INFO-F439 Methods in Bioinformatics

RNA structure prediction using positive and negative evolutionary information

Draguet Simon, Stützle Santangelo Alexander

May 30, 2025

Introduction

RNA play crucial roles in cell physiology, far beyond its classical image as a messenger between DNA and protein. RNA carries genetic information (mRNA), helps assemble proteins (tRNA & rRNA), regulates genes (miRNA) and even catalyses chemical reactions (ribozymes). Their activity has been shown to have impact gene expression, regulation and many essential cellular processes. Much like proteins, their functions are closely tied to their three dimensional structure. Because structure determines function, being able to predict RNA structure is vital for understanding how they work and will allow researchers to identify drug targets and engineer RNA molecules for biotechnology and medicine.

The CaCoFold algorithm (Cascade variation and covariation Constrained Folding algorithm) (Figure 1) enables the prediction of a consensus RNA secondary structure based on a multiple sequence alignment, thereby identifying the most conserved structural features among a set of relatively similar sequences. It was developed as an extenstion of the R-scape framework, leveraging R-scape's ability to identify statistically significant covariation signals to guide and constraint the folding process.

The objective of this project is to provide a description of the algorithm's operational principles, to assess the quality and reliability of its predictions, and to discuss the inherent limitations of the CaCoFold approach.

Algorithm initialization

R-scape takes as input a multiple sequence alignment (MSA) file and an optional E-value threshold used to determine which covarying pairs are statistically significant (Figure 1a). Using a G-test corrected by APC (Average Product Correction), R-scape is able to robustly quantify covariation between each pair of columns.

Pairs showing significant covariation below the E value threshold are considered positive pairs (Figure 1b), suggesting conserved base pairs. R-scape also identifies pairs of positions that display significant variation but no significant covariation; these are called negative pairs and are considered unlikely to form base pairs. All the positive base pairs will be included in the final CacoFold structure, and all negative base pairs are forbidden to appear.

MaxCov Cascade Algorithm

The MaxCov cascade algorithm is the first step of the CacoFold algorithm itself, it takes as input the positive and negative base pairs that R-scape has identified from the initial multiple sequence alignment. The purpose of the MaxCov cascade algorithm is to group the identified positive base pairs into nested subsets. In the context of the secondary structure of RNA, a nested structure is one in which base pairs do not cross each other when drawn in a 2D representation (Figure 1c).

It does so by using a modified Nussinov algorithm, a classic dynamic programming approach for the prediction of the structure of RNA. However the Nussinov algorithm has

been modified to group together in each layer the maximal subset of positive base pairs that are nested relative to each other.

The first layer (CO in Figure 1c) groups the maximal subset of nested positive base pairs with the smallest cumulative E-value (representing the significantly covarying base pairs). If other other positive base pairs did not fit into the first nested set, subsequent layers (C1, C2...Cn) of the MaxCov algorithm are performed. This continues until all positive basepairs have been grouped into nested subsets. Additionally for each layer a set of forbidden basepairs is generated which includes all the negative basepairs.

Cascade Folding Algorithm

For each layer that has been identified by the MaxCov cascade algorithm, the Cascade Folding algorithm receives a set of positive basepairs that were grouped together by MaxCov. It also receives a set of negative base pairs that are associated with each layer.

For each of these layers, the goal of the cascade folding algorithm is to compute the most probable constrained nested structure. A constrained nested structure is the most probable predicted structure given the constraints provided by evolutionary information. These constraints are derived from the positive and negative base pairs assigned to each layer (Figure 1d).

Depending on the layers, different probabilistic folding algorithms are used: The first layer (S0), which is designed to capture the main nested structures, uses the probabilistic RNA Basic Grammar (RBG) model. The RBG model is used to find the main nested structures of the RNA, it focuses on identifying the primary secondary structure elements that form a fold. For the subsequent layers (S+ = S1, S2... Sn) the algorithm uses the G6X model to incorporate the remaining basepairs that were not included in the main nested structure. These remaining pairs are often involved in more complex interactions such as pseudoknots, non-nested tertiary contacts, or base triplets.

At this stage, for each layer, a single constrained nested structure is constructed. This enables the identification of complex non-nested interactions by examining how the basepairs across this collection of layered structures relate to each other. The subsequent filtering step then processes these base pairs from each layer to the final structure.

Alternative Helix Filtering

The alternative helix filtering step is crucial for combining the structural elements identified in the different layers into a single complete structure. The goal is to remove redundancies and prioritise elements supported by evolutionary evidence.

First the nested structures from the S+ layers are analysed and broken down into alternative helices. An alternative helix is operationally defined as a set of contiguous basepairs, allowing for small disruptions like a one or two residue bulge or a 1x1 internal loop. The algorithm makes sure that each basepair belongs to one and only one helix. If at least one basepair in a helix is positive, then the helix is called positive. Positive helices are always retained (Figure 1e).

Any alternative helix without covariation support can be retained only if they are longer than 15 basepairs and they must overlap in no more than 50% of their bases with already any helix that has already been selected from previous layers.

Automatic display of the complete structure

Finally, the complete predicted structure can be displayed (Figure 1f). This last step takes the main nested structure (S0) predicted in the first layer of the cascade folding algorithm and the filtered set of alternative helices from the subsequent layers (S+). These structures are then integrated together with the main nested structure as the final RNA structure. In order to display the CacoFold structure the algorithm uses an adapted version of R2R for visualization. This tool is capable of automatically drawing consensus structures, including both nested and non-nested basepairs.

In the final structure all the significantly covarying basepairs are highlighted in green, and the consensus structure for the alignment is visualized, potentially annotating pseudoknots

