entries.

If \(v\) defines a subregion with valid substructure, \(\zeta(v)\) also defines a valid substructure. Let \(v' = (x_1 - 1, x_1, k_0)\). We only need to fill in the entries for \(B\) provided that \(v\) can be obtained from \(v'\) by applying \(\zeta\) function repeatedly. Intuitively, \(\zeta\) guides which recursion formula to use. And there are only \(O(m)\) such \(v\) values. In [13], they collect all these \(v\) and form a chain of triples labeled \(V\), \((x_1 - 1, x_1, k_0)\) is the first element of the chain \(V\) and the last element of \(V\) is an empty region. Building \(V\) requires only \(O(m)\) time. Then compute \(B\) for all possible subregions \((e, f, g)\) on \(T\) from the last element of \(V\) to the first element of \(V\). Therefore, the time complexity of the algorithm is \(O(mn^3)\) and it requires \(O(mn^3)\) space.

### 3.2 Our Method

In Han’s algorithm, during the calculation of \(B\), for each value of \(v\) on \(S\), they need to consider all possible subregions \((e, f, g)\) on \(T\) (where \(e < f \leq g\)). However, for a subregion \(v\) on \(S\), if we can estimate which \((e, f, g)\) on \(T\) would not align with \(v\) in the optimal semiglobal alignment, then we will be able to skip those subregions and would lead to a speedup in the computation of \(B\) practically. To estimate which \((e, f, g)\) on \(T\) would not align with a given \(v\) on \(S\) in the optimal semiglobal alignment, our method involves two steps. First we try to compute a lower bound of the optimal semiglobal alignment score \(e\) between \(S\) and \(T\). That means we can compute a value \(c\) such that the optimal semiglobal alignment score between \(S\) and \(T\) will definitely greater than or equal to \(c\). Then according the value of \(c\), we can estimate for a given \(v\) the set of \((e, f, g)\) which would not align with \(v\) in any of the optimal semiglobal alignments between \(S\) and \(T\) (i.e., there may be more than one alignment between \(S\) and \(T\) with same maximum score). The computation of the value of \(c\) has to be as fast as possible and the value of \(c\) should be closed to the optimal semiglobal alignment score in order to achieve a considerable speed up. Note that we guarantee that the regions skipped by our algorithm are not the answers (see Lemmas 4 and 7), thus our algorithm will not miss any answers. On the other hand, for those regions that are not skipped by our algorithm, we evaluate them in the same way as in Han’s algorithm, so we also will not produce more false positives.
3.2.1 Computation of Lower Bound

Let \( S_{[i_0 \ldots i_k]} \) be the query sequence with a simple pseudoknot structure \( M_{i_0, i_k} \). As shown in the Fig. 4, a pseudoknot can be divided into two parts: 1) \( S_{[i_0 \ldots i_2 - 1]} \) with a set of base pairs \( U \); and 2) \( S_{[i_2 \ldots i_k]} \) with a set of base pairs \( V \). \( U = \{(i, j) \in M_{i_0, i_2} | i_0 \leq i < x_1 \leq j \leq i_2 - 1 \} \) and \( V = \{(i, j) \in M_{i_2, i_k} | x_2 \leq i < x_2 \leq j \leq i_k \} \). According to the definition of simple pseudoknot, since \( U = M_L \) and \( V = M_R \) (defined in Section 2), \( U \cup V = M_{i_0, i_k} \) and both of \( U \) and \( V \) are regular structures. Also, there is no branching in the structure of \( U \) (or \( V \)). That means there does not exist two base pairs \((i_1, j_1), (i_2, j_2) \in U \) or \( V \) such that \( i_1 < j_1 < i_2 < j_2 \) or \( i_2 < j_2 < i_1 < j_1 \).

