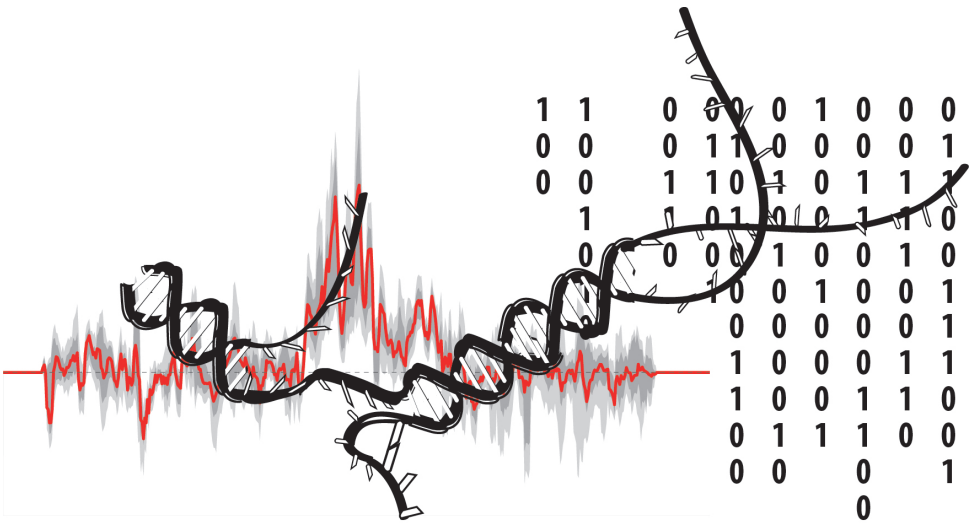
MIMEAnTo Manual

Version 1.0



April 28, 2016

In this manual, we provide a step-by-step description of how to use the MIMEAnTo software. MIMEAnTo analyzes data generated by the Mutational Interference Mapping Experiment (MIME, Smyth et al, Nature Methods 12,866–872 (2015)). MIME+MIMEAnTo allows to analyze non-coding RNA with regard to its function.
In MIME RNA is randomly mutated, selected-by-function (e.g. binding to a protein), physically separated, and sequenced using NGS. The mutation frequencies in the functionally selected vs. non-selected pools contain information about the function and structural commitment of each nucleotide within the analyzed RNA: thus, a mutation that does not affect selection is not required for the function of the RNA.
Unfortunately, instead of RNA copy numbers S, sequencing reads R, which are confounded by errors introduced during library preparation and sequencing, are typically recovered from the experiments (see Fig. 1, upper panel). Thus, sequencing error correction, as well as statistical assessment is essential. The following equation accommodates these considerations:

$$\frac{\mathrm{Kd}_{m,w}}{\mathrm{Kd}_{w,w}}(i,j) \approx \frac{\frac{R_{m,w}(i,j,\mathrm{unbound})}{R_{w,w}(i,j,\mathrm{unbound})} - \mathbb{E}_J\big(\kappa_{w\to m}(i)\big)}{\frac{R_{m,w}(i,j,\mathrm{bound})}{R_{w,w}(i,j,\mathrm{bound})} - \mathbb{E}_J\big(\kappa_{w\to m}(i)\big)}, \tag{1}$$

allowing the relative dissociation constant of mutation m at nucleotide position i, $\mathrm{Kd}_m/\mathrm{Kd}_w(i)$, to be recomputed in a jackknife-like fashion for each position $j \in J\backslash i$. Above, $R_{m,w}(i,j)$ denotes the number of reads that carry a mutation m at position i and a wild type residue at position $j \neq i$ and $\mathbb{E}_J\big(\kappa_{w\to m}(i)\big)$ denotes the expected probability of falsely detecting a wild type at position i as a mutant m.
MIMEAnTo will allow you to analyze your MIME data with statistical certainty using the equation above, after error $\mathbb{E}_J\big(\kappa_{w\to m}(i)\big)$ correction and quality filtering.

