BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 5 2010, pages 610–616
doi:10.1093/bioinformatics/btp680

Structural bioinformatics Advance Access publication December 16, 2009

RNAsnoop: efficient target prediction for H/ACA snoRNAs

Hakim Tafer[1,2,3,*], Stephanie Kehr[2,3], Jana Hertel[2,3], Ivo L. Hofacker[1] and Peter F. Stadler[1–6]

[1]Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria, [2]Bioinformatics Group, Department of Computer Science, [3]Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, [4]Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, [5]RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology, Perlikstraße 1,D-04103 Leipzig, Germany and [6]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM, USA

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Small nucleolar RNAs are an abundant class of non-coding RNAs that guide chemical modifications of rRNAs, snRNAs and some mRNAs. In the case of many 'orphan' snoRNAs, the targeted nucleotides remain unknown, however. The box H/ACA subclass determines uridine residues that are to be converted into pseudouridines via specific complementary binding in a well-defined secondary structure configuration that is outside the scope of common RNA (co-)folding algorithms.

Results: `RNAsnoop` implements a dynamic programming algorithm that computes thermodynamically optimal H/ACA-RNA interactions in an efficient scanning variant. Complemented by an support vector machine (SVM)-based machine learning approach to distinguish true binding sites from spurious solutions and a system to evaluate comparative information, it presents an efficient and reliable tool for the prediction of H/ACA snoRNA target sites. We apply `RNAsnoop` to identify the snoRNAs that are responsible for several of the remaining 'orphan' pseudouridine modifications in human rRNAs, and we assign a target to one of the five orphan H/ACA snoRNAs in *Drosophila*.

Availability: The C source code of `RNAsnoop` is freely available at http://www.tbi.univie.ac.at/~htafer/RNAsnoop

Contact: htafer@tbi.univie.ac.at

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 26, 2009; revised on November 30, 2009; accepted on December 6, 2009

1 INTRODUCTION

Box H/ACA snoRNA facilitates the conversion of Uracil to pseudouracil (Ψ) in a specific sequence context (Bachellerie *et al.*, 2002). The specificity for a particular target site is the consequence of the hybridization of snoRNA and target RNA, in most cases a ribosomal RNA. The target U is positioned by two specific interactions of the flanking target RNA sequence with the complementary sequence of the recognition loop of the snoRNA (Ni *et al.*, 1997), see Figure 1. The 'correct' secondary structures of snoRNAs are typically hard to predict. Thus, the exact structure of the interior loop, and hence the sequence motifs complementary to the binding site, are unknown. We employ here the idea of Thermodynamic Matchers (Höchsmann *et al.*, 2006) to determine the energetically optimal structure of an H/ACA snoRNA that is bound to a given putative target sequence. The implementation of Thermodynamic Matchers (Reeder *et al.*, 2007) is not directly applicable, however, since the snoRNA–target interaction corresponds to a complex pseudoknot (in the conceptual concatenation of snoRNA and mRNA) that is beyond the scope of existing RNA folding software.

The prediction of putative snoRNA target sites is an integral part of two programs [`snoGPS` (Schattner *et al.*, 2004) and `Fisher` (Freyhult *et al.*, 2008)] that attempt to detect H/ACA snoRNAs in genomic DNA. Both programs search for sequence complementarities between a list of possible target sites and the binding region of the snoRNA candidate. In these models, mismatches between the target and the snoRNA are not allowed. Furthermore, neither program provides information on the energetics of the interaction or the stability of the stems, two factors that were recently shown to be important for correctly predicting snoRNA–target interactions (Xiao *et al.*, 2009).

We present here a dynamic programming algorithm named `RNAsnoop`, that specifically captures the structure of the snoRNA–target interaction and is optimized for scanning speed. The thermodynamic considerations are combined with a machine learning component to increase the specificity of target predictions, which can be improved even further by including comparative information.

2 SINGLE-SEQUENCE RNASNOOP

2.1 Specialized folding algorithm

`RNAsnoop` implements a specialized co-folding algorithm that takes into account that stringent structural constraints must be satisfied for a functional interaction of a box H/ACA snoRNA stem–loop and its target. As input, `RNAsnoop` takes one of the typical two stem–loop components of a known or predicted H/ACA snoRNA. The closing stem, T is assumed to be known from the a priori prediction of the snoRNA structure. The part of the snoRNA sequence enclosed by T is allowed to interact with the target structure. Figure 1 outlines the general principle.

*To whom correspondence should be addressed.