Let \( T[1 \ldots n] \) be the target sequence with unknown structure. The high level idea of the heuristics to compute the lower bound of the optimal semiglobal alignment score is first to perform an optimal semiglobal structural alignment between the first part (i.e., \( S_{[i_0 \ldots i_2 - 1]} \) with a set of base pairs \( U \)) and the target \( T[1 \ldots n] \). Let \( T[e' \ldots f'] \) be the resulting subregions on \( T \) having optimal alignment score. We then perform an optimal prefix-global structural alignment between the second part (i.e., \( S_{[x_2 \ldots x_k]} \)) with a set of base pairs \( V \) and \( T[f' + 1 \ldots n] \). The lower bound of the score is set to be the sum of the optimal semiglobal structural alignment score for the first part and the optimal prefix-global structural alignment score for the second part. The optimal prefix-global structural alignment is defined as follows.

**Definition 3.** The optimal prefix-global structural alignment problem between \( S[1 \ldots m] \) and \( T[1 \ldots n] \) is to find a prefix \( T[1 \ldots p] \) where \( 0 \leq p \leq n \) (i.e., \( T[1 \ldots p] \) is an empty string when \( p < 1 \)) such that the score of the optimal part-global alignment between \( S[1 \ldots m] \) and the prefix \( T[1 \ldots p] \) is maximum. This alignment is referred as the optimal prefix-global structural alignment.

Considering the computation of the semiglobal structural alignment between the first part \( S_{[i_0 \ldots i_2 - 1]} \) with a set of base pairs \( U \) and the target \( T[1 \ldots n] \), due to the characteristics that there does not exist two base pairs \((i_1, j_1), (i_2, j_2) \in U \) such that \( i_1 < j_1 < i_2 < j_2 \) or \( i_2 < j_2 < i_1 < j_1 \), the time complexity of the algorithm is only \( O(mn^2) \) where \( m \) is the length of \( S \) and \( n \) is the length of \( T \). Let \( \text{Struct}_U(i, j) \) be the set of base pairs \((i', f') \in U \) such that \( i \leq i' < f' \leq j \). Define \( C(i, j, e, f) \) be the score of the optimal structural alignment between \( S_{[i \ldots j]} \) with substructure \( \text{Struct}_U(i, j) \) and the target \( T[e \ldots f] \). The score of the optimal semiglobal structural alignment between \( S_{[i_0 \ldots i_2 - 1]} \) and \( T[1 \ldots n] \) is \( \max_{i, j < e, f} C(i, j, x_1 - 1, e', f') \), where \( T[e' \ldots f'] \) is the resulting subregions on \( T \) having optimal semiglobal structural alignment score. \( C(i, j, e, f) \) can be computed as follows.

**Lemma 2.** If \((i, j) \) is a base pair, \( C(i, j, e, f) = \max \)

\[
\begin{align*}
&\text{MATCH}_{\text{both}}(\gamma) + \gamma(S[i, j], T[e, f]) + \gamma(S[j, i], T[f, e]), \\
&\text{MATCH}_{\text{single}}(\gamma) + \gamma(S[i, j], T[e, f]) + \gamma(S[j, i], T[f, e]), \\
&\text{DELETE}(\gamma) + \gamma(S[i, j], T[e, f]) + \gamma(S[j, i], T[f, e]), \\
&\text{INSERT}(\gamma) + \gamma(T[e', f']), \\
&\text{DELETE}(\gamma) + \gamma(T[e', f']).
\end{align*}
\]

The case, which \((i, j) \) is not a base pair, is also similar. Similarly, to fill the dynamic programming table, not all entries for all possible subranges of \( S \) needs to be filled. Let \( v = (i, j) \) representing a subregion \( S[i \ldots j] \), a function \( \xi(v) \) can be defined as follows to determine for which subregions in \( S \) we need to fill the corresponding \( C \) entries.

**Lemma 3.** Let \( D(k_0, n) \) be the set of base pairs \((k_0, e) \) with a set of base pairs \( \text{Struct}_V(k) \) and the target \( T[e' \ldots f'] \) based on the alignment between \( S_{[i_0 \ldots i_2 - 1]} \) and \( T[e' \ldots f'] \). The optimal prefix-global structural alignment between \( S_{[x_2 \ldots x_k]} \) with a set of base pairs \( V \) and \( T[f' + 1 \ldots n] \) given the alignment between \( S_{[i_0 \ldots i_2 - 1]} \) and \( T[e' \ldots f'] \) can be obtained from the entry \( D(k_0, n) \). \( D(k, g) \) can be computed as follows.

