Junction-Explorer Help File

Dongrong Wen, Christian Laing, Jason T. L. Wang and Tamar Schlick

Overview

RNA junctions are important structural elements of three or more helices in the organization of the global structure of RNA molecules [3, 4, 6]. A common motif among junctions is the coaxial stacking of helices [2, 3, 4, 9]. This motif occurs when two separate helical elements stack to form coaxial helices as a pseudo-continuous helix. In addition, analysis from solved crystal structures indicates that RNA junctions can be classified into families according to their 3D shape or topology [4, 6]. The information obtained from coaxial stacking and topology (family) prediction can help predict RNA three-dimensional structures and gain a better understanding of RNA tertiary interactions.

Given an RNA secondary structure, the Junction-Explorer web server can identify and locate the junctions on the RNA secondary structure. For each identified RNA junction, the web server is able to predict the presence of coaxial helical stacking and the topology (family) of the junction. Junction-Explorer employs a machine learning algorithm called random forests (RFs) for prediction [1, 7]. The RF classifier uses coaxial helical stacking and junction topology information from solved RNA 3D junctions as training data. Predictions are determined at the secondary structure level based on various features included in the classifier such as sequence, length, context, and thermodynamic parameters from RNA junctions [3, 4, 6, 9]. Junction-Explorer predicts coaxial stacks and topologies for both three and four-way junctions and only coaxial stacks for five and higher-order junctions.

Coaxial helical stacking

In an RNA junction, helices are numbered based on the 5′ to 3′ order and a helix must contain at least two consecutive base pairs. In Junction-Explorer, only the stacking between two neighboring helices is considered. Therefore, for a three-way junction, the web server predicts the stacking between Helix 1 and Helix 2, between Helix 2 and Helix 3, or between Helix 1 and Helix 3. “No stacking” is displayed if there is no coaxial stacking predicted in the junction. In a four-way junction, there are at most two coaxial stackings. For example, a stacking could be between Helix 2 and Helix3 and another stacking could be between Helix 1 and Helix 4. However, in a three-way junction, there is at most one coaxial stacking. The same representation scheme is used for five and higher-order junctions.

Junction topology

Lescoute and Westhof [6] compiled and analyzed the topology of three-way junctions in folded RNAs, classifying these junctions into three families *A*, *B*, and *C* (Figure 1, top). In most of the structured three-way junctions, two of the helices in the junctions stack coaxially. Laing and Schlick [4] analyzed RNA four-way junctions and classified them into nine families, namely *H, cH, cL, cK, π, cW, ψ, cX,* and *X*, according to coaxial stacking interactions and helical conformation signatures. Figure 1 (bottom) lists these nine families.

How to use the web server

Junction-Explorer accepts as input an RNA sequence along with its secondary structure in one of the three formats: Bpseq format, CT format and Vienna dot-bracket format. The user takes the following three steps when using the web server:

1. Paste an RNA sequence and its secondary structure represented in one of the three formats into the blank text field of the web server (or simply click any example button above the text field to retrieve an example RNA molecule).
2. Select the corresponding format option.
3. Click the "Submit" button.

After the user submits the RNA molecule, Junction-Explorer identifies and locates the junctions in the molecule and predicts the presence of coaxial helical stacking and topology (family) of each junction in the input structure. The tool presents as output a detailed report, listing the type, location, loops, presence of predicted helical coaxial stacking, and predicted topology (family) of each identified junction in the molecule. A graphical display of predicted results for each junction is also presented, which allows the user to visualize the stack and family configuration in the junction. A screenshot of a sample output is presented in Figure 2.

Usually, the web server displays the output on the web browser promptly. However, when the size of the input data is too large, processing the input structure becomes time-consuming. Under this circumstance, the web server provides a hyperlink instead. The user can access the predicted result through the hyperlink. When pseudoknots exist in the input structure, the web server uses the K2N tool [8] for pseudoknot removal to create a pseudoknot-free structure before performing junction identification and prediction.