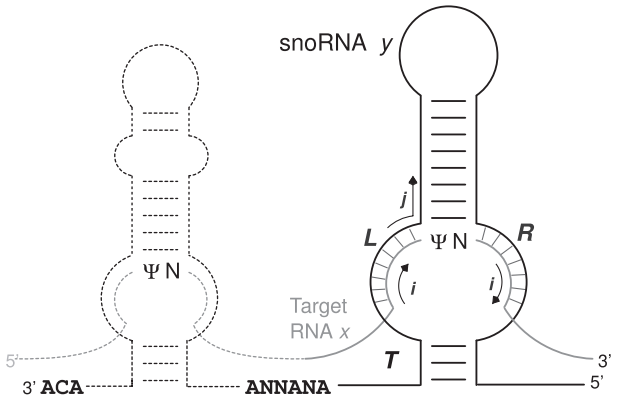

Fig. 1. Box H/ACA snoRNAs typically interact with both stem–loop structures with regions of a target RNA flanking the Uracil residue that is to be pseudouridylated. Computation of the interaction structure is performed separately for the two stems–loop components of a H/ACA snoRNA. The closing stem T at the root of each branch is assumed to be given from the structure prediction. The region inside of T is decomposed into the upper stem–loop structure with an energy contribution M, left and right interaction structures with their energy contribution L and R, respectively. Since RNAsnoop scans the target RNA in 5′–3′ direction, the snoRNA is read in 3′–5′ direction.

The interaction structure can be decomposed into the unbranched stem–loop 'above' the pseudouridylation site, and the left and right 'arms' of the binding site itself. The total energy of these components will be optimized by dynamic programming. In addition, the snoRNA–target interaction is influenced by the short closing stem of the interaction loop.

The upper stem–loop structure of the snoRNA (with sequence y) is simply modeled as an unbranched fold. The energies of its optimal substructures satisfy the recursion

$$M_{p,q} = \min \begin{cases} \mathcal{H}(y[p,q]) \\ \min\limits_{k,l} M_{p-k,q+l} + \mathcal{I}(y[p-k,p], y[q,q+l]), \end{cases} \quad (1)$$

where $\mathcal{H}(y[p,q])$ denotes the energy parameters (Lu *et al.*, 2006; Mathews *et al.*, 1999) for a hairpin loop formed by the subsequence $y[p,q] = y_p y_{p+1} \ldots y_q$ including the closing pair (y_p, y_q). Analogously, $\mathcal{I}(y[u,p], y[q,v])$ is the energy of an interior loop composed of the sequences $y[u,p]$ and $y[q,v]$, again including the delimiting base pairs (y_p, y_q) and (y_u, y_v).

Inspection of known snoRNA–rRNA interactions revealed that the interaction region can contain only single and tandem mismatches but no bulges. Therefore, we allow only stacked base pairs and symmetrical loops of lengths 2 and 4. Thus, the left part satisfies the recursion

$$L_{i,j} = \min_{k=1,2,3} L_{i-k,j+k} + \mathcal{I}(x[i-k,i], y[j,j+k]). \quad (2)$$

The index i runs along the target RNA x, while j refers to the position on the snoRNA y. To ensure that all interactions start inside the recursion matrix we set $L_{i,j} = 0$

The r.h.s. array R contains the optimal folding energies of the interaction structure up to positions i on the target and j on the snoRNA consisting of the l.h.s. binding region L, the snoRNA stem–loop M and the partial r.h.s. binding region $R_{i,j}$. It thus extends a r.h.s. binding region or refers to its first base pair. In the latter case, nucleotide x_{i-2} is the uracil that is pseudouridylated. The corresponding recursion reads

$$R_{i,j} = \min \begin{cases} \min\limits_{k,l \le 2} R_{i-k,j+l} + \mathcal{I}(x[i-k,i], y[j,j+l]) \\ \min\limits_{l \in [3,|y|-j]} L_{i-3,j+l+1} + M_{j+1,j+l} \\ \qquad \text{if} \quad x[i-2] = 'U'. \end{cases} \quad (3)$$

For each i, the best binding energy at target position i is $\max_j R_{i,j}$.

Space and time requirements for the M-matrix are limited by the size $|y|$ of the snoRNA stem–loop structure, which is a user-specified constant, typically 120 nt. Formally, the space and time complexity is $\mathcal{O}(|y|^2)$ and $\mathcal{O}(|y|^4)$, respectively. The space requirements for the L and R arrays are limited to $5 \times |y|$ independent of the target $|x|$ of the target RNA. This is possible because the length of interior loops in the recursions is restricted to not more than 4 and the transition from L to R recursion only looks back to $i-4$. The time complexity for L is $\mathcal{O}(|x| \cdot |y|)$, while for R we need $\mathcal{O}(|x| \cdot |y|^2)$ operations. The total run time is thus $\mathcal{O}(|x||y|^2 + |y|^4)$, i.e. we have a linear 'scanning algorithm' for long target RNAs.

Due to the difference in accessibility between sites with pseudouridine and uridine residues in both human and yeast (see Fig. 2 and Supplementary Fig. S1), we extended RNAsnoop so that accessibility information are considered in the folding step. Accessibility profiles as computed by RNAup (Mückstein *et al.*, 2006) or RNAplfold (Bernhart *et al.*, 2006; Bompfünewerer *et al.*, 2008) describe the energy necessary to open the secondary structure on an interval of the target sequence. The full implementation of RNA–RNA interactions is too expensive in terms of computational resources for a target search program. We therefore borrow the approach from RNAplex (Tafer and Hofacker, 2008a), which uses an affine approximation to speed up the computation of RNA–RNA interaction energies. A recent extension (Tafer and Hofacker, 2008b) shows that the accuracy can be improved substantially by incorporating precomputed accessibility profiles in the parameterization of the interaction energies. Here, we use the same idea to approximate the influence of the target site accessibility on the snoRNA–rRNA interactions, while preserving the linear run time of RNAsnoop.

