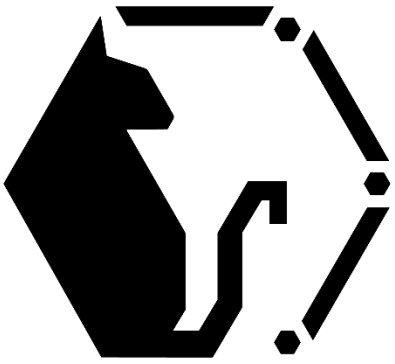
RiboCAT User Manuel

Joshua Hatterschide
William Cantara
Weixin Wu

Karin Musier-Forsyth Labs
The Ohio State University



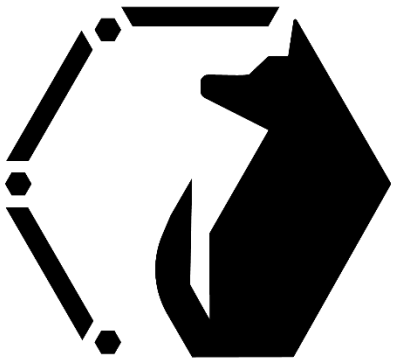
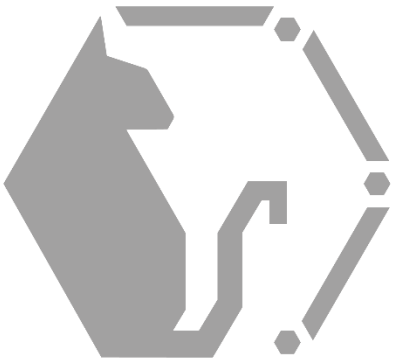
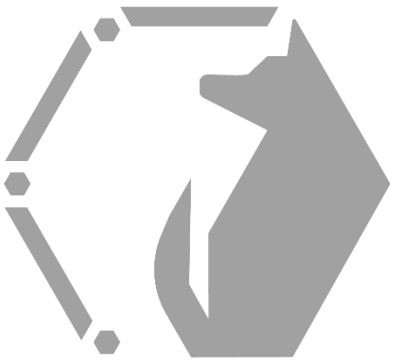


Table of Contents





Requirements and Installation

I. System Requirements

- Both RiboCAT and RiboDOG require at least Microsoft Excel 2010 for all functions to work properly.
- To access .fsa files using RiboCAT, Python 2.X (any version of python 2) is required, and this version must be made the systems Python path.
- Both RiboCAT and RiboDOG are fully functional on Mac and PC; however, Excel for Mac version >15 require a one-time installation step for functionality.

II. Adding Python 2.X to the Path

- Right click on the Start Menu button and click System.
- In the following window, click Advanced System Settings in the left-hand panel.
- In Advanced System Settings, click on the button that says Environment Variables.
- Find the PATH variable, click edit, and change it to the directory of your python.exe file for Python 2.X (e.g. "C:\Python27").
- For more information, see the following link: http://pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/

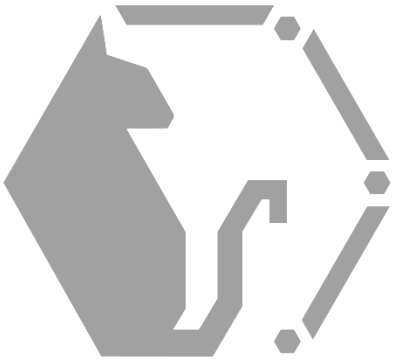
III. Excel Auto Save Suppression

- If users are experiencing significant lag in the Excel document outside of processing time, this could be the result of excessive auto saving.
- PC: Auto saving in Excel can be adjusted or completely turned off by going to File, Options, and then clicking on the "Save" tab. Here the time between auto saving can be adjusted as desired.
- Mac: Auto saving in Excel can be adjusted or completely turned off by clicking Excel, Preferences and the going to the "Save" tab. Here the time between auto saving can be adjusted as desired.

