Predicting pseudoknotted structures across two RNA sequences

Jana Sperschneider¹,*, Amitava Datta¹ and Michael J. Wise¹,²

¹School of Computer Science and Software Engineering and ²School of Chemistry and Biochemistry, University of Western Australia, Perth, Australia

ABSTRACT

Motivation: Laboratory RNA structure determination is demanding and costly and thus, computational structure prediction is an important task. Single sequence methods for RNA secondary structure prediction are limited by the accuracy of the underlying folding model, if a structure is supported by a family of evolutionarily related sequences, one can be more confident that the prediction is accurate. RNA pseudoknots are functional elements, which have highly conserved structures. However, few comparative structure prediction methods can handle pseudoknots due to the computational complexity.

Results: A comparative pseudoknot prediction method called DotKnot-PW is introduced based on structural comparison of secondary structure elements and H-type pseudoknot candidates. DotKnot-PW outperforms other methods from the literature on a hand-curated test set of RNA structures with experimental support.


Contact: janaspe@csse.uwa.edu.au

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on March 2, 2012; revised on August 1, 2012; accepted on September 18, 2012

1 INTRODUCTION

Macromolecules such as DNA, RNA and proteins have the ability to form diverse tertiary structures, which enable functionality and thus, life. For many decades, proteins were deemed the global apparatus of the host (Brierley et al., 2007). RNA viruses use pseudoknots for hijacking the replication process outside the loop. From an algorithmic point of view, even the simplest type of pseudoknot adds considerable computational demands due to crossing base pairs. Biologists have delivered a wealth of studies, which show that pseudoknots have an astonishing number of diverse functions and occur in most classes of RNA (Staple and Butcher, 2005). RNA viruses use pseudoknots for hijacking the replication apparatus of the host.

A limited number of RNA comparative structure prediction methods can handle pseudoknots due to the computational complexity. Several of these methods take a sequence alignment as an input. ILM is an algorithm that takes as an input either individual sequences or a sequence alignment (Ruan et al., 2004). A base pair score matrix is prepared initially and helices are added to the structure in an iterative fashion. In the approach hmatch, a maximum weighted matching algorithm with combined thermodynamic and covariance scores is used (Witwer et al., 2004). This program gives the option to be combined with RNAalifold.

*To whom correspondence should be addressed.
KNetFold is a machine learning method, which takes a sequence alignment as an input and outputs a consensus structure allowing pseudoknots (Bindewald and Shapiro, 2006). SimuFold takes an alignment as an input and simultaneously calculates a structure including pseudoknots, a multiple-sequence alignment and an evolutionary tree by sampling from the joint posterior distributions (Meyer and Miklos, 2007). Tfold combines stem stability, covariance and conservation to search for compatible stems and subsequently for pseudoknots for a set of aligned homologous sequences (Engelen and Tahi, 2010). Several comparative structure prediction methods including pseudoknots do not rely on an initial sequence alignment. The graph-theoretical approach comRNA computes stem similarity scores and uses a maximum clique finding algorithm to find pseudoknotted structures (Ji et al., 2004). SCARNA performs pairwise structural alignment of stem fragments with fixed lengths derived from the probability dot plot (Tabei et al., 2008).

In the following, a novel comparative approach for predicting structures including H-type pseudoknots called DotKnot-PW will be introduced. The input consists of two unaligned, evolutionarily related RNA sequences. Similarity scores between structure elements will be calculated. Statistically significant pairs will be used to find the set of conserved structure elements common to two sequences, which maximize a combined thermodynamic and similarity score. Using a hand-curated test set of pseudoknotted structures with experimental support, the prediction accuracy of DotKnot-PW will be compared with methods from the literature.

2 APPROACH

Pseudoknots are functional elements in RNA structures and therefore, the most promising approach for comparative prediction is a structure comparison with less focus on exact sequence matching. In fact, perfect conservation on the sequence level can be more of a curse than a blessing. Especially ncRNAs are known to evolve quickly and so-called consistent and compensatory base pairs in both sequences will give much more confidence for structure conservation than a sequence alignment. One strong point of the DotKnot method for single sequence pseudoknot prediction (Siperschneider and Datta, 2010; Siperschneider et al., 2011) is that the set of possible H-type pseudoknot candidates (and secondary structure elements) is explicitly computed and thus readily available for further investigation. The main steps in the pairwise pseudoknot prediction approach DotKnot-PW are as follows (Fig. 1):

1. Run DotKnot for two unaligned sequences Seq\(^x\) and Seq\(^y\). This returns secondary structure element and H-type pseudoknot candidate dictionaries.
2. Calculate pairwise base pair similarity scores for secondary structure elements and H-type pseudoknot candidates. Keep significant pairs that have a low estimated P-value.
3. Use significant pairs to calculate the set of conserved structure elements and pseudoknots for the two sequences that maximizes a combined free energy and similarity score.

