Evaluating the predictability of conformational switching in RNA

Björn Voss*, Carsten Meyer and Robert Giegerich

1International NRW Graduate School in Bioinformatics and Genome Research
and 2Institute for Bioinformatics, Bielefeld University, P.O. Box 100 131, 33501 Bielefeld, Germany

Received on September 25, 2003; revised and accepted on December 15, 2003
Advance Access publication February 12, 2004

ABSTRACT

Motivation: There are various cases where the biological function of an RNA molecule involves a reversible change of conformation. paRNAss is a software approach to the prediction of such structural switching in RNA. It is based on three hypotheses about the secondary structure space of a switching RNA molecule that can be evaluated by RNA folding and structure comparison. In the positive case, the predicted structural switching must be verified experimentally.

Results: After reviewing the strategy used in paRNAss, we present recent improvements on the algorithmic level of the approach, and the results of an evaluation procedure, comprising 1500 RNA sequences. It could be shown that the paRNAss approach performs well on known examples for conformational switching in RNA. The overall number of positive predictions was small, whereas for human 3′ UTRs, representing regulatory important regions, it was substantially higher than for arbitrary natural and random sequences.

Availability: paRNAss is available as a Web service at http://bibiserv.techfak.uni-bielefeld.de/parnass
Contact: bvoss@techfak.uni-bielefeld.de

Supplementary information: Detailed information on the analyses summarized in Table 1 can be found at http://bibiserv.techfak.uni-bielefeld.de/parnass/examples.html

1 INTRODUCTION

1.1 Conformational switching in RNA

RNA fulfils a broad range of functions in living cells. In messenger RNA (mRNA), the plain sequence of bases, the primary structure, is sufficient to determine the sequence of amino acids of the encoded protein. In other cases, e.g. in ribosomal RNA or transfer RNA, a certain three-dimensional structure is necessary for the correct function. This structure is not rigid, and sometimes even a significant change of shape is required. Such conformational switches have been proven or are suspected to be involved in several important processes: regulation of gene expression in prokaryotes by attenuation (Fayat et al., 1983), translational regulation of Escherichia coli ribosomal protein S15 (Philippe et al., 1995), regulation of self-cleavage activity of Hepatitis Delta Virus (Lazinski and Taylor, 1995), translocation process in protein biosynthesis (Wool et al., 1992) trans-splicing in trypanosomes (LeCuyer and Crothers, 1993), splicing of pre-mRNA by spliceosomes (Madhani and Guthrie, 1992) and others.

1.2 The secondary structure space of an RNA sequence

The tertiary structure of an RNA is very hard to determine, both in vitro and in silico. As the building blocks of the tertiary structure are given by the secondary structure and the latter is much easier to access, one commonly uses the secondary structure as an approximation of the tertiary structure. For this reason, the term structure space from now on refers to the secondary structure space.

The structure space of a given RNA sequence is described by a thermodynamic model, given by experimentally determined energy parameters (Mathews et al., 1999). Base pairing and base pair stacking are structure stabilizing factors associated with negative energy contributions, depending on the nucleotides involved. Single-stranded parts (loops) have destabilizing effects associated with positive energy contributions, mainly depending on the length of the single-stranded parts. Based on these energy contributions, a free energy value for each structure is calculated. The structure attaining minimal free energy is called the mfe-structure. Since the energy parameters are determined in vitro under conditions that are near but not identical to in vivo conditions, it is not guaranteed that the predicted mfe-structure corresponds to the native structure. As a consequence of this and due to the possibility of several important conformations, one is also interested in energetically near-optimal structures.

We consider two structures as neighbours if they differ by just opening or closing 1 bp. Neighbouring structures are very likely to achieve a similar energy value. This leads to the presence of many similar structures when looking at near-optimal structures within a certain percentage of the mfe value.