2.2 Machine learning component

Xiao *et al.* (2009) showed that the interaction energy is necessary but not sufficient to distinguish functional from non-functional snoRNA–rRNA interactions. Stability of the stems enclosing the pseudourydilation pocket as well as structural features relative to the stems and the interaction regions are equally relevant. In order to take those parameters into account we used a machine learning method [support vector machine (SVM)] to analyze the output of RNAsnoop. We developed two models depending on whether or not RNAsnoop considers the target site accessibility. We used the experimentally verified interactions from yeast (Schattner *et al.*, 2004) and human (Xiao *et al.*, 2009). When using the human interactions for testing we trained exclusively on the yeast dataset. Since the training dataset did not contain experimentally confirmed non-functional interactions, we augmented it by adding artificial ones. For each snoRNA-stem involved in a verified interaction, we let RNAsnoop run against yeast 28S and 18S sequence. All hits that had an interaction energy smaller than the one of the experimentally validated interaction and that do not target a known

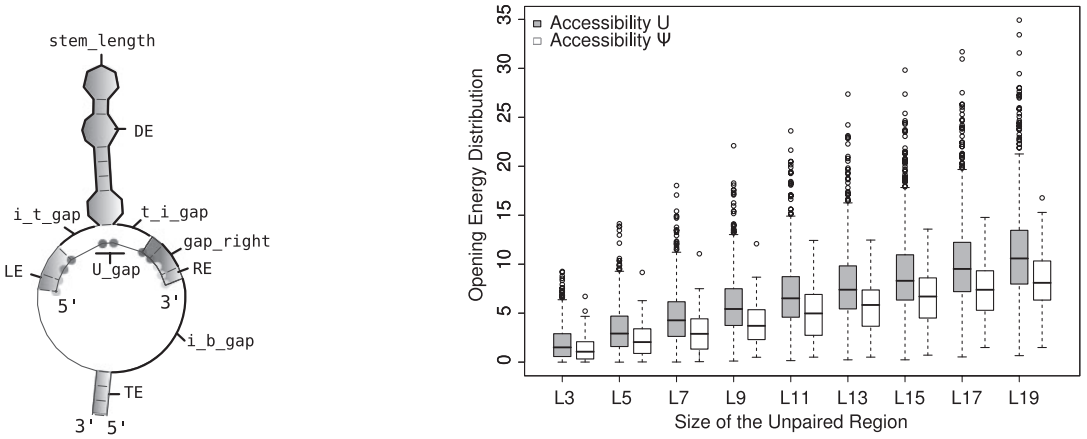

Fig. 2. Features considered in the SVM model. (Left) Structural (black bold lines) and energy features (shaded regions). `TE`: lower stem energy, `LE`: 5′ interaction energy, `DE`: upper stem energy, `RE`: 3′ interaction energy, For each nucleotide in the target, its local opening energy is represented by a gray circle, where light gray represents low local opening energy and dark gray high local opening energy. The target total opening energy (`OE`) is the sum of all local opening energies, `YE`: YE = LE + RE + TE + DE, `XE`: XE = LE + RE + DE, `dYE`: dYE = YE + OE, `t_i_gap`: number of nucleotides between the 5′ end of the upper stem and 3′ end of the 5′ interaction on the snoRNA, `U_gap`: number of nucleotides between the 3′ end of the 5′ interaction and the 5′ end of the 3′ interaction on the mRNA, `i_b_gap`: number of nucleotides between the end of the lower stem and the 3′ end of the 5′ interaction on the snoRNA, `i_t_gap`: number of nucleotides between the 5′ end of the 5′ interaction and the 5′ end of the snoRNA stem, `stem_length`: length of the upper stem, `stem_asymmetry`: difference in the number of nucleotides located in loops between the 5′ and 3′ side of the upper stem, `gap_right`: number of gaps in the 3′ interaction on the mRNA. (Right) Box-plots showing the accessibility distribution for all known uridines (gray, 1062 datapoints) and pseudouridines (white, 92 datapoints) sites in human 28S and 18S rRNAs. The target accessibility was computed by using `RNAup` on the whole length sequences of 28S and 18S rRNAs. The target size was varied between 3 and 19 nt in steps of 2 nt and was centered around the (pseudo)uridine site.

pseudouridylation site were considered non-functional. The final training dataset contained 43 positive and 103 negative interactions.

For both models we derived a set of 29 features to pass to the SVM, and then selected a subset following the approach described by Chen and Lin (2006). Features that were included at the end are described in some detail in Figure 2. We used different feature set depending on whether accessibility is taken into account or not.

For the case where the target accessibility was neglected, only five features are used, four of which describe the geometry of the interaction itself (`t_i_gap`, `U_gap`, `i_t_gap` and `gap_right`) and the length of the intervening stem `stem_length`.