The key point of the DotKnot-PW approach is how to score the similarity of stems, secondary structure elements and H-type pseudoknot candidates derived from sequences Seq\(^x\) and Seq\(^y\).

Related work has been done for stem finding in unaligned sequences, where stem candidates are assigned a matching score across unaligned sequences, e.g. in SCARNA. Another point is how to assess the significance of a similarity score using P-values. These points will be explained in detail in the following section.

3 MATERIALS AND METHODS

For two unaligned RNA sequences Seq\(^x\) and Seq\(^y\), the single sequence prediction method DotKnot (Siperschneider and Datta, 2010; Siperschneider et al., 2011) returns two stem dictionaries \(D(x)\) and \(D(y)\) derived from the probability dot plot. It also returns secondary structure element dictionaries \(D^s(x)\), \(D^s(y)\) and \(D^p(y)\) and H-type pseudoknot candidate dictionaries \(D_s(x)\) and \(D_s(y)\) (Fig. 1). To detect conserved structure elements for the two sequences, a pairwise structural comparison is performed. Instead of a full structure-to-structure alignment, which takes \(O(n^3)\) time and \(O(n^3)\) space, pairwise base pair similarity scores are calculated using the RIBOSUM85-60 matrix for base pair substitutions (Klein and Eddy, 2003).

3.1 Base pair similarity scores of stems

For two given stems \(s(x)\) and \(s(y)\) with fixed lengths in sequences Seq\(^x\) and Seq\(^y\), respectively, the base pair similarity score \(sim(s(x), s(y))\) is calculated using an unweighted local structure alignment of the base pairs with the RIBOSUM85-60 matrix. As an example, consider the following optimal unweighted local structure alignment of the two stems with base pair similarity score of \(sim(s(x), s(y)) = 22.04\) using the RIBOSUM85-60 matrix.

\[

t(x) = \text{UUCUAUC} \ldots \text{GAUGAGA} \\
s(x) = ((((((\ldots)))))) \\
\]

\[

t(y) = \text{--UUGUAC} \ldots \text{GUACAA}-- \\
s(y) = ((((((\ldots)))))) \\
\]

To evaluate the significance of base pair similarity scores instead of the raw score, one has to find out what the underlying probability distribution is. Similar to the case of unweighted local sequence alignments (Karlin and Altschul, 1990), it is assumed here that the base pair similarity scores follow an extreme value distribution. However, the main difference is that a comparison between fixed-length stem fragments is made. It is important to remember that parameters \(\lambda\) and \(K\) describe the extreme value distribution of optimal local alignment scores in the asymptotic limit of long sequences (Altschul et al., 2001). Here, the parameters for the generalized extreme value distribution are pre-calculated using maximum
likelihood of fitting to a distribution of the histogram of a large sample of random base pair similarity scores. The maximum likelihood fitting was performed using the \texttt{zscore} package of the R statistical language for a range of stem lengths (see Supplementary Material). The \( P \)-value is defined as the probability to obtain a score greater than or equal to the observed score strictly by chance. A stem \( s(x) \) in sequence Seq and a stem \( s(y) \) in sequence Seq’ are a significant pair if the score \( \text{sim} [s(x), s(y)] \) has an estimated \( P \)-value less than \( \alpha \). Stem pairs with a \( P \)-value larger than \( \alpha \) are not considered in the following.

### 3.2 Base pair similarity scores of interrupted stems

For two interrupted stems, the base pair similarity score is calculated by deleting bulges and internal loops and scoring stems as consecutive base pairs. Base pair similarity scores for regular and interrupted stems are also calculated if the difference in number of base pairs is less than 5. For example, a stem with one bulge might be a conserved match with a regular stem. A stem \( s(x) \) in sequence Seq and a stem \( s(y) \) in sequence Seq’ are a significant pair if the score \( \text{sim} [s(x), s(y)] \) has an estimated \( P \)-value less than \( \alpha \).