**Lemma 4.** Let \( \tau(k, g) \) be the set of base pairs \((k, g) \) in \( V \) or \( -1 \) otherwise. Let \( \tau(j) = f \) where \( i_0 \leq j \leq x_2 - 1 \) and \( e' \leq f \leq g \) if \( S_{[i_j]} \) is aligned with \( T[f] \) in the resulting optimal alignment between \( S_{[i_0 \ldots i_2 - 1]} \) and \( T[e' \ldots f'] \). Otherwise, set \( \tau(j) = -1 \) if \( S_{[i_j]} \) is aligned with space in the resulting optimal alignment. \( D(k, g) = \max \)

\[
\begin{align*}
&\text{MATCH}_{\text{both}}(\gamma) + \gamma(S[i, j], T[e, f]) + \gamma(S[j, i], T[f, e]), \\
&\text{MATCH}_{\text{single}}(\gamma) + \gamma(S[i, j], T[e, f]) + \gamma(S[j, i], T[f, e]), \\
&\text{DELETE}(\gamma) + \gamma(S[i, j], T[e, f]) + \gamma(S[j, i], T[f, e]), \\
&\text{INSERT}(\gamma) + \gamma(T[e', f']), \\
&\text{DELETE}(\gamma) + \gamma(T[e', f']).
\end{align*}
\]
Lemma 4. The resulting lower bound is always less than or equal to the score of the optimal semiglobal structural alignment between S and T.

Proof. Let T'[e'...f'] where e ≤ e', f ≤ f be the resulting region with the optimal semiglobal structural alignment with S[1...x2-1]. Let T'[f'+...g] where g ≤ b be the resulting region with the optimal prefix-global structural alignment with S[x2...m]. The lower bound equals to the corresponding score of the resulting structural alignment between S[1...x2-1] ∪ S[x2...m] and T'[e'...f'] ∪ T'[f'+...g']. Since it is one of the candidates of the optimal semiglobal structural alignment between S[1...m] and T[1...n], the resulting lower bound is always less than or equal to the score of the optimal semiglobal structural alignment between S and T. □

3.2.2 Pruning

After computing the lower bound of the optimal structural alignment between S and T, for each subregions (e, f, g) on T (where e < f ≤ g), we will try to estimate which (e, f, g) on T would not align with v in the optimal alignment. Then subsequent steps of computation of the alignment which involves (e, f, g) aligned with v can be skipped and so it would lead to a speedup in the computation of B.

Considering the computation of semiglobal structural alignment between S[1...m] and T[1...n], for a subregion R_x = R(S, (i, j, k)) on S and another subregion R_y = R(T, (e, f, g)) on T, B(R_x, R_y) is defined as the score of the optimal semiglobal structural alignment between R_x and R_y. In the following, we will compute the maximum possible alignment score between S and T provided that R_x is aligned with R_y. The situation that R_x is aligned with R_y means that S[1...i] is aligned with T[1...e] and S[j...k] is aligned with T[f...g]. Formally, R_y is defined as aligned with R_y if for the resulting optimal alignment of S and T: S'[1...r] and T'[1...r] where r ≥ m, n, there exists a, b, c such that after removing the spaces, S'[1...a] and T'[1...a] will be equal to S[1...v] and T[1...v] and S'[b...c] and T'[b...c] will be equal to S[j...k] and T[f...g], respectively.