For the model with accessibility, 11 features are used. In addition to features describing the geometry of the interaction (`t_i_gap`, `U_gap`, `i_b_gap`, `i_t_gap` and `gap_right`) and of the upper stem (`stem_length` and `stem_asymmetry`), we utilize the four energy values `YE`, `DE`, `XE` and `dYE` defined in the caption of Figure 2.

Training and test datasets can be found in Supplementary Tables T3 and T4.

2.3 Performance

2.3.1 Accuracy We compared the prediction accuracy of `RNAsnoop`, `snoGPS` and `fisher` on the human (Xiao *et al.*, 2009) and yeast (Schattner *et al.*, 2004) datasets of experimentally confirmed/rejected snoRNA–rRNA interactions. For a given snoRNA involved in a confirmed interaction, we determined how many target sites were predicted to bind with a better score/energy than the experimentally reported one. Table 1 summarizes these rank values for the confirmed interactions in yeast. We clearly see that `fisher` is less sensitive, detecting only 16 of the 44 interactions in yeast. Still, these 16 interactions were all ranked first, indicating that `fisher` has a high specificity. In comparison, `RNAsnoop` and `snoGPS` detect 43 and 41 of the 44 verified interactions in yeast, and 11 and 10, respectively, in human. We remark that `RNAsnoop` did not identify the interaction of snR82 with LSU-U2349, because `RNAsnoop` predicts the adjacent position LSU-U2351 as preferred target. On average, `RNAsnoop` ranks the confirmed interactions higher in the list than `snoGPS`. This trend is also seen in the ROC curve in Figure 3, where `RNAsnoop` shows a higher prediction accuracy than `snoGPS`.

In human, `RNAsnoop` performs better than `snoGPS`. In particular, the SVM version successfully rejects the four non-functional snoRNA–rRNA interactions and successfully ranks 11 out of the 12 confirmed interactions first (Table 2). Still, one of the confirmed interaction was rejected by the SVM.

Further, we looked at the false positive rate of `RNAsnoop`. To this aim, we considered the putative targets of orphan snoRNA HBI-36 (Cavaillé *et al.*, 2000), a brain-specific snoRNA found in all vertebrates (Gardner *et al.*, 2009) returned by `RNAsnoop`. We downloaded from `BIOMART` (Haider *et al.*, 2009) the unprocessed transcript sequences that are expressed in brain and that have homologs in chicken. We did not limit ourselves to exons as it was proven that at least C/D snoRNA can bind intronic regions and subsequently influence the splicing process (Bazeley *et al.*, 2008). We downloaded a total of 9429 unspliced sequences, summing a total of 0.75 Gb, roughly a quarter of total human genome. For each sequence, we used `RNAplfold` to compute the local accessibility (Bompfünewerer *et al.*, 2008). Based on this `RNAsnoop` returned a total of 1 278 134 putative targets (515 751 hits for the 5′ stem and 762 383 hits for the 3′ stem) with a SVM

Table 1. Prediction comparison of RNAsnoop (abbreviated as RNAsn.), snoGPS and Fisher for the known snoRNA–rRNA interactions in yeast

snoRNA	Target	Position	snoGPS	fisher	RNAsn.	RNAsn. A	snoRNA	Target	Position	snoGPS	fisher	RNAsn.	RNAsn. A
snR11	25S	2416	3	–	12	14	snR10	25S	2923	2	1	28	26
snR161	18S	632	6	1	8	5	snR46	25S	2865	1	1	1	1
snR161	18S	766	1	–	11	2	snR49	18S	120	3	1	1	1
snR189	18S	466	2	1	1	1	snR49	18S	211	2	–	5	5
snR189	25S	2735	1	–	1	1	snR49	18S	302	1	–	5	4
snR191	25S	2258	1	–	5	2	snR49	25S	990	4	–	–	1
snR191	25S	2260	99	–	8	1	snR5	25S	1004	3	1	1	1
snR3	25S	2129	4	–	1	1	snR5	25S	1124	1	–	8	1
snR3	25S	2133	1	–	1	1	snR8	25S	960	68	–	3	5
snR3	25S	2264	2	–	3	1	snR8	25S	986	55	1	2	3
snR31	18S	999	1	1	1	1	snR80	18S	759	–	–	2	2
snR32	25S	2191	1	1	1	1	snR80	25S	776	–	–	2	2
snR33	25S	1042	1	1	1	1	snR81	25S	1052	57	1	2	1
snR34	25S	2826	2	–	1	1	snR82	25S	2349	1	1	–	–
snR34	25S	2880	1	–	1	1	snR82	25S	2351	1	–	1	2
snR35	18S	1191	1	–	1	1	snR82	25S	1110	–	–	2	4
snR36	18S	1187	12	1	7	2	snR83	18S	1290	1	–	58	7
snR37	25S	2944	1	–	2	2	snR83	18S	1415	4	–	1	1
snR42	25S	2975	1	1	4	1	snR84	25S	2266	1	–	2	2
snR43	25S	966	1	–	1	1	snR85	18S	1181	1	1	1	1
snR44	18S	106	1	–	2	2	snR86	25S	2314	13	–	3	1
snR44	25S	1056	2	1	1	2	snR9	25S	2340	33	–	18	19

RNAsn. A, accessibility version of RNAsnoop.