Data Preparation

The python script pythonscript name found in the directory `MIMEAnTo/scripts` will parse SAM files (the base counts after alignment of the NGS reads to a reference sequence) to the required format:

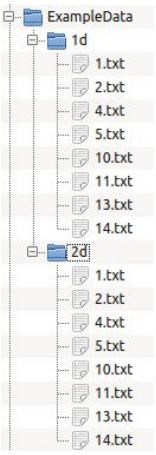


Figure 1: Exemplary input data structure.

The directory for the data of an experiment contains two subdirectories: `1d` and `2d`. They contain the single (base counts at position i) and covariation data (base counts for the pair of positions (i,j)) respectively in a textfile for each sample, given the name convention: *X.txt*, where X is the barcode for the sample (a unique number).
The single variation data (1d) contains the occurrence of nucleotides (#) for each position in the following format, separated by tabulator:

```
position    #A    #C    #G    #U
```

The covariation data (2d) contains the co-occurrence counts for each pair of nucleotides (e.g. #AG = number of reads where at the first position the nucleotide A is found and at the second position nucleotide G) for two different sites within the sequence in the format:

```
pos1  pos2  #AA  #AC  #AG  #AU  #CA  #CC  #CG  #CU  #GA  #GC  #GG  #GU  #UA  #UC  #UG  #UU
```

The first row is the header of the table which will be skipped internally while reading the file.

Step I: Initialization of a project

In the first step of the analysis-pipeline, the project has do be initialized with the following data:
Note: All entered information will be saved in the `project.txt` in the given result directory and is saved in every step of the pipeline. It can also always be saved in between with button **save project**.

result directory (mandatory). This defines where all (sub-)result files, figures and settings will be saved. If the directory is entered, an existing project file can be loaded, either from the same directory (if the project is already existing) or from another directory, see Fig. 2. In the latter case a new project file will be written to the new result directory. If no project file is loaded and a new project is started from scratch, the form has to be filled and can be saved into a project file later on.