Given subregions R_x = R(S, (i, j, k)), R_y = R(T, (e, f, g)) and B(R_x, R_y), let −R_x = S[i + 1...j - 1] ∪ S[k + 1...m] and −R_y = T[e + 1...f - 1] ∪ T[g + 1...n], given R_x is aligned with R_y, when we consider the case that i > 1 and k < m, the maximum semiglobal alignment score between S and T would be equal to the sum of the optimal suffix-global alignment score between R_x and R_y and the optimal prefix-global alignment score between −R_x and −R_y. The optimal suffix-global structural alignment is defined as follows:

Definition 4. Optimal suffix-global structural alignment between S[1...m] and T[1...n] is to find a suffix T[q...n] where 1 ≤ q ≤ n (i.e., T[q...n] is an empty string when q > n) such that the score of the optimal structural alignment between S[1...m] and the suffix T[q...n] is maximum.

Let max_s(R_x, R_y) be the maximum alignment score between S and T provided that R_x is aligned with R_y. Since the optimal suffix-global alignment score between R_x and R_y is always less than or equal to B(R_x, R_y), if we can compute the upper bound of the optimal prefix-global alignment score between −R_x and −R_y (let it be E(−R_x, −R_y)), then max_s(R_x, R_y) is less than or equal to B(R_x, R_y) + E(−R_x, −R_y). If the value B(R_x, R_y) + E(−R_x, −R_y) is smaller than the lower bound computed previously, we can definitely know that in any of the optimal alignments, R_x should not be aligned with R_y and then the subsequent steps of computation of the alignments which involve R_x aligned with R_y can be skipped and so it would lead to a speedup in the computation of B.

In the following, we will provide a heuristic algorithm to compute E(−R_x, −R_y). The method is time efficient and the value is proved to be greater than or equal to the optimal suffix-global alignment score between −R_x and −R_y.

As shown in Fig. 5, the region −R_x can be further decomposed into two parts: 1) S[i + 1...j - 1] with a set of base pairs U^{i,j-1}_k; 2) S[k + 1...m] with a set of base pairs V^{m}_{k+1}. U^{i,j-1}_k = \{(i', j') ∈ M | i + 1 ≤ i' < x_i ≤ j' ≤ j - 1\} and V^{m}_{k+1} = \{(i', j') ∈ M | k + 1 ≤ i' ≤ j' \leq m\}. The upper bound of the optimal prefix-global alignment score (E(−R_x, −R_y)) between −R_x (i.e., S[i + 1...j - 1] ∪ S[k + 1...m]) and −R_y (i.e., T[e + 1...f - 1] ∪ T[g + 1...n]) is set to be the sum of 1) the optimal alignment score between S[i + 1...j - 1] with a set of base pairs U^{i,j-1}_k; and 2) the optimal prefix-global alignment score between S[k + 1...m] with a set of base pairs V^{m}_{k+1} and T[g + 1...n].
Let \( F(i, j, e, f) \) be the optimal alignment score between \( S[i \ldots j] \) with a set of base pairs \( U_i^j \) and \( T[e \ldots f] \). Direct computation of \( F(i, j, e, f) \) will require at least \( O(n^4) \). Also, if the computation of \( F(i, j, e, f) \) is required for each subregion \( R_x \), then the total time required will be further blow up. Thus, in order to compute \( F(i, j, e, f) \) efficiently, we need to make use of the triple list \( V \) in the Han’s algorithm mentioned in Section 3.1. \( V \) is a chain of triples \((i, j, k)\) which includes the subregions on \( S \) we need to come across during the computation of \( B \). There are \( O(m^3) \) elements in the chain. \( V \) can be first computed in the beginning according to the function \( \zeta \) in Section 3.1. In order words, we only need to obtain the value of \( F(i, j, e, f) \) for all \( e \) and for those \( i, j \) such that there exists \((i', j', k')\) in \( V \) such that \( i = i' + 1 \) and \( j = j' - 1 \). Thus there are only \( O(m) \) number of \((i, j)\). Let \( R \) be the set of \((i, j)\) for which we need to compute the value of \( F \).