### 3.3 Base pair similarity scores of multiloops

Calculating the base pair similarity score for two multiloop structures is complex due to the variety of inner loop elements, which may be regular or interrupted stems. A multiloop \( S^M_i(x) \) can be decomposed into an outer stem \( s_i(x) \) and a set of inner structure elements \( S_i(x) = [s_i(x), \ldots, s_k(x)] \), where \( k \geq 2 \). The base pair similarity score \( \text{sim}[s_i(x), s_j(y)] \) for the outer stems of two multiloops can be easily obtained from the previously calculated base pair similarity scores. If the outer stem is a conserved match, a local alignment on the set of inner structure elements is used to find the base pair similarity score. Here, gaps are allowed in the local alignment of inner structure elements; however, no gap penalty is used. Let two sets of inner structure elements \( S_j(x) = [s_i(x), \ldots, s_k(x)] \) and \( S_j(y) = [s_i(y), \ldots, s_k(y)] \) be given. Let \( H(i, j) \) be the maximum similarity score between a suffix of \( S_j(x) \) and a suffix of \( S_j(y) \). The optimal local alignment is calculated as follows:

\[
H(i, 0) = 0, \quad 0 \leq i \leq n \\
H(0, j) = 0, \quad 0 \leq j \leq m \\
H(i, j) = \max \left\{ 0, H(i-1, j), H(i-1, j-1) + \text{sim}[s_i(x), s_j(y)] \right\}
\]

A multiloop \( S^M_j(x) \) in sequence Seq’ and a multiloop \( S^M_j(y) \) in sequence Seq’ are a significant pair if the similarity score \( \text{sim} [S^M_i(x), S^M_j(y)] \) has an estimated \( P \)-value less than \( \alpha \).

### 3.4 Base pair similarity scores of H-type pseudoknots

A H-type pseudoknot has two pseudoknot stems \( S_1 \) and \( S_2 \). The prerequisite for a conserved pseudoknot pair is that both core H-type pseudoknot stem pairs \([S_1(x), S_1(y)]\) and \([S_2(x), S_2(y)]\) are significant. The base pair similarity score for two H-type pseudoknots \( p(x) \) and \( p(y) \) in sequences Seq and Seq’, respectively, is the sum of base pair similarity scores for the core pseudoknot stems as well as the base pair similarity score from a gapped local alignment of the recursive secondary structure elements in the loops (as described for multiloops). A pseudoknot \( p(x) \) in sequence Seq and a pseudoknot \( p(y) \) in sequence Seq’ are a significant pseudoknot pair if the similarity score \( \text{sim}[p(x), p(y)] \) has an estimated \( P \)-value less than \( \alpha \).

### 3.5 Dissimilarity and weight of significant structure elements pairs

The base pair similarity score calculated in the previous sections might not be powerful enough to distinguish true positive conserved structure element pairs from false-positive structure element pairs due to the finite lengths of stems and exclusion of loop sequences in the alignment. Therefore, a dissimilarity score is also used to confirm whether a pair is significant. The dissimilarity for two given structure elements \( s(x) \) and \( s(y) \) in sequences Seq and Seq’ is defined as:

\[
\text{dissim}[s(x), s(y)] = \sum_{k=1}^{2} \text{dissim}_k [s(x), s(y)]
\]

where \( \text{dissim}_k \) is the difference in the stem lengths and \( \text{dissim}_2 \) is the difference in the number of loop lengths. As an example, consider the pseudoknot pair \( p_1(x) \) and \( p_1(y) \) in sequences Seq and Seq’, respectively, with stems \( S_1 \), \( S_2 \) and loops \( L_1 \), \( L_2 \), \( L_3 \). The pseudoknot pair has similarity of 6.

\[
p(x) \quad \text{UCUCACAGAAUGGAUGUCUUGCUGCUAUAAUAGAUAGAGAAGGUUAUAGCAG} \\
p(y) \quad \text{UCUCACAGAAUGGAUGUCUUGCUGCUAUAAUAGAUAGAGAAGGUUAUAGCAG}
\]

A weight is assigned to a significant pair, which is a combination of the free energy, covariation and dissimilarity. The overall weight \( s \) of a significant structure element pair \([s(x), s(y)]\) in sequences Seq and Seq’ is a combination of the free energy weights \( w[s(x)] \) and \( w[s(y)] \), base pair similarity score \( \text{sim}[s(x), s(y)] \) and dissimilarity \( \text{dissim}[s(x), s(y)] \):

\[
s[s(x), s(y)] = \alpha \times \text{sim}[s(x), s(y)] - \beta \times \{ w[s(x)] + w[s(y)] \} - \gamma \times \text{dissim}[s(x), s(y)]
\]

Only structure element pairs with positive score \( s \) are allowed in the following dynamic programming algorithm. Here, \( \alpha \) and \( \beta \) are set to 0.5 and \( \beta \) is set to 1.