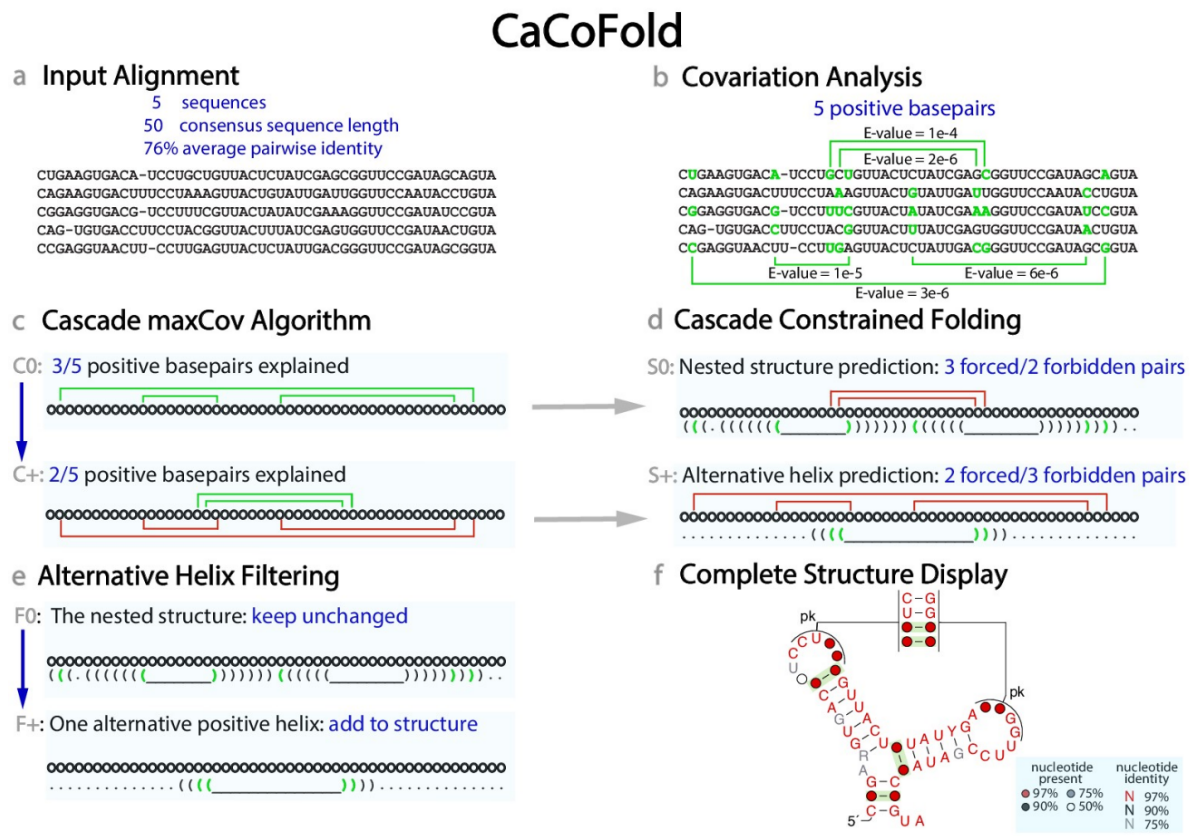

Figure 1: The CacoFold Algorithm, the illustration comes from the paper. **(a)** Toy alignment of 5 sequences. **(b)** The statistical analysis identifies 5 significantly covarying position pairs in the alignment (E-value < 0.05). **(c)** In this example, the maxCov algorithm requires 2 layers to explain all the covarying pairs. In the layer C0, three positive basepairs are depicted together. In the C+ layers, these basepairs are excluded (in red) and the remaining two can be grouped together. **(d)** The algorithm finds the most probable fold constrained by the assigned positive base pairs (green parentheses) and the exclusion of the other positive base pairs. **(e)** The S+ alternatives that do not meet the criteria explored in the alternative helix filtering section are removed. **(f)** The final structure is displayed by an adaptation of the R2R program.

("pk") and triplets ("tr") for the alternative helices according to their overlap with the main nested structure.

Materials and methods

Unlike protein structures, which include quality indicators such as pLDDT with AlphaFold, the quality of RNA structures predicted by CaCoFold can only be assessed through visual inspection. Under these conditions, it is necessary to select RNAs with well-characterized and relatively simple structures to evaluate the algorithm's ability to accurately reconstruct RNA. Therefore, for this project, we selected tRNA (RFAM code RF00005), which delivers amino acids during genetic translation; 5S rRNA (RF00001), a component of the large ribosomal subunit; and the glutamine riboswitch (RF01739), which regulates glutamine synthetase in cyanobacteria (Klähn et al., 2018).

To generate analyzable and critically assessable data beyond simple visual comparisons, we produced various RNA structures for each RNA type by modifying the dataset — either by altering the number of sequences or their identity, i.e., by randomly selecting sequences from a specific dataset.

In practice, the initial data were sourced from the multiple sequence alignments (MSAs) available on RFAM for each RNA type. The sequences were extracted and converted to FASTA format. Using these MSAs seems to

ensure the generation of a relatively reliable reference structure, as it incorporates a broad and curated set of aligned sequences.

On one hand, a given percentage of the sequences was randomly selected to create new MSAs using Clustal Omega (Madeira et al., 2024) with default parameters. These new alignments were then used with CaCoFold to generate different structures of the same RNA type, based on varying amounts of sequence data — and thus varying levels of informational content.

On the other hand, ten additional datasets were randomly created, each containing only 50% of the original sequences, and processed in the same manner. The resulting structures were intended to evaluate the impact of sequence identity on structural predictions — in other words, to assess whether prioritizing sequence quantity or sequence diversity leads to more accurate RNA structure prediction.

Results

Although this algorithm does not provide a direct numerical quality score, as previously mentioned, the following analyses rely on several indirect indicators: the number of significant positive base pairs detected, their respective positions, their e-values, and other related metrics.

Influence of the number of sequences

At first glance, a greater number of sequences should provide more information about base-pair associations, potentially leading to the detection of more positive base pairs, or at least more significant ones.

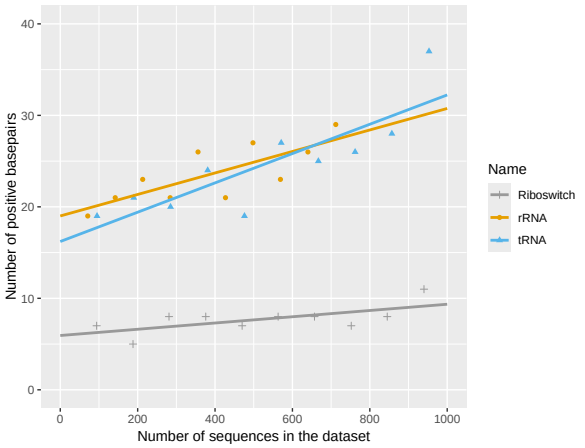

Figure 2: Number of positive base pairs per sequence in the dataset, considering an E-value threshold of 0.05. Each group of points is associated with its corresponding regression line.

This initial hypothesis is relatively well supported here. As shown in Figure 2, the number of positive pairs tends to increase with the number of sequences, regardless of the RNA type considered in this study.

However, it's worth noting that this trend is not uniform: the number of pairs identified in the various structures of the glutamine riboswitch varies very little compared to the other RNAs. This is likely due to the relative simplicity of its secondary structure compared to the two others (see Appendix).

Conversely, the increase in sequence number does not appear to lead to higher statistical significance in pair detection, as illustrated by Figures 3. For all RNA types, structures based on fewer sequences do not consistently yield less significant pairs, and having more sequences does not necessarily result in greater significance.

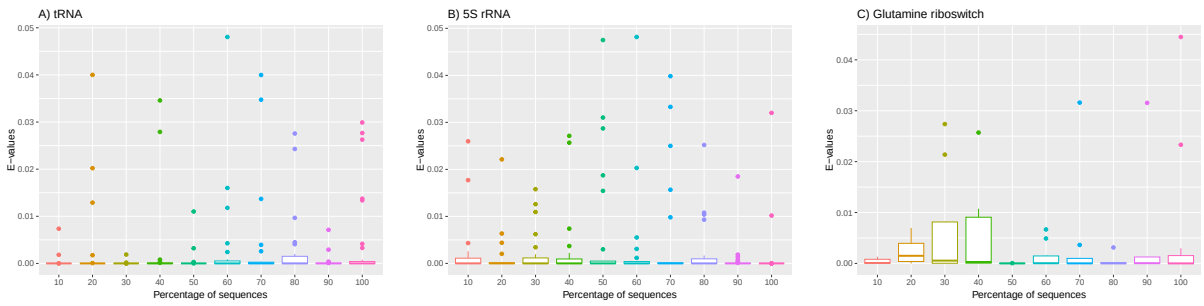

Figure 3: E-values distributions per percentage of sequences considered.