IV. Excel for Mac Versions 15 and Up

- Unfortunately, changes made to VBA in Excel versions ≥15 for Mac had detrimental impacts on accessing files outside of Excel. As a result, some functions of RiboDOG are currently not functional on this platform, and this one-time one-file installation is required for full RiboCAT functionality:

1. Go to https://research.cbc.osu.edu/musier-forsyth.1/tools/ and download the file "RiboCATCallPython.scpt"
2. Open Finder > Hold the Alt key and click Go in the Finder menu bar > Click Library
3. Click Application Scripts (if it does not exist, create this folder)
4. Click com.microsoft.Excel (if it does not exist, create this folder, this is **case sensitive**)
5. Copy "RiboCATCallPython.scpt" to the com.microsoft.Excel folder
6. RiboCAT should now be fully functional



RiboCAT Instructions

I. Data Import

- Press the *Browse/Import* button, select a sequencing, minus, and plus FSA file, and press open. The data will then appear to the right of the buttons. If you already have a project saved in a ".ribocat" file, this can be opened here as well.

- To transfer the data to the "StdAlign" sheet, press the "Copy Data" button. This should pull up a user form. Select the file name that corresponds to each trace and the correct dyes used in experiments (See dye table). Then press "Okay." The data should now be copied to the "StdAlign" sheet under the "Size Standard Raw Data" and "Experimental Raw Data" headers.

Abbreviated Dye Table

Dye 1	Blue Dyes
Dye 2	Green Dyes
Dye 3	Yellow Dyes
Dye 4	Red Dyes
Dye 5	Orange Dyes

- *Example:* an experiment may use NED dye (a Dye 3) for the experimental data (Minus, Plus, and Sequencing), and LIZ dye (a Dye 5) for the size standards which would be included in all three capillaries.

- See the *Applied Biosciences Standard Dye Sets* Table for more dye information.

- *Optional:* Press the "Clear Data" button to clear the data on the "DataImport" sheet.

- *Optional:* Press the "Clear RiboCAT" button to clear the data from the entire project. Be sure to only use this function after data has been exported using the "Export Data" function on "ScaleNorm" sheet. (See *Reactivity Calculation and Sequence Alignment* for more information on Data Export).

II. Signal Alignment

- On the "StdAlign" sheet, press the *Pick Size Standards* button to complete the size standard picking process.
    - This will attempt to pick the peaks by first ensuring the largest two size standards from the list are picked correctly. Click "No" if the peak in question is incorrect. The program will then attempt to identify the next most likely peak in descending order.

- If the program is incapable of identifying the largest two peaks, these can be entered manually, and the remaining peaks should be able to be identified programmatically.
    - To do this scroll left and identify the largest two size standard peaks from each trace and enter their X-values into the "Size Standard Peak Picking" table. Then press the *Pick Size Standards* button.

- In the case of messy size standard traces, difficult peaks in the size standard list may be picked

Applied Biosystems Standard Dye Sets for Genotyping Applications

Dye Set	DS-02	DS-20[1]	DS-30	DS-31	DS-32	DS-33	DS-34	DS-40
Filter Set	E5	A	D	D	F	G5	C	S
Blue Dyes	dR110	5-FAM™	6-FAM™	6-FAM	5-FAM	6-FAM	6-FAM	6-FAM
Green Dyes	dR6G	JOE™	HEX™	VIC®	JOE	VIC	TET™	dR6G
Yellow Dyes	dTAMRA™	TAMRA™	NED™	NED	NED	NED	HEX	
Red Dyes	dROX™	ROX™	ROX	ROX	ROX	PET®	TAMRA	
Orange Dyes	LIZ®					LIZ		LIZ

[1] Dye primer matrix standards.

manually and optimized to the local maximum by pressing the *Optimize Manually Picked Size Standards* button.

- The size standard list can be altered to accommodate other size standard sets.
    - This can be done by changing the molecular weight values in the size standard position table (make sure that modified lists begin in row 8 and do not contain any gaps).
    - *Note:* The alignment algorithm fits a 9th order polynomial to the size standard molecular weights and migration times. Therefore, the closer the number of size standards gets to 9, the worse the fit will be.
- When confident in the size standard peak assignments, select the *Align Data* button to align the Size Standard and Experimental data.
- Finally, check the quality of the data and in the bottom right charts, select an X_{nt} analysis range, and enter it into the "Range" table above the experimental alignment. Then press the *Truncate Data* button.
- *Size Standard Picking Equations:*