### 3.6 Finding best set of significant structure elements

Let \( p_1, \ldots, p_n \) be the number of structure elements in the first sequence Seq’ and \( p_1', \ldots, p_n' \) be the number of structure elements in the second sequence Seq’. Each structure element has a left and right endpoint in the sequence and is a stem, interrupted stem, multiloop or H-type pseudoknot. Structure elements can also be represented as nodes in a graph. In each sequence, the structure elements are ordered by their right endpoints. An edge is drawn between two structure elements in the first and the second sequence if their base pair similarity score has a \( P \)-value less than \( \alpha \). Given the set of edges between nodes \( p_1, \ldots, p_n \) and \( p_1', \ldots, p_n' \), the goal is to find the set of edges with maximum weight that are non-crossing. This relates to finding the set of non-overlapping structure elements in the two sequences that maximize the score under the requirement that the interval ordering is preserved. A set of structure elements in the first and second sequence, which preserves the interval ordering is called a feasible structure element alignment and must satisfy the following two requirements. Each structure element can be aligned with at most one other structure element in the other sequence. The order of structure elements must be preserved with respect to the alignment. That is, if structure elements \( p_i \) and \( p'_i \) in the first sequence are aligned with \( p_j \) and \( p'_j \) in the second sequence, respectively, the pairs may never overlap: \( p'_i < p_i \wedge p'_j < p_j \) (Fig. 2).

Given nodes \( p_1, \ldots, p_n \) in the first sequence Seq’ and \( p_1', \ldots, p_n' \) in the second sequence Seq’ let \( f(i, a) \) be the maximum sum of edge weights for nodes between 1 and \( i \) in the first sequence and 1 and \( a \) in the second sequence such that the edges are non-crossing (\( i \leq n \) and \( a \leq m \)). The nodes that maximize the sum of edge weights are called an optimal
Predicting pseudoknotted structures

Fig. 2. A set of edges with positive scores is given between nodes \(p_1, \ldots, p_7\) in the first sequence and \(p_1, \ldots, p_6\) in the second sequence. The goal is to find the best set of non-overlapping structure elements in the two sequences such that the interval ordering is preserved. The optimal structure element alignment, which preserves the interval ordering includes structure elements \(p_1, p_4, p_7\) in the first sequence and \(p_1, p_4, p_6\) in the second sequence.

For two unaligned RNA sequences \(Seq^x\) and \(Seq^y\), the single sequence prediction method DotKnot returns structure element dictionaries derived from the probability dot plot. Let \(n\) and \(m\) be the number of structure elements in sequences \(Seq^x\) and \(Seq^y\), respectively. Calculating the similarity scores and the optimal structure element alignment takes \(O(nm)\) time. Furthermore, nested structures are taken into account for significant outer stem pairs, which have estimated \(P\)-value less than \(\alpha\). For each significant outer hairpin loop pair, the optimal structure element alignment of inner elements is computed.

### 3.7 Time requirements

For two unaligned RNA sequences \(Seq^x\) and \(Seq^y\), the single sequence prediction method DotKnot returns structure element dictionaries derived from the probability dot plot. Let \(n\) and \(m\) be the number of structure elements in sequences \(Seq^x\) and \(Seq^y\), respectively. Calculating the similarity scores and the optimal structure element alignment takes \(O(nm)\) time. Furthermore, nested structures are taken into account for significant outer stem pairs, which have estimated \(P\)-value less than \(\alpha\). Let \(a\) be the number of significant stem pairs, where both stems are hairpin loops. For each significant outer hairpin loop pair, the optimal structure alignment of inner elements is computed. In the worst case, this increases time requirements to \(O(a \times nm)\). The number of structure elements depends on the base composition of the sequence. Empirically, \(n\) and \(m\) can be observed to grow linearly with the length of the sequence for uniform base distribution (see Supplementary Material). In practice, DotKnot-PW can be expected to run in the order of minutes for sequences shorter than 500 nt.

### 4 RESULTS

Many pseudoknot prediction programs have been evaluated using all the entries in the PseudoBase database (van Batenburg et al., 2000). There are several caveats in this approach. First, the sequences given in PseudoBase are those which exactly harbor the pseudoknot. However, in practice structure prediction algorithms will be applied to longer sequences without prior knowledge of the pseudoknot location. Second, long-range pseudoknot entries appear in a truncated version in the database. Third, some classes of pseudoknots have a large number of entries (such as short H-type pseudoknots in the 3'-untranslated regions of plant viruses), whereas more complex types of pseudoknots only have one representative (such as long-range rRNA pseudoknots). Therefore, a hand-curated dataset of pseudoknot structures will be used here.

