# ShapeSorter – a method for detecting conserved RNA structure features supported by SHAPE evidence

Volodymyr Tsybulskyi & Irmtraud M. Meyer

April 8, 2022

To whom correspondence should be addressed.
Email: irmtraud.meyer@cantab.net

**This PDF file includes:**

- Supplementary Tables S1 and S2
- Supplementary Figure S1 and S2

## 1 Supplementary Tables

### 1.1 Test and training set of ShapeSorter

| Name | Length | N. Seqs. | Av. PPID | Source |
|---|---|---|---|---|
| 5S rRNA, E. coli | 120 | 713 | 58 | 1 |
| Adenine riboswitch, V. vulnificus | 71 | 134 | 63 | 1 |
| Fluoride riboswitch, P. syringae * | 66 | 288 | 53 | 1 |
| 5' pseudoknot domain, HIV-1 * | 500 | 131 | 93 | 1 |
| RNase P, B. subtilis | 401 | 115 | 64 | 1 |
| Signal recognition particle RNA, human | 301 | 92 | 76 | 1 |
| tRNA(phe), E. coli | 76 | 955 | 47 | 1 |

**Table S1. Test set of ShapeSorter.** For each alignment of the test set that we compiled, we specify the length (in nucleotides), the number of sequences in the alignment (N. Seqs.) and the average pairwise percent identity (Av. PPID). The source indicates the method paper which first introduced the reference sequence, RNA structure and corresponding SHAPE-probing data: 1 - ShapeKnots test set, 2 - ShapeKnots training set, 3 - ProbFold training set, 4 - PPfold training set, 5 - GTFold training set, 6 - RNAstructure training set. Please see the section on alignments for more details how these alignments were compiled. Sequences with a pseudo-knotted reference RNA structure are denoted by an asterisk ($*$) behind their name.

| **Name** | **Length** | **N. Seqs.** | **Av. PPID** | **Source** |
|---|---|---|---|---|
| 16S (small subunit ribosomal RNA), E. coli | 1542 | 100 | 73 | 3,4,5 |
| 23S (bacterial large subunit ribosomal RNA), E. coli | 2904 | 103 | 69 | 3,4,6 |
| 5' domain of 16S rRNA, E. coli | 530 | 100 | 68 | 2 |
| 5' domain of 16S rRNA, H. volcanii | 473 | 87 | 75 | 2 |
| 5' domain of 23S rRNA, E. coli | 511 | 103 | 68 | 2 |
| 5SRNA, E. coli | 170 | 713 | 58 | 3 |
| Adenine riboswitch, V. vulnificus | 121 | 134 | 54 | 3 |
| cyclic-di-GMP riboswitch, V. cholerae | 135 | 156 | 61 | 3 |
| cyclic-di-GMP riboswitch, V. cholerae | 97 | 156 | 61 | 2 |
| Glycine riboswitch, F. nuleatum | 198 | 45 | 33 | 3 |
| Group II intron, O. iheyensis * | 412 | 407 | 55 | 2 |
| Group I Intron, T. thermophila * | 425 | 13 | 37 | 2 |
| IRES domain, Hepatitis C virus * | 336 | 80 | 77 | 2,4 |
| Lysine riboswitch, T. maritime * | 174 | 48 | 43 | 2 |
| M-Box riboswitch, B. subtilis | 154 | 158 | 64 | 2 |
| Pre-Q1 riboswitch, B. subtilis * | 34 | 36 | 70 | 2 |
| Ribonuclease, E. coli | 198 | 115 | 63 | 3 |
| SAM I riboswitch, T. tengcongensis * | 118 | 457 | 58 | 2 |
| SARS corona virus pseudoknot * | 82 | 57 | 58 | 2 |
| TPP riboswitch, E. coli | 79 | 110 | 62 | 2 |
| tRNA (phenylalanine), yeast | 116 | 955 | 47 | 3 |
| tRNA (asparagine), yeast | 75 | 955 | 43 | 2 |

