Supplementary material to: A comparative method for finding and folding RNA secondary structures within protein-coding regions

Jakob Skou Pedersen*, Irmtraud Margret Meyer*, Roald Forsberg,
Peter Simmonds and Jotun Hein
* These authors contributed equally to this work

Introduction

The supplementary material contains three main sections, each of them supplies detailed information on a subject of the main text: the two variants of RNA-DECODER called RNA-DECODER-TWO-STEP and RNA-DECODER-EXTENDED, performance results, and evolutionary information and prediction performance.

RNA-DECODER-TWO-STEP and RNA-DECODER-EXTENDED

We have devised two alternative RNA-ss fold prediction programs called RNA-DECODER-TWO-STEP and RNA-DECODER-EXTENDED which are based on RNA-DECODER. Both methods were developed in response to the occurrence of structural changes between the individual sequences of the alignment. They employ RNA-DECODER to exploit the comparative information of the alignment, but return single sequence predictions. A detailed description of both methods is given below.

RNA-DECODER-EXTENDED

The idea behind RNA-DECODER-EXTENDED is to adjust the prediction made by RNA-DECODER to encompass the individual differences of each sequence.

RNA-DECODER-EXTENDED first makes a prediction for the alignment using the CYK as algorithm as described for RNA-decoder. An individual prediction is then made for each sequence of the alignment by discarding all predicted non-consensus base pairs and by extending the remaining stems with neighboring consensus base pairs until a non-consensus base-pair is found (adopted from Knudsen and Hein [1]). The six consensus base pairs are `G-C, C-G, A-U, U-A, U-G` and `G-U`.

RNA-DECODER-TWO-STEP

The idea behind RNA-DECODER-TWO-STEP is to exploit the comparative information to delineate the conserved core sites of an (evolving) RNA structure. Knowing these simplifies the remaining folding problem for each individual sequence.

RNA-Decoder-Two-Step first makes a partial prediction $y^{\mathcal{S}\star}_{part}$ for the alignment by running the CYK algorithm and retaining only base-pairing columns whose posterior probability is at least 0.8. In the second step, the CYK algorithm is used on each individual sequence x_i while keeping the annotation of all those positions fixed that are covered by the partial prediction. The resulting label sequence predicted for x_i is then:

$$y_i^{\mathcal{S}\star} = \underset{y_i^{\mathcal{S}}}{\operatorname{argmax}}\, P(y_i^{\mathcal{S}} | x_i, y^{\mathcal{C}}, y^{\mathcal{S}\star}_{part}, M).$$

Performance results

This section reports various detailed performance results.

Performance of RNA-Decoder by structure

The results for the five-fold cross-evaluation on the HCV 1a and the HCV 1a & 1b set can be found in the left column of Table 1. Correctly predicting the fourth and fifth of the structural elements seems to be easier than the rest. This may be due to two reasons: first, structures 4 and 5 are the shortest structures with the lowest fraction of non-consensus base pairs (see Table 1 of main text) and, second, training the model on a training set which comprises the longer and more diverse structural elements 1–3 should increase its predictive power. The high fraction of non-consensus base pairs within structure 2 (HCV 1a set) and structures 2 and 3 (HCV 1a & 1b set) is probably also responsible for the large difference between the single-nucleotide and the pair performance.

Performance of RNA-Decoder-Two-Step and RNA-Decoder by structure

We repeated the cross-evaluation experiments on datasets HCV 1a and HCV 1a & 1b using RNA-Decoder-Extended and RNA-Decoder-Two-Step. As can be seen from the middle column of Table 1, using RNA-Decoder-Extended increases the single-nucleotide and pair sensitivity, however, usually at the expense of a lowered specificity (this lowering of the specificity is stronger for the HCV 1a set than for the more diverged HCV 1a & 1b set).

The results of RNA-Decoder-Two-Step are shown in the right column of Table 1. When comparing its performance to that of RNA-Decoder and RNA-Decoder-Extended, one can see that it generally improves the specificity with respect to RNA-Decoder. However, the sensitivity is not generally improved, but overall lowered with respect to RNA-Decoder (compare e.g. the three different predictions for structure 3 of the HCV 1a set).

