App Introduction

Detection of Unknown Events with Tunable Thresholds (DUETT) provides a quantitative and automated method to detect events in cotranscriptional SHAPE-Seq data. Events come in two categories: swing events are sudden changes in SHAPE-Seq reactivity either going from low to high (upswing event) or high to low (downswing event); linear ramps are long events that persist for multiple positions to create a slope upwards (upramp) or downwards (downramp). All events are detected down each column in the data matrix; DUETT detects swing events using equations inspired by PIR controllers to test for proportional, integral, and relative changes in reactivity. DUETT performs ramp detection performing linear regression and testing for noise, steepness, and uniformity.

The provided example data illustrates this app’s use. The provided data comes from a SHAPE-Seq experiment that measures RNA folding as a function of transcript length. See Example Data Descriptions below.

This app is interactive meaning any change you make on the front should update on the user interface (after pressing the update button). If you make a change that breaks the code/figure, you can either reload the app (losing all progress) or revert the change. This app comes with an automated way to fit parameters though it can take upwards of several minutes.

Consult the manuscript [citation] for more detailed and rigorous explanations.

App Startup

To start the app, make sure R and RStudio are installed. Both can be found for free online. Double click on ui.R and RStudio pops up. Press Ctrl-Shift-Enter to run the app and a user interface pops up with the examples already displayed. To exit or restart the app, simply exit the RStudio pop-up (RStudio does not need to be restarted).

File Inputs and Outputs

For the file input, look in the example\_data folder and the “SRP\_wt\_rho\_table.txt” file. This tab-separated file has 125 columns and 96 rows. In principle, any 2D matrix can be loaded. Note that both column row names should not exist in the txt file.

An input data file can be loaded using the browse button in Input data file. If you do not input anything, then the “SRP\_wt\_rho\_table.txt” file will be loaded instead. You can input multiple files/replicates to view events that share consensus across replicates. If using multiple files, all files must have the same data dimensions.

You can specify your output file name, and the default name will be example\_output if you specify nothing. Once you have created your plot (explained below), you can export results as a table .csv file or export the figures in a pdf. If you are using multiple replicates, you can check the “Print replicate info” box to print out the results from each individual replicate in addition to the shared consensus.

Plotting Parameters

To familiarize yourself with this app, start with the example data. First, press “Update plot” at the bottom of the right-most panel. This will update the figure and takes several seconds. There are two figures in this app: first is a heatmap showing the data with events highlighted in red/blue and the second shows the reactivity and detected events in selected columns. Upswing/downswings are shown in red/blue boxes, linear ramps are red/blue lines, and concurrent events are shown with a green dotted line.

This app supports basic visual changes:

* Figure width and figure height adjust the figure sizes in inches.
* Check diverging data if the data has negative values. This also changes the color scheme from grayscale to blue-white-red diverging scale where white corresponds to zero.
* Check Log the colors to convert the colors to log-scale. This allows lower magnitude values to show up as more colorful.
* Number rows/column should be self-explanatory
* Numbering interval controls the intervals for both the row/columns
* Numbering offset adds (or subtracts) a set integer value from row labels
* Resize axis labels and resize boxes scales the row/column labels and the boxes, respectively. One corresponds to no change.
* y-axis range specifies the y-axis in the column details. Specify two numbers, the min and max of the y-axis separated by a comma.
* Columns to display allows the user to specify what columns to show below, and Show all columns automatically shows all columns regardless of the box above. To specify columns, use numbers only and separate with a comma (without spaces)

PIR Parameters

The PIR thresholds are responsible for detecting swing events. DUETT calculates PIR values for each transcript length and a swing event is detected if all three PIR values exceed the PIR threshold.

*Pre-processing*

Before describing the parameters, some data processing is required to attenuate noise. We take the mean of a sliding window down each column and set aside these new values as “pre-event values.” The values are stored in the last position of the sliding window. Taking the mean smooths out noisy events, and the original data is still retained as “post-event” data. For all three parameters, the pre-event values will be compared to the post-event data. The size of the sliding window is specified with *Window size*.

*Proportional (P)*

The proportional (*P*) parameter is the absolute change from pre-event data to post-event data (the name is confusing because it is adapted from PID controllers), shifted by one position. For example, if the pre-event value is 1.2, and the next transcript length has a post-event value of 1.6, then the differential change is 0.4. P is necessary to filter out low magnitude changes that are due to noise.

*Relative (R)*

Relative (*R*) is the relative change and is the same as P except divided by the magnitude of the pre-event value. For example, if the pre-event value is 4, and the next position has a post-event value of 5, then the proportional change is 0.2 (increase by 20%). R is necessary to filter out proportionally-low changes in high magnitude measurements. For example, a nucleotide with high reactivity will have noise that causes large absolute fluctuations, creating high values in P. However, the relative change is small and the noise falls below the R threshold.

There is another wrinkle with relative change. An upswing and a downswing event of equal magnitudes do not appear the same with relative change. For example, going from 4 to 6 is a relative change of +50%, but going from 6 to 4 is a P change of -33%. Instead, the R threshold is specified in the upswing direction and is automatically adjusted for the downswing direction. This equation allows equal magnitude upswings/downswings appear consistent in terms of an R threshold:

*Integral (I)*

Integral (*I*) change integrates the values of P across multiple positions. By default, I is integrated over additional positions of the same length as window size, but this can be overridden in I length. I helps to remove sharp transient changes or anomalies that persist over very few positions. This is based on the assumption that real events persist for several transcript lengths.

