BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 23 2012, pages 3058–3065
doi:10.1093/bioinformatics/bts575

Structural bioinformatics

Advance Access publication October 8, 2012

Predicting pseudoknotted structures across two RNA sequences

Jana Sperschneider[1,*], Amitava Datta[1] and Michael J. Wise[1,2]

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

Associate Editor: Ivo Hofacker

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.

Availability: `DotKnot-PW` and the RNA structure test set are available at the web site http://dotknot.csse.uwa.edu.au/pw.

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 players in the cell until RNA entered the spotlight. For example, RNA structures have been found to be catalytically active, which was assumed to be the privilege of proteins. Furthermore, small RNAs are known to regulate gene expression and RNA viruses employ a plethora of structure elements to invade the host cell.

To gain insight into macromolecule function, one must investigate the structure. The first step in RNA folding is stable base pairing that leads to a secondary structure. As RNA structure formation is of hierarchical nature, secondary structure is the basis for the tertiary fold that produces the functional structure. Especially for RNAs, structure determination by experimental means is an intricate and expensive task. Computational RNA structure prediction is therefore an invaluable tool for biologists. Comparative structure prediction is considered the most reliable approach for computational RNA structure prediction. Single sequence structure prediction is always limited by the accuracy of the underlying folding model. Three main streams have been identified for comparative RNA secondary structure prediction: (i) predict a structure from a pre-computed sequence alignment; (ii) simultaneously compute an alignment and a structure and (iii) alignment-free methods (Gardner and Giegerich, 2004).

Tools for multiple sequence alignments such as `ClustalW` (Thompson *et al*., 1994) are readily available and thus, structure prediction from an alignment is a tempting approach [e.g. `RNAalifold` (Hofacker *et al*., 2002)]. Such methods heavily depend on the sequence conservation and quality of the underlying alignment. However, ncRNAs are conserved rather on the structure level than on the sequence level. The gold standard of RNA comparative structure prediction is the Sankoff approach as it does not rely on a high-quality sequence alignment and captures the structural conservation of ncRNAs. Sankoff (1985) introduced a theoretical dynamic programming algorithm for simultaneous folding and aligning for a set of N sequences that takes $O(n^{3N})$ time and $O(n^{2N})$ space. Practical variants have been derived which more or less retain the Sankoff principle by sacrificing optimality. Alignment-free methods aim to avoid the pragmatic restrictions made in a practical Sankoff approach as well as the reliance on a high-quality alignment [e.g. `CARNAC` (Perriquet *et al*., 2003)]. Note that all of these comparative structure prediction methods exclude the prediction of RNA pseudoknots.

RNA pseudoknots are crossing structure elements with diverse functions. The principle of pseudoknot formation is that bases within a loop region pair with complementary unpaired bases outside the loop. From an algorithmic point of view, even the simplest type of pseudoknot adds considerable computational demands due to crossing base pairs. In fact, the majority of comparative RNA structure prediction methods exclude pseudoknots. 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 (Brierley *et al*., 2007).

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 `hxmatch`, 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). Simulfold 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, covariation 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 (Sperschneider and Datta, 2010; Sperschneider *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.

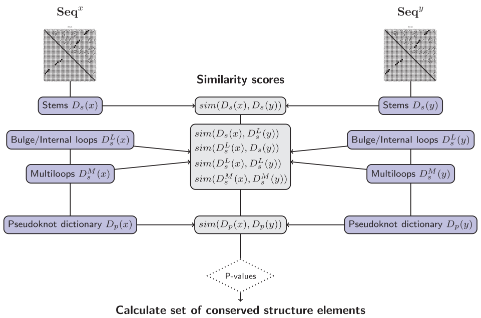

Seq^x
Seq^y
Similarity scores
Stems $D_s(x)$
$sim(D_s(x), D_s(y))$
Stems $D_s(y)$
Bulge/Internal loops $D_s^L(x)$
$sim(D_s(x), D_s^L(y))$
$sim(D_s^L(x), D_s(y))$
$sim(D_s^L(x), D_s^L(y))$
$sim(D_s^M(x), D_s^M(y))$
Bulge/Internal loops $D_s^L(y)$
Multiloops $D_s^M(x)$
Multiloops $D_s^M(y)$
Pseudoknot dictionary $D_p(x)$
$sim(D_p(x), D_p(y))$
Pseudoknot dictionary $D_p(y)$
P-values
Calculate set of conserved structure elements