*To whom correspondence should be addressed.
The structures one is really interested in are the local energy minima, having the lowest free energy with respect to all their neighbours. Since the size of the structure space is exponential in the length of the sequence (Waterman, 1995), it is very likely that a large number of such local energy minima are present in the structure space. Thinking of the structure space as a landscape suggests the notion of a valley for a local minimum and all its neighbouring structures that can be reached by opening (or closing) base pairs and thereby always increasing the free energy. We call the set of all structures that belong to one valley a family of structures. A structure having two neighbours from different valleys corresponds to a saddle point in the landscape.

The shape of the structure space gives hints about the structural properties of the RNA molecule. The existence of one valley associated with a near-optimal free energy value being much ‘deeper’ than all other valleys hints at a function through structure property. If the landscape is more rugged and not bearing a significant feature, the structure of the RNA might be less important (mRNAs).

In the case of a conformational switch, we propose that the structure space satisfies the following hypotheses:

(1) There must be a local minimum in the structure space close to the overall energetic minimum. Their structures must be significantly different, in order to represent two positions of the switch with different regulatory function.

(2) The structures residing in these minima must be clearly separated by an energy barrier (Flamm et al., 2002) to ensure that each conformation is stable and switching can only be triggered by external events.

(3) The structure space must not provide another local minimum close to the overall energetic minimum as we assume that the RNA automatically finds the alternative position once the change is triggered.

Again, the folding space can contain tens or hundreds of local minima. Only by checking all three conditions do we achieve reliable predictions. The application of the first and second conditions filters out RNAs whose energy landscape provides only one valley, whereas the third condition rejects those with three or more.

2 ALGORITHM

2.1 A run through a paRNAss experiment

We first provide an overview; algorithmic details will follow in Section 3. A paRNAss experiment takes five steps (examples mentioned below refer to the Spliced Leader of Leptomonas collosoma):

Step 1: sampling the structure space. Using an RNA folding program, we draw a sample set \( S = \{s_1, \ldots, s_p\} \) from the folding space of our target RNA. If Hypotheses 1 and 3 are fulfilled, this set should contain members of two families of structures. In the other case, this set may contain one or more than two families, which leads to classification as a non-switch in the subsequent steps.

Step 2: pairwise distance calculation. For all \( s_i, s_j \in S \), we calculate their pairwise distance, \( d_{\text{EB}}(s_i, s_j) \). We do so for the energy barrier distance, \( \delta_{\text{EB}} \), and at least one other distance measure, \( \delta_2 \), on the structure space and plot the results in a \( \delta_{\text{EB}}, \delta_2 \) coordinate system. If both elements in a pair are from the same structure family, their distance should be small. Conversely, if they are from different families, their distance should be large. Thus, the plotted distance diagram should exhibit two clouds of points, one in the lower left, one in the upper right (Fig. 1).

Step 3: clustering. We use a standard clustering algorithm to split \( S \) into two disjoint clusters, \( C_1 \) and \( C_2 \), based on the pairwise distances under either \( \delta_{\text{EB}} \) or \( \delta_2 \).

Step 4: consensus structure calculation. For each cluster \( C_i \), we derive a consensus structure \( c_i \) by first taking all base pairs present in the majority of the members of \( C_i \), and then reapplying the folding algorithm with these base pairs fixed.
predictability of conformational switching in RNA

2.2 Recent improvements of paraRNAss

The previously described approach was first proposed in Giegerich et al. (1999). Due to a lack of sufficiently powerful algorithms, for the pairwise distance calculation and especially for generating the structure sample, the results of this early version were not as expected. The following improvements and the availability of more computational power increased the reliability and speed of this strategy and made possible the evaluation procedure described in Section 4.

2.2.1 Energy model

The use of an up-to-date energy model (Mathews et al., 1999; Walter et al., 1994) is essential to obtain biologically relevant foldings. paraRNAss now takes advantage of the newest versions of RNAsubopt and RNAfold [Vienna RNA package (Hofacker et al., 1994, Version 1.4)].

2.2.2 Restricting the search space to canonical structures

Due to the calculation of pairwise distances, the operating time of paraRNAss is quadratic in the size of the sample set. To increase the performance, it is therefore favourable to have a smaller sample set. This is achieved by calculating only canonical structures, meaning structures without isolated (unstacked) base pairs. Since isolated base pairs are energetically unstable, the restriction to canonical structures reduces the size of the sample set without losing relevant structures. Furthermore, this restriction does not increase only the performance but also the distinctiveness power as in some cases the occurrence of isolated base pairs in mainly unpaired regions has a negative effect on the distance measures, which may imply uncertainties in the interpretation of the distance plots.