Linear Approximation Equation:

$$X_{o,i} \cong m_{i+1}^{i+2}\left(X_{nt,i} - X_{nt,i+1}\right) + X_{o,i+1} \quad (1)$$

Min/Max Constraint Equations:

$$min = \langle A_{i+1}, A_{i+2}\rangle/5 \quad (2)$$

$$max = 2.5\langle A_{i+1}, A_{i+2}\rangle \quad (3)$$

Here, $X_{o,i}$ is the migration time of a size-standard peak i, $X_{nt,i}$ is the nucleotide length of size-standard i, and A_i is the amplitude of size-standard peak i.

- *Alignment Equations:*

Polynomial Fit Equation:

$$X_o = B + \sum_{j=1}^{9} M_j {X_{nt}}^j \quad (4)$$

Here, X_o is a migration time data point, B and M_j are the polynomial constants determined by fitting this equation size-standard peak data, and X_{nt} is the resulting nt based X-axis value.

III. Preprocessing

- On the "PreProc" sheet, *Execute Smoothing*, *Correct Baseline*, and apply the *Signal Decay Direction* in that order.
- The smoothing and baseline correction windows can be adjusted as desired. However, excessive smoothing and baseline correction can result in loss of information from the data.
- *Preprocessing Equations:*

Smoothing Equations Window Size 1 and 2:

$$SA_i = \langle A_{i-1}, 2 * A_i, A_{i+1}\rangle \quad (5)$$

$$SA_i = \langle A_{i-2}, 2 * A_{i-1}, 4 * A_i, 2 * A_{i+1}, A_{i+2}\rangle \quad (6)$$

Baseline Correction Equation:

$$BA_i = A_i - \{MIN(A_{i-ws}, \dots, A_{i+ws})\} \quad (7)$$

Here, A_i is the intensity of data point i, $SA_i$is the intensity of the smoothed data point i, and BA_i is the intensity of the background subtracted data point i.

Termination Probability Equation:

$$P_{term}(i) = \frac{I(i)}{P_{unk} + \sum_{j=i}^{k} I(j)} \quad (8)$$

Unknown Probability Equation:

$$\sum_{i=1}^{k/2} P_{term}(i) - \sum_{i=1+k/2}^{k} P_{term}(j) \approx 0 \quad (9)$$

Here, the probability of termination (P_{term}) at a given nt is equal to the intensity of that nt divided by the sum of intensities for every nt in the RNA. This calculation must include the probabilities for the data within the user-specified range up to nt k ($\sum_{j=i}^{k} I(j)$), as well as the unknown probabilities for the data beyond nt k (P_{unk}). The algorithm determines the value for P_{unk} that minimizes the difference between the sum of probabilities in the first ($\sum_{i=1}^{k/2} P_{term}(i)$) and second halves ($\sum_{i=1+k/2}^{k} P_{term}(j)$) of the electropherogram.

IV. Peak Picking

- Press the *Sequencing Peak Pick* and *Experimental Peak Pick* buttons.
- RiboCAT will automatically run checks to inform the user of erroneously picked peaks. To view these checks and edit the peak list, press the "Open Peak Editor" button.

- The "Minus" and "Plus" pages of the Peak Editor contain lists of peaks that have been identified as either too close together or too far apart. The "Mismatch" page alerts the user of peaks that have been identified in one trace, but not in the other.

- To view the peaks to change, the user can click on a peak in the list and select "Go to Peak." This will center the screen on the peak of interest. The user may then decide whether to add or remove peaks from the region.

    - *Note:* All warnings on the Mismatch page should be alleviated before proceeding to Gaussian fitting. This is because it can be assumed that a peak must be added or removed if the traces are in disagreement at a certain X_{nt} location. However, after corrections have been made, it is likely that a few checks may remain on the Minus and Plus pages.