Fig. 1. For two unaligned RNA sequences Seq^x and Seq^y, DotKnot-PW produces structure element dictionaries derived from the probability dot plot. Similarity scores and *P*-values are computed to detect conserved elements

3 MATERIALS AND METHODS

For two unaligned RNA sequences Seq^x and Seq^y, the single sequence prediction method DotKnot (Sperschneider and Datta, 2010; Sperschneider *et al.*, 2011) returns two stem dictionaries $D_s(x)$ and $D_s(y)$ derived from the probability dot plot. It also returns secondary structure element dictionaries $D_s^L(x), D_s^L(y)$ and $D_s^M(x), D_s^M(y)$ and H-type pseudoknot candidate dictionaries $D_p(x)$ and $D_p(y)$ (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 $O(n^4)$ 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 $s_i(x)$ and $s_j(y)$ with fixed lengths in sequences Seq^x and Seq^y, respectively, the base pair similarity score $sim[s_i(x), s_j(y)]$ is calculated using an ungapped local structure alignment of the base pairs with the RIBOSUM85-60 matrix. As an example, consider the following optimal ungapped local structure alignment of the two stems with base pair similarity score of $sim[s_1(x), s_2(y)] = 22.04$ using the RIBOSUM85-60 matrix.

$s_1(x)$

```
UCUCUAUC.......GAUAGAGA
((((((((...... .))))))))
```

$s_1(y)$

```
--UUGUAC.......GUACAA--
--((((((.......))))))--
```

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 ungapped 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 λ 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 fitting of a distribution to the histogram of a large sample of random base pair similarity scores. The maximum likelihood fitting was performed using the `ismev` 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 $s_i(x)$ in sequence Seq^x and a stem $s_j(x)$ in sequence Seq^y are a significant pair if the score $sim[s_i(x), s_j(y)]$ has an estimated *P*-value less than α. Stem pairs with a *P*-value larger than α 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 $s_i(x)$ in sequence Seq^x and a stem $s_j(y)$ in sequence Seq^y are a significant pair if the score $sim[s_i(x), s_j(y)]$ has an estimated *P*-value less than α.

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 $s_i^M(x)$ can be decomposed into an outer stem $s_i^o(x)$ and a set of inner structure elements $S_i(x) = [s_1(x),\ldots, s_k(x)]$, where $k \geq 2$. The base pair similarity score $sim[s_i^o(x), s_j^o(y)]$ for the outer stems of two multi-loops 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 $S_i(x) = [s_1(x),\ldots, s_n(x)]$ and $S_j(y) = [s_1(y),\ldots, s_m(y)]$ be given. Let $H(i, j)$ be the maximum similarity score between a suffix of $S_i(x)$ and a suffix of $S_j(y)$. The optimal local alignment is calculated as follows:

$$H(i,\ 0) = 0,\ 0 \leq i \leq n$$

$$H(0,\ j) = 0,\ 0 \leq j \leq m$$

$$H(i,j) = \max \begin{cases} 0 \\ H(i-1,j) \\ H(i-1,j-1) + sim[s_i(x), s_j(y)] \\ H(i,j-1) \end{cases}$$

A multiloop $s_i^M(x)$ in sequence Seq^x and a multiloop $s_j^M(y)$ in sequence Seq^y are a significant pair if the similarity score $sim[s_i^M(x), s_j^M(y)]$ has an estimated *P*-value less than α.

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 $p_i(x)$ and $p_j(y)$ in sequences Seq^x and Seq^y, 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 $p_i(x)$ in sequence Seq^x and a pseudoknot $p_j(y)$ in sequence Seq^y are a significant pseudoknot pair if the similarity score $sim[p_i(x), p_j(y)]$ has an estimated *P*-value less than α.