Therefore, although an increase in positive base pairs is observed, this does not lead to better structural accuracy. There are no major differences between the tRNA structures, and only the 5S rRNA structures with 40% and 90% of the sequences, as well as the glutamine riboswitch structures with 20% and 60%, deviate from the expected structure. The fact that erroneous structures appear both at low and high sequence percentages suggests that a larger number of sequences does not guarantee structural quality.

As previously mentioned, most of the predicted structures are relatively similar. However, it is interesting to assess the degree of this similarity. In this project, it is evaluated based on the position of base pairs (Fig. 8) and their occurrence across all structures of the same RNA type (Fig. 9).

For 5S rRNA, recurring association patterns can be observed, which seem to shift depending on the percentage of sequences used, but only one of these pairs is not unique, appearing in just two structures. These patterns are less clear in tRNA, where there is greater variability in the positions of positive pairs, and again only one non-unique pair. Finally, in the glutamine riboswitch, associations appear to occur preferentially in the early residues, but all are unique.

These observations highlight the high sensitivity of CaCoFold to the input data. All structures were generated from the same sequence dataset, with only a variable fraction used in each case. One might have expected the resulting structures to share the same base pairs with differing levels of significance — which is not the case.

In conclusion, while CaCoFold generally produces reasonably faithful structures regardless of the number of sequences used, these structures are never entirely accurate. We can never be sure that the associations displayed will be observed experimentally.

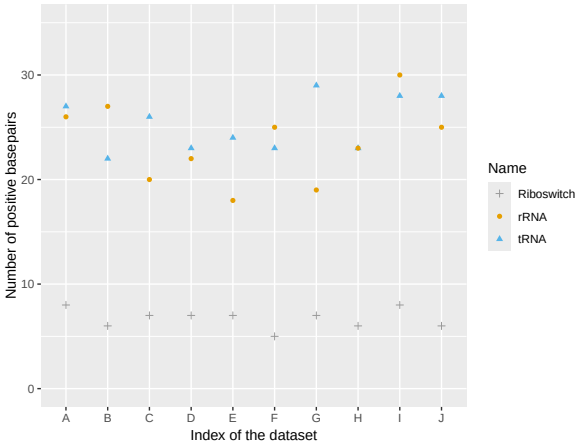

Figure 4: Number of positive base pairs per dataset, considering an E-value threshold of 0.05. Each dataset is created by randomly selecting 50% of the original sequences.

Influence of the sequences

As previously mentioned, the influence of sequences on structure is assessed here by varying the sequences constituting a dataset of fixed size.

First, Figure 4 shows considerable variability in the number of base pairs across datasets, except for the glutamine riboswitch, as already observed. This initial finding suggests that some sequences promote the emergence of a greater number of significant base pairs.

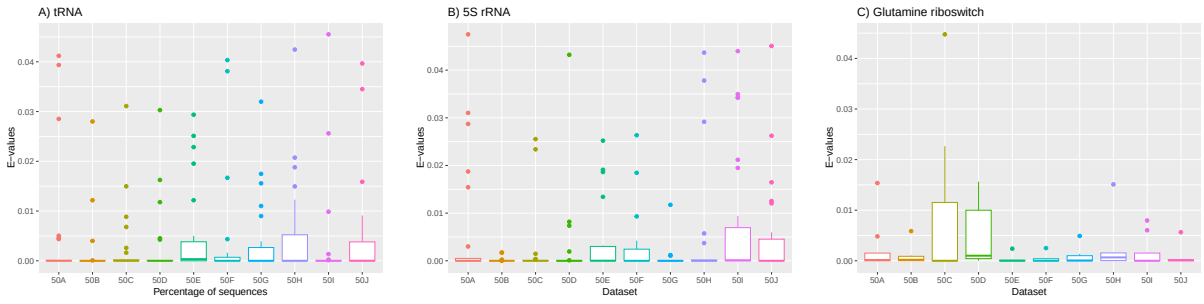

Figure 5: E-values distributions per random 50% dataset of the sequences considered.

However, it is important to note that a high number of pairs is not necessarily indicative of better structural quality: in the case of tRNA, structure B is closer to the reference structure than structure G, despite containing fewer positive pairs. As previously observed, there is no clear correlation between the number of pairs and their statistical significance (Figure 5).

By analyzing the positions of base pairs (Figure 10) and their occurrences (Figure 11), one can infer that the more complex a structure is, the less impact the presence or absence of certain sequences has on its formation. Indeed, the most complex RNA in this study—the 5S rRNA—shows highly concentrated base-pair positions in specific regions, along with a larger number of pairs shared across multiple structures. This limited variability is also evident in the structural outcomes, which show minimal differences.

In contrast, for tRNA, the distribution of pair positions is more diffuse, without a corresponding increase in the number of shared pairs. Moreover, the only pair found in two structures differs from the one in Figure 9A, making it difficult to identify a meaningful association. Finally, similar observations apply to the glutamine riboswitch when varying the sequences.

Influence of the diversity

It was observed that several regions from the same gene appear in certain datasets. Therefore, it would be relevant to investigate whether sequence diversity affects the final structure. To do this, we simply analyzed the genes that appear multiple times in the datasets with the lowest and highest numbers of positive base pairs.

In the case of tRNA, datasets B and G present an interesting contrast: dataset B is closer to the reference structure, and as shown in Figure 6, it appears to include slightly more diverse gene sequences. However, this apparent correlation does not hold for the riboswitch (fig 7), where structure I is completely incorrect, despite not having genes that are significantly more overrepresented than in structure B, which is relatively accurate.

Therefore, the quality of a predicted structure does not seem to be directly linked to the repeated presence of specific regions from the same gene. Nonetheless, it would be worth investigating whether the presence of genes from phylogenetically related species contributes to the observed structural differences.

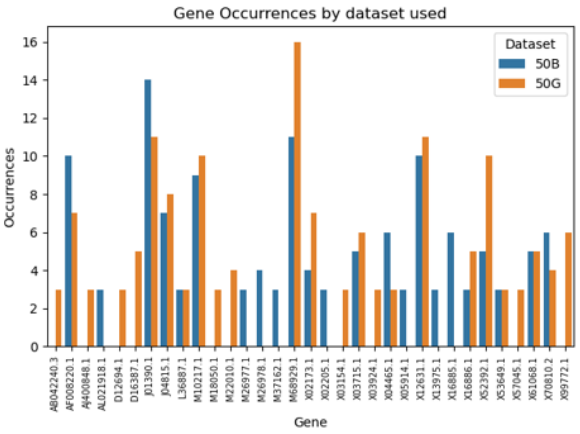

Figure 6: Genes occuring more than 2 times between tRNA datasets B and G.