2.2.3 An improved distance measure based on RNA secondary structure alignment

Good performance of the distance measure is essential for the paraRNAss approach. Many of the commonly used methods are not free from artefacts. Recently, a promising distance measure, $d_{TAD}$, based on tree alignments has become available, implemented in the tool RNAforester (Höchsmann et al., 2003). While paraRNAss still provides other distance measures, we now prefer $d_{TAD}$.

3 METHODS

3.1 Generating the structure sample

We use RNAsubopt (Vienna RNA package 1.4) to enumerate the complete set of (sub)optimal structures within an energy threshold of the mfe value. We no longer use MFOLD (Zuker et al., 1999) because it calculates a heuristic (and therefore possibly biased) subset of all feasible structures. It might therefore miss relevant suboptimal structures and hence produce artefacts in the distance plots. Parameters for the generation of the structure sample are the folding temperature, $T$ (in °C), the suboptimality threshold, $P$ (in kcal/mol) and a bound, $N$, on the number of structures. If the number of structures generated under the given settings of $T$ and $P$ exceeds $N$, $N$ of them are randomly chosen as the sample set. Typical values are $T = 37$, $P = 3$, $N = 50$.

3.2 Metrics for pairwise structure comparison

parRNAss provides four alternative approaches for structure comparison.

The morphological distance, $d_{MD}$, is a slightly modified version of a formula suggested by Zuker (1989). Here structures are represented as sets of base pairs, $(i, j) \in s$ means that residues $i$ and $j$ form a base pair in $s$. For two structures $s_1, s_2$, we define

$$d_{MD}(s_1, s_2) = \max \left\{ d_{MD}^{s_1, s_2}, d_{MD}^{s_2, s_1} \right\},$$

Figure 3. The validation plot shows the distances of all sample structures from the two consensus structures $c_1$ and $c_2$. This particular plot is based on the energy barrier distance.
where
\[ d'_{\text{MD}}(s_1, s_2) = \sum_{(i_1, j_1) \in s_1} \min_{(i_2, j_2) \in s_2} \max \left\{ \frac{|i_1 - i_2|}{|j_1 - j_2|} \right\}. \]

\( d_{\text{MD}} \) is strictly positive and symmetric but does not satisfy the triangle inequality. Although it is not a metric in the mathematical sense, it behaves quite reasonably as a distance measure.

The tree alignment distance, \( d_{\text{TAD}} \), is based on the extended RNA forest representation introduced in Höchsmann et al. (2003). The nodes of trees representing RNA structures are either P(air)- or B(ase)-nodes. The following edit operations are applicable: base replacement, base deletion, pair replacement and pair deletion, scored by cost contributions either \( P(\text{air}) \)- or \( B(\text{ase}) \)-nodes. The following edit operation (2003). The nodes of trees representing RNA structures are their string representation with dots and parentheses, e.g. \(( ( . ( . . ) ) ) ) . \), as used with the Vienna RNA package. We define
\[ d_{\text{SD}}(s_1, s_2) = d_w(y_1, y_2), \]
where \( y_i \) is the string representation of \( s_i \), and \( d_w \) is an edit distance (i.e. the score of an optimal alignment) on strings. Being defined via the edit distance model, \( d_{\text{SD}} \) is a metric. This distance measure is provided with RNAdistance (Vienna RNA package 1.4).