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 $s_i(x)$ and $s_j(y)$ in sequences Seq^x and Seq^y is defined as:

$$dissim[s_i(x), s_j(y)] = \sum_{k=1,\ldots,2} dissim_k[s_i(x), s_j(y)]$$

where $dissim_1$ is the difference in the stem lengths and $dissim_2$ 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^x and Seq^y, respectively, with stems S_1, S_2 and loops L_1, L_2, $L3$. The pseudoknot pair has dissimilarity of 6.

$p_1(x)$

```
UCUCUAUCAGAAUGGAUGUCUUGCUGCUAUAAUAGAUAGAGAAGGUUAUAGCAG
((((((((................[[[[[[[[[[.)))))))).]]]]]]]]]]
```

$p_1(y)$

```
UUGUACAGAAUGGUAAGCCAAGUGUCAAUAGGAGGUACAAGCAACCUAUUGCAU
((((((................[[[.[[[[[[[.))))))....]]]]]]]]]]
```

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 $[s_i(x), s_j(y)]$ in sequences Seq^x and Seq^y is a combination of the free energy weights $w[s_i(x)]$ and $w[s_j(y)]$, base pair similarity score $sim[s_i(x), s_j(y)]$ and dissimilarity $dissim[s_i(x), s_j(y)]$:

$$\begin{aligned} s[s_i(x), s_j(y)] &= \alpha \times sim[s_i(x), s_j(y)] - \beta \times \{w[s_i(x)] + w[s_j(y)]\} \\ &\quad -\gamma \times dissim[s_i(x), s_j(y)] \end{aligned}$$

Only structure element pairs with positive score s are allowed in the following dynamic programming algorithm. Here, α and γ are set to 0.5 and β is set to 1.

3.6 Finding best set of significant structure elements

Let $p_1^x, \ldots, p_n^x$ be the number of structure elements in the first sequence Seq^x and $p_1^y, \ldots, p_m^y$ be the number of structure elements in the second sequence Seq^y. 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 α. Given the set of edges between nodes $p_1^x, \ldots, p_n^x$ and $p_1^y, \ldots, p_m^y$, 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^x and p_j^x in the first sequence are aligned with p_a^y and p_b^y in the second sequence, respectively, the pairs may never overlap: $p_i^x < p_j^x \wedge p_a^x < p_b^x$ (Fig. 2).

Given nodes $p_1^x \ldots, p_n^x$ in the first sequence Seq^x and $p_1^y, \ldots, p_m^y$ in the second sequence Seq^y, 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

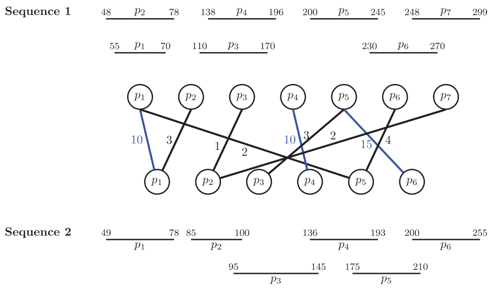


Fig. 2. A set of edges with positive scores is given between nodes $p_1,\ldots,p_7$ in the first sequence and $p_1,\ldots,p_6$ in the second sequence. The goal is to find the best set of non-overlapping structure elements in the two sequences such that the interval ordering is preserved. The optimal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

structure element alignment for the two sequences. The optimal structure element alignment is calculated using dynamic programming. For a given structure element p_i with start point a_i and end point b_i, let pre(i) be the non-overlapping predecessor. For each structure element, its predecessor is pre-computed using the sorted list of structure elements. The recursion for calculating the optimal structure element alignment is as follows:

$$f(0,a)=0$$

$$f(i,0)=0$$

$$f(i,a)=\max\begin{cases} f(i-1,a) \\ s(i,a)+f[\mathrm{pre}(i),\mathrm{pre}(a)] \\ f(i,a-1) \end{cases}$$