When it comes to pseudoknots, many structures have been published based on a secondary structure predicted by free energy minimization. These predicted secondary structures are used as a working model and refined using experimental techniques such as chemical and enzymatic probing. However, the native structure remains unsolved unless tertiary structure determination methods such as X-ray crystallography are used. Testing structures that are based on computer predictions with no experimental support creates a bias in the benchmark and will be avoided in this evaluation.

A total of 16 pseudoknotted reference structures from different RNA types were collected, which have strong experimental support. For each reference structure, a supporting set of 10 evolutionarily related sequences was obtained from the RFAM database (Gardner et al., 2010). Note that for the vast majority of supporting sequences, no experimentally determined structures are available. The average pairwise sequence identities vary from 55% to 99%. Given a reference structure, the performance of prediction algorithms is evaluated in terms of sensitivity (S), i.e. the percentage of base pairs in the reference structure, which are predicted correctly, as well as positive predictive value (PPV), i.e. the percentage of predicted pairs, which are in the reference structure. The Matthews correlation coefficient (MCC) is also reported and is in the range from \(-1\) to \(1\), where \(1\) corresponds to a perfect prediction and \(-1\) to a prediction that is in total disagreement with the reference structure. The performance of each method for predicting the reference structure was evaluated as described in Gardner and Giegerich (2004).

DotKnot-PW was compared with methods that are freely available and use standard input and output formats. The comparative methods are CARNAC, Tfold and hxmatch (with the -A option using RNAalifold). All of these methods return structure predictions for only the reference sequence with regards to the support set of evolutionarily related sequences. Tfold and hxmatch take a sequence alignment as the input. ClustalW with the default parameters was used to produce the initial sequence alignment. DotKnot-PW and CARNAC take a set of unaligned sequences as the input. Furthermore, prediction results for the reference sequence (not the supporting sequences) were obtained from the single sequence methods DotKnot (Sperschneider and Datta, 2010; Sperschneider et al., 2011), ProbKnot (Bellaousov and Mathews, 2010), IPKnot (Sato et al., 2011) and RNAfold (Hofacker et al., 1994). Note that all methods except CARNAC and RNAfold allow pseudoknot prediction.

The results are shown in Table 1. DotKnot-PW has the highest average MCC of 0.75 for the test sequences. For each reference structure with the 10 support sequences from the corresponding RFAM family, 10 predictions are returned ordered by the combined free energy and similarity score. If only the
<table>
<thead>
<tr>
<th>Type</th>
<th>ID</th>
<th>Info</th>
<th>Family</th>
<th>Prediction of common structure</th>
<th>Reference structure prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DotKnot-PW, First (Average)</td>
<td>CarNAc</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CARNAC</td>
<td>Tfold</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hxmatch</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DotKnot</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ProbKnot</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IPknot</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RNA fold</td>
</tr>
<tr>
<td>Frameshift</td>
<td>SARS-CoV (Plant et al., 2005)</td>
<td>NMR</td>
<td>82 nt</td>
<td>92.3 (79.2)</td>
<td>38.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 PK</td>
<td>62% MCC</td>
<td>0.96 (0.84)</td>
</tr>
<tr>
<td></td>
<td>VMV (Pennell et al., 2008)</td>
<td>NMR</td>
<td>68 nt</td>
<td>100 (100)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 PK</td>
<td>89% MCC</td>
<td>0</td>
</tr>
<tr>
<td>Ribozyme</td>
<td>HDV (Ferre-D’Amare et al., 1998)</td>
<td>X-ray</td>
<td>87 nt</td>
<td>93.8 (90.7)</td>
<td>21.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 PK</td>
<td>74% MCC</td>
<td>0.97 (0.94)</td>
</tr>
<tr>
<td></td>
<td>glmS-Ba (Klein and Ferre-D’Amare, 2006)</td>
<td>SC, MG</td>
<td>151 nt</td>
<td>76.4 (71.7)</td>
<td>7.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 PKs</td>
<td>55% MCC</td>
<td>0.85 (0.81)</td>
</tr>
<tr>
<td>Untranslated regions</td>
<td>BCV (Williams et al., 1999)</td>
<td>SP, MG</td>
<td>57 nt</td>
<td>55.3 (51)</td>
<td>26.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 PK</td>
<td>78% MCC</td>
<td>1 (0.96)</td>
</tr>
<tr>
<td></td>
<td>BaMV (Lin et al., 2007)</td>
<td>SP, MG</td>
<td>170 nt</td>
<td>59.5 (53.8)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 PK</td>
<td>99% MCC</td>
<td>0.64 (0.58)</td>
</tr>
<tr>
<td>Telomerase</td>
<td>The-telo (Theimer and Feigon, 2006)</td>
<td>SP</td>
<td>159 nt</td>
<td>81.6 (61.1)</td>
<td>68.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 PK</td>
<td>72% MCC</td>
<td>0.86 (0.65)</td>
</tr>
<tr>
<td>IRES</td>
<td>PSIV (Pfingsten et al., 2006)</td>
<td>X-ray</td>
<td>194 nt</td>
<td>65.5 (63.8)</td>
<td>39.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 PKs</td>
<td>57% MCC</td>
<td>0.66 (0.68)</td>
</tr>
<tr>
<td></td>
<td>CSFV (Kolupaeva et al., 2000)</td>
<td>SP, MG</td>
<td>244 nt</td>
<td>39 (44.9)</td>
<td>45.1</td>
</tr>
</tbody>
</table>