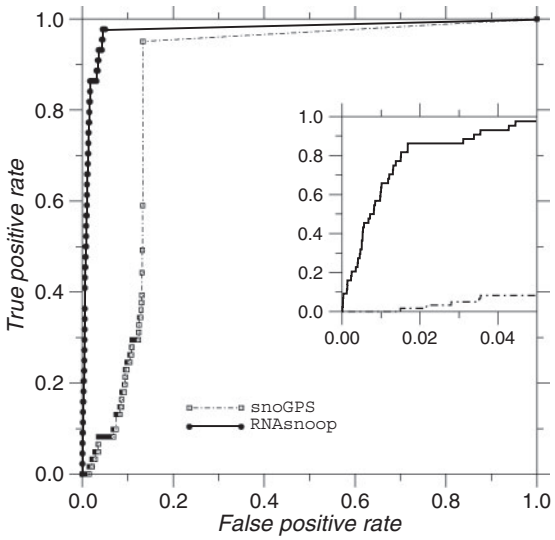

Fig. 3. ROC curve for RNAsnoop and snoGPS on the yeast data set (Schattner *et al.*, 2004). RNAsnoop was used without the SVM functionality.

Table 2. Prediction performance in human for snoGPS, RNAsnoop (RNAsn.), RNAsnoop with accessibility (RNAsn. A) and the SVM in human

snoRNA	Target	Position	Type	snoGPS	RNAsn.	RNAsn. A	SVM
ACA19_1	28S	3709	+	1	1	1	1
ACA19_2	28S	3618	+	25	2	1	1
ACA19_1	18S	863	–	10	1	4	–
ACA19_1	18S	866	–	10	–	–	–
ACA24_1	18S	863	+	–	1	1	1
ACA24_2	18S	612	–	86	3	6	–
ACA28_1	18S	815	+	1	4	1	1
ACA28_2	18S	866	+	–	2	4	1
ACA42_1	18S	572	–	3	4	19	–
ACA42_2	18S	109	+	1	1	1	1
ACA50_1	18S	34	+	1	1	1	–
ACA50_2	18S	105	+	2	1	1	1
ACA62_1	18S	34	+	3	24	1	1
ACA62_2	18S	105	+	2	1	1	1
ACA67_1	18S	572	+	2	2	1	1
ACA67_2	18S	109	+	1	1	1	1

The numbers represent the rank of the interaction for the corresponding snoRNA stem. In column 'Type', +, – represent experimentally confirmed or rejected interactions, respectively. When using the human interactions for testing, we trained the SVM exclusively on the yeast dataset.

P-value >0.5. This corresponds to one hit every 586 nt. At a *P*-value of 0.844, the false positive rate drops to 0.00001 predictions per nucleotide (Supplementary Fig. S6). While the number of false positives diminishes uniformly with increasing *P*-values, the energy dependency of the false positives is sigmoid shaped. The false positives rate grows slowly between −40 kcal/mol and −30 kcal/mol, then increases strongly between −30 kcal/mol and −20 kcal/mol before reaching a plateau between −20 kcal/mol and −10 kcal/mol. A false positive rate of 0.00001 predictions per nucleotide is reached for an energy of −28 kcal/mol (Supplementary Fig. S6).

2.3.2 Run time We compared the run time of RNAsnoop with that of snoGPS and RNAhybrid. We modified fisher to turn it

into a target finder; the resulting run time, however, was so high that we decided not to evaluate it further. RNAhybrid uses a dynamic programming algorithm to find putative miRNA–targets and has a run time of $\mathcal{O}(|x| \cdot |y|)$. Since the run time of RNAsnoop is linear in the target size but quadratic in the snoRNA size, we varied the length of both sequences. Since H/ACA snoRNA stems vary greatly in length (Bally *et al.*, 1988; Torchet *et al.*, 2005), we incremented the snoRNA stem size in steps of 30 nt from 60 up to 420 nt, keeping the target RNA length fixed to 5000 nt. Conversely, the target length was varied between 1000 and 256 000 nt with a snoRNA stem length set to 200. We set the threshold for each program so that they returned at most one hit. Independently of the snoRNA or target sequence size, snoGPS and RNAsnoop have a similar run time. They are around 15 times faster than RNAhybrid (Supplementary Figs S2 and S3).

3 A COMPARATIVE VERSION

The use of alignments in the target search can further help to find real snoRNA–RNA interactions. On one hand, the absence of conserved target site in closely related species may indicate that the proposed interaction does not occur in nature. The presence of compensatory mutations between the snoRNA binding bucket and the target site, on the other hand, can lend further credibility to single-sequence target predictions (Chen *et al.*, 2007).