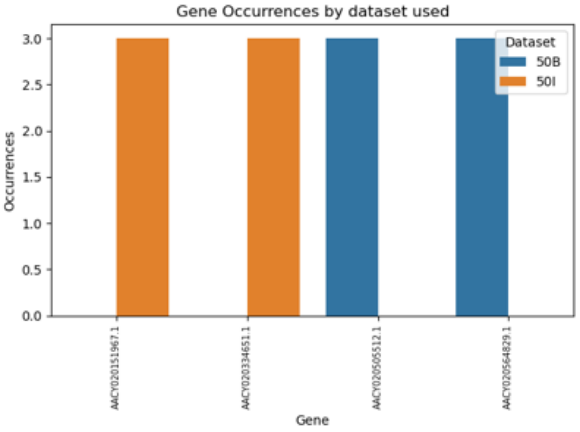

Figure 7: Genes occuring more than 2 times in glutamine riboswitch datasets B and I.

Discussion

Based on the various results presented earlier, it was determined that CaCoFold does not produce identical secondary structures when the input sequences differ, even slightly. This observation is particularly concerning, given that an RNA molecule's function is tightly linked to its precise secondary and tertiary structure. Any alteration—whether in the identity of a nucleotide or in the base-pairing interactions—can compromise this essential function.

Therefore, if CaCoFold consistently predicts different base-pairings depending on the input sequences, we cannot confidently assert that the resulting structure accurately reflects the true secondary structure of the RNA being studied.

Moreover, the absence of spatial constraints, such as distances and angles between residues, limits any meaningful comparison between the predicted 2D structures and experimentally derived 3D models. Visual comparison remains the only option, which is often challenging due to the complexity of RNA structures and the limitations of projecting them onto two dimensions.

That said, CaCoFold remains a useful tool in cases where no experimental 3D structure is available. In such scenarios, it can provide a preliminary structural hypothesis or a rough template of an unknown RNA molecule.

However, the data collection method used in this project can be called into question. The initial assumption was that using the full set of sequences from RFAM would lead to accurate structure prediction. Yet, the resulting models often failed to represent the best possible structures. A potentially better approach would have been to manually curate a set of high-quality, well-sequenced entries, rather than relying solely on the RFAM database.

Conclusion

This project explored the functioning and performance of the CaCoFold algorithm for RNA secondary structure prediction based on multiple sequence alignments. While the tool offers an interesting approach by incorporating both positive and negative evolutionary constraints, the results reveal a high sensitivity to input data. This limits its reliability for precise structural predictions without experimental validation. Nonetheless, CaCoFold remains a valuable tool for generating structural hypotheses when no experimental 3D data is available.

Remark

All data and code used in this project are available at `https://github.com/sdraguet/-INFO-F439-MiB-RNA-structure-predictor`.

An artificial intelligence has been used to correct grammatical errors in this report, as well as to resolve certain problems in the code used.

References

Klähn, S., Bolay, P., Wright, P. R., Atilho, R. M., Brewer, K. I., Hagemann, M., Breaker, R. R., & Hess, W. R. (2018). A glutamine riboswitch is a key element for the regulation of glutamine synthetase in cyanobacteria. Nucleic Acids Research. https://doi.org/10.1093/nar/gky709

Madeira, F., Madhusoodanan, N., Lee, J., Eusebi, A., Niewielska, A., Tivey, A. R. N., Lopez, R., & Butcher, S. (2024). The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. Nucleic Acids Research, 52(W1), W521–W525. https://doi.org/10.1093/nar/gkae241

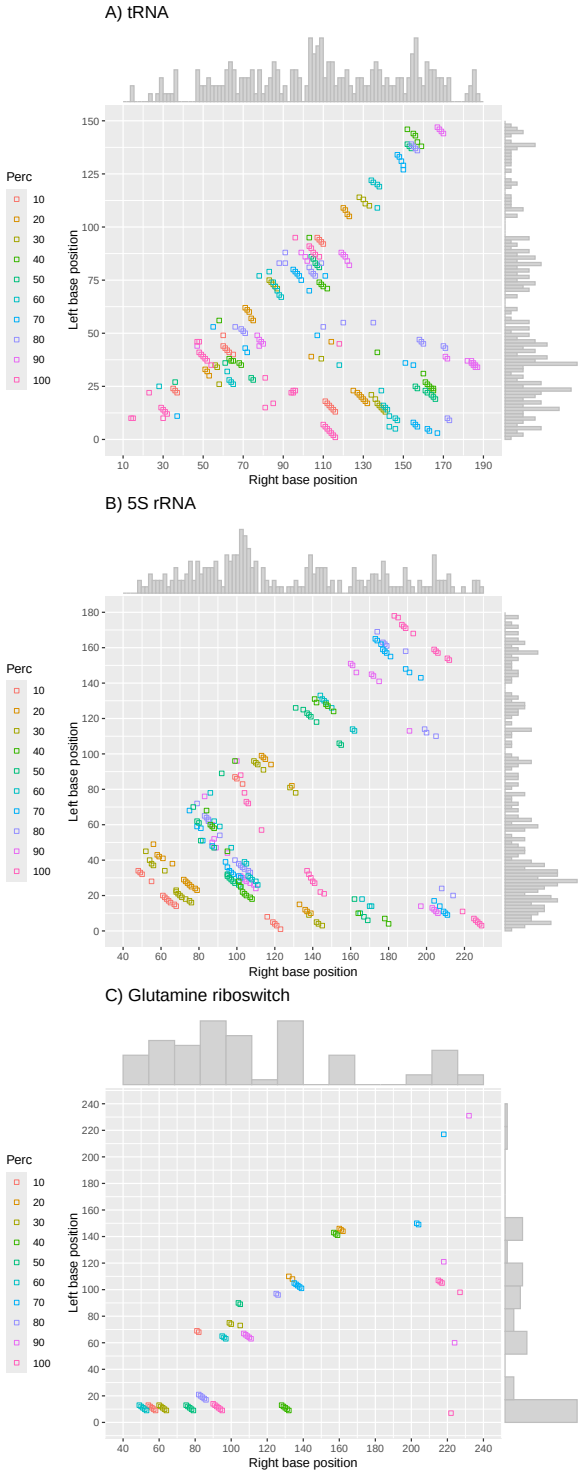

Figure 8: Positions of positive base pairs relative to the percentage of sequences, along with their distribution.