- For larger data sets, it can be useful to keep track of peak changes on scratch paper in a table similar to the one below:

Previous Peak	Check Value	*Add*	*Remove*
98.2	100.1	99.2	
114.0	114.5		114.5
130.1	130.5		130.1
144.3	145.9	X	X
…	…	…	…

- *Peak Picking Equations:*

Peak – Sharpening Equation:

$$Y_{Sharp} = Y_{Preproc} - Y''_{Preproc} \quad (10)$$

The sharpened amplitude (Y_{Sharp}) is calculated as the difference between the amplitude of preprocessed data ($Y_{Preproc}$) and the second derivative of the preprocessed data ($Y''_{Preproc}$). Peaks are then identified as the local maxima within a moving window of ~1.1 nt on the X-axis of the peak-sharpened data. Importantly, the peak-sharpened data is used only to identify the X-axis position of each peak. All further analysis is based on the unsharpened, preprocessed data.

Peak Distribution Check Criteria

$$0.6 < X_{nt,i} - X_{nt,i-1} < 1.5 \quad (11)$$

Although peak separation will vary throughout the length of the capillary, analysis of many datasets revealed that peaks are rarely separated by less than 0.6 or more than 1.5 X-axis units. Therefore, the Minus and Plus checking systems were implemented to alert the user of locations where the identified peaks are not within 0.6-1.5 units of separation. Additionally, peaks are occasionally identified in one trace that are not identified in the other. Therefore, the Mismatch checking system was also incorporated to notify the user of peaks in one trace that have no match in the other.

V. Gaussian Fitting

- On the "GaussFit_Minus" and "GaussFit_Plus" sheets press the *Calculate Gaussians* buttons.

- Assess the quality of the Gaussian fittings for both traces. Quality can be determined by the extent of overlap between the Gaussian sums and the raw data. Then select a continuous X_{nt} range corresponding to quality fitting.

- Enter this into the Min/Max X_{nt} table on the "Scale_Norm" sheet.

Gaussian Equation:

$$y = Ae^{-(x-P)^2/\sigma^2} \quad (12)$$

The parameters A, P, and σ of the Gaussian function are unique to each peak of a CE trace and represent the amplitude, X-axis position, and width of the peak at half-height respectively. Values for these parameters are calculated using a moving window that includes the peak of interest, as well as one peak to either side. The A, P, and σ parameters are varied for the peak of interest and the y-values of the three peaks are summed at every data point within the window and compared to the corresponding preprocessed data. The optimum value for a parameter is selected as the value that gives the lowest error between Gaussian approximations and the preprocessed data within the window. Peak areas are then calculated by integrating the Gaussian functions.

VI. Reactivity Calculation and Sequence Alignment

- On the "Scale_Norm" sheet, press the *Scaling* and *Normalization* buttons to complete the reactivity calculations.

- Make sure to create a sequence file that is formatted as a text file containing only the RNA sequence on the first line with no spaces.

- Press the Browse *button* next to the "Seq. File" cell to find and open the sequence file corresponding to the RNA being analyzed.

- Enter the ddN used in the sequencing trace, and chose an offset value. (For most probing methods including SHAPE the offset should be -1). Then press *Align Sequence*.

    - *Note*: The calculated RMSD displays the average difference between the positions of

the experimental peaks and the sequencing peaks they are paired with. In this context, a good RMSD is usually <1.

- *Reactivity Calculation Equations:*

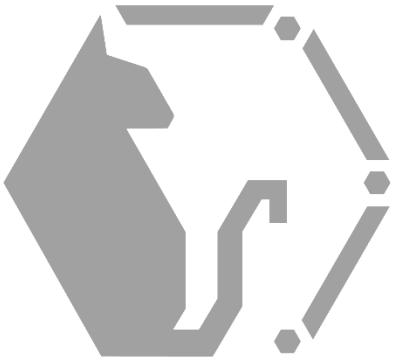
Scaling Equation:

$$\alpha = \frac{\langle A_{(20\%,+)} \rangle}{\langle A_{(20\%,-)} \rangle} \quad (13)$$

Normalization Equation:

$$R(nt) = \frac{A_{bs}(nt)}{\langle Top\ 10\%\ A_{bs} \rangle} \quad (14)$$

Here, a scaling factor (α) that is calculated by dividing the average of the lowest 20% ($A_{(20\%,+)}$) of plus peaks by the average of the lowest 20% of minus peaks ($A_{(20\%,-)}$), and The normalized reactivity values are determined by first subtracting the minus-peak areas from the plus, and then dividing the resulting background-subtracted values ($A_{bs}(nt)$) by the average of the top 10% of these values ($\langle Top\ 10\%\ A_{bs} \rangle$). Outliers are excluded from the top 10% calculation if they are greater than 1.5 times the interquartile range.