**Table S2. Training set of ShapeSorter.** For each alignment of the test set that we compiled, we here specify the length (in nucleotides), the number of sequences in the alignment (N. Seqs.) and the average pairwise percent identity (Av. PPID). The source indicates the method paper which first introduced the reference sequence, RNA structure and corresponding SHAPE-probing data: 1 - ShapeKnots test set, 2 - ShapeKnots training set, 3 - ProbFold training set, 4 - PPfold training set, 5 - GTFold training set, 6 - RNAstructure training set. Please see the section on alignments for more details on how these alignments were compiled. Sequences with a pseudo-knotted reference RNA structure are denoted by an asterisk ($*$) behind their name.

# 2 Supplementary Figures

## 2.1 Performance of ShapeSorter and the other programs on the training set

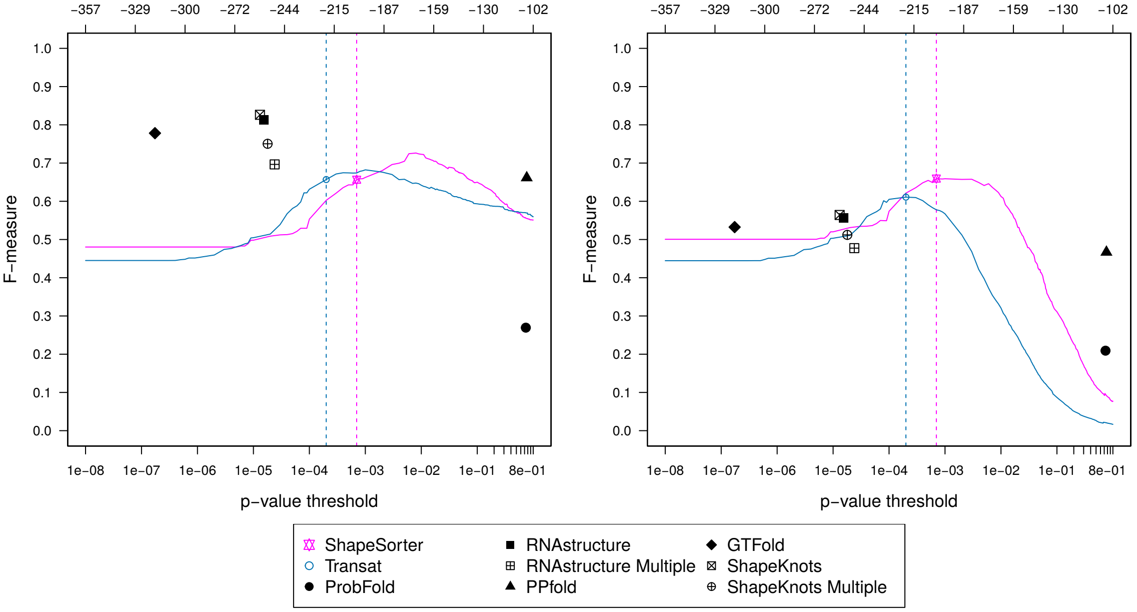

**Figure S1. Predictive performance of ShapeSorter and other programs in terms of $F_{measure}$ for training set** for nucleotides (left) and base-pairs (right). The symbols and dashed vertical lines for ShapeSorter (pink dashed line) and Transat (blue dashed line) are positioned at the p-values that corresponds to the respective p-value thresholds. These values correspond to the p-values that maximise the $MCC$ for base-pairs for this training set, see Figure S2 below. The symbols for all other programs apart from ShapeSorter and Transat are positioned at the average MFE-value of their respective, predicted RNA secondary structures, see the x-axis at the top which shows free energies in kcal/Mol.