Noise Parameters

These noise parameters are included in the PIR parameters panel and can be used for final noise filtering (applied only to swing events). From experience, even with meticulous PIR parameter tuning, there will always be some events that are misrepresented due to noise.

First, some real event may have gaps that were not detected. Instead of a continuous series of detected events, there will be gaps. We assume that noise causes real events to occasionally have gaps in an otherwise continuous series of detected events. This is solved by the *event gap* parameter, which fills in gaps of the specified length between detected events. To disable, set the parameter to 0.

Second, some false events appear as detected swing events. We assume that noise occasionally creates short instances of detected events. We solve this by removing any event series that have a length at or below the *noise length* parameter. This parameter can be disabled by setting to 0.

Linear Ramp Parameters

The linear ramp parameters control detection of linear ramp events. Linear regression is conducted over a specified *ramp length* and if the fitted line passes three conditions (*p-value*, , and *DWS*), then the entire window is marked with a linear ramp. Linear ramps take on the form of where is reactivity, is consecutive integers from 1 to the ramp length, and is the ramp slope

*Ramp length*

Ramp length controls the window size to fit the linear ramp. This parameter should generally be long to distinguish it from swing events detected by the PIR parameters. A rule of thumb is to set the ramp length to your expectations of the shortest ramp in the data. It should be noted that longer ramp lengths suffer from over-elongating the ramps; a short ramp will be detected as longer because a large proportion will still allow it to pass the three conditions below.

*Ramp p-value*

Ramp *p*-value asks if the noise around the linear fit is too high. A lower p-value means there is less variance around the line and that the linear ramp is more likely to truly exist, rather than because of pure noise.

Because many lines are being fitted, it is suggested that you choose a *p*-value by using multiple hypothesis correction, like the Bonferroni correction. A Bonferroni correction would divide your original *p*-value (usually 0.05) by the number of fitted lines:

But less conservative corrections also exist.

*Ramp coefficient*

The coefficient sets the minimum absolute value of the slope. Naturally, linear ramps with a zero slope are not events and the minimum value should be set according to expectations.

*Ramp Durbin-Watson statistic (DWS)*

DWS accounts for the shape of the data values around the line. Not only do we expect a Gaussian distribution around the fitted lines, but we also expect this distribution to persist down the length of the line. This is better described visually:

|  |
| --- |
|  |

In this toy example, both scenarios have fitted linear ramps, but the left should not qualify because the residuals (distance between data and line) are not uniformly distributed down the line. DWS looks for scenarios on the left (specifics: it looks for autocorrelation in residuals) and ranges from 0-4 where 2 is the ideal scenario of uniformly distributed noise and 0 represents a poor distribution (left scenario). DWS values above 2 suggest a positive autocorrelation and is highly unlikely in SHAPE-Seq datasets. It is suggested to leave DWS at the standard 1.

Example Data Descriptions

The example data file is “SRP\_wt\_rho\_table.txt”. The data is from the published manuscript:

Kyle E Watters, Eric J. Strobel, Angela M. Yu, John T. Lis, & Julius B. Lucks. Cotranscriptional folding of a riboswitch at nucleotide resolution. *Nature Structural and Molecular Biology*. (2016)

Cotranscriptional SHAPE-Seq measures RNA folding by exposing increasing transcript lengths of RNA to SHAPE molecules. SHAPE preferentially binds to open/unpaired regions of RNA and subsequent Seq assays measure the location and presence of attached SHAPE. This results in reactivity patterns that uncover RNA structural properties – highly reactive positions indicate regions of unstructured RNA and lowly reactive positions indicate constrained regions due to structure or interaction with other RNAs, ligands, or proteins

The data is a 2D matrix where columns are nucleotide positions and rows are RNA transcript lengths. As we go down the rows, the transcript length increases, and folding behavior can change. The first row has the minimum RNA length of 30 and each row adds the next position, up to the max of 125.

Automated parameter optimization

DUETT provides a method to automatically select PIR thresholds for any dataset. This functionality can be found in a pop-up after pressing the “Optimize thresholds” button. DUETT uses a heuristic to identify a PIR threshold combination that balances lenient with stringent thresholds. Starting with low PIR thresholds where both noise and real events are detected. DUETT scans over combinations of increasing PIR threshold values (within a user-defined range) and records the number of detected events. We expect a sharp decrease in the number of detected events as threshold values increase, followed by a leveling off, forming an elbow. DUETT identifies the vertex of the elbow—representing detection of true events—by finding the PIR threshold values that correspond to a detection output closest to the origin. If needed, the automatically identified thresholds serve as a starting point for manual tuning.

Fair warning, this process can take a long time (the accompanying manuscript’s data takes ~15 minutes with three replicates). After it is done, DUETT will print out a pdf of the optimally found PIR threshold values. To further tune thresholds, exit the optimization pop-up and manually input the information from the pdf. Sorry, I cannot figure out how to automatically transfer the optimization values to the user-interface.

Author Information

This app was created by Albert Xue of the Bagheri Lab at Northwestern University.

Current version updated on 09/25/18

You can reach him at [xuebert@u.northwestern.edu](mailto:xuebert@u.northwestern.edu)