Published online 31 January 2010

Nucleic Acids Research, 2010, Vol. 38, No. 7 ***e103***
doi:10.1093/nar/gkq021

DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model

Jana Sperschneider* and Amitava Datta

School of Computer Science and Software Engineering, The University of Western Australia, Perth, WA 6009, Australia

Received October 26, 2009; Revised December 6, 2009; Accepted January 11, 2010

ABSTRACT

RNA pseudoknots are functional structure elements with key roles in viral and cellular processes. Prediction of a pseudoknotted minimum free energy structure is an NP-complete problem. Practical algorithms for RNA structure prediction including restricted classes of pseudoknots suffer from high runtime and poor accuracy for longer sequences. A heuristic approach is to search for promising pseudoknot candidates in a sequence and verify those. Afterwards, the detected pseudoknots can be further analysed using bioinformatics or laboratory techniques. We present a novel pseudoknot detection method called DotKnot that extracts stem regions from the secondary structure probability dot plot and assembles pseudoknot candidates in a constructive fashion. We evaluate pseudoknot free energies using novel parameters, which have recently become available. We show that the conventional probability dot plot makes a wide class of pseudoknots including those with bulged stems manageable in an explicit fashion. The energy parameters now become the limiting factor in pseudoknot prediction. DotKnot is an efficient method for long sequences, which finds pseudoknots with higher accuracy compared to other known prediction algorithms. DotKnot is accessible as a web server at http://dotknot.csse.uwa.edu.au.

INTRODUCTION

RNA is a versatile nucleic acid, which is no longer seen as the passive intermediate between DNA and proteins. Numerous functional RNAs with an astonishing variety have been uncovered in the past decade. For example, non-coding RNAs participate in a wide range of cellular processes, are able to regulate gene expression and can act as catalyst (1,2). Macromolecule function is closely connected to its 3D folding and structure prediction from the base sequence is thus of great importance. RNA structure formation is understood to be hierarchical and, therefore, secondary structure prediction is the foundation for determining the tertiary folding (3,4).

There are two streams in computational RNA secondary structure prediction: comparative approaches and single sequence methods. In general, comparative approaches give accurate results for a set of well-conserved sequences (5). However, comparative methods rely on the quality of multiple alignments and are not always feasible for RNA structure prediction, due to a lack of reliable data sets (6).

When only a single sequence is given, the most popular approach for RNA structure prediction is free energy minimization. In the minimum free energy (MFE) model, continuous base pairs contribute enthalpic terms and loop regions are purely entropic. RNA comprises various secondary structure elements, i.e. stems, hairpin loops, bulge loops, internal loops and multiloops. Much experimental work has been done to determine their free energy parameters (7,8). The key concept that allows for dynamic programming is that all of these motifs are non-crossing and self-contained in terms of their free energy. The MFE secondary structure based on the additive free energy model can be predicted in $O(n^3)$ time and $O(n^2)$ space using dynamic programming (9,10). MFE prediction has been extended in several ways (11). Suboptimal structures with free energy close to the MFE can be calculated (12,13). Using the dynamic programming principle, the full equilibrium partition function for RNA secondary structure is computed in $O(n^3)$ time and $O(n^2)$ space (14). From the partition function, probabilities for base pairs and structure elements are derived. The main advantage of the MFE algorithm is its guarantee to find an optimal structure with regards to the underlying energy model. However, the inability to predict crossing structure elements, so-called pseudoknots, is a major drawback.

*To whom correspondence should be addressed. Tel: +61 8 6488 3449; Fax: +61 8 6488 1089; Email: janaspe@csse.uwa.edu.au

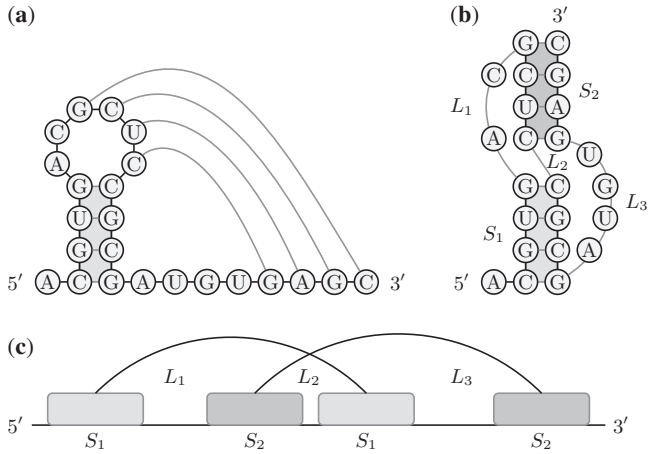

Figure 1. A simple H-type pseudoknot. (**a**) Unpaired bases within a hairpin loop bond with unpaired bases in a single-stranded region outside the loop. (**b**) The resulting pseudoknot has two stems S_1 and S_2 and three loops L_1, L_2 and L_3. (**c**) A pseudoknot has at least two crossing stems, which can be displayed as intervals on the line.

Pseudoknots are functional structure elements, which have been reported in most classes of RNA (15). A pseudoknot forms when unpaired bases in a loop pair with complementary bases in a single-stranded region outside the loop (Figure 1). Pseudoknots play key roles in viral genome replication and regulation of protein synthesis (16). Pseudoknots are also found in the cell where they participate in processes such as splicing, ribosomal frameshifting, telomerase activity and ribosome function (17–19). In many cases, they assist in the overall 3D folding (20,21) and should not be excluded from computational structure prediction.

RNA secondary structure prediction with arbitrary pseudoknots under a basic energy model is NP-complete (22,23). Restricted classes of pseudoknots can be included in the dynamic programming algorithm for prediction of the MFE structure, resulting in high computational complexity. In dynamic programming, there is always a trade-off between the generality of pseudoknots, which can be predicted, and runtime. Rivas and Eddy (24) cover a broad class of pseudoknots including kissing hairpins in their algorithm, which requires $O(n^6)$ time and $O(n^4)$ space. More restricted pseudoknots are included in other dynamic programming algorithms, which have runtime of $O(n^5)$ using $O(n^4)$ or $O(n^3)$ space (22,23,25). All of these algorithms are only feasible for short RNA sequences. The most practical method is pknotsRG, which computes the MFE structure with canonical simple recursive pseudoknots in $O(n^4)$ time and $O(n^2)$ space (26). Dynamic programming does guarantee to find a structure with minimum free energy with respect to the underlying energy model. However, the energy model for pseudoknots used in dynamic programming is only a simple parameterization adopted from the affine multiloop energy model (24–26). It was first introduced by Rivas and Eddy (24) because of the lack of experimentally measured parameters for pseudoknot energies. Predictive accuracy of MFE folding is always limited by the underlying energy model, and hence pseudoknot prediction results are poor for longer sequences.

Due to the computational complexity of dynamic programming for pseudoknot prediction, heuristic approaches were developed as an alternative. Heuristic methods do not necessarily return the MFE structure; however, they can include a wide class of pseudoknots and more advanced energy models in reasonable runtime. RNA secondary structure prediction including pseudoknots has been approached using genetic algorithms (27,28), stochastic context-free grammars (29,30), kinetic folding simulations (31,32) and maximum weighted matching in a folding graph (33). Iterative stem adding procedures have also been developed (34–36). In several heuristic algorithms, the underlying energy model for secondary structures is simple base pair maximization, neglecting loop entropies (33,34). This may lead to unreliable results, especially for longer sequences. Another drawback is that most heuristic methods reported in the literature employ the same affine pseudoknot energy model as dynamic programming.