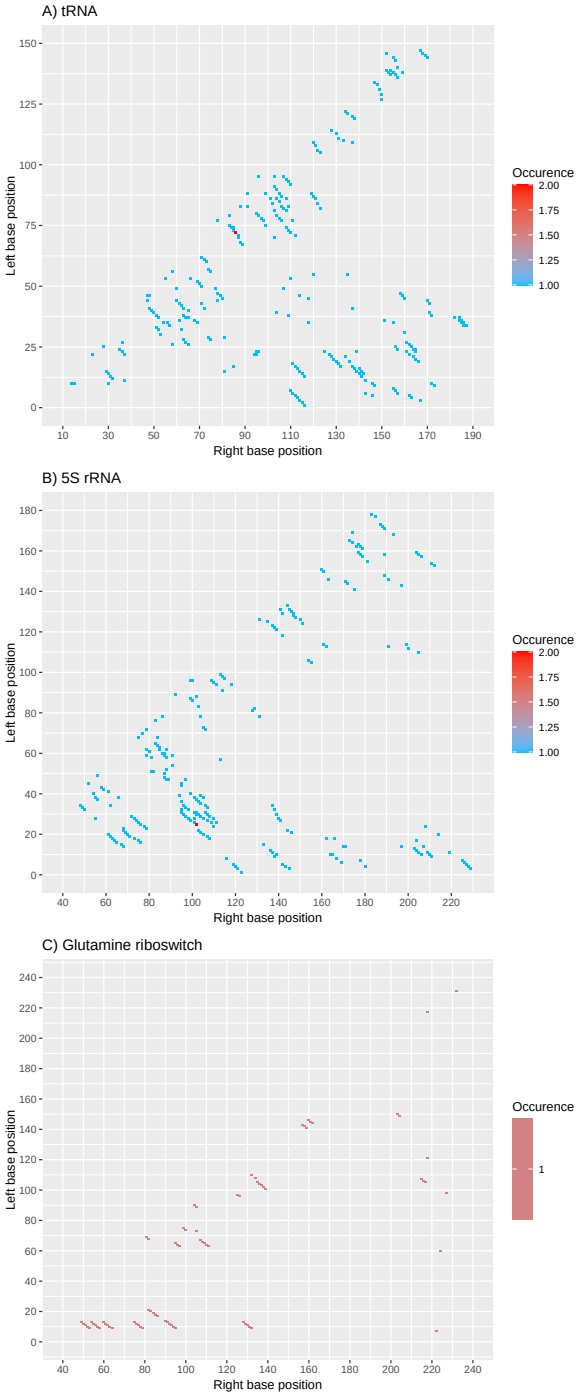

Figure 9: Occurrences of positive base pairs across datasets of different sizes. The base pairs (86,72) in A and (102,25) in B are the only pairs that occur more than once — in both cases, they appear exactly twice.

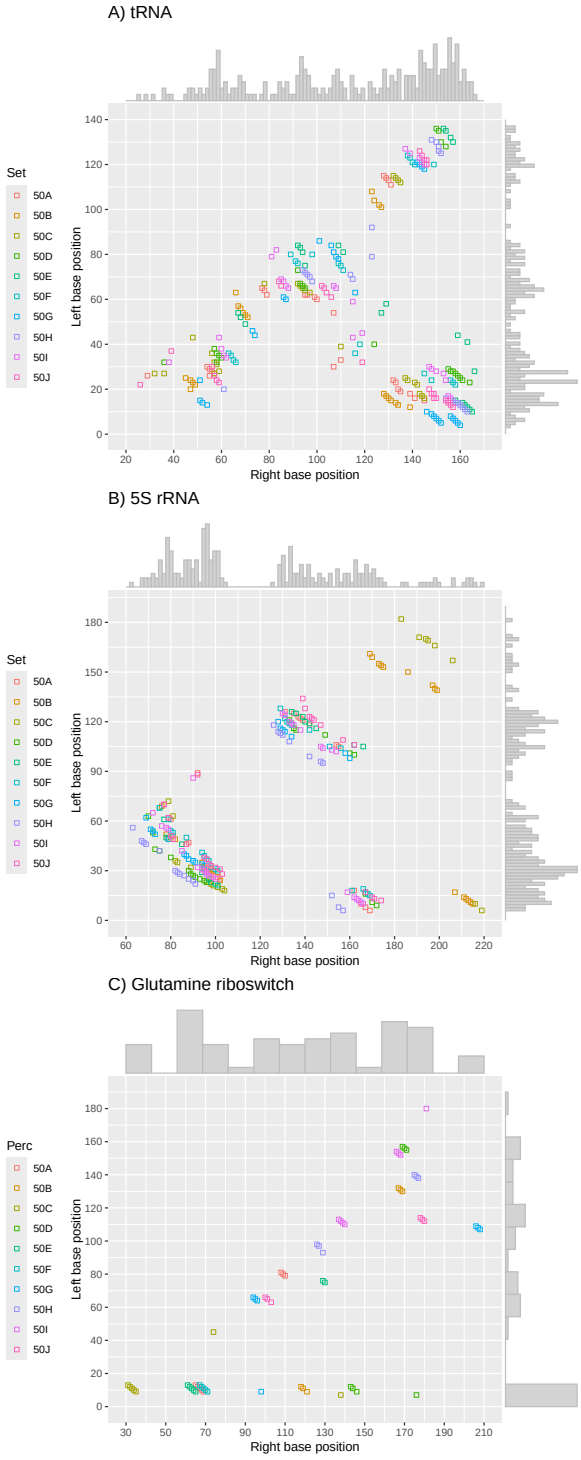

Figure 10: Positions of positive base pairs relative to the 50% sequences dataset used, along with their distribution.

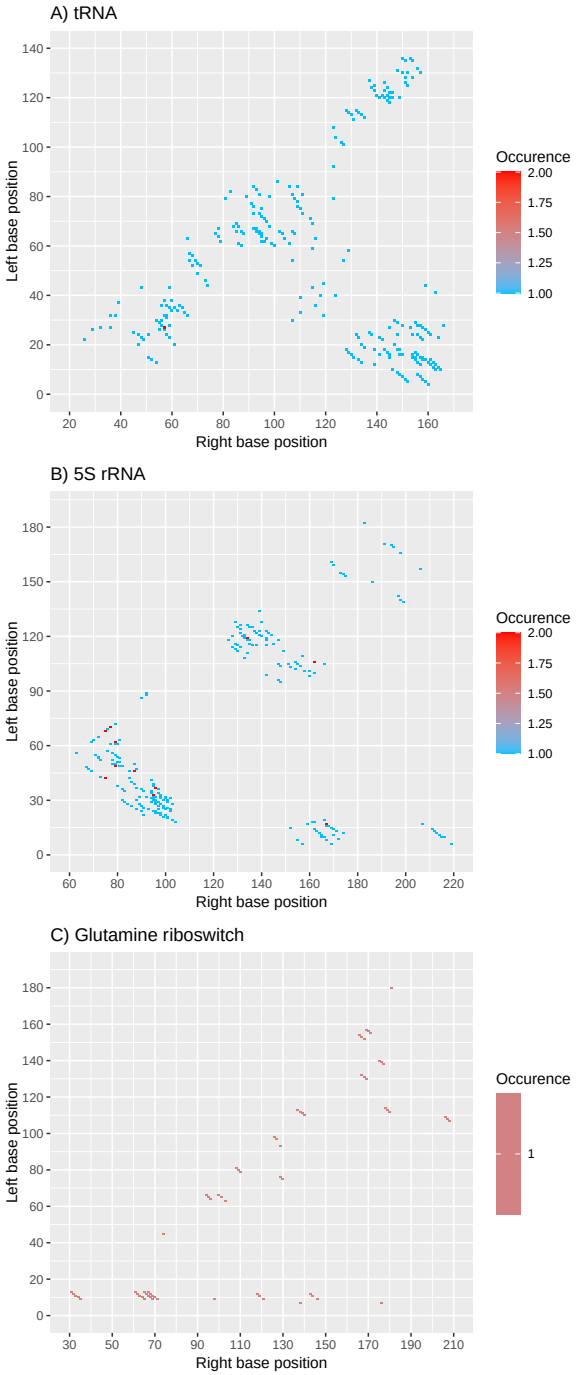

Figure 11: Occurrences of positive base pairs across different 50% sequences datasets used. Among the pairs that occur more than once, everyone of them appear exactly twice.