The alignment extension of RNAsnoop is based on the same approach used in RNAalifold (Bernhart *et al.*, 2008; Hofacker *et al.*, 2002), where a thermodynamic energy minimization folding algorithm is coupled with a simple scoring model to assess evolutionary conservation. As in the single sequence algorithm, the upper stem is modeled as an unbranched fold by a slightly modified RNAalifold algorithm. The interaction part uses the same approach as RNAalifold, with the sole difference that only interior loops are allowed between the snoRNA and its target.

For an efficient analysis of data, we provide and recommend the perl script SNOOPY. It uses both the SVM as well as the homology information to predict putative target interactions. SNOOPY takes as input a snoRNA alignment and a target alignment. In a first step SNOOPY uses mLocARNA to obtain sequence/structure alignments of the snoRNAs (Will *et al.*, 2007). If the sum of scores of mLocARNA pairwise alignments for a sequence is <2500, then the sequence is discarded. Duplicates and sequences belonging to species that are present in only one of the two alignments are also removed. SNOOPY preselects possible targets in a user-defined reference organism by means of the single-sequence version of RNAsnoop and one of the two SVM models. For each reported targets, SNOOPY extracts the corresponding slice from the alignments and then realigns the corresponding subsequences with Clustalw (Thompson *et al.*, 1994). Target sequences for which the pairwise alignment score is below a threshold, or which do not exhibit a U residue at the previously predicted site, are removed together with the snoRNA sequences from the same organisms. Whenever the number of retained sequences is above a user-defined threshold, the alignment version of RNAsnoop is applied. Finally, SNOOPY reports for each snoRNA alignment a user-specified number of putative interactions. These interactions can be ranked either by their SVM score or by the single sequence interaction energy for the reference organism.

4 APPLICATIONS

In order to test the usability of RNAsnoop, we consider the problems of finding snoRNAs associated with 'orphan' pseudouridylation sites in human rRNAs. Although the role of snoRNAs in locating target uridine residues was discovered more than a decade ago, there are still a few pseudouridylation sites in human rRNAs (Maden and Wakeman, 1988; Ofengand and Bakin, 1997) for which the responsible snoRNAs have not yet been determined. We used the single sequence version of RNAsnoop to predict the possible snoRNAs that may pseudouridylate these orphan sites. For this we used all the known human H/ACA sequences reported in snoRNA-LBME-db (Lestrade and Weber, 2006) and tested them against the 11 reported orphan sites in the human LSU and SSU. Based on the currently available snoRNA data, eight orphan sites can be mapped to existing snoRNA stems. Interestingly, two orphan snoRNAs (ACA38B, ACA51), and two stems, for which no function was reported, were among the predictions. Additionally, four stems with known targets were predicted to target four of the orphan sites. The predicted interactions are listed in Table 3, Figure 4 and Supplementary Figure S4.

We used SNOOPY to assign putative targets to the five orphan snoRNAs found in *Drosophila* (Or-aca1, Or-aca2, Or-aca3, Or-aca4 and Or-aca5). For each orphan snoRNAs reported in Flybase (Ashburner and Drysdale, 1994), we searched for homologous sequences in the 11 other *Drosophila* species by using blast (Altschul *et al.*, 1990). For each species, the sequence with the highest homology with *D.melanogaster* was selected. The sequences were then aligned with mLocARNA, a variant of the Sankoff algorithm. For each snoRNA, the full-length alignment was then divided into a 5′ and 3′ stem alignments.

The rRNA alignments were retrieved from the arb-silva database (Pruesse *et al.*, 2007). In order to get the best possible alignments, we realigned them with Clustalw, Muscle (Edgar, 2004), and RNAsalsa (Stocsits *et al.*, 2009). The quality of the alignments was assessed by determining how well the conserved pseudouridylation sites in *D.melanogaster* and *Homo sapiens* were aligned in the 12 drosophilid rRNA sequences. Based on this quality measure, RNAsalsa was found to perform best (Supplementary Tables T1 and T2). Alignments of snRNAs were taken from Marz *et al.* (2008).

Of the five orphan snoRNAs, only Oaca-4 was reported to have a target. We predict that the first stem modifies U2499 on the 28S rRNA (Fig. 5 and Supplementary Fig. S5). This target site is interesting since it was reported to be pseudouridylated (Giordano *et al.*, 1999), but no corresponding snoRNA is known. Moreover, in human and yeast, this position which correspond to U3674 in human and U2191 in yeast, is conserved and pseudouridylated (Lestrade and Weber, 2006). U3674, finally, remains an orphan site in human.

Interestingly, both the target and binding buckets are completely conserved from *D.melanogaster* to *D.willistoni*, see Figure 5. On the other hand, 6 out of the 12 bp found in the upper stem exhibit compensatory mutations.