For case I, \((i, j, k)\) forms a base pair with another base \((i_1, j_1, k_1)\) in \( V \), then the total time required will be further blow up. For case IV, \( F \) is computed according to either \((i + 1, j + 1, k)\) or \((i + 1, j, k - 1)\) depending on which pair is in the \( R \). Therefore, for all cases, during the computation of \( F \), the entries we refer to are always valid.

To conclude, in order to fill the dynamic programming table, not all entries for all possible subranges \((i, j)\) of \( S \) need to be filled. For any given subrange \( w = (i, j) \), a function \( \Psi(w) \) can be defined as follows to determine for which subregions in \( S \) we need to fill the corresponding \( F \) entries.

\[
\Psi(w) = \begin{cases} 
(i + 1, j - 1), & \text{if } (i, j) \text{ base pair in } U_i^j, \\
(i, j - 1), & \text{else if } \exists k \text{ s.t. } (i - 1, j, k) \in V, \\
(i + 1, j), & \text{else if } \exists k \text{ s.t. } (i, j + 1, k) \in V, \\
(i + 1, j - 1), & \text{else if } \exists k \text{ s.t. } (i, j, k - 1) \in V,
\end{cases}
\]

\( R \) can be obtained by using \( \Psi(w) \). Thus, to compute all the \( F \) for all \((i, j, k)\) in \( V \), the computation can be performed as preprocessing and the total time required is \( O(m^3) \).

Similarly, the optimal prefix-global alignment score \((G(k + 1, g + 1))\) between \( S[k + 1 \ldots m] \) with a set of base pairs \( V_{k+1}^m \) and \( T[g + 1 \ldots n] \) for all \( k \) and \( g \) can be computed in time \( O(mn) \). The upper bound of the optimal prefix-global alignment score between \(-R_x \) and \(-R_y \) is set to the sum of \( F(i + 1, j - 1, e + 1, f - 1) \) and \( G(k + 1, g + 1) \).

**Lemma 7.** The resulting upper bound is always greater than or equal to the score of the optimal prefix-global alignment between \(-R_x \) and \(-R_y \).

**Proof.** Consider \( S[1 \ldots m] \) and \( T[1 \ldots n] \), let \( E(\neg R_x, \neg R_y) \) be the optimal prefix-global alignment score between \(-R_x \) and \(-R_y \), where \( \neg R_x = R(S, (i, j, k)) \) and \( R_y = R(T, (e, f, g)) \). Since \( \neg R_x = S[i + 1 \ldots j - 1] \cup S[k + 1 \ldots m] \) and \( \neg R_y = T[e + f \ldots j - 1] \cup T[g + 1 \ldots n] \), let \( S'[1 \ldots r] \cup S''[1 \ldots r] \) and \( T'[1 \ldots r] \cup T''[1 \ldots r] \) be the resulting optimal prefix-global alignment between \( \neg R_x \) and \( \neg R_y \). Note that \( E(\neg R_x, \neg R_y) \) is the sum of the alignment score between \( S'[1 \ldots r] \) and \( T'[1 \ldots r] \), and the alignment score between \( S''[1 \ldots r] \) and \( T''[1 \ldots r] \). Because \( S'[1 \ldots r] \) and \( T'[1 \ldots r] \) is one of the candidates when computing the optimal alignment score between \( S[1 \ldots j - 1] \) and \( T[e + f \ldots j - 1] \), \( F(i + 1, j - 1, e + 1, f - 1) \) is always greater than or equal to the alignment score between \( S'[1 \ldots r] \) and \( T'[1 \ldots r] \). Similarly, because \( S''[1 \ldots r] \) and \( T''[1 \ldots r] \) is one of the candidates when computing the optimal prefix-global alignment score between \( S[k + 1 \ldots m] \) and \( T[g + 1 \ldots n] \),
G(k + 1, g + 1) is always greater than or equal to the alignment score between $S^w[1...r']$ and $T^w[1...r']$. Therefore, the sum of $F(i + 1, j - 1, e + 1, f - 1)$ and $G(k + 1, g + 1)$ needs only constant time. If the value $B(R_x, R_y) + E(-R_x, -R_y)$ is smaller than the lower bound computed previously, we can definitely know that in any of the optimal alignments, $R_x$ would not aligned with $R_y$ and then the subsequent steps of computation of the alignment which involves $R_x$ aligned with $R_y$ can be skipped and so it would lead to a speedup in the computation of $B$.