Appendix

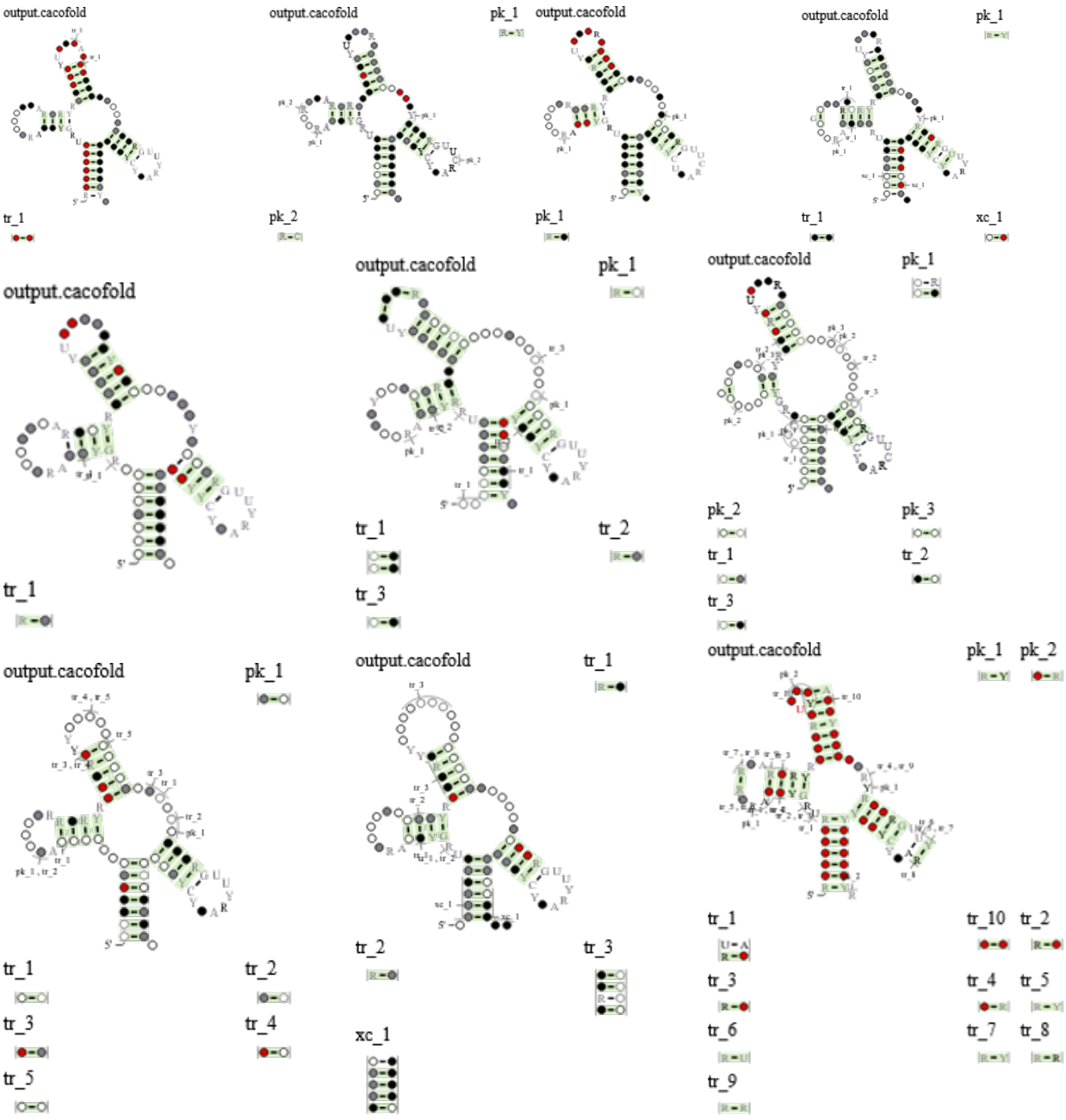

Figure 12: Structures of tRNA determined by CaCoFold, with an increasing percentage of sequences (10, 20, 30, etc.)