Furthermore, nested structures are taken into account for significant outer stem pairs, which have estimated P-value less than α. 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 α. 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 structure 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 1. Prediction results using a test set of different RNA classes

Type	ID	Info	Family		Prediction of common structure				Reference structure prediction			
					DotKnot-PW, First (Average)	CARNAC	Tfold	hxmatch	DotKnot	ProbKnot	IPknot	RNA fold
Frameshift	SARS-CoV (Plant *et al*., 2005)	NMR	RF00507	*S*	92.3 (79.2)	38.5	57.7	38.5	92.3	69.2	73.1	73.1
		82 nt	82 nt	PPV	100 (90)	90.9	93.8	58.8	100	72	79.2	73.1
		1 PK	62%	MCC	**0.96** (0.84)	0.59	0.73	0.47	**0.96**	0.7	0.76	0.73
	VMV (Pennell *et al*., 2008)	NMR	RF01840	*S*	100 (100)	0	50	42.9	100	50	50	50
		68 nt	55 nt	PPV	100 (100)	0	50	54.5	100	41.2	41.2	41.2
		1 PK	89%	MCC	**1** (**1**)	0	0.49	0.48	**1**	0.44	0.44	0.44
Ribozyme	HDV (Ferre-D'Amare *et al*., 1998)	X-ray	RF00094	*S*	93.8 (90.7)	21.9	21.9	12.5	93.8	40.6	62.5	37.5
		87 nt	90 nt	PPV	100 (97.1)	100	30.4	57.1	100	48.1	80	42.9
		1 PK	74%	MCC	**0.97** (0.94)	0.46	0.24	0.26	**0.97**	0.43	0.7	0.39
	glmS-Ba (Klein and Ferre-D'Amare, 2006)	SC, MG	RF00234	*S*	76.4 (71.7)	7.3	50.9	43.6	63.6	76.4	72.7	69.1
		151 nt	178 nt	PPV	95.5 (92.9)	44.4	87.5	80	67.3	82.4	85.1	82.6
		2 PKs	55%	MCC	**0.85** (0.81)	0.18	0.66	0.59	0.65	0.79	0.78	0.75
	EC-RNaseP (Harris *et al*., 2001)	X-ray	RF00010	*S*	55.3 (51)	26.8	29.3	46.3	53.7	74	69.9	60.2
		377 nt	380 nt	PPV	68.7 (64.4)	86.8	65.5	83.8	56.4	77.8	86	64.9
		2 PKs	64%	MCC	0.61 (0.57)	0.48	0.44	0.62	0.55	0.76	**0.77**	0.62
Untranslated regions	BCV (Williams *et al*., 1999)	SP, MG	RF00165	*S*	100 (95.6)	55.6	61.1	0	100	55.6	94.4	55.6
		63 nt	63 nt	PPV	100 (96.7)	100	84.6	0	81.8	66.7	100	66.7
		1 PK	78%	MCC	**1** (0.96)	0.74	0.71	-0.01	0.9	0.6	0.97	0.6
	BaMV (Lin *et al*., 2007)	SP, MG	RF00290	*S*	59.5 (53.8)	0	*	50	40.5	64.3	50	40.5
		170 nt	145 nt	PPV	69.4 (62.4)	0	*	65.6	41.5	69.2	67.7	45.9
		1 PK	99%	MCC	0.64 (0.58)	0	*	0.57	0.4	**0.66**	0.58	0.43
	HPeV1 (Nateri *et al*., 2002)	MG	RF00499	*S*	95.3 (86)	11.6	37.2	79.1	83.7	88.4	86	79.1
		116 nt	112 nt	PPV	95.3 (88.2)	100	100	91.9	83.7	92.7	97.4	94.4
		1 PK	88%	MCC	**0.95** (0.87)	0.34	0.61	0.85	0.83	0.9	0.91	0.86
Telomerase	Tthe-telo (Theimer and Feigon, 2006)	SP	RF00025	*S*	81.6 (61.1)	68.4	60.5	57.9	73.7	65.8	71.1	60.5
		159 nt	160 nt	PPV	91.2 (70.7)	86.7	79.3	84.6	70	55.6	67.5	54.8
		1 PK	72%	MCC	**0.86** (0.65)	0.77	0.69	0.7	0.72	0.6	0.69	0.57
Riboswitch	preQ1 (Rieder *et al*., 2010)	NMR	RF00522	*S*	100 (73.4)	55.6	*	55.6	55.6	55.6	55.6	55.6
		34 nt	44 nt	PPV	100 (100)	100	*	100	100	55.6	100	71.4
		1 PK	80%	MCC	**1** (0.84)	0.74	*	0.74	0.74	0.53	0.74	0.61
	SAM-I (Montange and Batey, 2006)	X-ray	RF00162	*S*	73.5 (53.8)	55.9	55.9	64.7	47.1	73.5	64.7	64.7
		94 nt	102 nt	PPV	100 (83.7)	79.2	100	91.7	57.1	100	100	88
		1 PK	73%	MCC	**0.85** (0.66)	0.66	0.74	0.77	0.51	**0.85**	0.8	0.75
IRES	PSIV (Pfingsten *et al*., 2006)	X-ray	RF00458	*S*	65.5 (63.8)	39.7	94.8	77.6	70.7	70.7	63.8	72.4
		194 nt	198 nt	PPV	66.7 (73)	100	96.5	91.8	69.5	67.2	78.7	71.2
		2 PKs	57%	MCC	0.66 (0.68)	0.63	**0.96**	0.84	0.7	0.69	0.71	0.72
	CSFV (Kolupaeva *et al*., 2000)	SP, MG	RF00209	*S*	39 (44.9)	45.1	68.3	74.4	62.2	75.6	65.9	68.3
		244 nt	272 nt	PPV	42.1 (53.1)	97.4	80	85.9	68	80.5	68.4	71.8
		1 PK	83%	MCC	0.4 (0.49)	0.66	0.74	**0.8**	0.65	0.78	0.67	0.7