**Lemma 8.** The space complexity of our method is $O(mn^3)$.

**Proof.** In our method, we need to compute the tables $B, C, D, E, F,$ and $G$. For each table, the number of entries we need to store during calculation is $O(mn^3)$. Thus, the overall space complexity is $O(mn^3)$. □

We remark that this space complexity is the same as that of Han’s algorithm. Note also that if a region cannot be pruned, we still go through the same calculation as in Han’s algorithm, thus the worst case time complexity for evaluating a region is still $O(mn^3)$, the same as that of Han’s algorithm.

**4 Experimental Results**

We implemented our algorithm for simple pseudoknots in C++. By inputting a query RNA sequence $S$ with its secondary structure, the program scan along the reference target sequence $T$ and output the score for every region of $T$. A higher score indicates that the reference target sequence is similar to the query RNA sequence $S$ in terms of both structure and sequence. In the database Rfam 10.1, there are 71 pseudoknotted families. Among them around 40 families have simple pseudoknot structures. To evaluate the effectiveness of our algorithm, we randomly selected around 20 families with simple pseudoknot structures for the experiment. For each family, we randomly pick a seed member as the query sequence $S$. The target sequence $T$ is a random sequence (with equally distribution of A, C, G, and T) of length 50,000 and the members of the family are embedded at the random positions on $T$. Since the PAL [13] program is not available, we also implemented their method for comparison on the performance. We applied the sliding window technique for both Han’s method and our method. Table 1 and Fig. 7 show the comparison of the running time between our algorithm and Han’s. Our speed up gradually increases when the query length increases. The speed up is five times when the query length is 30. And the speed up is seven times when the query length is 50. The speed up is 8.5 to 11.0 times when the query length is 65 or more. The experiment was performed using a machine with 24 G memory and a quad-core 2.6 GHz CPU. It shows that our method is efficient especially for long query sequence.

To evaluate the efficiency of our method, we further examine the sensitivity of our algorithm and compare it with the existing method called RNATOPS [15] and BLAST. RNATOPS is a profile-based RNA structure search program that can detect RNA pseudoknots in genomes. RNATOPS is based on structures and BLAST is based on sequence, while our and Han’s algorithm are based on both sequence and structure. For each family, we embedded the corresponding members into the random sequence and search it with the RNA query pattern. To understand the sensitivity of the software, say for a specific family, there are $X$ seed members embedded and we check the top $X$ reported regions with highest scores and see how many of them are the real members. Table 2 shows the sensitivity of our method, Han’s method, RNATOPS, and BLAST. The sensitivity of our method is the same as that of Han’s method. As shown in the table, the sensitivity of our method ranges from 90-100 percent with an average of 97 percent, which outperforms both RNATOPS (average 9 percent) and BLAST (average 68 percent). Fig. 6 shows an example of the distribution of our scores of the false hits (i.e., the regions which do not contain any real member) and the true hits (i.e., the regions which contain the real members) for the family RF01118. One can see that the scores of the real hits are usually higher than those of the false hits. It shows that our method could identify the real members from the target sequence.

2. If there are too many seed members in the family, we will randomly pick some of them so that the total length of all the seed members is around 2,500 (i.e., 5 percent of the total length of the target sequence).
sequence on the genome $T$ (i.e., computation of $C$ and $D$); 2) preprocessing steps for pruning (i.e., computation of the matrix $E$). Table 3 shows the time required for each of the preprocessing steps. In summary, the total time required for both of the preprocessing steps is less than 2 percent of the total running time of our method.