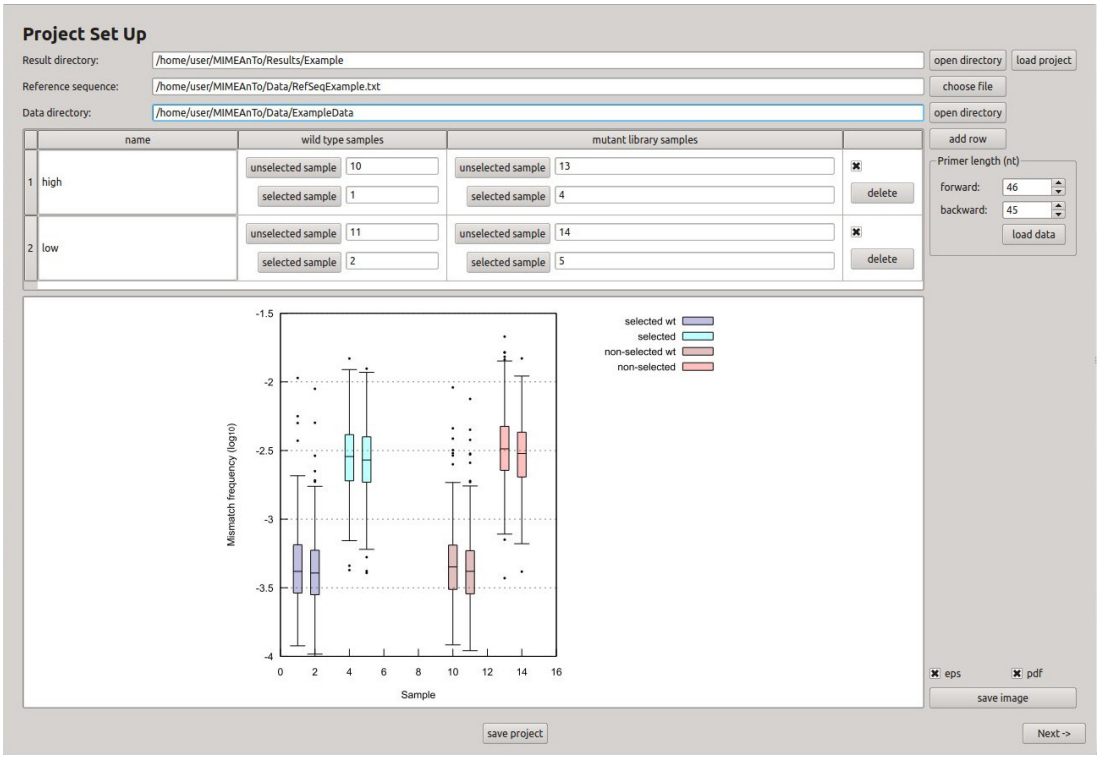


Figure 2: Step I of the analysis pipeline: Data and results directories are defined, or an existing project can be loaded. Details for the respective experiments are entered and a first visual depiction is presented, allowing to explore the data.

reference sequence (mandatory). A reference sequence must be given (see Fig. 2), either in fasta format (without ambiguous nucleotides) or in a text file containing "position,basenumber", without any header, where the nucleotides are provided by numbers following the convention:
A = 1 C = 2 G = 3 U = 4

data directory (mandatory). The data directory gives the path to the subdirectories `1d` and `2d`, see also Fig. 1.

data table. Data to enter in each row:

- name (mandatory): name for each sample set (a sample set consists of selected and non-selected pools from a wild type- and mutant library respectively). The name can be chosen freely, but we recommend to indicate the different conditions of each sample set, e.g. a particular protein concentration used in the experiment.
  - wild type samples (mandatory): data for the selected and non-selected pools from the wild type library. The barcode X has to be entered for each sample file.
  - data from mutant libraries: Analogous to the wild type.

Sample sets can be added with the button **add row**, removed with the **delete** button and inactivated with the checkbox. Inactivation can be undone in the next step.
If new data is submitted or changes of a loaded project are made, the **load data** button needs to be pushed. Pushing this button invokes a first plausibility check showing boxplots of the mismatch frequencies summarized over all positions for each library (using the provided `1d` data). This boxplot allows the user to assess whether the data looks reasonable at first sight (e.g. mutant libraries should have a higher mismatch frequency than their corresponding wild type counterpart).
Clicking the button **save image** will save the plot *mutRateBoxPlot* into the subdirectory `pathtoresults/plots`.

Step II: Error Estimation

To more precisely approximate the *true* mutation frequency in the mutant libraries after selection, error corrections have to be performed. Moreover, errors present random *noise*, and since the *actual* mutation frequencies in the respective samples represent the *signal* to be analyzed, error quantification will guide the subsequent

analysis by allowing to estimate a signal-to-noise ratio $D_{m,w}(i,j)$ for each mutation m and pair of positions (i,j) of the analyzed RNA.
The per position i and mutation m expected errors $\mathbb{E}_J\big(\kappa_{w\to m}(i)\big)$ are estimated by a jackknife-like re-sampling procedure (using the `2d` data of the wild type libraries). The parameter **percentage of maximum coverage threshold** (default 50%) is a quality criterium for estimating the expected error, by determining the depth of re-sampling, i.e. only position pairs (i,j) exceeding a minimum coverage are regarded for the error calculations.
With the **estimate error** button, the errors for all samples are computed and plotted. Or, if already computed and saved before, they can be loaded with the **load error** button.
The **save errors** button saves the calculated errors to text files, located in the subdirectory `pathtoresults/errors`.
The plot contains two tabs: In the first tab, the mean (+IQR) error rates ($\log_{10}$) per position are shown. The graphic on the top depicts the errors for the *selected* pools whereas the bottom graphic shows the errors for the *non-selected* pools. The interquartile ranges are shown as transparent background shading and are calculated from the re-sampling procedure. In the second tab, the **coefficient of variation** between all samples belonging to either the selected (solid lines) or non-selected pools (dashed lines) are depicted (the coefficient of variation between the samples that appeared in the same plot in the first tab is calculated). The two plots *errorEstimates* and *coefficientOfVariation* can be saved with **save images** in the subdirectory `pathtoresults/plots`.