A different algorithmic framework is used in heuristic pseudoknot detection programs (37–40). Here, one attempts to find promising pseudoknot candidates in a sequence as a first step. These potential pseudoknots are subsequently analysed and verified. After pseudoknot detection, the remaining sequence can be folded using free energy minimization in $O(n^3)$ time and $O(n^2)$ space. Pseudoknot detection has two main advantages over dynamic programming. First, it is computationally much more efficient and, therefore, practical for scanning long RNA sequences for pseudoknots. Second, the underlying framework is less restrictive and allows for easy incorporation of sophisticated energy rules for pseudoknots or even comparative information. Pseudoknot detection delivers accurate pseudoknot prediction results for longer sequences in many cases (37,39).

Pseudoknot energy models have been studied in more detail in the past years and reliable energy parameters are in high demand. It is widely accepted that pseudoknot energy cannot be estimated with an additive model as used in MFE folding. There is strong interference between opposite loops and stems (L_1 and S_2, L_3 and S_1), which come in close contact in the 3D fold. In a simple hairpin type (H-type) pseudoknot, loops L_1 and L_3 span across the deep narrow (major) groove and the shallow wide (minor) groove, respectively. The corresponding loop entropies are, therefore, not equivalent and depend on the structure and length of the opposite stem (41,42).

The lack of loop entropy parameters is a critical issue in pseudoknot prediction (43). For H-type pseudoknots with interhelix loop size ≤ 1 nt, loop entropy values were derived using several fitted parameters (41). Gaussian chain approximation based on polymer physics for pseudoknot loop entropies was also proposed (44). Lattice-based models were developed to take into account volume exclusion effects (45–47). The most

successful models are based on the atomic coordinates of the RNA backbone. Using polymer statistical mechanics, Cao and Chen (42) calculated loop entropy parameters for pseudoknots with interhelix loop size ≤ 1 nt. Recently, their so-called virtual bond model has been extended to pseudoknots with interhelix loop size ≤ 6 nt (48). Inclusion of such a pseudoknot energy model is not straightforward in dynamic programming in reasonable runtime due to dependencies between opposite loops and stems. Calculation of the full partition function under the virtual bond model takes $O(n^6)$ time and $O(n^2)$ space, making the approach only feasible for sequences shorter than say 150 nt (42). However, pseudoknot energy parameters derived from the virtual bond model are readily available in tabular form for many stem and loop length combinations (42,48). It is straightforward to incorporate such energy parameters in a pseudoknot detection approach, making prediction using much improved energy models feasible for long sequences.

We present DotKnot, a pseudoknot prediction method that incorporates stem–loop correlated pseudoknot energy parameters from the virtual bond model (42,48). The workflow is similar to the detection approach used in KnotSeeker (39), with two main improvements. First, a secondary structure partition function calculation in $O(n^3)$ time and $O(n^2)$ space is the basis for finding a set of promising structure elements with high confidence (14). This set includes heuristically derived bulge loops, internal loops and multiloops with low free energy. Second, pseudoknot candidates are assembled using the set of promising structure elements and, therefore, loop entropy parameters can readily be used for pseudoknot energy evaluations (42,48). There is no dynamic programming kernel for verification of pseudoknot candidates. This is a major step towards successful pseudoknot prediction as the vast majority of methods in the literature use simple approximations of pseudoknot energies. However, improving the folding model behind the algorithmic framework is clearly the key to accurate prediction (43).

DotKnot predicts the class of recursive H-type pseudoknots where one of the pseudoknot stems can be interrupted by bulge or internal loops. All of the three pseudoknot loops are allowed to form internal secondary structures. Using the set of structural elements derived from the probability dot plot as a construction kit, we could predict a broad class of pseudoknots, including kissing hairpins, in an efficient manner. However, our knowledge about energy parameters and folding mechanisms become the limiting factor for a pseudoknot search tool such as DotKnot. Given an RNA sequence, DotKnot returns only the (possibly empty) set of detected pseudoknots. These detected pseudoknots can subsequently be verified using laboratory techniques or comparative information. The remaining non-crossing sequence can then be folded in $O(n^3)$ time and $O(n^2)$ space using state-of-the-art RNA secondary structure prediction algorithms to obtain a global folding.

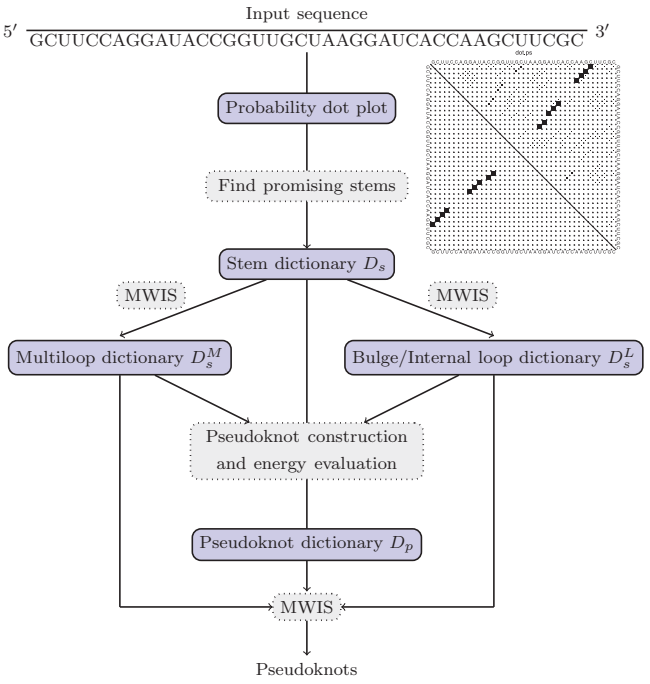

Figure 2. Workflow used by DotKnot for detecting pseudoknots in a sequence using the probability dot plot. MWIS stands for maximum weight independent set calculation (56).

MATERIALS AND METHODS

The detailed workflow for DotKnot is shown in Figure 2. Given an RNA sequence, DotKnot finds sequence fragments with pseudoknot folding potential. These candidates are analysed in terms of their free energy values and credibility in the folded sequence. The output is a (possibly empty) set of detected pseudoknots, which can then be examined closely.

Stems in the probability dot plot

RNA structure forms through complementary base pairing, resulting in stabilizing stems and destabilizing loop regions. The basic building blocks for a pseudoknot are two crossing stems. We use the probability dot plot derived from the partition function as a guide for finding such building blocks. Given an RNA sequence, the partition function Q is defined as the weighted sum over the set of all possible secondary structures S, i.e. $Q(T) = \sum_{s \in S} e^{-\Delta G_s / RT}$ where R is the universal gas constant and T the temperature. Once the partition function is known, probabilities for base pairs and structure elements can be calculated. The software RNAfold returns the probability dot plot representing both base pair probabilities and stack probabilities (11). The stack probability P_{ij} for a base pair (i,j) is defined as the probability that pair (i,j) and the subsequent pair $(i+1, j-1)$ are formed simultaneously (49).