Performance of RNA-Decoder and Pfold as function of prediction confidence

RNA-Decoder assigns posterior probabilities to its label predictions at each site. We here evaluate how well the posterior probability of a site measures the confidence we can have in the predicted label. Table 2 reports the performance among all sites which are assigned a certain minimum posterior probability (see left column "min. pp.") as well as the fraction of the alignment covered (see columns "% data"). We perform this evaluation for both the 1a and the 1a & 1b sets.

	RNA-Decoder											
	HCV 1a set											
					RNA-Decoder-Extended				RNA-Decoder-Two-Step			
str.	sn_s	sp_s	sn_p	sp_p	sn_s	sp_s	sn_p	sp_p	sn_s	sp_s	sn_p	sp_p
1	0.96	0.89	0.96	0.89	1.0	0.86	1.0	0.86	1.0	0.96	1.0	0.96
2	0.76	0.86	0.44	0.5	0.72	0.78	0.4	0.44	0.74	0.88	0.44	0.52
3	0.78	0.97	0.75	0.94	0.83	0.97	0.80	0.94	0.5	1.0	0.5	1.0
4	1.0	1.0	1.0	1.0	1.0	0.93	1.0	0.93	1.0	0.88	1.0	0.88
5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
all	0.88	0.93	0.79	0.84	0.89	0.89	0.8	0.80	0.83	0.94	0.75	0.85
	HCV 1a & 1b set											
					RNA-Decoder-Extended				RNA-Decoder-Two-Step			
str.	sn_s	sp_s	sn_p	sp_p	sn_s	sp_s	sn_p	sp_p	sn_s	sp_s	sn_p	sp_p
1	0.86	0.83	0.84	0.81	0.98	0.82	0.96	0.8	1.0	0.86	1.0	0.86
2	0.62	0.86	0.36	0.5	0.7	0.88	0.4	0.5	0.18	0.9	0.16	0.8
3	0.43	0.65	0.25	0.39	0.43	0.65	0.25	0.39	0.5	1.0	0.5	1.0
4	0.86	0.92	0.79	0.85	0.93	0.87	0.86	0.8	1.0	0.88	1.0	0.88
5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
all	0.73	0.85	0.61	0.71	0.79	0.84	0.66	0.70	0.69	0.91	0.68	0.90
	Polio set											
					RNA-Decoder-Extended				RNA-Decoder-Two-Step			
str.	sn_s	sp_s	sn_p	sp_p	sn_s	sp_s	sn_p	sp_p	sn_s	sp_s	sn_p	sp_p
CRE	0.57	1.0	0.57	1.0	0.57	1.0	0.57	1.0	1.0	0.91	1.0	0.91

Table 1: Prediction performance of RNA-Decoder, RNA-Decoder-Extended, and RNA-Decoder-Two-Step on the annotated structures of the HCV 1a and the combined HCV 1a & 1b. We report the performance in terms of sensitivity and specificity for pairs of base pairing nucleotides (sn_p and sp_p) as well as for single nucleotides (sn_s and sp_s). Please refer to the main text for a definition of these performance measures.

	HCV 1a set					HCV 1a & 1b set				
min. pp.	sn_s	sp_s	sn_p	sp_p	% data	sn_s	sp_s	sn_p	sp_p	% data
0.0	0.88	0.93	0.79	0.84	1.0	0.73	0.85	0.61	0.71	1.0
0.5	0.90	0.94	0.84	0.87	0.89	0.76	0.88	0.74	0.84	0.82
0.6	0.90	0.95	0.83	0.88	0.83	0.76	0.89	0.74	0.86	0.77
0.7	0.92	0.98	0.86	0.92	0.68	0.77	0.93	0.76	0.92	0.68
0.8	0.89	1.0	0.89	1.0	0.46	0.84	0.93	0.83	0.92	0.60
0.9	0.92	1.0	0.92	1.0	0.34	0.92	0.92	0.91	0.91	0.53

Table 2: Prediction performance and data coverage (% data) of RNA-Decoder as function of the minimum posterior probability (min. pp.) for the HCV 1a as well as the combined HCV 1a & 1b data sets. Note that requiring a minimum posterior probability of 0.0 leads to the inclusion of the entire alignment. Please refer to the main text for a definition of the performance measures sn_p, sp_p, sn_s and sp_s.

Requiring a high minimum posterior probability leads to a good performance with very high specificity on both the 1a and the 1a & 1b sets. 68% of both data sets are covered by posterior probabilities of at least 0.7, and have specificities which are in the nineties. Requiring a high minimum posterior probability is seen to especially improve the performance on the 1a & 1b set. This is probably due to variations in the structure between some sequences in this set, which therefore affect both performance and posterior probability of a subset of the sites.

