RNAsnoop: efficient target prediction for H/ACA snoRNAs

Hakim Tafer\(^1,2,3,\ast\), Stephanie Kehl\(^2,3\), Jana Herte\(2,3\), 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, Härtselstrasse 16-18, D-04107 Leipzig, \(^4\)Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, \(^5\)RNomics Group, Fraunhofer Institut für 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 a support vector machine (SVM)-based machine learning approach to distinguish true binding sites from spurious solutions and a system to evaluate 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.

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 28, 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 (Hochsmann 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 \[\text{snoGPS} \] (Schattner et al., 2004) and \[\text{Fisher} \] (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.

\(^\ast\)To whom correspondence should be addressed.
The index snoRNA consisting of the l.h.s. binding region

\[ M \]

loop

Since satisfies the recursion pairs and symmetrical loops of lengths 2 and 4. Thus, the left part

the interaction region can contain only single and tandem

y

2006; Mathews et al.

interaction structure up to positions \( i \) the delimiting base pairs (\( \text{loop composed of the sequences} \)). Analogously, \( T(y[p,q]) \) is the energy of an interior loop composed of the sequences \( y[p,q] \) and \( y[q,v] \), again including the delimiting base pairs \( (y[p,q]) \) and \( (y[q,v]) \).

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

\[ L_{i,j} = \min_{k=1,2,3} \left( L_{i-k,j+k} + T(s[i-k],y[j,j+k]) \right) \]

The index \( i \) runs along the target RNA \( s \), while \( j \) refers to the position on the snoRNA \( y \). To ensure that all interactions start inside the interaction 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 \( R_{j} \). It thus extends a r.h.s. binding region or refers to its first base pair. In the latter

case, nucleotide \( y_{i,j} \) is the uracil that is pseudouridylated. The corresponding recursion reads

\[ R_{i,j} = \min_{k,l \geq 2} \left( R_{i-k,j+l} + T(s[i-k],y[j,j+l]) \right) \]

For each \( i \), the best binding energy at target position \( i \) is max \( R_{i,j} \).

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 \( O(|y|^2) \) and \( O(|y|^3) \), respectively. The space requirements for the \( L \) and \( R \) arrays are limited to \( 5 \times |y| \) independent of the target \( |s| \) 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 \( i \)–\( j \). The space complexity for \( L \) and \( R \) respectively is \( O(|y|^3) \). Hence, the total run time is thus \( O(|y|^2 + |y|^3) \), 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 RNAsnoop (Mückstein et al., 2006) or RNAsnoop (Bernhart et al., 2006; Bomphrey 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 RNAsnoop (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 pseudouridylation 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 validated 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 scan 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
T3 and T4. 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.

The 3′ (white, 92 datapoints) sites in human 28S and 18S rRNAs. The target accessibility was computed by using i_b_gap

\[ \text{i_b_gap} \]

U_gap

\[ \text{U_gap} \]

values for the confirmed interactions in yeast. We clearly see that than the experimentally reported one. Table 1 summarizes these rank

many target sites were predicted to bind with a better score/energy snoRNA involved in a confirmed interaction, we determined how confirmed/rejected snoRNA–rRNA interactions. For a given snoRNA-validated interaction, we considered the putative targets of orphan snoRNA 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_b_gap, 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 BICMBART (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 CD snoRNA can bind intronic regions and subsequently influence the splicing process (Bazely et al., 2008). We 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.) and snoGPS and Fisher for the known snoRNA–rRNA interactions in yeast