It has been shown that choosing base pairs with high probability can improve secondary structure prediction (50). There are also several approaches that discuss the exclusion of base pairs with low probability for improving runtime. Here, the common technique is to use a cut-off

value for base pair probabilities in order to determine only significant base pairs (51). However, for pseudoknot stems we cannot simply dismiss base pairs with low probability. The partition function calculation is based on the ensemble of secondary structure elements, which are non-crossing interactions. In general, the pseudoknot stems are visible in the probability dot plot as they are members of the folding space. One can expect that at least one of the pseudoknot stems will have low base pair and stack probabilities. By default, RNAfold only displays probabilities larger than $1 \times E^{-5}$. For pseudoknot detection in the dot plot, we use a cutoff probability of $1 \times E^{-11}$ to make sure that all pseudoknot stems can be found for long sequences.

Given the probability dot plot, stems are assembled according to certain criteria. First, a stem must have at least three base pairs. Second, one can expect that the stack probabilities in a stable stem do not rise or drop sharply (49). Helix elongation is an energetically favourable process; however, wobble base pairs can be destabilizing (8,52,53). Furthermore, there may be other stable secondary structure elements competing for base pairs. Therefore, we demand that the absolute percentage increase or decrease of stack probabilities for subsequent base pairs (i, j), $(i+1, j-1)$ in a stem has to be smaller than a certain threshold δ.

One has to keep in mind that base pair probabilities are not independent. Therefore, stem probabilities can never be calculated by simply multiplying the base pair probabilities. However, the average probability of participating base pairs in a stem can be used as a confidence indicator (54). In our approach, we calculate the confidence c for a stem as the average stack probabilities. For each stem, an absolute weight is introduced in addition. The confidence c is based on the energy model for secondary structures, excluding pseudoknots. Therefore, pseudoknot stems tend to have low average probabilities especially for longer sequences, which does not necessarily correspond to their dominance in native RNA structures. We assign two additional weights for a stem based on a local energy evaluation. The simple stacking model employs the favourable base pair stacking parameters in a stem; however, it excludes the destabilizing energy contributed by the hairpin loop. The more sophisticated free energy model includes all entropy and enthalpy parameters derived by the Turner group (8). The tool RNAeval is used to evaluate the local free energies of the stem candidates according to the two energy models introduced above, taking into account dangling ends on both sides (11). DotKnot stores two stem weights $w_{stack}(s_i)$ (simple stacking model) and $w(s_i)$ (free energy model) for a stem s_i. Only stems s_i satisfying the following conditions are kept: $w_{stack}(s_i) < 0.0$ kcal/mol and $w(s_i) < 4.0$ kcal/mol. The resulting stems are stored in a stem dictionary $D_s = \{s_1, s_2, \ldots, s_n\}$. Each stem s_i has a unique key (a_i, b_i), where a_i is the start position and b_i the end position of the stem in sequence S. For each stem s_i, we store its length and the following values in the stem dictionary: $c(s_i)$,$w_{stack}(s_i)$ and $w(s_i)$.

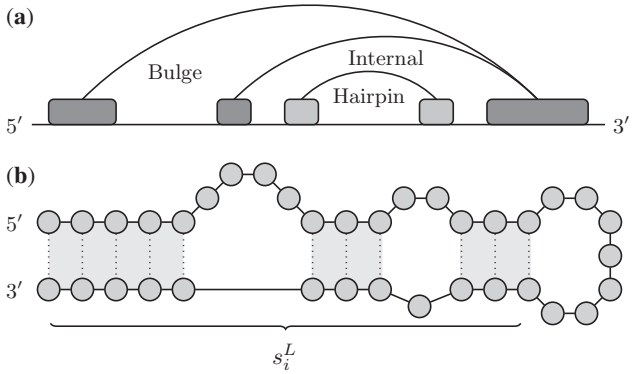

Figure 3. (**a**) Example for a stem s_i^L interrupted by a bulge loop and an internal loop. (**b**) The corresponding secondary structure is shown.

Finding bulges, internal loops and multiloops

In RNA structures, stems are often interrupted by bulge loops or internal loops. In the following, a stem interrupted by bulge loops or internal loops is denoted as s_i^L (Figure 3). Naive construction of interrupted stems s_i^L from the stem dictionary D_s would be inefficient due to a large number of possible stem combinations. Therefore, we employ a heuristic strategy for finding interrupted stems with low free energy.

Given the stem dictionary D_s, interrupted stems are constructed using maximum weight independent set (MWIS) calculations with the confidence indicators as stem weights (Supplementary Algorithm 1). Using confidence c instead of local free energy gives better results in the MWIS calculation. First, it implicitly penalizes the formation of long bulge or internal loops. Second, the confidence is a relative measurement based on the whole folding ensemble. Only stems s_i with confidence $c(s_i) \geq 1 \times E^{-3}$ are considered as base pairs below this threshold are unlikely to participate in non-crossing secondary structure formation (55). This cutoff step also significantly reduces runtime for longer sequences. Each stem $s_i \in D_s$ has two (left and right) endpoints a_i and b_i and is represented as an interval on the line. First, the sorted endpoints list for all stems is constructed and scanned from left to right. If a right endpoint is discovered, a candidate list of stems nested in the interval $s_i = [a_i : b_i]$ is returned for the corresponding stem s_i. A MWIS calculation on the candidate list returns the set of internal stems with maximum weight in linear time (56). The outer stem s_i can, therefore, become an interrupted stem s_i^L with bulges and internal loops or the exterior stem of a multiloop. The sum of confidences serves as the updated weight for outer stem s_i. After the whole endpoints list has been scanned, stems interrupted by bulge loops or internal loops are stored in a new dictionary D_s^L. Multiloops are stored in a separate dictionary D_s^M. As a last step, the free energies of the interrupted stems and multiloops are evaluated with the tool RNAeval (11). Only structure elements with negative free energy are stored.

So far, DotKnot derived a set of regular stems, interrupted stems and multiloops in a heuristic fashion from the probability dot plot (Figure 2). There is no guarantee

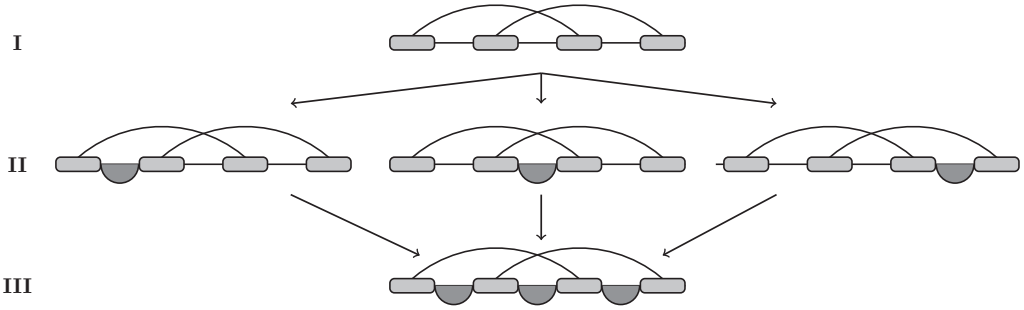

Figure 4. Construction of a recursive H-type pseudoknot. On the first level, two stems form a core H-type pseudoknot. On the second level, recursive secondary structure elements may form in each of the three loops. On the third level, the recursive H-type pseudoknot is assembled.

to find the structure element with local minimum free energy for a given sequence stretch. However, dictionaries D_s^L and D_s^M will contain structures with free energy value close or equal to the local minimum free energy. These elements will be of great benefit for the elimination of false positive pseudoknots.