The fact that no credible targets have been predicted for the remaining four orphan snoRNAs is not unexpected. First, snoRNAs have also been implicated in modifying 'non-canonical targets' such as mRNAs (Bazeley *et al.*, 2008; Kishore and Stamm, 2006; Uliel *et al.*, 2004), some cause cleavage of pre-rRNAs (Fayet-Lebaron *et al.*, 2009), and Taft *et al.* (2009) recently showed that Or-aca5 is

Table 3. Predicted snoRNAs targeting the orphan pseudouridines in human ribosomal RNAs

rRNA	Position	snoRNA	Stem	Function	SVM-score	Energy
18S	681	ACA55	2	18S-36	0.76	−34.32
18S	918	ACA13	1	18S-1248	0.81	−35.90
28S	1523	SNORA38B*	1	–	0.66	−18.08
28S	1849	–	–	–	–	–
28S	3674	–	–	–	–	–
28S	3747	ACA52	2	–	0.87	−28.94
28S	3749	–	–	–	–	–
28S	3863	U71c	2	18S-406	0.53	−19.14
28S	4266	ACA64	1	–	0.75	−32.00
28S	4323	ACA51*	2	–	0.63	−20.39
28S	4501	ACA10	1	28S-4491	0.54	−15.00

No snoRNAs were found for position 1849, 3674 and 3749 on rRNA 28S. ACA51 and SNORA38B are orphan snoRNAs while ACA52-2 and ACA64-1 are orphan stems.

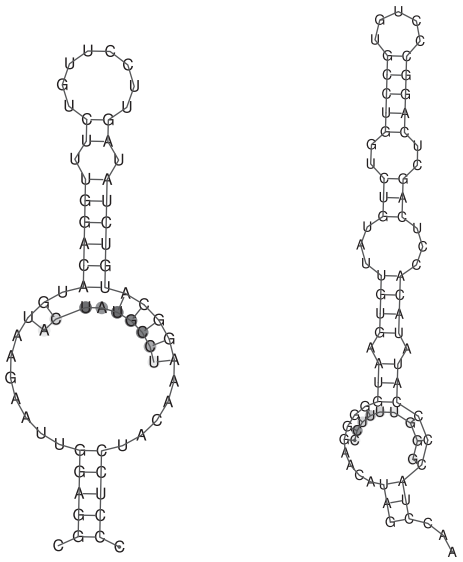

Fig. 4. Structure of the interactions between human Ψ orphan sites and orphan snoRNAs returned by RNAsnoop. From left to right: SNORA38B-1:28S−1523, ACA51-2:28S−4323, where, i.e. ACA51-2:28S−4323, means that the second stem of ACA51 binds to position 4323 on rRNA 28S. The single nucleotide opening energy for the target is gray coded and is represented as circles on top of the corresponding nucleotide. Structures drawings were produced automatically by RNAsnoop.

processed by *Dicer*, suggesting a function in the RNA interference pathway.

5 DISCUSSION

We presented here RNAsnoop, a tool specifically designed to predict complex H/ACA snoRNA–RNA interactions that are outside the scope of conventional RNA–RNA prediction tools. In contrast with previous tools, it uses a dynamic programming approach coupled with a nearest-neighbor energy model to identify putative targets. This allows RNAsnoop to capture structural and

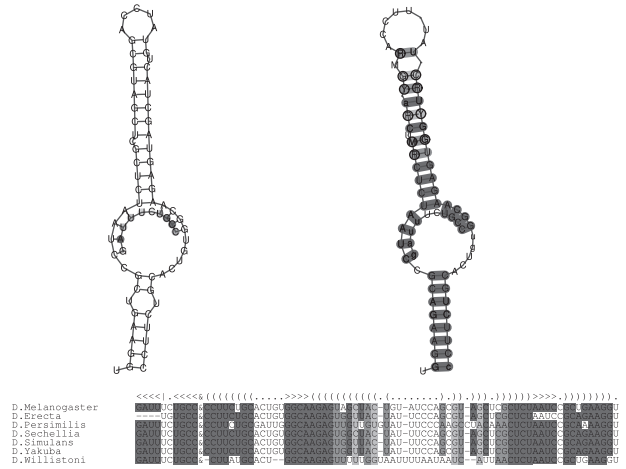

Fig. 5. Structure of the interactions between Or-aca4 and its putative target. (Left) Single sequence structure and (right) multiple sequences structure. (Below) Alternative representation of the multiple sequences structure. The consensus structure is represented in dot bracket format on top of the target and snoRNA alignments. The angle brackets represent intermolecular base pairs and the braces represent intramolecular base pairs. The & column separates the alignment of the snoRNA sequences on the right side and the corresponding slice of the target sequences alignment on the left side. For the multiple sequences and alignment figures, the shade in the order light, middle and dark gray indicate 1–3 different types of base pairs.

energetic features essential for correctly predicting snoRNA–target interactions (Xiao *et al.*, 2009). Coupled with a SVM classification, SNOOPY achieves good performance ranking first 11 out of 12 confirmed snoRNA–mRNA interactions in human and excluding all experimentally rejected interactions. These good results should, however, not be overestimated as both the training and test datasets are small and were extracted from only two species.

The run time of RNAsnoop is comparable with that of snoGPS, and scales linearly with the length of the target sequence. Together with the improved accuracy, this makes RNAsnoop not only suitable for target search in rRNA and snRNA sequences or in specific putative mRNA candidates, but also for large-scale genome-wide surveys.