(continued)
As an example, consider the S15 mRNA pseudoknot that binds to specific proteins in the autoregulation mechanism of ribosomal protein S15 synthesis (Philippe et al., 1995). For the reference sequence S15 and 10 support sequences from the corresponding RFAM family, DotKnot-PW returns pairwise predictions ordered by the combined free energy and similarity score. The top two pairwise predictions with the highest scores are shown in Figure 3.

5 DISCUSSION

We presented DotKnot-PW for prediction of structures common to two RNA sequences, including H-type pseudoknots. Both DotKnot and DotKnot-PW have been designed as dedicated pseudoknot prediction tools. In the following, important aspects of pseudoknot prediction will be discussed.

5.1 The underlying folding model

Single sequence prediction methods are always limited by the underlying RNA folding model. This may be the set of free energy parameters used by free energy minimization methods or the underlying methodological framework such as maximum expected accuracy methods. DotKnot-PW shows excellent results on H-type pseudoknots with short interhelix loops. For this type of pseudoknots, DotKnot-PW uses free energy pseudoknot parameters by Cao and Chen (2006, 2009) based on polymer statistical mechanics. Improvements of the accuracy of free energy parameters, both for secondary structures and pseudoknots, will lead to more accurate prediction methods. However, one has to keep in mind that the algorithms themselves must be designed in such a fashion that novel parameters can be efficiently incorporated. The heuristic framework of DotKnot-PW has been designed such that it can incorporate sophisticated free energy parameters for pseudoknots, secondary structures and coaxial stacking. In the future, DotKnot-PW could also use contributions from basic tertiary structure elements such as base triples around the pseudoknot junction or stem-loop interactions.

5.2 The type of pseudoknot

Pseudoknot prediction algorithms come in two flavors: either they can predict a certain, restricted class of pseudoknots or they do not have a restriction on the type of pseudoknot that can be predicted. For methods using free energy parameters, the inclusion of general types of pseudoknots might be more of a
Table 1, H-type pseudoknots common to two sequences. As shown in designed as an algorithm for finding the structure including 3064 DotKnot-PW structure elements. This approach called was presented based on the pairwise comparison of DotKnot. Here, an extension of the single sequence prediction method 5.4 Gaining confidence with multiple sequences for RNA families when compared with the single sequence predic- 6C O N C L U S I O N tion tool and should be applied to RNA sequences where pseudoknotted interactions are suspected in the structures. Prediction accuracy will inevitably decrease for sequences, which are longer than say 400 nt for any single sequence structure prediction method (Reeder et al., 2006). To achieve reliable results, short sequences should be folded using DotKnot and predictions should be compared with results from other methods from the literature. To gain confidence in predictions, subsequent comparative prediction using DotKnot-PW and other comparative methods is highly recommended. Ideally, experimental veri- fication of computationally predicted pseudoknots should be sought.

ACKNOWLEDGEMENTS

We thank the reviewers for their time and thoughtful comments that helped to considerably improve the quality of the manuscript.

Funding: International Postgraduate Research Scholarship (IPRS) from the Australian Government (to J.S.).

Conflict of Interest: none declared.

REFERENCES


Predicting pseudoknotted structures