Construction of candidate pseudoknots

One major advantage of the detection approach is that the pseudoknot prediction target class is predefined and transparent. In contrast, dynamic programming algorithms construct pseudoknots through their recursion scheme, which in some cases leads to unspecific pseudoknot target classes (24,57). In this work, recursive H-type pseudoknots will be considered similar to the class of pseudoknots predicted by pknotsRG (26). A recursive H-type pseudoknot has two crossing stems S_1 and S_2, resulting in three loops L_1, L_2 and L_3. All of the three loops are allowed to form internal secondary structures; however, loop–loop interactions are not allowed. In this work, we also include pseudoknots where one of the stems S_1 or S_2 is interrupted by bulges or internal loops. This leads to a more comprehensive prediction class as in pknotsRG, where only bulges of size 1 nt are considered (26). In general, the three dictionaries D_s, D_s^L and D_s^M allow for the construction of complicated pseudoknot folds. For example, it is straightforward to construct kissing hairpins from the stem dictionary. The pseudoknot energy parameters become the bottleneck in this approach, not necessarily the complexity of pseudoknots.

The three main steps for constructing recursive H-type pseudoknots are shown in Figure 4. First, so-called core H-type pseudoknots form through simple combination of two crossing stems. They become recursive H-type pseudoknots when additional secondary structure elements fold in each of the three loops. Note that recursive pseudoknots are not allowed in the loops as this may lead to sterically infeasible configurations. The overall recursive H-type candidate pseudoknot is assembled in a third step.

First level: assembling core H-type pseudoknots

On the first level, two crossing stems are combined to form a core H-type pseudoknot (Figure 4). Core H-type pseudoknots are the building blocks for more complex pseudoknots. The pseudoknot stems can either be regular stems taken from the dictionary D_s or interrupted stems from the dictionary D_s^L (Figure 5). Here, we only allow pseudoknots with at most one interrupted stem because for more complex and less rigid pseudoknots we would have to employ a highly assumptive energy model. Certain loop length restrictions are applied because more meaningful results can be expected from prediction of shorter and well-studied pseudoknots. Loop L_1 and L_3 are both required to have at least 1 and 2 nt, respectively. Interhelix loop L_2 can have a size of 0 nt. All three loops are restricted to a maximum length, as there are no reliable energy parameters for very long pseudoknots. Loops L_1 and L_3 can have a length of up to 100 nt, whereas interhelix loop L_2 is restricted to a maximum length of 50 nt. During construction of the pseudoknot candidates, we discovered that crossing stems may compete for a base pair. This leads to an overlap at loop L_2. In such a case, one of the stems is truncated according to certain rules (Supplementary Figure S2 and Supplementary Algorithm 2).

There is little knowledge about the 3D folding of complex pseudoknots with interrupted stems, which may lead to bending or distortions of the RNA A-helix. Loops L_1 and L_3 need to cross the major and minor groove of stems S_2 and S_1, respectively. To disallow sterically infeasible configurations for long interrupted stems, we make the following assumptions (41). For interrupted stems with more than 10 bp (including bulges and internal loops), the loops bridging the stems must have a minimum length: loop L_1 must be ≥ 2 nt and loop L_3 must be ≥ 6 nt.

After the first level of pseudoknot construction, free energy values are evaluated for each core H-type pseudoknot, which allows to filter unlikely pseudoknots. Only pseudoknots with low free energy (and therefore likely to form) will remain for the next step, which involves recursive secondary structure formation in the three pseudoknot loops. For all core H-type pseudoknots $p_1, \ldots, p_n$, free energy ΔG is calculated as

$$\Delta G(p_i) = w_{\text{stack}}(S_1) + w_{\text{stack}}(S_2) - T\Delta S_{L_1,L_2,L_3}$$

where $T\Delta S_{L_1,L_2,L_3}$ is the purely entropic free energy for loops L_1, L_2 and L_3. Three different pseudoknot energy models are employed for the loop entropy calculation, each of that comprise pseudoknots with certain characteristics. The length of loop L_2 determines which energy

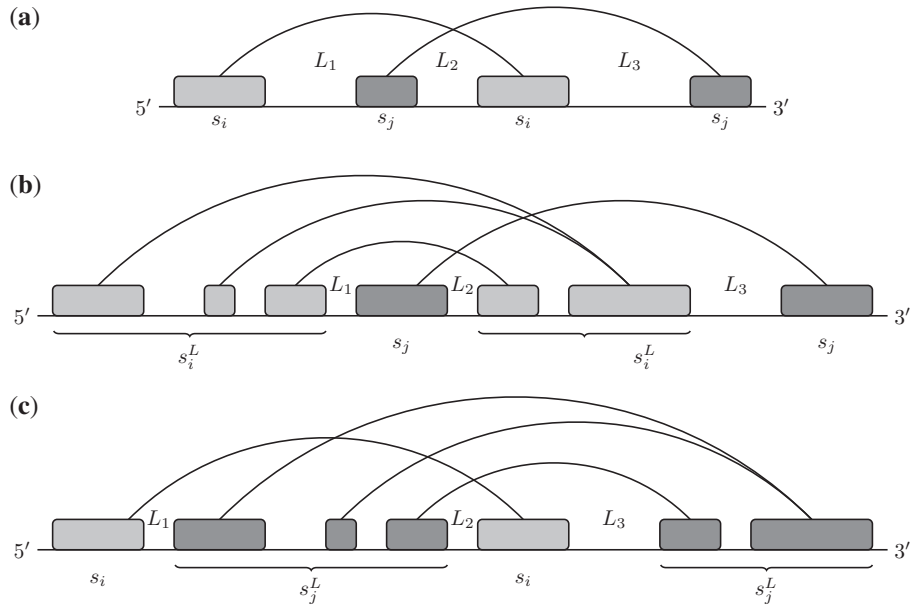

Figure 5. On the first level of pseudoknot construction two stems form a crossing structure. (**a**) Two regular stems s_i and s_j are crossing. (**b**) Stem s_i^L and stem s_j are crossing. (**c**) Stem s_i and stem s_j^L are crossing.

model is used for the respective pseudoknot candidate (Table 1). For details on pseudoknot loop entropy calculation using the virtual bond models CC06 and CC09 see (42) and (48), respectively. For pseudoknots with long interhelix loop L_2, there is no physical loop entropy model available and we have to employ heuristics (24–26). This heuristic energy model (LongPK) is also used for pseudoknots with one interrupted stem regardless of the length of loop L_2. Stems interrupted by long bulge or internal loops are likely to result in bending rather than rigid formations (48). Therefore, the loop entropy calculation becomes intricate. We calculate the loop entropy as $T\Delta S_{L_1,L_2,L_3} = \alpha + \beta \times L$, where L is the number of unpaired nucleotides in the three pseudoknot loops with $\alpha = 7.0$ kcal/mol and $\beta = 0.1$ kcal/mol.

For pseudoknots with regular stems and loop length L_2 of 0 or 1 nt, one can generally assume that the two pseudoknot stems are coaxially stacked. This leads to a stabilizing effect for the two base pairs at the interhelix junction. Here, coaxial stacking is calculated using the Turner energy model, multiplied by an estimated weighting parameter $g < 1$ and added to the free energy of a pseudoknot (7,8,24). For pseudoknots with interrupted stems and absent loop L_2, we also add the appropriate coaxial stacking energy multiplied by an estimated weighting parameter $g < 1$. After energy evaluation, only pseudoknots with negative free energy are stored in the pseudoknot dictionary D_p. Additionally, we demand that the free energy of a core H-type pseudoknot needs to be lower than the free energies $w(S_1)$ and $w(S_2)$ of the pseudoknot stems S_1 and S_2.