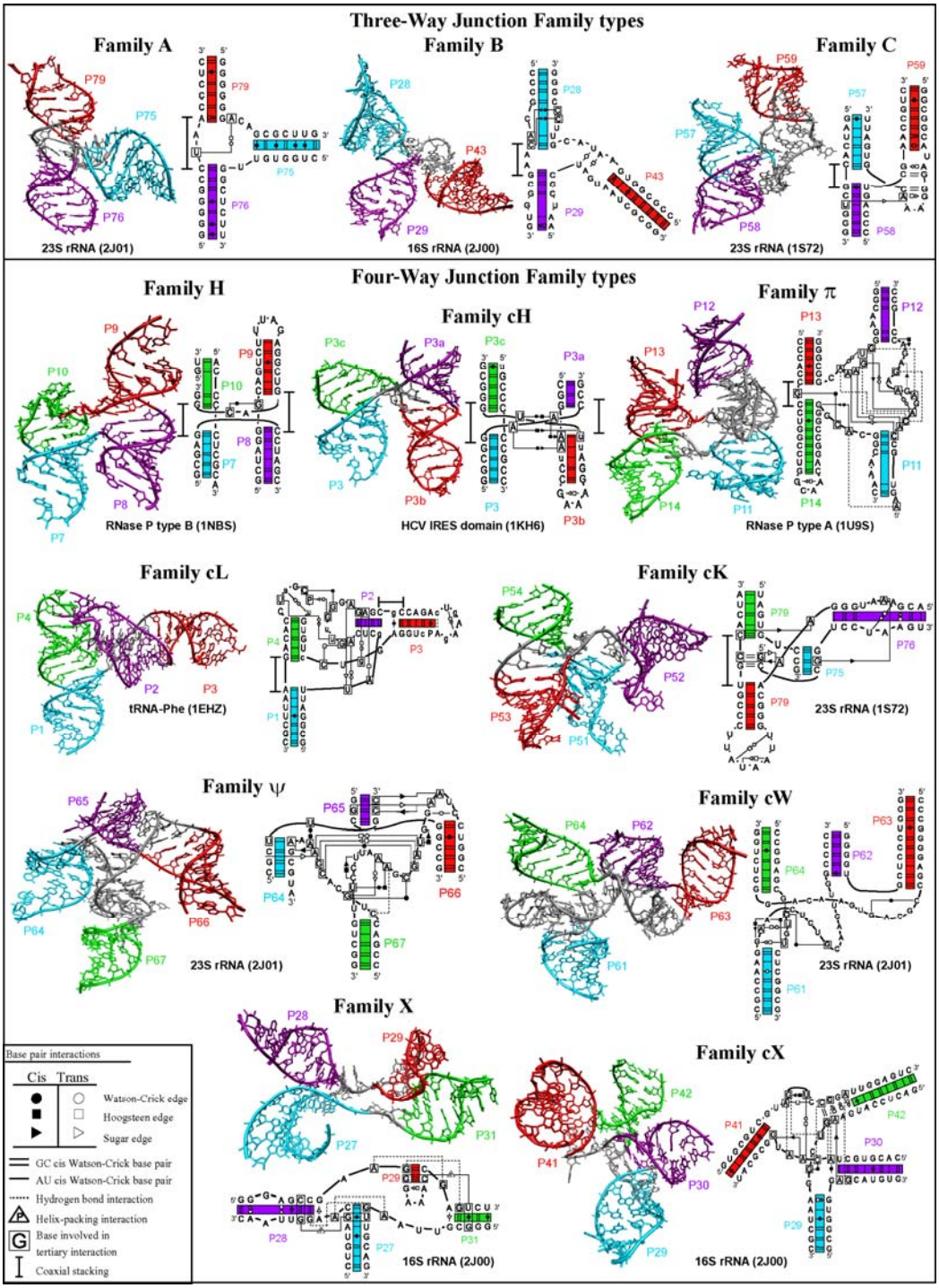


Figure 1. Examples of three and four-way junction families presented in three-dimensional and network diagram representations. RNA three and four-way junctions can be classified into three and nine families respectively according to their coaxial stacking and topology. The network interaction symbology follows the Leontis-Westhof notation [5] (see inset boxes).

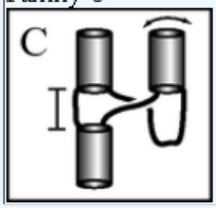
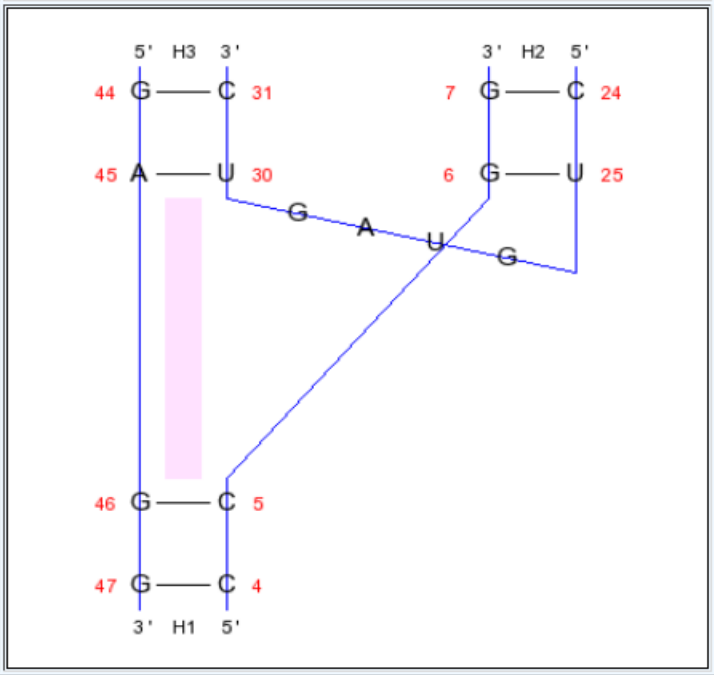
Junction Type	3-way junction
Junction Location	Helix 1 4 C-G 47 5 C-G 46 Helix 2 6 G-U 25 7 G-C 24 Helix 3 30 U-A 45 31 C-G 44
Junction Loops	J12(0): - J23(4): GUAG J31(0): -
Coaxial Stacking Prediction	Stacking between Helix 1 and Helix 3
Topology Prediction	Family C 
Prediction Visualization	