The energy barrier distance, \( d_{\text{EB}} \), is designed to capture the minimal amount of energy necessary for the molecule to shift between two structures. A transition path from \( s_1 \) to \( s_2 \) is given by a sequence of intermediate structures. Let \( e(s) \) denote the free energy of \( s \).
\[ e(p) = \max \{ e(s) - e(s_1) \mid s \text{ is the intermediate structure in } p \} \]
As the number of possible paths is excessively large, our implementation uses a greedy heuristic to approximate \( d_{\text{EB}} \): a list is built that holds for each base pair that has to be closed (‘close’ in short) the base pairs that have to be opened (‘opens’ in short). The close with the least number of opens is chosen, and its opens are performed and deleted from the open list of the remaining closes. Then the close is performed and eliminated from the list. This procedure is repeated until no close is left. Finally possible opens not connected to a close are performed. The energy barrier distance satisfies the axioms of a metric. It is always non-negative as the starting structure of a path is also in the set of intermediate structures. In the plots, we show the two values \( d'_{\text{EB}}(s_1, s_2) \) and \( d'_{\text{EB}}(s_2, s_1) \) instead of \( d_{\text{EB}}(s_1, s_2) \). This gives some extra information about the different energy levels of \( s_1 \) and \( s_2 \) but does not affect the general appearance of the distance plot.

### 3.3 Clustering and structure prediction

The clustering is performed using Ward’s method (Ward, 1963) and takes two parameters: \( D \) names the distance measure (\( d_{\text{TAD}}, d_{\text{EB}}, d_{\text{MD}}, d_{\text{SD}} \)) upon which the clustering is to be based. Clusters for different distance measures should be obtained and compared. \( C \) specifies the number of clusters to be generated. Normally, \( C = 2 \). For each cluster, the consensus is derived using RNAfold as explained in Section 2.1. Again, parameter \( T \) indicates the folding temperature. These two steps generate outputs for the dendrograms and for the squiggles plots of the predicted structures, as well as a string representation of structures in the Vienna style.

### 3.4 Structure validation

A validation step is necessary because the existence of three (or even more) families of structures could also lead to a distance plot with only two separate clouds if the distances of all inter-family pairs of structures are similar. Additionally, the two consensus structures could represent the two possible un-knotted projections of a pseudoknot. Structure validation takes the string representations of the predicted consensus conformations. It uses \( d_{\text{EB}} \) to calculate the distances of each sample structure to each of the two predicted structures, \( c_1 \) and \( c_2 \). These are plotted in a \( d_{c_1}, d_{c_2} \) coordinate system. In the positive case, the two structure families show up as clouds near the \( d_{c_1} \) axis and far from the \( d_{c_2} \) axis and vice versa. This proves that all structures in the sample are actually close to one of the predicted structures (Hypothesis 3), while the structures themselves are sufficiently different (Hypothesis 1 and 2). In the case of three families in the sample set, as described above, an additional cloud near the bisecting line would appear in the validation plot. To exclude the possibility that the two consensus
Data-mining the literature delivered information on 20 known structural switches in RNA, two of which are described as tertiary structural switches. For the remaining 18 sequences, the part containing the switch was extracted and analysed to test paRNAss on true positive and false negative predictions. The results are shown in Table 1. More detailed information on each analysis can be found on the paRNAss Web page. For 11 sequences, paRNAss successfully predicted a switch. In one case (E.coli hok mRNA), paRNAss predicted correct alternate structures, but the validation results were negative, which may be due to the fact that this switch is triggered by sequence truncation (not a conformational switch in our point of view).

Altogether, paRNAss predicted 12 out of 18 switches.

### 4.1 Known switches

#### 4.1.1 Experiment 1: true positives and false negatives

Data-mining the literature provided information on 20 known structural switches in RNA, two of which are described as tertiary structural switches. For the remaining 18 sequences, the part containing the switch was extracted and analysed to test paRNAss on true positive and false negative predictions. The results are shown in Table 1. More detailed information on each analysis can be found on the paRNAss Web page. For 11 sequences, paRNAss successfully predicted a switch. In one case (E.coli hok mRNA), paRNAss predicted correct alternate structures, but the validation results were negative, which may be due to the fact that this switch is triggered by sequence truncation (not a conformational switch in our point of view).

Altogether, paRNAss predicted 12 out of 18 switches.

#### 4.1.2 Experiment 2: switch degradation

To get an idea of the stability of switches under mutations (selection pressure), the sequences of the pheS-pheT operon from E.coli and the Spliced Leader from L.collosoma were randomized 1, 3, 4, 5, 10, 15, 20, 25, 30 and 35%.