Evolutionary information and prediction performance

This section analyzes the effect of alignment size on prediction performance. The measure of prediction performance chosen is the posterior probability of correctly predicting a site to be stem-pairing or not. This is a continuous measure that allows for fine-grained analysis.

Performance as function of total tree length

We performed the following experiment in order to investigate the relationship between the performance and the amount of evolutionary information available. First, a large number of alignment subsets were sampled from each of the HCV 1a and HCV 1a & 1b data sets for structure 4. Each sample contains the reference sequence and a uniformly distributed number of additional sequences, chosen at random. The posterior probability of the correct label (either pairing or not) was then calculated along the alignment. The performance was summarized by calculating the average of these posterior probabilities (called APPCL), as well at the standard deviation. The amount of evolutionary information available in each sample was measured by the total tree length (TTL) of the spanning phylogenetic tree. Finally, the samples were binned according to TTL, and the average APPCL and average standard deviation within each bin were plotted (see Figure 1).

Note that the left figure (HCV 1a data set) effectively corresponds to the short-TTL part of the right figure (combined HCV 1a & 1b set), as samples with a TTL of less than 0.6 include

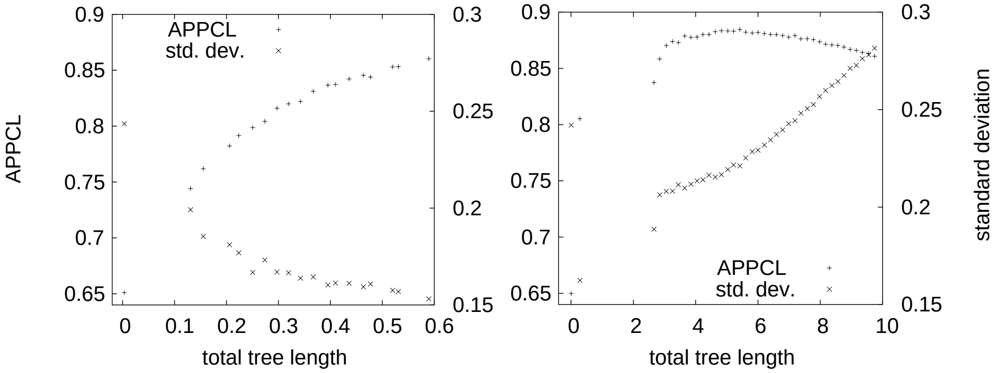

Figure 1: RNA-ss prediction performance as function of the amount of evolutionary information in the input alignment for structure 4. The performance is measured in terms of the average posterior probability of the correct label along the alignment (APPCL, see left y-axis) as well as its standard deviation (see right y-axis). A perfect prediction would have an APPCL of 1.0 and a standard deviation of 0.0. The evolutionary information is measured by the TTL. The left figure is based on 1000 sampled subsets of the HCV 1a set whose TTL of 0.59 was sub-divided into 25 bins, while the right is based on 10,000 sampled subsets of the HCV 1a & 1b set whose TTL of 9.84 was sub-divided into 50 bins. The points are not necessarily equi-distant since the average within a bin need not coincide with the middle bin value.

only HCV 1a sequences. The samples having a TTL of more than 2.5 include sequences from the HCV 1b data set. The gap of data for TTLs between 0.6 and 2.5 is due to the long branch separating the HCV 1a from the HCV 1b clades.

The APPCL for the HCV 1a data set increases with TTL (left figure), while the variation among the predictions decreases. The rate of increase in performance is high initially and then decreases.

The APPCL initially decreases when the first HCV 1b sequences are included (right figure). This is probably due to the introduction of sequences with structures which have evolved relative to the HCV 1a sequences. The following increase can be attributed to an increasing confidence in the predictions of the non-evolving positions (see also Figure 5 of main text). The final slow decrease is probably due to a growing confidence in the wrong pair predictions for the bulge positions and a corresponding decline for the last stem-pairing positions (see also Figure 5 of main text). The increase in standard deviation with TTL summarizes the growing discrepancy in the posterior probability of the correct label along the sequence.

References

[1] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, 31(13):3423–3428, Jul 2003.