With the checkbox **join** the option is given, to compute a single error estimate for all selected vs. non-selected pools for further evaluations. In general, this is recommended if the **coefficient of variation** between the pools (second tab) is low. Joining the errors may then achieve a better statistical foundation for the error estimate. If the coefficient of variation is above a certain threshold (here 50%), e.g. errors vary significantly between selected pools, a recommendation is given that the errors should be considered independently (as plotted in the first tab).
As in the step before, sample sets can be activated and deactivated. Only the activated samples are shown in the plot (after pushing **estimate error** again) and are considered for error correction and signal-to-noise quantification in the next step.

Step III: Effect (Kd) Quantification

In the third step, raw Kd values can either be imported, if already computed and saved previously (button **load raw Kd values**), or calculated (button **compute raw Kd values**). Only the activated sample sets are considered for the calculations.
After loading, or *de novo* calculation of the raw output (Kd values), parameter settings are enabled and the user can filter the results according to the following quality criteria (submitted with button **apply quality criteria**):

% of maximal coverage: Determines the depth of re-sampling, i.e. only position-pairs (i,j) exceeding a minimum coverage are regarded for Kd estimation. The minimum coverage is assessed with respect to the position with the maximal read coverage (default: 50%). This quality criteria will be important for filtering when the coverage is in general very high ($\geq 1 \cdot 10^6$ per position).

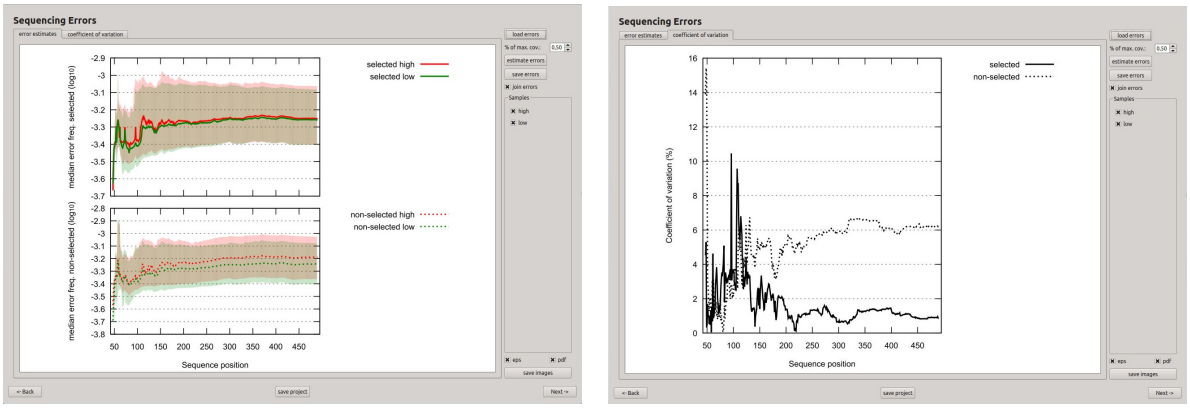


Figure 3: Analysis of (sequencing- and reverse transcriptase induced) errors introduced during the experimental procedure. The left figure (first tab) shows the mean errors (lines) and their corresponding interquartile range (shaded area) in selected- (top) vs. non-selected pools (bottom). The right figure compares errors between experimental conditions in terms of the coefficient of variation (variation between the errors plotted together in the left figure).