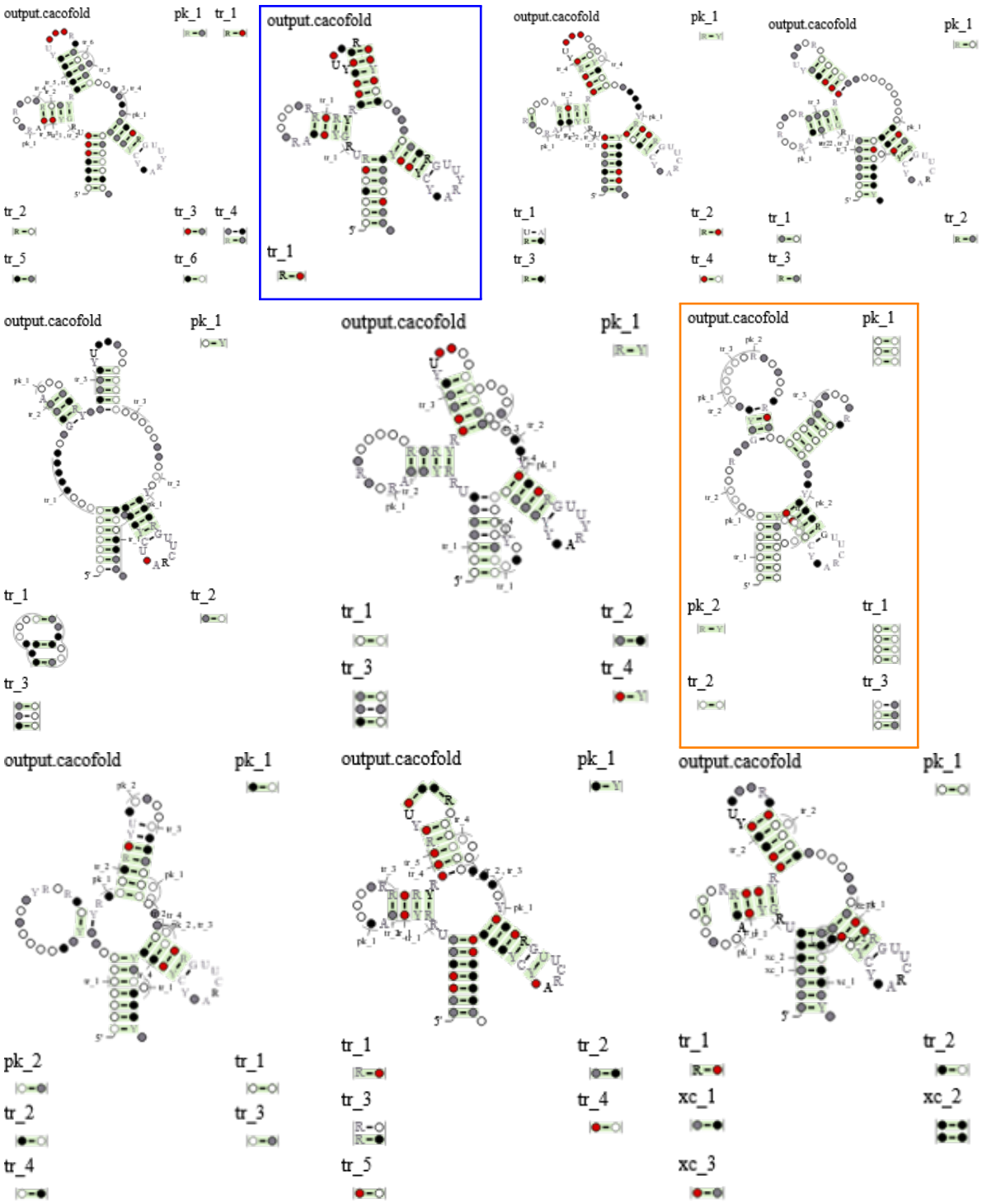

Figure 13: Structures of tRNA determined by CaCoFold, with different sequence datasets (from A to I). The structures B and G, mentioned in the report, are highlighted in blue and orange respectively.

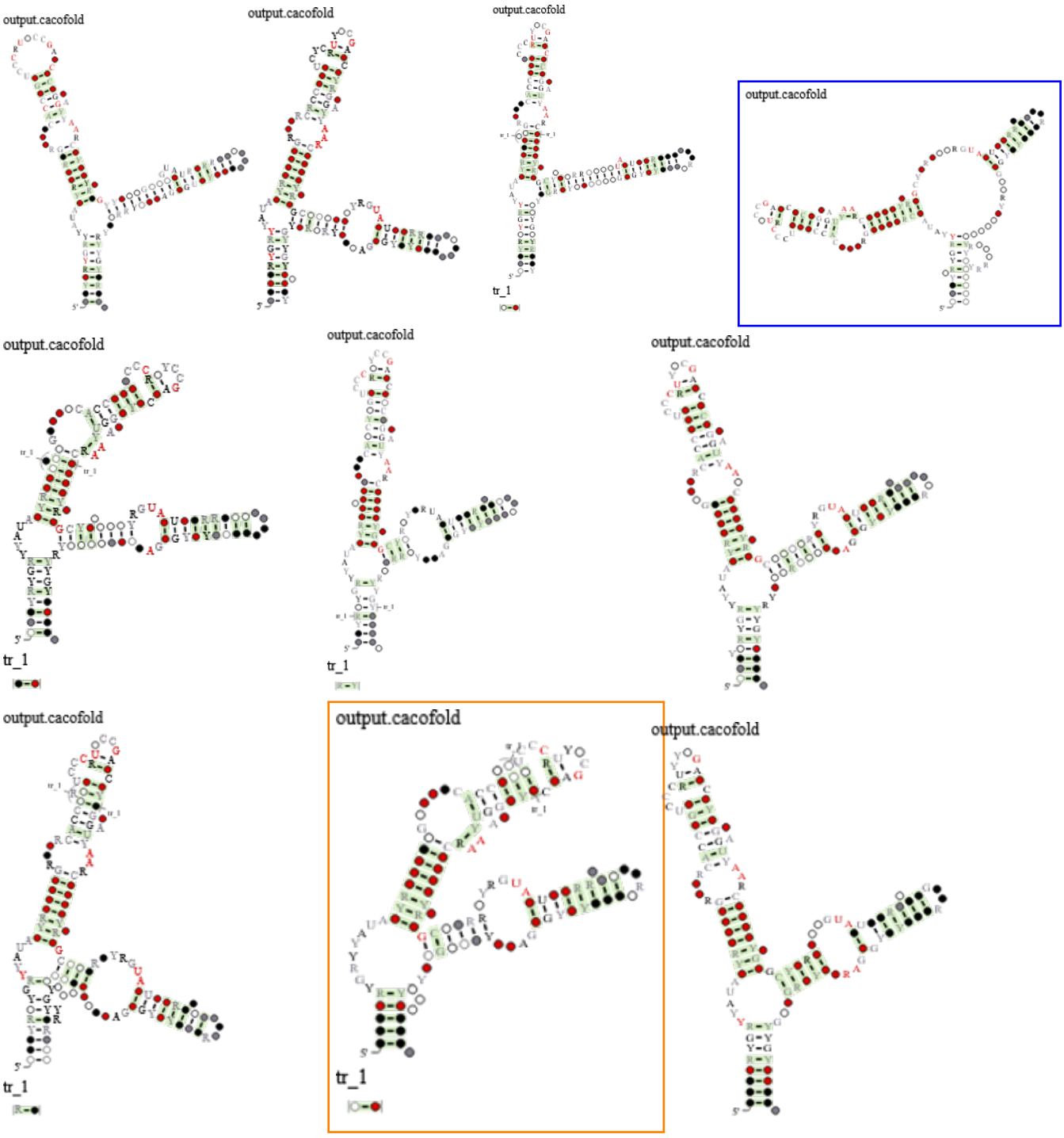

Figure 14: Structures of 5S rRNA determined by CaCoFold, with an increasing percentage of sequences (10, 20, 30, etc.). The structures containing 40% and 90% of the sequences, mentioned in the report, are indicated in blue and orange respectively.

Figure 15: Structures of 5S rRNA determined by CaCoFold, with different sequence datasets (from A to I).