Table 2 summarizes the outcome of the measurements for the randomized switch sequences. Even the introduction of one change (<2% rand.) in the sequences is capable

---

Table 2. Summary of the Predictability of Conformational Switching in RNA

<table>
<thead>
<tr>
<th>Sequence</th>
<th>nt</th>
<th>Type</th>
<th>pk</th>
<th>LDI</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attenuator</td>
<td>73</td>
<td>mRNA</td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Spliced Leader</td>
<td>56</td>
<td>mRNA</td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>E.coli DstA (Lease and Belfort, 2000)</td>
<td>87</td>
<td>mRNA</td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>E.coli S15</td>
<td>74</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>5′ UTR of E.coli htbB mRNA (Nahvi et al., 2002)</td>
<td>203</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>5′ UTR of MS2 RNA genome (Groeneveld et al., 1995)</td>
<td>73</td>
<td>mRNA</td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Tetrahymena group I intron (Cao and Woodson, 1998)</td>
<td>108</td>
<td>mRNA</td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>HDV ribozyme</td>
<td>154</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>T4 td gene intron (Semrad and Schroeder, 1998)</td>
<td>164</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>HIV-1 leader (Huthoff and Berkhout, 2001)</td>
<td>336</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>HIV-2 leader (Dirac et al., 2002)</td>
<td>544</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>3′ UTR of AMV RNA (Olsthoorn et al., 2002)</td>
<td>145</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>E.coli a operon (Schlax et al., 2001)</td>
<td>140</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>E.coli hok (Franck et al., 1997)</td>
<td>143</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>5′ UTR of E.coli thmR (Winkler et al., 2002b)</td>
<td>165</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>S-box leader of B.subtilis metE (Epstein et al., 2003)</td>
<td>247</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Bacillus subtilis ribD leader (Winkler et al., 2002a)</td>
<td>304</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>B.subtilis ypaA leader (Winkler et al., 2002a)</td>
<td>349</td>
<td>mRNA</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
</tbody>
</table>

nt, nucleotides; LDI, long distance interaction; pd, pseudoknot present in at least one structure; uRNA, untranslated RNA.
4.2 Non-switches

A set of native and randomized mRNA sequences was analysed in comparison with sequences from human 3’ UTRs. As well as random sequences, mRNAs, especially the coding regions, are assumed to rarely possess regulatory function, whereas 3’ UTRs are proven to be important regulatory regions. Since switches are regulatory elements, they are supposed to occur more often in such regulatory regions. We restricted the length of the analysed sequences to 80 nt due to the facts that numerous known switches are in this length range and that the used approach performs well, meaning that the produced plots are easily interpretable in the positive case as well as in the negative case.

4.2.1 Experiment 3: arbitrary natural sequences, random sequences, human 3’ UTRs

The set of ‘arbitrary’ natural sequences was built of 110 sequences, corresponding to parts (80 nt) of 26 E.coli mRNAs. Therefore, each mRNA was cut into pieces of length 80, and four to six of these were chosen (equally distributed over the whole sequence). Exon 2 of the Caenorhabditis elegans cAMP-dependent protein kinase catalytic subunit C gene (M37114) and bases 1596–1675 of the Drosophila melanogaster egghead gene (NM_080313) mRNA were randomized (5, 10, 15, 20, 25, 30 and 35%) to form the test set of random sequences. For the set of human 3’ UTRs, the first 80 bases of the first 400 entries in the human 3’ UTR section of the UTRdb [version of 2003/02/12 (Pesole et al., 2002)] were extracted.

The results of the analysis under the above-mentioned conditions using paRNAss are summarized in Table 3. paRNAss predicted one parnacl in the set of 110 parts of E.coli mRNAs. This means that a parnacl is predicted with a frequency of 0.9%. For the set of random sequences derived from natural sequences, there were six parnacl predictions in 420 sequences, leading to a parnacl frequency of 1.4%. The analysis of the 400 parts of human 3’ UTRs led to 14 parnacl predictions. This corresponds to a parnacl prediction rate of 3.5%.