Funding: European Union under the auspices of the FP-6 SYNLET and the FP-7 QUANTOMICS project (in part); the Austrian GEN-AU project 'Regulatory Noncoding RNA' (in part).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Ashburner,M. and Drysdale,R. (1994) Flybase–the *Drosophila* genetic database. *Development*, **120**, 2077–2079.

Bachellerie,J.P. *et al.* (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.

Bally,M. *et al.* (1988) SnR30: a new, essential small nuclear RNA from *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **16**, 5291–5303.

Bazeley,P.S. *et al.* (2008) snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions. *Gene*, **408**, 172–179.

Bernhart,S.H. *et al.* (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.

Bernhart,S. *et al.* (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614–615.

Bompfünewerer,A.F. *et al.* (2008) Variations on RNA folding and alignment: Lessons from benasque. *J. Math. Biol.*, **56**, 119–144.

Cavaillé,J. *et al.* (2000) Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc. Natl Acad. Sci. USA*, **97**, 14311–14316.

Chen,Y.-W. and Lin,C.-J. (2006) Combining SVMs with various feature selection strategies. In Guyon,I. *et al.* (eds) *Feature Extraction, Foundations and Applications*, Studies in Fuzziness and Soft Computing. Springer, Berlin/Heidelberg, pp. 315–324.

Chen,C.L. *et al.* (2007) Exploration of pairing constraints identifies a 9 base-pair core within box C/D snoRNA-rRNA duplexes. *J. Mol. Biol.*, **369**, 771–783.

Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113–132.

Fayet-Lebaron,E. *et al.* (2009) 18S rRNA processing requires base pairings of snR30 H/ACA snoRNA to eukaryote-specific 18S sequences. *EMBO J.*, **28**, 1260–1270.

Freyhult,E. *et al.* (2008) `Fisher`: a program for the detection of H/ACA snoRNAs using MFE secondary structure prediction and comparative genomics — assessment and update. *BMC Res. Notes*, **1**, 49.

Gardner,P.P. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, 136–140.

Giordano,E. *et al.* (1999) minifly, a *Drosophila* gene required for ribosome biogenesis. *J. Cell Biol.*, **144**, 1123–1133.

Haider,S. *et al.* (2009) BioMart central portal–unified access to biological data. *Nucleic Acids Res.*, **37**, 23–27.

Höchsmann,T. *et al.* (2006) Thermodynamic matchers: strengthening the significance of RNA folding energies. In Markstein,P. and Xu,Y. (eds) *Computational Systems Bioinformatics, CSB 2006*. World Scientific, Singapore, pp. 111–121.

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

Kishore,S. and Stamm,S. (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science*, **311**, 230–232.

Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**, 158–162.

Lu,Z.J. *et al.* (2006) A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res.*, **34**, 4912–4924.

Maden,B.E.H. and Wakeman,J.A. (1988) Pseudouridine distribution in mammalian 18 S ribosomal RNA. A major cluster in the central region of the molecule. *Biochem. J.*, **249**, 459–464.

Marz,M. *et al.* (2008) Evolution of spliceosomal snRNA genes in metazoan animals. *J. Mol. Evol.*, **67**, 594–607.

Mathews,D.H. *et al.* (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Mückstein,U. *et al.* (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182.

Ni,J. *et al.* (1997) Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. *Cell*, **89**, 565–573.

Ofengand,J. and Bakin,A. (1997) Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryotes, prokaryotes, archaebacteria, mitochondria and chloroplasts. *J. Mol. Biol.*, **266**, 246–268.

Pruesse,E. *et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–7196.

Reeder,J. *et al.* (2007) `Locomotif`: from graphical motif description to RNA motif search. *Bioinformatics*, **23**, i392–i400.

Schattner,P. *et al.* (2004). Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.*, **32**, 4281–4296.

Stocsits,R.R. *et al.* (2009) Accurate and efficient reconstruction of deep phylogenies from structured RNAs. *Nucleic Acids Res.*, **37**, 6184–6193.

Tafer,H. and Hofacker,I.L. (2008a) `RNAplex`: a fast tool for RNA-RNA interaction search. *Bioinformatics*, **24**, 2657–2663.

Tafer,H. and Hofacker,I.L. (2008b) RNAplex a fast interaction tool incorporating target site accessibility. In *ISMB 2008*, poster, Toronto, Canada.

Taft,R.J. *et al.* (2009) Small RNAs derived from snoRNAs. *RNA*, **15**, 1233–1240.

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.

Torchet,C. *et al.* (2005) The complete set of H/ACA snoRNAs that guide rRNA pseudouridylations in *Saccharomyces cerevisiae*. *RNA*, **11**, 928–938.

Uliel,S. *et al.* (2004) Small nucleolar RNAs that guide modification in trypanosomatids: repertoire, targets, genome organisation, and unique functions. *Int. J. Parasitol.*, **34**, 445–454.

Will,S. *et al.* (2007) Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comp. Biol.*, **3**, e65.

Xiao,M. *et al.* (2009) Functionality and substrate specificity of human box H/ACA guide RNAs. *RNA*, **15**, 176–186.