RiboDOG Instructions

I. Data Summarization

- Open an empty RiboDOG file, select a sequence file, and enter the number of primers used in experiments. Additionally, select the desired orientation of the data from the drop down menu.

- Select the RiboCAT files for each primer, and the press the *Import Data* button.

II. Post-analysis of X_{nt} values

- Due to the nature of the signal alignment used in RiboCAT, the X_{nt} values for each nucleotide should be highly reproducible between replicates of RNA-probing experiments.

- Thus, the data summarization by RiboDOG can be used as a way to double check the identity of the peaks identified for each nucleotide.

- This can be done easily by looking at the Range column of the chart. As these values are roughly based off of nucleotide length, ranges that are $\gtrapprox 1$ signify places where the peak picking may have differed between replicates.

- An example of this method can be seen below with example data:

nt	Replicate 1	Replicate 2	Replicate 3	Range
50	250.1	250.3	250.2	0.2
51	249.1	249.3	249.3	0.2
52	248.2	248.2	248.3	0.2
53	247.2	247.3	247.4	0.3
54	246.1	245.2	246.5	1.3
55	245.1	244.1	245.3	1.2
56	244.0	243.0	244.5	1.5
57	242.9	241.9	243.2	1.3
58	242.0	241.0	242.3	1.3
59	241.0	240.0	241.2	1.1
60	239.9	239.0	240.1	1.1

In this example data, nucleotide 54 and all of the nucleotides after it show ranges of $\gtrapprox 1$. Therefore, there may be an extra peak in one of the replicates, or a missing peak in one of the replicates.

nt	Replicate 1	Replicate 2	Replicate 3	Range
50	250.1	250.3	250.2	0.2
51	249.1	249.3	249.3	0.2
52	248.2	248.2	248.3	0.2
53	247.2	247.3	247.4	0.3
54	246.1	Add Peak	246.5	0.4
55	245.1	245.2	245.3	0.3
56	244.0	244.1	244.5	0.5
57	242.9	243.0	243.2	0.3
58	242.0	241.9	242.3	0.4
59	241.0	241.0	241.2	0.2
60	239.9	240.0	240.1	0.1

By adding a space for a peak in Replicate 2, we see that this fixes this range and all of those after. Therefore, it is likely that a peak must be added at this position.

- Press "Recalculate Statistics" on the "Import Data" sheets to see the effects that shifting columns has on the range.

III. Trace Visualizer

- To open the Trace Visualizer, click the "Open Trace Visualizer" button on the first page.

- The Trace Visualizer has two tabs with distinct functions. The "Visualize" tab allows users to select traces from different replicates of the same primer to compare them as this can be very helpful in determining whether discrepancies in X_{nt} values indicate extra or missing peaks.

- The "Edit" tab allows users to see how the addition or subtraction of a peak would affect the Gaussian fitting by pressing the "Update Preview" button. It also allows the users to incorporate these changes by pressing the "Incorporate Alterations" button. To accomplish this, RiboDOG re-runs the RiboCAT functions of Gaussian Fitting, Scaling, Normalization, and Sequence Alignment. It will update the altered ".ribocat" file with the results of all of these recalculations, and re-import the data into RiboDOG.

    - *Note:* The Trace Visualizer is the only function of RiboCAT or RiboDOG that does not work identically on Mac and PC operating systems for a few reasons. 1) It is not possible for user forms to be opened in Mac as "modeless," meaning the user cannot freely scroll around the workbook while the form is open. For this reason, when run on a Mac, it will automatically take the user to the region of the chart that pertains to the

information entered. 2) It is also not possible to load charts or images into user forms in Mac. To account for this, when run on a Mac, RiboDOG will pull the charts up in the "Preview" application to be viewed. While this still allows Mac users all of the functionality of PC users, it can be slightly slower as Preview is opening.

IV. Data Export

- To generate a .shape file simply return the RiboDOG home page, and click the "Generate RNA Structure File" button after analysis of the summarized data is complete.