(continued)

Table 1. Continued

Type	ID	Info	Family		Prediction of common structure				Reference structure prediction			
					DotKnot-PW, First (Average)	CARNAC	Tfold	hxmatch	DotKnot	ProbKnot	IPknot	RNA fold
mRNA	S15 (Philippe *et al.*, 1995)	SP, MG	RF00114	*S*	100 (81.8)	52.9	52.9	41.2	100	58.8	58.8	58.8
		74 nt	112 nt	PPV	100 (83.4)	52.9	40.9	46.7	100	47.6	52.6	52.6
		1 PK	77%	MCC	**1** (0.82)	0.52	0.46	0.43	**1**	0.52	0.55	0.55
	repZ (Asano and Mizobuchi, 1998)	SP, MG	RF01087	*S*	73.8 (71.4)	9.5	31	83.3	71.4	57.1	66.7	57.1
		149 nt	149 nt	PPV	86.1 (80.2)	44.4	61.9	85.4	85.7	70.6	82.4	64.9
		1 PK	90%	MCC	0.8 (0.76)	0.2	0.43	**0.84**	0.78	0.63	0.74	0.6
tmRNA	Ec-tmRNA (Nameki *et al.*, 1999)	NMR, SC	RF00023	*S*	37.5 (50)	6.7	34.6	27.9	75	56.7	66.3	50
		363 nt	377 nt	PPV	45.9 (58.4)	100	67.9	61.7	79.6	60.8	76.7	47.7
		4 PKs	57%	MCC	0.41 (0.54)	0.26	0.48	0.41	**0.77**	0.59	0.71	0.49
	Average			*S*	**77.7** (70.5)	31	50.4	49.7	74	64.5	67	59.5
	Average			PPV	**85.1** (80.9)	73.9	74.2	71.2	78.8	68	78.9	64.6
	Average			MCC	**0.81** (0.75)	0.45	0.6	0.59	0.76	0.66	0.72	0.61