Second level: recursive structure formation

Secondary structure elements often form in pseudoknot loops, resulting in a recursive pseudoknot. After finding stable core H-type pseudoknots, the three loops L_1, L_2 and L_3 are examined for likely secondary structure elements. It is a valid assumption that the three loops form recursive elements independently and can be treated separately (Figure 4). From an algorithmic point of view, it is efficient to find recursive structure elements using a MWIS calculation (Supplementary Algorithm 3). Given a core H-type pseudoknot, three candidate lists hold all possible secondary structure elements from dictionaries D_s, D_s^L and D_s^M contained in each of the loops L_1, L_2 and L_3. A standard MWIS calculation with free energy weights for each of the three lists returns the set of secondary structure elements with best local free energy for each loop. The results are combined to form a recursive H-type pseudoknot.

For each recursive H-type pseudoknot, we need to evaluate the free energy with respect to the recursive structure elements. As described in Table 1, the set of pseudoknots is divided into the three different classes according to the length of loop L_2. To account for recursive structure elements, the loop entropies need to be recalculated. Following the notation in Cao and Chen (48), we first calculate the effective loop lengths. The effective loop length l_i^{eff} for a pseudoknot loop L_i ($i = 1, 2, 3$) with internal structure elements is the number of unpaired nucleotides outside those internal structure elements plus the number of internal structure elements.

Table 1. Three different energy models used for pseudoknot energy evaluation

ID	Loop L_2	Stems S_1, S_2	Entropy $T\Delta S$
CC06	0 nt $\leq \lvert L_2 \rvert \leq$ 1 nt	regular	Virtual bond model
CC09	2 nt $\leq \lvert L_2 \rvert \leq$ 6 nt	regular	Virtual bond model
LongPK	7 nt $\leq \lvert L_2 \rvert \leq$ 50 nt	regular	Heuristic model
	0 nt $\leq \lvert L_2 \rvert \leq$ 50 nt	interrupted	Heuristic model

For all pseudoknots with recursive structure elements, the loop entropy is recalculated using the effective loop lengths. Internal secondary structure elements add free energy values as given by the Turner energy model (8). We keep only pseudoknots p_i in the pseudoknot dictionary D_p, which satisfy the following two conditions. First, a recursive H-type pseudoknot must have free energy $\Delta G(p_i) < -5.25$ kcal/mol. Second, the normalized pseudoknot free energy must fulfill $\Delta G(p_i)/l_i \leq \varepsilon$ where l_i denotes the length of pseudoknot p_i. Setting the threshold to $\varepsilon = -0.25$ is not too restrictive; however, it helps us to eliminate pseudoknots with high free energy (58). For example, for the 3′-UTR TMV sequence with length 214 nt, we have 445 candidate stems in dictionary D_s and 2026 pseudoknot candidates before filtering. After the length-normalized filtering step, only 274 pseudoknot candidates remain.

Verification of pseudoknot candidates in the sequence

The set of recursive H-type pseudoknots stored in dictionary D_p will certainly contain false positive pseudoknots. Therefore, a MWIS calculation with local free energy weights is employed using the structure elements from all four dictionaries: D_s, D_s^L, D_s^M and D_p. Stems $s_i \in D_s$ need to have confidence $c(s_i) > 1 \times \mathrm{E}^{-3}$ to participate (55). Furthermore, stems are allowed to contain nested structure elements, including pseudoknots. The MWIS procedure with nesting is described in detail in KnotSeeker (39). Note that we do not include the free energy gain of −1.5 kcal/mol for the outer loop as in KnotSeeker, because favourable nested structures are already included through dictionaries D_s^L and D_s^M.

RESULTS

We evaluated our algorithm on a set of pseudoknotted and pseudoknot-free sequences of different RNA types (Table 2). Given a sequence, DotKnot is a method that predicts only pseudoknots. Therefore, predictive accuracy is measured for base pairs belonging to a pseudoknot. Two measurements are used for comparison of DotKnot and other selected algorithms from the literature. For each published pseudoknot in a sequence we report sensitivity $S = 100 \times (\mathrm{TP}/\mathrm{TP} + \mathrm{FN})$ and the positive predictive value $\mathrm{PPV} = 100 \times (\mathrm{TP}/\mathrm{TP} + \mathrm{FP})$. True positive (TP) corresponds to the number of correctly predicted base pairs in the predicted pseudoknot, False negative (FN) to the number of base pairs in the published pseudoknot that were not predicted and False positive (FP) to the number of incorrectly predicted base pairs in the predicted pseudoknot. A pseudoknot is said to be predicted by an algorithm if it is a crossing structure element and at least one of the two pseudoknot stems is partially predicted. Furthermore, the ratio $r =$ (number of correctly predicted pseudoknots)/(number of predicted pseudoknots) is reported. We compare DotKnot to two dynamic programming methods, namely pknots (24) and pknotsRG (26), the pseudoknot detection tool KnotSeeker (39) and the heuristic approach HotKnots (35). HotKnots returns a number of sub-optimal scenarios; however, we only evaluate predictive accuracy for the best solution.

Table 2. RNA types and sequences used for pseudoknot prediction

Type of RNA	Sequence ID	Reference
5S rRNA	5SColi, 5SDmobilis	(59)
tRNA	DA0260, DC0010, DY4441	(60)
miRNA	Dros-mel, ath-MIR156a, Human-mi	(61)
Ribozymes	HDV, HDVanti	(62,63)
IRES	CrPV	(64)
3′-UTR	BCV, MHV, NeRNV, TMV	(65–67)
tmRNA	EColi-tmRNA, LP-tmRNA	(68)
Aptamer	HIV1-1.3a	(69)
Viral tRNA-like	LRSVbeta, TYMV	(70,71)
Frameshifting	BWYV, JEV, MEV, VMV, SARS-CoV	(72–75)
Telomerase	Human-telo, Tetra-telo	(76)
5′-UTR	HPeV1	(77)
mRNA	T2, T4	(78)

Note that 5S rRNA, tRNA and miRNA are pseudoknot-free.

5S rRNA, tRNA and miRNA are pseudoknot-free types of RNA. For the 5S rRNA and miRNA sequences chosen in our test set, DotKnot does not introduce any spurious pseudoknots. For two of the tRNA sequences, DotKnot predicts false positive pseudoknots. KnotSeeker, pknotsRG and HotKnots also predict false positive pseudoknots for some of the tRNA sequences. Minimum free energy prediction is known to have low accuracy for tRNAs. This is due to modified bases as well as coaxially stacked helices, which determine the characteristic 3D cloverleaf fold of tRNAs (24,26). Coaxial energies are implemented by pknots, which might explain why it does not predict false positive pseudoknots for the tested tRNA sequences.

Several of the test sequences contain pseudoknots where one of the core pseudoknot stems is interrupted by bulges or internal loops: *Escherichia coli* tmRNA, *Legionella pneumophila* tmRNA, the SARS frameshifting pseudoknot, human telomerase and *Tetrahymena* telomerase. For all of these pseudoknots, DotKnot delivers the best results in terms of sensitivity and specificity. For example, the *Tetrahymena* telomerase RNA (TER) contains a pseudoknot with a conserved central GA bulge in one of its stems, which is vital for telomerase function (76). DotKnot perfectly predicts this bulged pseudoknot, while all other methods do not predict a pseudoknot structure. The biological relevance of pseudoknots with bulged residues should not be underestimated. Therefore, a prediction algorithm that can handle pseudoknots with interrupted stems such as DotKnot is highly desirable.