Figure 2. The screenshot showing a predicted report generated by Junction-Explorer.

Input format

The user can input an RNA sequence and its secondary structure in one of the following three formats.

1. Bpseq format: There may be several header lines, each beginning with a "#". A header line beginning with the "#" is for comments, and the header line is not considered as part of the input RNA. Then multiple lines follow the header lines, where each line comprises three columns. The first column contains the position number of a nucleotide. This position number must start with 1. The second column contains the nucleotide name (A, C, G, or U). The third column contains the position number of the base with which the nucleotide is paired. If the nucleotide is not paired with any base, the third column is 0. A space must be used to separate two neighboring columns.

2. CT format: Here the first line constitutes the header of the format, which contains the length and name of the input molecule. In the CT format, each line consists of 6 columns. The first and sixth columns contain the position number of a nucleotide (base). This position number must start with 1. The third (forth, respectively) column contains the position number minus one (plus one, respectively). The second column contains the nucleotide name (A, C, G, or U). The fifth column contains the position number of the base with which the nucleotide is paired. If the nucleotide is not paired with any base, the fifth column is 0. A tab must be used to separate two neighboring columns.

3. Vienna dot-bracket format: Here the first line constitutes the header, which starts with the ">" character followed by the name of the input molecule. The second line contains the input sequence from 5′ to 3′. The third line contains the secondary structure of the input sequence, where a base pair is represented by an opening and closing bracket and an unpaired base is represented by a dot.

If a secondary structure has pseudoknots, it must be input in Bpseq or CT format, but not Vienna dot-bracket format.

Output format

The web server displays a table for each identified junction, listing the following

information concerning the junction. (If no junction is identified, the web server displays a message indicating so.) An example of predicted results is shown in Figure 2.

1. Junction Type: This field shows the type of the junction, which can be three-way, four-way, five-way or higher-order depending on how many helices are involved.
2. Junction Location: This field shows the nucleotides and their positions for the helices involved in the junction. These positions define the location of the junction. For each helix, only the two consecutive base pairs that are closest to the junction are displayed.
3. Junction Loops: This field shows the size and the nucleotides in the loop region between every two neighboring helices. For example, "J23 (4): GUAG" means that the loop region between Helix 2 and Helix 3 contains four nucleotides G, U, A and G. As another example, "J12 (0): -" indicates that the loop region between Helix 1 and Helix 2 has zero nucleotide.
4. Coaxial Stacking Prediction: This field shows the predicted outcome for coaxial stacking, as explained in the Coaxial Helical Stacking section above.
5. Topology Prediction: This field shows the predicted outcome for junction family, as explained in the Junction Topology section above.
6. Prediction Visualization: This field presents a graphical display of the predicted coaxial stacking of helices and predicted topology for the junction.

References

1. Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5-32.
2. Kim, S.H., Sussman, J.L., Suddath, F.L., Quigley, G.J., McPherson, A., Wang, A.H., Seeman, N.C. and Rich, A. (1974) The general structure of transfer RNA molecules. *Proc Natl Acad Sci U S A*, 71, 4970-4974.
3. Laing, C., Jung, S., Iqbal, A. and Schlick, T. (2009) Tertiary motifs revealed in analyses of higher-order RNA junctions. *J Mol Biol*, 393, 67-82.
4. Laing, C. and Schlick, T. (2009) Analysis of four-way junctions in RNA structures. *J Mol Biol*, 390, 547-559.
5. Leontis, N. B. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, 7, 499-512.
6. Lescoute, A. and Westhof, E. (2006) Topology of three-way junctions in folded RNAs. *RNA*, 12, 83-93.
7. Liaw, A. and Wiener, M. (2002) Classification and regression by random

forest. *R News*, 2, 18-22.

8. Smit, S., Rother, K., Heringa, J. and Knight, R. (2008) From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal. *RNA*, 14, 410-416.
9. Xin, Y., Laing, C., Leontis, N.B. and Schlick, T. (2008) Annotation of tertiary interactions in RNA structures reveals variations and correlations. *RNA*, 14, 2465-2477.