Note: Each reference structure is given by its ID (see Supplementary Material for dot-bracket notation). The following column gives the method of experimental support (NMR, NMR spectroscopy; X-ray, X-ray crystallography; SC, sequence comparison; MG, mutagenesis; SP, structure probing), length of the sequence and number of pseudoknots. For each reference structure, the corresponding RFAM family ID, average sequence length and average pairwise sequence identity is shown. The * symbol means that the method failed to run. The 'first' prediction for DotKnot-PW is the pairwise prediction with highest combined free energy and similarity score.

pairwise prediction with highest combined free energy and similarity score is taken, `DotKnot-PW` has an improved average MCC of 0.81. `Tfold` and `hxmatch` have average MCC of 0.6 and 0.59, respectively. `CARNAC` has average MCC of 0.45 with much higher average specificity than sensitivity.

The prediction results for single sequence structure prediction for each of the reference sequences with experimentally determined structures are also shown in Table 1. Note that this does not include the prediction for the support sequences from `RFAM`, as no experimentally determined structures are available. All single sequence pseudoknot prediction methods show improved results over using `RNAfold`. `DotKnot` has the highest average MCC of 0.76, followed by `IPKnot` and `ProbKnot`.

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

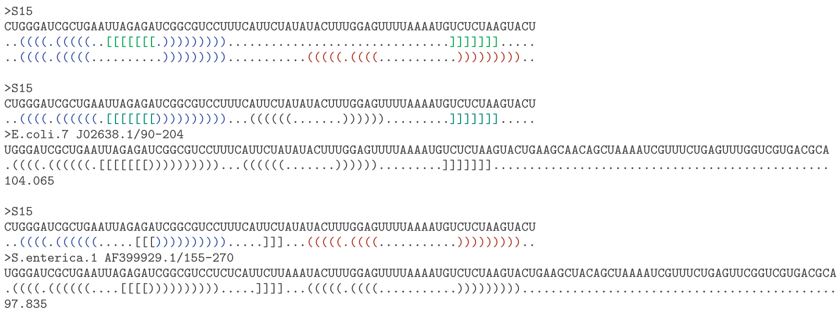
We presented `DotKnot-PW` for prediction of structures common to two RNA sequences, including H-type pseudoknots. Both `DotKnot` and `DotKnot-PW` have been designed as dedicated pseudoknot prediction tools. In the following, important aspects of pseudoknot prediction will be discussed.

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



```
>S15
CUGGGAUCGCUGAAUUAGAGAUCGGCGUCCUUUCAUUCUAUAUACUUUGGAGUUUUAAAAUGUCUCUAAGUACU
..((((.(((((..[[[[[[[.))))))))).......................]]]]]]].....
..((((.(((((..........)))))))))...........(((((.((((...........)))))))))..

>S15
CUGGGAUCGCUGAAUUAGAGAUCGGCGUCCUUUCAUUCUAUAUACUUUGGAGUUUUAAAAUGUCUCUAAGUACU
..((((.((((((.[[[[[[[))))))))))...((((((.......)))))).........]]]]]]].....
>E.coli.7 J02638.1/90-204
UGGGAUCGCUGAAUUAGAGAUCGGCGUCCUUUCAUUCUAUAUACUUUGGAGUUUUAAAAUGUCUCUAAGUACUGAAGCAACAGCUAAAAUCGUUUCUGAGUUUGGUCGUGACGCA
.((((.((((((.[[[[[[[))))))))))...((((((.......)))))).........]]]]]]]...............................................
104.065

>S15
CUGGGAUCGCUGAAUUAGAGAUCGGCGUCCUUUCAUUCUAUAUACUUUGGAGUUUUAAAAUGUCUCUAAGUACU
..((((.((((((.....[[[)))))))))).....]]]...(((((.((((...........)))))))))..
>S.enterica.1 AF399929.1/155-270
UGGGAUCGCUGAAUUAGAGAUCGGCGUCCUCUCAUUCUUAAAUACUUUGGAGUUUUAAAAUGUCUCUAAGUACUGAAGCUACAGCUAAAAUCGUUUCUGAGUUCGGUCGUGACGCA
.((((.((((((....[[[[)))))))))).....]]]]...(((((.((((...........)))))))))............................................
97.835
```