For complex pseudoknot foldings such as the hepatitis delta virus (HDV) double pseudoknot configuration, DotKnot gives the best prediction results out of all tested algorithms. The CrPV IRES has a long pseudoknot, which contains another nested pseudoknot. For this pseudoknot, DotKnot also gives the best prediction in terms of sensitivity and PPV.

DotKnot has excellent accuracy for simple H-type pseudoknots as those reported in many viral 3′-UTRs and frameshifting regions. We found that the energy models CC06 and CC09 used for predicting such

Table 3. Summary of pseudoknot detection results on various RNA sequences

Sequence			DotKnot			KnotSeeker			pknotsRG			pknots			HotKnots		
ID	nt	PK	*S*	PPV	*r*	*S*	PPV	*r*	*S*	PPV	*r*	*S*	PPV	*r*	*S*	PPV	*r*
5SColi	120	0	–	–	0/0	–	–	0/0	–	–	0/0	–	–	0/0	–	–	0/0
5SDmobilis	133	0	–	–	0/0	–	–	0/0	–	–	0/0	–	–	0/0	–	–	0/0
DA0260	75	0	–	–	0/1	–	–	0/1	–	–	0/0	–	–	0/0	–	–	0/1
DC0010	73	0	–	–	0/0	–	–	0/1	–	–	0/0	–	–	0/0	–	–	0/0
DY4441	73	0	–	–	0/2	–	–	0/0	–	–	0/1	–	–	0/0	–	–	0/0
Dros-mel	81	0	–	–	0/0	–	–	0/0	–	–	0/0	–	–	0/0	–	–	0/0
ath-MIR156a	123	0	–	–	0/0	–	–	0/0	–	–	0/0	–	–	0/0	–	–	0/0
Human-mi	110	0	–	–	0/0	–	–	0/1	–	–	0/0	–	–	0/1	–	–	0/0
HDV	87	1	**93.8**	**100**	1/1	90.6	93.6	1/1	90.6	93.6	1/1	81.3	81.3	1/1	0	0	0/0
HDVanti	91	1	**100**	**96.1**	1/1	84	75	1/1	0	0	0/0	44	33.4	1/1	0	0	0/0
CrPV	190	2	**52.3**	**57.5**	2/2	20.5	32.1	2/2	0	0	0/0	*	*	*	0	0	0/0
			78.6	**73.4**		**78.6**	**73.4**		0	0		*	*	*	0	0	
BCV	345	1	100	81.2	1/1	**100**	**90**	1/1	**100**	**90**	1/1	*	*	*	0	0	0/0
MHV	315	1	100	81.2	1/1	**100**	**90**	1/2	**100**	**90**	1/3	*	*	*	0	0	0/0
NeRNV	287	5	**100**	**100**	4/7	100	88.9	4/4	0	0	1/2	*	*	*	0	0	0/0
			77.8	**100**		77.8	87.4		0	0		*	*	*	0	0	
			90	**100**		**90**	**100**		0	0		*	*	*	0	0	
			0	0		0	0		0	0		*	*	*	0	0	
			100	**100**		**100**	**100**		**100**	**100**		*	*	*	0	0	
TMV	214	5	**100**	**100**	5/5	77.8	87.5	5/5	0	0	0/0	*	*	*	0	0	0/0
			100	**100**		81.8	90		0	0		*	*	*	0	0	
			88.9	**100**		**88.9**	**100**		0	0		*	*	*	0	0	
			95.7	86.9		100	95.8		0	0		*	*	*	0	0	
			100	**100**		**100**	**100**		0	0		*	*	*	0	0	
EColi-tmRNA	363	4	**100**	**100**	3/5	**100**	**100**	2/3	0	0	0/0	*	*	*	0	0	0/0
			100	**100**		0	0		0	0		*	*	*	0	0	
			0	0		**28.6**	**40**		0	0		*	*	*	0	0	
			68.4	**81.3**		0	0		0	0		*	*	*	0	0	
LP-tmRNA	362	4	**90**	**90**	3/8	0	0	0/2	0	0	0/0	*	*	*	0	0	0/0
			0	0		0	0		0	0		*	*	*	0	0	
			25	**26.7**		0	0		0	0		*	*	*	0	0	
			70.6	**100**		0	0		0	0		*	*	*	0	0	
HIV1-1.3a	37	1	**100**	**100**	1/1	**100**	**100**	1/1	**100**	**100**	1/1	**100**	**100**	1/1	**100**	**100**	1/1
LRSVbeta	221	1	**85.3**	**87.9**	1/1	14.7	62.5	1/1	0	0	0/0	*	*	*	0	0	0/0
TYMV	85	2	0	0	1/2	0	0	1/1	0	0	1/1	0	0	1/1	0	0	0/0
			62.5	55.6		100	80		62.5	50		**100**	**88.9**		0	0	
BWYV	69	1	50	57.1	1/1	37.5	50	1/1	0	0	0/1	0	0	0/0	**100**	**88.9**	1/1
JEV	138	1	100	78.3	1/3	**100**	**90**	1/2	0	0	0/0	0	0	0/1	0	0	0/0
MEV	138	1	**100**	**72**	1/2	0	0	0/1	0	0	0/1	0	0	0/0	0	0	0/0
VMV	79	1	**100**	**82.3**	1/1	**100**	**82.3**	1/1	0	0	0/0	100	60.9	1/1	0	0	0/0
SARS-CoV	244	1	**92.3**	**100**	1/2	92.3	92.3	1/3	38.5	66.7	1/1	*	*	*	38.5	55.6	1/1
Human-telo	210	1	**61.3**	**51.3**	1/1	35.5	35.5	1/2	0	0	0/1	*	*	*	0	0	0/0
Tetra-telo	159	1	**100**	**100**	1/1	0	0	0/0	0	0	0/0	0	0	0/0	0	0	0/0
HPeV1	709	1	54.5	54.5	1/7	**100**	**100**	1/8	54.5	54.5	1/4	*	*	*	0	0	0/0
T2	946	1	**100**	**100**	1/9	**100**	**100**	1/5	**100**	**100**	1/1	*	*	*	0	0	0/0
T4	1340	1	**100**	**100**	1/10	**100**	**100**	1/8	0	0	0/1	*	*	*	0	0	0/0

For each pseudoknot, the best results in terms of both sensitivity *S* and positive predictive value PPV are marked in bold. The * symbol indicates that we were not able to run the algorithm due to the high time and space requirements. PK corresponds to the number of pseudoknots in the sequence as reported in the literature. We use pknots 1.05 with coaxial energies, pknotsRG 1.3 and HotKnots 1.2 without suboptimal solutions.

pseudoknots give better results than the heuristic pseudoknot energy parameters employed by the other algorithms. For example, the NeRNV and TMV 3′-UTRs both have five pseudoknots where four are simple H-type pseudoknots with interhelix loop ≤ 1 nt. For these pseudoknots, DotKnot gives the most accurate predictions, which we claim is due to the improved energy parameters by Cao and Chen (42,48).