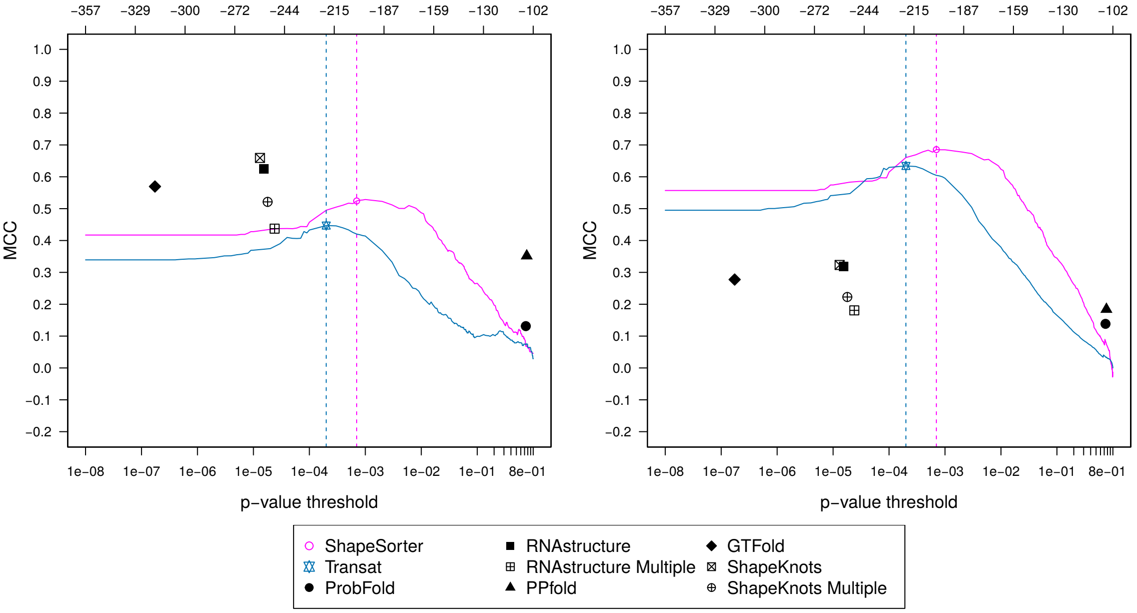

**Figure S2. Predictive performance of ShapeSorter and other programs in terms of *MCC* for training set** for nucleotides (left) and base-pairs (right). The symbols and dashed vertical lines for ShapeSorter (pink dashed line) and Transat (blue dashed line) are positioned at the p-values that corresponds to the respective p-value thresholds. These values correspond to the p-values that maximise the *MCC* for base-pairs for this training set, *i.e.* see the max of the pink and blue curves for base-pairs on the right. The symbols for all other programs apart from ShapeSorter and Transat are positioned at the average MFE-value of their respective, predicted RNA secondary structures, see the x-axis at the top which shows free energies in kcal/Mol.

# 3 Supplementary Formulae

The following explains how the equations (4) in the manuscript change, in case we are dealing with the innermost base-pair at $(i, j)$ of a helix (which therefore has no futher, inner neighbouring base-pair to stack with) or if we are dealing with an un-paired sequence position at $i$ which does not have an unpaired, previous sequence position at $i - 1$. As defined earlier, $\mathcal{A}$ denotes the RNA alphabet.

$$\begin{aligned} P_{exp}((i,j) \mid \theta_{stack}) = & \sum_{x_{i-1}\in\mathcal{A}} \sum_{x_{j+1}\in\mathcal{A}} P((x_{i-1}, x_{j+1}), (x_i, x_j)) \cdot \\ & P_{exp}^{pair}(i) \cdot P_{exp}^{pair}(j+1) \cdot \\ & P_{exp}^{pair}(i \mid i-1) \cdot P_{exp}^{pair}(j+1 \mid j) \\ P_{exp}((i,j) \mid \theta_{single}) = & \sum_{x_{i-1}\in\mathcal{A}} P(x_i) \cdot P(x_j) \cdot P_{exp}^{single}(i) \cdot \\ & P_{exp}^{single}(j) \cdot P_{exp}^{single}(i \mid i-1) \cdot \\ & P_{exp}^{single}(j+1 \mid j) \end{aligned} \tag{1}$$