Fig. 3. Pairwise prediction results for the S15 mRNA pseudoknot (RFAM family RF00114) with the top two combined free energy and similarity scores. The reference structure is shown at the top and folds into two conformations in dynamic equilibrium: a H-type pseudoknot or a series of hairpins. For the pairwise prediction with highest score, the pseudoknot structure is returned. For the second-best pairwise prediction, the alternative hairpin loop structure is returned

curse than a blessing, as no reliable free energy parameters for complex pseudoknots are available. DotKnot-PW has restrictions on the type of pseudoknot that can be predicted. However, this does not always lead to poor prediction results in practice. For example, DotKnot-PW shows the best result for the HDV ribozyme, which is a complex double nested pseudoknot.

5.3 The trouble with benchmarking

The results from the benchmark for structure prediction in Table 1 must be interpreted with care. First, the tested methods can be run with different parameters, possibly producing better results. However, as a typical user has no prior knowledge about the structure, the default parameters for each method are used. Of course, a comprehensive benchmark should include a larger number of structures to obtain a more reliable evaluation. However, in this study, the focus has been on a test set where the structures are supported experimentally. Many structures have been published, which were determined using computational tools and this will inevitably create a bias in a benchmark, and thus they were excluded here.

5.4 Gaining confidence with multiple sequences

Here, an extension of the single sequence prediction method DotKnot was presented based on the pairwise comparison of structure elements. This approach called DotKnot-PW is designed as an algorithm for finding the structure including H-type pseudoknots common to two sequences. As shown in Table 1, DotKnot-PW can greatly improve structure predictions for RNA families when compared with the single sequence prediction using DotKnot. In some cases, a comparative approach might have lower sensitivity than a single sequence prediction; however, this should not generally be judged as ‘inferior’. For example, ncRNAs might preserve some integral base pairs throughout evolution and only these will be detected by a comparative approach, which returns the set of base pairs common to a set of evolutionarily related sequences. DotKnot-PW uses a set of unaligned sequences as the input; therefore, no expert user intervention is required. In the future, DotKnot-PW will be extended to include intramolecular kissing hairpins. Furthermore, constrained folding will be implemented to predict a structure subject to constraints, e.g. enforce certain base pairs or regions, which must remain unpaired.

6 CONCLUSION

DotKnot-PW has been designed as a dedicated pseudoknot prediction 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 verification 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

Altschul,S.F. *et al.* (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.*, **29**, 351–361.

Asano,K. and Mizobuchi,K. (1998) An RNA pseudoknot as the molecular switch for translation of the repZ gene encoding the replication initiator of IncI alpha plasmid ColIb-P9. *J. Biol. Chem.*, **273**, 11815–11825.

Bellaousov,S. and Mathews,D.H. (2010) ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA*, **16**, 1870–1880.

Bindewald,E. and Shapiro,B.A. (2006) RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA*, **12**, 342–352.

Brierley,I. *et al.* (2007) Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat. Rev. Microbiol.*, **5**, 598–610.

Cao,S. and Chen,S.J. (2006) Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res.*, **34**, 2634–2652.

Cao,S. and Chen,S.J. (2009) Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA*, **15**, 696–706.

Engelen,S. and Tahi,F. (2010) `Tfold`: efficient in silico prediction of non-coding RNA secondary structures. *Nucleic Acids Res.*, **38**, 2453–2466.

Ferre-D'Amare,A.R. *et al.* (1998) Crystal structure of a hepatitis delta virus ribozyme. *Nature*, **395**, 567–574.

Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.

Gardner,P.P. *et al.* (2010) Rfam: Wikipedia, clans and the ''decimal'' release. *Nucleic Acids Res.*, **39** (Database issue), D141–D145.

Harris,J.K. *et al.* (2001) New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. *RNA*, **7**, 220–232.

Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.

Hofacker,I.L. *et al.* (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.

Ji,Y. *et al.* (2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, **20**, 1591–1602.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U S A.*, **87**, 2264–2268.

Klein,D.J. and Ferre-D'Amare,A.R. (2006) Structural basis of glmS ribozyme activation by glucosamine-6-phosphate. *Science*, **313**, 1752–1756.

Klein,R.J. and Eddy,S.R. (2003) RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**, 44.

Kolupaeva,V.G. *et al.* (2000) Ribosomal binding to the internal ribosomal entry site of classical swine fever virus. *RNA*, **6**, 1791–1807.

Lin,J.W. *et al.* (2007) Chloroplast phosphoglycerate kinase, a gluconeogenetic enzyme, is required for efficient accumulation of Bamboo mosaic virus. *Nucleic Acids Res.*, **35**, 424–432.

Meyer,I.M. and Miklos,I. (2007) SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.*, **3**, e149.

Montange,R.K. and Batey,R.T. (2006) Structure of the *S*-adenosylmethionine riboswitch regulatory mRNA element. *Nature*, **441**, 1172–1175.

Nameki,N. *et al.* (1999) An NMR and mutational analysis of an RNA pseudoknot of *Escherichia coli* tmRNA involved in trans-translation. *Nucleic Acids Res.*, **27**, 3667–3675.

Nateri,A.S. *et al.* (2002) Terminal RNA replication elements in human parechovirus 1. *J. Virol.*, **76**, 13116–13122.

Pennell,S. *et al.* (2008) The stimulatory RNA of the Visna-Maedi retrovirus ribosomal frameshifting signal is an unusual pseudoknot with an interstem element. *RNA*, **14**, 1366–1377.

Perriquet,O. *et al.* (2003) Finding the common structure shared by two homologous RNAs. *Bioinformatics*, **19**, 108–116.

Pfingsten,J.S. *et al.* (2006) Structural basis for ribosome recruitment and manipulation by a viral IRES RNA. *Science*, **314**, 1450–1454.

Philippe,C. *et al.* (1995) Molecular dissection of the pseudoknot governing the translational regulation of *Escherichia coli* ribosomal-protein S15. *Nucleic Acids Res.*, **23**, 18–28.

Plant,E.P. *et al.* (2005) A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol.*, **3**, e172.

Reeder,J. *et al.* (2006) Beyond Mfold: recent advances in RNA bioinformatics. *J. Biotechnol.*, **124**, 41–55.

Rieder,U. *et al.* (2010) Folding of a transcriptionally acting preQ1 riboswitch. *Proc. Natl. Acad. Sci. U S A.*, **107**, 10804–10809.

Ruan,J. *et al.* (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, **20**, 58–66.

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

Sato,K. *et al.* (2011) `IPknot`: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, **27**, i85–i93.

Sperschneider,J. and Datta,A. (2010) `DotKnot`: pseudoknot prediction using the probability dot plot under a refined energy model. *Nucleic Acids Res.*, **38**, e103.

Sperschneider,J. *et al.* (2011) Heuristic RNA pseudoknot prediction including intramolecular kissing hairpins. *RNA*, **17**, 27–38.

Staple,D.W. and Butcher,S.E. (2005) Pseudoknots: RNA structures with diverse functions. *PLoS Biol.*, **3**, e213.

Tabei,Y. *et al.* (2008) A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, **9**, 33.

Theimer,C.A. and Feigon,J. (2006) Structure and function of telomerase RNA. *Curr. Opin. Struct. Biol.*, **16**, 307–318.

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

van Batenburg,F.H.D. *et al.* (2000) `PseudoBase`: a database with RNA pseudoknots. *Nucleic Acids Res.*, **28**, 201–204.

Williams,G.D. *et al.* (1999) A phylogenetically conserved hairpin-type 3' untranslated region pseudoknot functions in coronavirus RNA replication. *J. Virol.*, **73**, 8349–8355.

Witwer,C. *et al.* (2004) Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 66–77.