In terms of computational performance, DotKnot is very efficient due to the sparseness of the probability dot plot, the resulting low number of pseudoknot candidates and the implementation using dictionaries in Python. DotKnot runs in the order of seconds for all of the test sequences except T2 and T4, which take several minutes. For T4 with 1340 nt, we have 6567 candidate stems in dictionary D_s and 7534 pseudoknot candidates before filtering. After the length-normalized filtering step, only 100 pseudoknot candidates remain for verification. Overall, it takes DotKnot <5 min to predict the correct pseudoknot in this sequence on our reference machine (Intel QC 2.66 GHz, 4 GB RAM). This is significantly faster than HotKnots, which takes 29 min, and pknotsRG, which takes 31 min. KnotSeeker is even faster than DotKnot and takes <2 min for the T4 sequence, because it does not rely on a partition function calculation. However, DotKnot is a more powerful prediction algorithm than

KnotSeeker due to the inclusion of pseudoknots with one interrupted stem.

CONCLUSION

We presented DotKnot, a program that detects recursive H-type pseudoknots given an RNA sequence. Pseudoknot detection is a promising and efficient approach for determining the folding of an RNA. Using pseudoknot detection tools such as DotKnot, KnotSeeker or HPknotter, one can find likely pseudoknots in a sequence with high accuracy (37,39). The structure of the detected pseudoknots can subsequently be investigated using laboratory or bioinformatics techniques. The remaining non-crossing sequence can be folded using secondary structure prediction algorithms in $O(n^3)$ time and $O(n^2)$ space. DotKnot and other pseudoknot detection approaches are very time efficient, even allowing scanning of long regions in viral genomes.

DotKnot assembles pseudoknot candidates from a set of structural building blocks. This set contains stems, bulge loops, internal loops and multiloops. In general, complex pseudoknots can be constructed. However, there is a trade-off between the generality of predictable pseudoknots and the biological relevance of the result. Therefore, we restrict DotKnot to the prediction of recursive H-type pseudoknots where one of the pseudoknot stems can contain bulges and internal loops. For these recursive H-type pseudoknots, we are confident that the pseudoknot energy parameters used give a good approximation. For more complex pseudoknot folds such as those with loop–loop interactions, we would have to employ a highly assumptive energy model, thus sacrificing predictive accuracy. In the future, DotKnot will be extended to the prediction of kissing hairpins and other biologically relevant classes of pseudoknots.

DotKnot uses stack probabilities from the probability dot plot as the basis for finding pseudoknot building blocks. At this stage, only a single sequence is accepted as an input. It is well-known that functional pseudoknots are highly conserved in nature, for example in virus families. The pseudoknot detection framework used in DotKnot allows for incorporation of comparative information using aligned probability dot plots. One can expect that reliable alignments will greatly improve confidence in the predicted pseudoknots. Especially for complex pseudoknot foldings such as those found in bacterial tmRNA or in the telomerase RNA component, inclusion of comparative information will be invaluable.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

J.S. would like to thank Elena Rivas, Eric Westhof and the participants of the RNA workshop 2009 in Benasque, Spain, for the inspiring meeting and discussions.

FUNDING

Funding for open access charge: The University of Western Australia.

REFERENCES

1. Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
2. Mello,C.C. and Conte,D. Jr (2004) Revealing the world of RNA interference. *Nature*, **431**, 338–342.
3. Tinoco,I. and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.
4. Shapiro,B.A., Yingling,Y.G., Kasprzak,W. and Bindewald,E. (2007) Bridging the gap in RNA structure prediction. *Curr. Opin. Struct. Biol.*, **17**, 157–165.
5. Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**.
6. Wilm,A., Higgins,D.G. and Notredame,C. (2008) R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res.*, **36**
7. Serra,M.J. and Turner,D.H. (1995) Predicting thermodynamic properties of RNA. *Methods Enzymol.*, **259**, 242–261.
8. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
9. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
10. Lyngsø,R.B., Zuker,M. and Pedersen,C.N. (1999) Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics*, **15**, 440–445.
11. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, **125**, 167–188.
12. Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.
13. Wuchty,S., Fontana,W., Hofacker,I.L. and Schuster,P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.
14. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
15. Brierley,I., Pennell,S. and Gilbert,R.J. (2007) Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat. Rev. Microbiol.*, **5**, 598–610.
16. Brierley,I., Gilbert,R.J. and Pennell,S. (2008) RNA pseudoknots and the regulation of protein synthesis. *Biochem. Soc. Trans.*, **36**, 684–689.
17. Giedroc,D.P., Theimer,C.A. and Nixon,P.L. (2000) Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Mol. Biol.*, **298**, 167–185.
18. Staple,D.W. and Butcher,S.E. (2005) Pseudoknots: RNA structures with diverse functions. *PLoS Biol.*, **3**, e213.
19. Giedroc,D.P. and Cornish,P.V. (2009) Frameshifting RNA pseudoknots: structure and mechanism. *Virus Res.*, **139**, 193–208.
20. Fechter,P., Rudinger-Thirion,J., Florentz,C. and Giege,R. (2001) Novel features in the tRNA-like world of plant viral RNAs. *Cell. Mol. Life Sci.*, **58**, 1547–1561.
21. Hammond,J.A., Rambo,R.P., Filbin,M.E. and Kieft,J.S. (2009) Comparison and functional implications of the 3D architectures of viral tRNA-like structures. *RNA*, **15**, 294–307.
22. Akutsu,T. (2000) Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl. Math.*, **104**, 45–62.
23. Lyngsø,R.B. and Pedersen,C.N. (2000) RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.*, **7**, 409–427.
24. Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.