To further analyze the effectiveness of our pruning method, we selected the family RF01118 and examined the percentage of table entries pruned for each subregion $R(v)$ on the query $S$. As shown in Fig. 8, in the beginning of the algorithm, the size of $R(v)$ is small and so it is expected that not many table entries are pruned. While the algorithm continues and the size of $R(v)$ increases, the pruning works more effectively and the number of table entries pruned increases dramatically. When the size of $R(v)$ reaches 20 percent of the query, there are over 50 percent of the table entries pruned. This statistics therefore could explain the resulting speed up of our algorithm.

### 5 CONCLUSIONS

We have designed an algorithm which first estimates the lower bound of the optimal alignment score and then applies an efficient pruning method for speeding up the dynamic programming algorithm for solving the RNA structural alignment problems. Experimental results show that the algorithm is effective and can achieve around 8.5 to 11 times speed up when the query length is longer than 65. On the other hand, in practice, the user sometimes would like to set a threshold so that only the subregions on the genome of which the resulting scores are higher than the threshold will be outputted. Our pruning method can work well with the inputted threshold too. The entries are then skipped if the estimated optimal alignment score for the entries is lower than the inputted threshold instead of the computed lower

### TABLE 2

Comparison of the Sensitivity of Our Method, Han’s Method, RNATOPS, and BLAST

<table>
<thead>
<tr>
<th>Family ID</th>
<th># of members embedded $(X)$</th>
<th># of members in top $X$ ans</th>
<th>Our and Han’s method</th>
<th>RNATOPS</th>
<th>BLAST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>sensitivity</td>
<td>false positive rate</td>
<td>sensitivity</td>
</tr>
<tr>
<td>RF00165</td>
<td>30</td>
<td>27</td>
<td>90%</td>
<td>10%</td>
<td>4</td>
</tr>
<tr>
<td>RF00390</td>
<td>25</td>
<td>24</td>
<td>96%</td>
<td>4%</td>
<td>2</td>
</tr>
<tr>
<td>RF00505</td>
<td>75</td>
<td>75</td>
<td>100%</td>
<td>0%</td>
<td>1</td>
</tr>
<tr>
<td>RF00507</td>
<td>23</td>
<td>21</td>
<td>91%</td>
<td>9%</td>
<td>3</td>
</tr>
<tr>
<td>RF01072</td>
<td>40</td>
<td>34</td>
<td>85%</td>
<td>15%</td>
<td>7</td>
</tr>
<tr>
<td>RF01074</td>
<td>12</td>
<td>12</td>
<td>100%</td>
<td>0%</td>
<td>0</td>
</tr>
<tr>
<td>RF01076</td>
<td>38</td>
<td>38</td>
<td>100%</td>
<td>0%</td>
<td>5</td>
</tr>
<tr>
<td>RF01080</td>
<td>16</td>
<td>15</td>
<td>94%</td>
<td>6%</td>
<td>1</td>
</tr>
<tr>
<td>RF01081</td>
<td>14</td>
<td>14</td>
<td>100%</td>
<td>0%</td>
<td>2</td>
</tr>
<tr>
<td>RF01087</td>
<td>37</td>
<td>35</td>
<td>95%</td>
<td>5%</td>
<td>2</td>
</tr>
<tr>
<td>RF01089</td>
<td>25</td>
<td>24</td>
<td>96%</td>
<td>4%</td>
<td>2</td>
</tr>
<tr>
<td>RF01093</td>
<td>38</td>
<td>38</td>
<td>100%</td>
<td>0%</td>
<td>3</td>
</tr>
<tr>
<td>RF01095</td>
<td>4</td>
<td>4</td>
<td>100%</td>
<td>0%</td>
<td>0</td>
</tr>
<tr>
<td>RF01097</td>
<td>15</td>
<td>15</td>
<td>100%</td>
<td>0%</td>
<td>2</td>
</tr>
<tr>
<td>RF01100</td>
<td>4</td>
<td>4</td>
<td>100%</td>
<td>0%</td>
<td>1</td>
</tr>
<tr>
<td>RF01102</td>
<td>8</td>
<td>8</td>
<td>100%</td>
<td>0%</td>
<td>0</td>
</tr>
<tr>
<td>RF01104</td>
<td>8</td>
<td>8</td>
<td>100%</td>
<td>0%</td>
<td>1</td>
</tr>
<tr>
<td>RF01109</td>
<td>12</td>
<td>12</td>
<td>100%</td>
<td>0%</td>
<td>2</td>
</tr>
<tr>
<td>RF01118</td>
<td>44</td>
<td>43</td>
<td>98%</td>
<td>2%</td>
<td>1</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td>97%</td>
<td>3%</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 3