5 DISCUSSION

5.1 Known switches

5.1.1 Test on true positive and false negative predictions

The results show that paRNAss reliably predicts structural RNA switches. The six switches paRNAss failed to recognize were either bearing pseudoknots (AMV in one structure, E.coli α-operon mRNA in both) or long distance interactions (LDI; Table 1). Currently available standard RNA folding programs are not able to calculate the secondary structure with pseudoknots because of the computational complexity. Some algorithms exist that are capable of computing specific types of pseudoknots (Giegerich and Reeder, 2003; Rivas and Eddy, 1999) and could be used to fill the gap. Nevertheless, paRNAss succeeded in two cases (E.coli S15 mRNA and HDV ribozyme) as the pseudoknot conformation shows sufficient dissimilarity even when approximating the pseudoknot by an un-knotted structure. Long distance interactions are difficult to handle with paRNAss because the large contribution of the switching part to the overall distance gets shadowed by numerous small contributions of non-switching parts, thus preventing the outcome of two distinct families in Step 1 of the paRNAss approach.

The analysis of thiM revealed two conformations, of which only one is in correspondence with the originally proposed conformations. A more detailed analysis of the original experiments showed that these results could also be explained with the conformations predicted by paRNAss. Furthermore, the energy difference of the original conformations is in the range of 10–15 kcal/mol, whereas the newly predicted structures differ ∼2 kcal/mol in free energy. Similar results were achieved for the S-box. In the case of ypaA and ribD, the free energies of the published conformations are ∼6, e.g. 10 kcal/mol above the mfe. Since the number of suboptimal structures for these sequences can only be calculated up to 4 kcal/mol due to computational limitations, these conformations are inaccessible. An algorithm capable of calculating only local minima could solve this problem as it would reduce the number of suboptimal structures dramatically.
5.1.2 Switch degradation  Taken together, the results for Attenuator and Spliced Leader show that even the introduction of 1 nt change is capable of destroying the property of being a switch, notwithstanding the fact that its conformations may persist in the space of suboptimal structures (Flamm et al., 2001). This is in accordance with a comparable analysis in Schuster et al. (1994), where it was shown that there is some probability that even a single mutation substantially alters the secondary structure. Furthermore, the authors state that sequences of length 100 with Hamming distances greater than 3 (corresponding to 3% randomization) are very unlikely to have identical or closely related structures. In our experiment, the switching property gets almost completely lost for 1% (Attenuator), respectively 5% (Spliced Leader), being in the same range as the before-stated value. The reason for the immediate appearance of this effect in the case of the Attenuator can be seen when taking a look at the two alternate structures of the native sequence in Figure 4. Together, they span about 90% of the sequence, and so each mutation is very likely to affect at least one structure. Furthermore, the two structures share 80% of the bases, and so both of them are affected in most cases. To give an example: changing the G-7 to a C and eliminating the base pairs G-7 was involved alters the free energy values as follows. The free energy of structure 1 increases from $-21.20$ to $-21.03$ kcal/mol, whereas structure 2 loses $-7.1$ kcal/mol, increasing the free energy from $-20.93$ to $-13.83$ kcal/mol. The corresponding structure is still in the folding space and it is still the member with the lowest free energy in its family, but it now falls beyond our threshold of 3 kcal/mol. This seems to be a very drastic example but shows one possible reason for the observed volatility. Generally, we assume that several mechanisms are accountable for these results: (1) one structure looses free energy and gets lost in the structure space (see example above); (2) one of the two prominent structures gains energy, kicking the other out of the threshold; (3) a new structure, which was buried in the suboptimals, gets prominent, revealing a third local minimum. Another fact is that the Attenuator is a conformational switch as well as a coding RNA. For that reason selective pressure is present in two different ways, and the evolutionary possibilities of improving the switch are restricted. In the case of the Spliced Leader, these mechanisms seem to intervene more smoothly. Additionally, the structures contain larger 5’ unpaired regions, which could serve as buffers for mutational events. The few pannacs that are predicted for higher randomization rates imply a small possibility of obtaining a conformational switch by chance.