<table>
<thead>
<tr>
<th>snoRNA Target</th>
<th>Position</th>
<th>snoGPS</th>
<th>fisher</th>
<th>RNAsn.</th>
<th>RNAsn. A</th>
<th>RNAsn. RNAsn. A</th>
</tr>
</thead>
<tbody>
<tr>
<td>snR11</td>
<td>25S 2416</td>
<td>3</td>
<td>–</td>
<td>12</td>
<td>14</td>
<td>snR10 25S 2923</td>
</tr>
<tr>
<td>snR161</td>
<td>18S 632</td>
<td>6</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>snR46 25S 2665</td>
</tr>
<tr>
<td>snR161</td>
<td>18S 766</td>
<td>1</td>
<td>–</td>
<td>11</td>
<td>2</td>
<td>snR49 18S 120</td>
</tr>
<tr>
<td>snR189</td>
<td>18S 466</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>snR49 18S 211</td>
</tr>
<tr>
<td>snR189</td>
<td>25S 2735</td>
<td>1</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>snR49 18S 302</td>
</tr>
<tr>
<td>snR191</td>
<td>25S 2258</td>
<td>1</td>
<td>–</td>
<td>5</td>
<td>2</td>
<td>snR49 25S 990</td>
</tr>
<tr>
<td>snR191</td>
<td>25S 2260</td>
<td>99</td>
<td>–</td>
<td>8</td>
<td>1</td>
<td>snR5 25S 1004</td>
</tr>
<tr>
<td>snR3</td>
<td>25S 2129</td>
<td>4</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>snR5 25S 1124</td>
</tr>
<tr>
<td>snR3</td>
<td>25S 2133</td>
<td>1</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>snR5 25S 960</td>
</tr>
<tr>
<td>snR3</td>
<td>25S 2264</td>
<td>2</td>
<td>–</td>
<td>3</td>
<td>1</td>
<td>snR8 25S 996</td>
</tr>
<tr>
<td>snR31</td>
<td>18S 999</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>snR80 18S 759</td>
</tr>
<tr>
<td>snR5</td>
<td>25S 2191</td>
<td>1</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>snR80 25S 776</td>
</tr>
<tr>
<td>snR33</td>
<td>25S 1042</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>snR80 25S 1052</td>
</tr>
<tr>
<td>snR34</td>
<td>25S 2826</td>
<td>2</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>snR82 25S 2349</td>
</tr>
<tr>
<td>snR34</td>
<td>25S 2880</td>
<td>1</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>snR82 25S 2551</td>
</tr>
<tr>
<td>snR35</td>
<td>18S 1191</td>
<td>1</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>snR82 25S 1110</td>
</tr>
<tr>
<td>snR36</td>
<td>18S 1187</td>
<td>12</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>snR83 18S 1290</td>
</tr>
<tr>
<td>snR37</td>
<td>25S 2944</td>
<td>1</td>
<td>–</td>
<td>2</td>
<td>2</td>
<td>snR83 18S 1415</td>
</tr>
<tr>
<td>snR42</td>
<td>25S 2975</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>snR84 25S 2266</td>
</tr>
<tr>
<td>snR43</td>
<td>25S 966</td>
<td>1</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>snR85 18S 1181</td>
</tr>
<tr>
<td>snR44</td>
<td>18S 106</td>
<td>1</td>
<td>–</td>
<td>1</td>
<td>2</td>
<td>snR86 25S 2314</td>
</tr>
<tr>
<td>snR44</td>
<td>25S 1056</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>snR89 25S 2340</td>
</tr>
</tbody>
</table>

RNAsn. A, accessibility version of RNAsnoop.

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

| snoRNA Target | Position | Type | snoGPS | RNAsn. | RSA 

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

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. 6). 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. 6).

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 $O(\lambda^3)$ [17,35]. Since the run time of RNAhybrid is linear in the target size and 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 500 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).

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-dic (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 snoRNA 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 RNAsaLisa (Stoccsits 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 RNA sequences. Based on this quality measure, RNAsaLisa was found to perform best (Supplementary Tables T1 and T2). Alignments of snoRNAs 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 (Payette-Lebaron et al., 2009), and Taft et al. (2009) recently showed that Or-ac5 is
Table 3. Predicted snoRNAs targeting the orphan pseudouridines in human ribosomal RNAs

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

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 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 GENAU project ‘Regulatory Noncoding RNA’ (in part).

Conflict of Interest: none declared.

REFERENCES