Breakdown of the Running Time

<table>
<thead>
<tr>
<th>Family ID</th>
<th>Query length</th>
<th>Lower bound calculation (in second)</th>
<th>Pre-processing for pruning (in second)</th>
<th>Subtotal</th>
<th>% of total running time</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF00505</td>
<td>66</td>
<td>16.1</td>
<td>16.85</td>
<td>32.95</td>
<td>1.60%</td>
</tr>
<tr>
<td>RF00507</td>
<td>85</td>
<td>35.7</td>
<td>36.8</td>
<td>72.5</td>
<td>1.78%</td>
</tr>
<tr>
<td>RF01076</td>
<td>72</td>
<td>24.54</td>
<td>25.81</td>
<td>50.35</td>
<td>1.89%</td>
</tr>
<tr>
<td>RF01087</td>
<td>148</td>
<td>234.7</td>
<td>261.23</td>
<td>495.93</td>
<td>1.93%</td>
</tr>
<tr>
<td>RF01089</td>
<td>116</td>
<td>100</td>
<td>103.8</td>
<td>203.8</td>
<td>1.94%</td>
</tr>
<tr>
<td>RF01118</td>
<td>124</td>
<td>163.5</td>
<td>175.8</td>
<td>339.3</td>
<td>1.92%</td>
</tr>
</tbody>
</table>

**Fig. 8.** Percentage of table entries pruned for different size of subregion $R(v)$ on the query $S$ in family RF01118.
bound. By setting the threshold as the 40 percent of the maximum score for the query (i.e., the maximum score is the score when the query is aligned with the sequence exactly the same as the query), we found that the speed up can further increase to 17 times to 28 times. Although the method presented in this paper is designed for simple pseudoknots, a similar technique can also be applied for the structural alignment algorithms for other pseudoknot types.

ACKNOWLEDGMENTS

This research was supported in part by the Seed Funding Programme for Basic Research (201011159115) of HKU and by the General Research Fund (GRF) of the Hong Kong Government (HKU 719611E).

REFERENCES


Christopher Ma received the master’s degree from the Department of Computer Science at the University of Hong Kong in 2012.

Thomas K.F. Wong received the PhD degree in computer science from the University of Hong Kong in 2011. Currently, he is working as a postdoctoral fellow at the same university. His research interest includes bioinformatics.

T.W. Lam received the PhD degree in computer science from the University of Washington, in 1988. Currently, he is working as a professor at the University of Hong Kong. His research interests include bioinformatics, text indexing, and algorithm design.

W.K. Hon received the PhD degree from the University of Hong Kong in 2005. Currently, he is working as an associate professor in the Department of Computer Science at National Tsing Hua University, Taiwan. He visited Purdue University from 2004 to 2006. His research interests include data compression, text indexing, and algorithm design.

K. Sadakane received the BS, MS, and PhD degrees from the Department of Information Science, University of Tokyo, in 1995, 1997, and 2000, respectively. He was a research associate in the Graduate School of Information Sciences, Tohoku University from 2000 to 2003, and an associate professor at Faculty of Information Science and Electrical Engineering, Kyushu University from 2003 to 2009. Since 2009, he has been an associate professor at the National Institute of Informatics. His research interest includes information retrieval, data structures, and data compression.

S.M. Yiu received the PhD degree in computer science from the University of Hong Kong in 1996 and is currently working as an assistant professor at the same university. His research interests include bioinformatics and computational biology. He is a member of the IEEE and the IEEE Computer Society.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.