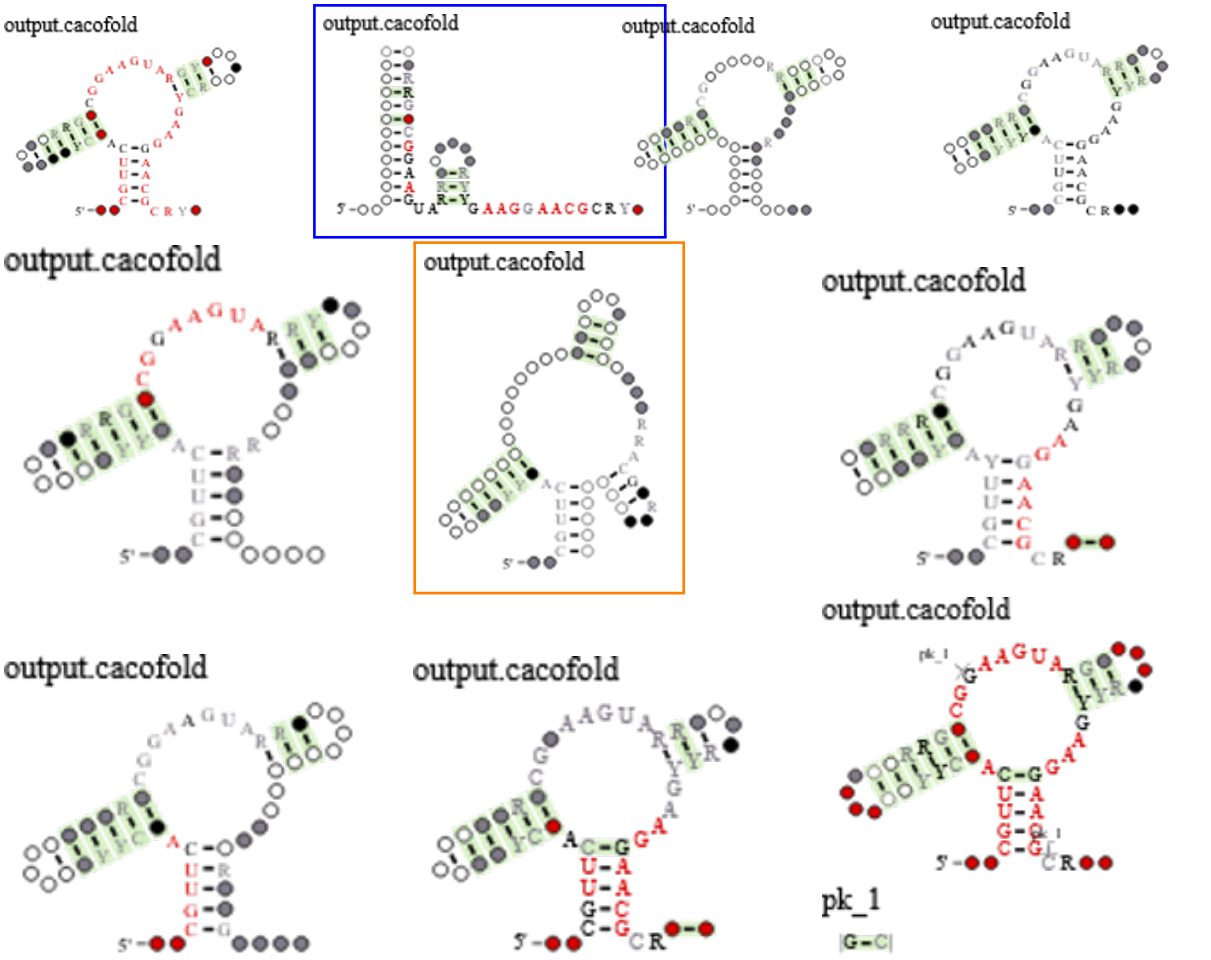

Figure 16: Structures of glutamine riboswitch determined by CaCoFold, with an increasing percentage of sequences (10, 20, 30, etc.). Structures containing 20% and 60% of the sequences, mentioned in the report, are highlighted in blue and orange respectively.

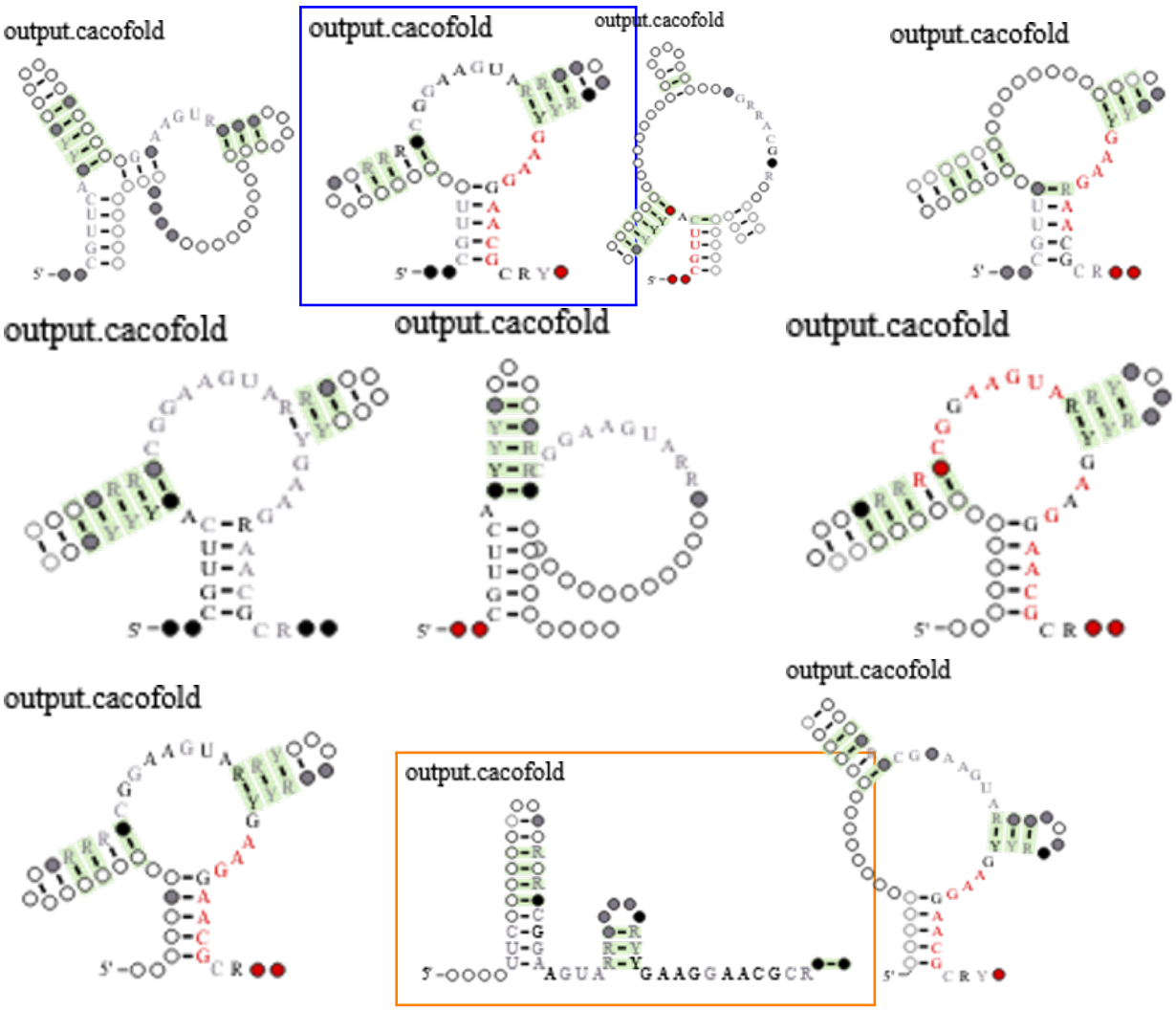

Figure 17: Structures of glutamine riboswitch determined by CaCoFold, with different sequence datasets (from A to I). Structures B and I, mentioned in the report, are highlighted in blue and orange respectively.