25. Dirks,R.M. and Pierce,N.A. (2003) A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, **24**, 1664–1677.
26. Reeder,J. and Giegerich,R. (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, **5**, 104.
27. Gultyaev,A.P., van Batenburg,F.H. and Pleij,C.W. (1995) The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.*, **250**, 37–51.
28. van Batenburg,F.H., Gultyaev,A.P. and Pleij,C.W. (1995) An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.*, **174**, 269–280.
29. Brown,M. and Wilson,C. (1996) RNA pseudoknot modeling using intersections of stochastic context free grammars with applications to database search. *Pac. Symp. Biocomput.*, 109–125.
30. Cai,L., Malmberg,R.L. and Wu,Y. (2003) Stochastic modeling of RNA pseudoknotted structures: a grammatical approach. *Bioinformatics*, **19**, i66–i73.
31. Abrahams,J.P., van den Berg,M., van Batenburg,E. and Pleij,C. (1990) Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Res.*, **18**, 3035–3044.
32. Xayaphoummine,A., Bucher,T., Thalmann,F. and Isambert,H. (2003) Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc. Natl. Acad. Sci. USA*, **100**, 15310–15315.
33. Tabaska,J.E., Cary,R.B., Gabow,H.N. and Stormo,G.D. (1998) An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, **14**, 691–699.
34. Ruan,J., Stormo,G.D. and Zhang,W. (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, **20**, 58–66.
35. Ren,J., Rastegari,B., Condon,A. and Hoos,H.H. (2005) HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, **11**, 1494–1504.
36. Chen,X., He,S., Bu,D., Zhang,F., Wang,Z., Chen,R. and Gao,W. (2008) FlexStem: improving predictions of RNA secondary structures with pseudoknots by reducing the search space. *Bioinformatics*, **24**, 1994–2001.
37. Huang,C.H., Lu,C.L. and Chiu,H.T. (2005) A heuristic approach for detecting RNA H-type pseudoknots. *Bioinformatics*, **21**, 3501–3508.
38. Huang,X. and Ali,H. (2007) High sensitivity RNA pseudoknot prediction. *Nucleic Acids Res.*, **35**, 656–663.
39. Sperschneider,J. and Datta,A. (2008) KnotSeeker: heuristic pseudoknot detection in long RNA sequences. *RNA*, **14**, 630–640.
40. Theis,C., Reeder,J. and Giegerich,R. (2008) KnotInFrame: prediction of -1 ribosomal frameshift events. *Nucleic Acids Res.*, **36**, 6013–6020.
41. Gultyaev,A.P., van Batenburg,F.H. and Pleij,C.W. (1999) An approximation of loop free energy values of RNA H-pseudoknots. *RNA*, **5**, 609–617.
42. Cao,S. and Chen,S.J. (2006) Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res.*, **34**, 2634–2652.
43. Chen,S.J. (2008) RNA folding: conformational statistics, folding kinetics and ion electrostatics. *Annu. Rev. Biophys.*, **37**, 197–214.
44. Aalberts,D.P. and Hodas,N.O. (2005) Asymmetry in RNA pseudoknots: observation and theory. *Nucleic Acids Res.*, **33**, 2210–2214.
45. Lucas,A. and Dill,K.A. (2003) Statistical mechanics of pseudoknot polymers. *J. Chem. Phys.*, **119**, 2414–2421.
46. Kopeikin,Z. and Chen,S.J. (2005) Statistical thermodynamics for chain molecules with simple RNA tertiary contacts. *J. Chem. Phys.*, **122**, 094909.
47. Kopeikin,Z. and Chen,S.J. (2006) Folding thermodynamics of pseudoknotted chain conformations. *J. Chem. Phys.*, **124**, 154903.
48. Cao,S. and Chen,S.J. (2009) Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA*, **15**, 696–706.
49. Bompfünewerer,A.F., Backofen,R., Bernhart,S.H., Hertel,J., Hofacker,I.L., Stadler,P.F. and Will,S. (2008) Variations on RNA folding and alignment: lessons from Benasque. *J. Math. Biol.*, **56**, 129–144.
50. Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.
51. Will,S., Reiche,K., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, 680–691.
52. Serra,M.J., Lyttle,M.H., Axenson,T.J., Schadt,C.A. and Turner,D.H. (1993) RNA hairpin loop stability depends on closing base pair. *Nucleic Acids Res.*, **21**, 3845–3849.
53. Giese,M.R., Betschart,K., Dale,T., Riley,C.K., Rowan,C., Sprouse,K.J. and Serra,M.J. (1998) Stability of RNA hairpins closed by wobble base pairs. *Biochemistry*, **37**, 1094–1100.
54. Hamada,M., Tsuda,K., Kudo,T., Kin,T. and Asai,K. (2006) Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics*, **22**, 2480–2487.
55. Hofacker,I.L. and Stadler,P.F. (1999) Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comput. Chem.*, **23**, 401–414.
56. Hsiao,J.Y., Tang,C.Y. and Chang,R.S. (1992) An Efficient Algorithm for Finding a Maximum Weight 2-Independent Set on Interval-Graphs. *Inf. Process. Lett.*, **43**, 229–235.
57. Condon,A., Davy,B., Rastegari,B., Zhao,S. and Tarrant,F. (2004) Classifying RNA pseudoknotted structures. *Theor. Comput. Sci.*, **320**, 35–50.
58. Freyhult,E., Gardner,P.P. and Moulton,V. (2005) A comparison of RNA folding measures. *BMC Bioinformatics*, **6**, 241.
59. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y.S., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M. *et al.* (2002) The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 15.
60. Sprinzl,M., Horn,C., Brown,M., Ioudovitch,A. and Steinberg,S. (1998) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **26**, 148–153.
61. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
62. Ferre-D'Amare,A.R., Zhou,K.H. and Doudna,J.A. (1998) Crystal structure of a hepatitis delta virus ribozyme. *Nature*, **395**, 567–574.
63. van Batenburg,F.H., Gultyaev,A.P., Pleij,C.W., Ng,J. and Oliehoek,J. (2000) Pseudobase: a database with RNA pseudoknots. *Nucleic Acids Res.*, **28**, 201–204.
64. Schüler,M., Connell,S.R., Lescoute,A., Giesebrecht,J., Dabrowski,M., Schroeer,B., Mielke,T., Penczek,P.A., Westhof,E. and Spahn,C.M. (2006) Structure of the ribosome-bound cricket paralysis virus IRES RNA. *Nat. Struct. Mol. Biol.*, **13**, 1092–1096.
65. Williams,G.D., Chang,R.Y. and Brian,D.A. (1999) A phylogenetically conserved hairpin-type 3′ untranslated region pseudoknot functions in coronavirus RNA replication. *J. Virol.*, **73**, 8349–8355.
66. Koenig,R., Barends,S., Gultyaev,A.P., Lesemann,D.E., Vetten,H.J., Loss,S. and Pleij,C.W. (2005) Nemesia ring necrosis virus: a new tymovirus with a genomic RNA having a histidylatable tobamovirus-like 3′-end. *J. Gen. Virol.*, **86**, 1827–1833.
67. van Belkum,A., Abrahams,J.P., Pleij,C.W. and Bosch,L. (1985) Five pseudoknots are present at the 204 nucleotides long 3′ noncoding region of tobacco mosaic virus RNA. *Nucleic Acids Res.*, **13**, 7673–7686.
68. Williams,K.P. (2000) The tmRNA website. *Nucleic Acids Res.*, **28**, 168.
69. Tuerk,C., MacDougal,S. and Gold,L. (1992) RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc. Natl. Acad. Sci. USA*, **89**, 6988–6992.
70. Solovyev,A.G., Savenkov,E.I., Agranovsky,A.A. and Morozov,S.Y. (1996) Comparisons of the genomic cis-elements and coding regions in RNA beta components of the hordeiviruses barley stripe mosaic virus, lychnis ringspot virus, and poa semilatent virus. *Virology*, **219**, 9–18.

71. Matsuda,D. and Dreher,T.W. (2004) The tRNA-like structure of Turnip yellow mosaic virus RNA is a 3′-translational enhancer. *Virology*, **321**, 36–46.
72. Su,L., Chen,L., Egli,M., Berger,J.M. and Rich,A. (1999) Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot. *Nat. Struct. Biol.*, **6**, 285–292.
73. Firth,A.E. and Atkins,J.F. (2009) A conserved predicted pseudoknot in the NS2A-encoding sequence of West Nile and Japanese encephalitis flaviviruses suggests NS1′ may derive from ribosomal frameshifting. *Virol. J.*, **6**, 14.
74. Pennell,S., Manktelow,E., Flatt,A., Kelly,G., Smerdon,S.J. and Brierley,I. (2008) The stimulatory RNA of the Visna-Maedi retrovirus ribosomal frameshifting signal is an unusual pseudoknot with an interstem element. *RNA*, **14**, 1366–1377.
75. Plant,E.P., Perez-Alvarado,G.C., Jacobs,J.L., Mukhopadhyay,B., Hennig,M. and Dinman,J.D. (2005) A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol.*, **3**, e172.
76. Theimer,C.A. and Feigon,J. (2006) Structure and function of telomerase RNA. *Curr. Opin. Struct. Biol.*, **16**, 307–318.
77. Nateri,A.S., Hughes,P.J. and Stanway,G. (2002) Terminal RNA replication elements in human parechovirus 1. *J. Virol.*, **76**, 13116–13122.
78. Du,Z. and Hoffman,D.W. (1997) An NMR and mutational study of the pseudoknot within the gene 32 mRNA of bacteriophage T2: insights into a family of structurally related RNA pseudoknots. *Nucleic Acids Res.*, **25**, 1130–1135.