5.2 Non-switches  Regarding the predictions for the set of mRNAs and random sequences as false positive predictions, implies an overall error rate of 1.3%. This must be seen as an upper bound as we cannot exclude the possibility that some are true (but unproven) positives. In the case of the human 3’ UTRs, the frequency of pannac predictions is 3.5%, which is substantially (2.7-fold) higher. Since these regions often possess regulatory functions, we assume that these predictions are good candidates for structural RNA switches.
5.3 Computational efficiency

The computational effort of switch prediction depends on the sequence length, the number of folds considered and the method used for structure comparison. If the tree alignment distance is used for the pairwise comparison of 50 structures, this dominates the overall computation time. As an orientation, we give measurements for the smallest and largest problem instances in our set of 18 known switches, made on an UltraSparcIII 900 MHz processor. The Spliced Leader (56 nt, 36 structures) requires 96 s when using the tree alignment distance, and 5 s without it. The HIV-2 leader (544 nt, 50 structures) requires 158 min when using the tree alignment distance and 7 min without it.

6 CONCLUSION

The results of the evaluation show that the strategy used by paRNAss is able to reveal conformational switches in RNA if no pseudoknots or LDIs are involved. Approximately 66% of the published switches contain such a feature in at least one of their conformations. Nonetheless, the detection of a conformational switch even works for some switches with a pseudoknot or LDI. The predicted consensus structures for these can of course be just rough approximations. The degradation of structural switches could be shown for even one point-mutation. This gives hints on the effort needed to evolve a conformational RNA switch and further leads to the assumption that structural switching in RNA is rather seldom. This is in agreement with the low amount of predicted parnacs for arbitrary natural and random sequences. Furthermore, it was shown that regulatory regions, like 3' UTRs, are more likely to contain a parnac. This is in good correspondence with the regulatory function of bistable RNAs and should encourage experimenters to check for correctness of these parnacs (see Table 4).

Until now the focus of paRNAss is on the prediction of structural RNA switching. Other approaches exist to reveal more general properties of the energy landscape of an RNA. Kitagawa et al. (2003) studied the conformational landscape of human snRNA. Their analysis is based on the newly introduced ‘tree representation distance’. We reviewed their analyses of the human splicesosomal snRNAs using the new version of paRNAss. The experiments on U1, U4, U5 and U6 snRNA delivered comparable results, whereas for the U2 snRNA the distance plot ($d_E$, $d_{TAD}$) shows three separate clouds of points, as displayed in Figure 5. Since each point in the distance plot corresponds to a pair of structures, the appearance of three distinct clouds must be due to three, four or even more families of structures in the sample set. A look at the structures in the sample set revealed three families. Consensus structures of these three families are shown in Figure 6. This result is in clear contrast with the result reported by Kitagawa et al. (2003), as they propose a ‘uni-valley’ shape of the folding space and therefore only one prominent structure. Possible reasons for this discrepancy may lie with the use of MFOLD, producing a heuristic subset of all feasible structures, and the coarse grained ‘tree representation distance’.

The distance plot in Figure 5 also documents how much the discriminative power of paRNAss has increased due to the algorithmic improvements reported here. A clear separation of three structure families has not been observed before. A further improvement of the approach could be achieved if an algorithm for the calculation of local minima were available as this would allow analyses going much deeper in the structure space.

Due to these recent findings and the results of our evaluation, we are confident that the algorithmic improvements and the incorporation of the tree alignment distance into the paRNAss approach make structural switch prediction in RNA more reliable and might further enable deeper insights in the folding space of RNA.

1This naming is somewhat misleading, as RNA structures are treelike and all distance measures work on tree representations.
The major obstacle we faced during the design of our evaluation was the small number of proven conformational switches. On one hand this can be seen as a drawback to our evaluation procedure, but on the other hand we think that this expresses the explicit need for a bioinformatics approach in this area of research. \( \text{padRNAs} \) is such an approach and we are convinced that it will be helpful in revealing more conformational RNA switches and in turn will get improved based on these findings.

**REFERENCES**