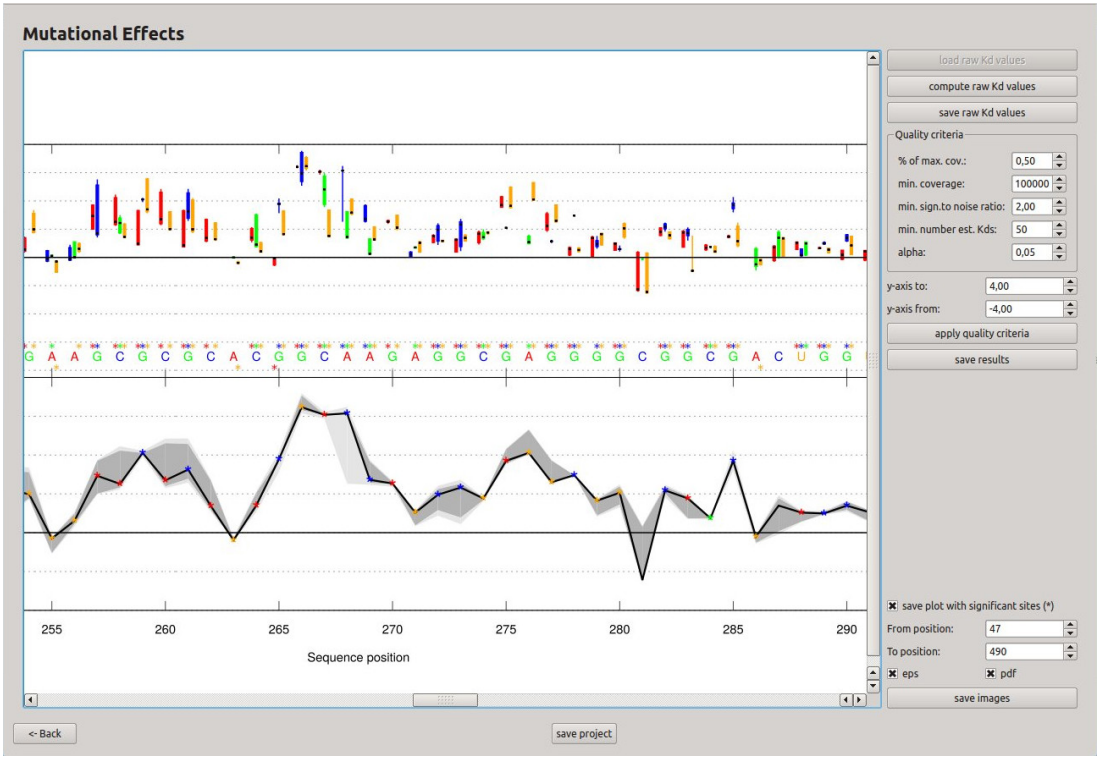


Figure 4: blubblub

minimum coverage: Determines the depth of re-sampling, i.e. only position-pairs exceeding a minimum coverage are regarded for Kd estimation. Here, the minimum coverage is determined in absolute numbers (default:100,000). This criterium will be important for filtering when the coverage is rather moderate ($\approx 1 \cdot 10^5$ per position). A rule-of-thumb to select the 'minimum coverage criterium' can be derived as follows: Let us assume that the mutation rate in the library is μ on average. We would like to evaluate only mutations m at position i that are present in at least 100 reads. Thus, the 'minimum coverage criterium' should be [minimum coverage] $\geq \frac{100}{\mu}$.

minimum signal-to-noise ratio: A signal-to-noise ratio $D_{m,w}(i,j)$ (default: 2) is computed for each pair of positions (i,j) and for each mutation m. We recommend to use signal-to-noise ratios $D_{m,w}(i,j)$ of at least 2 in MIMEAnTo (the *true* signal is at least twice the expected error). Note that a signal-to-noise ratio below 2 would mean that one tries to read *meaning* into *noise*. The value of the minimum signal-to-noise > 2 should be guided by careful consideration between data *quality* and *quantity*. I.e. the higher $D_{m,w}(i,j)$, less evaluable positions are to be expected, where one can predict an exact Kd value. To this end, $D_{m,w}(i,j) > 10$ may be unrealistic in most applications. A strong MIME signal should furthermore be robust against a range of $D_{m,w}(i,j)$ values.

minimal number of estimatable Kds: This criterium is important for statistical ascertainment (default: 50): P-values will only be computed for mutation m at position i, if the number of re-samplings exceeds the threshold provided here.

alpha: Level of significance (default: 0.05). Kd estimates are marked as significant if $P < \alpha$

Note: In order to statistically evaluate whether a mutation m at position i significantly affects binding raw Kd_m/Kd_w values are computed for all positions i that have a sufficient signal-to-noise ratio. "Sufficient" is defined as follows: If the ratio is below a user-supplied threshold (above) both in the selected and non-selected samples, the corresponding Kd estimate $\mathrm{K}_{m,w}(i,j)$ is discarded. If the signal is below the threshold at either the selected- or non-selected samples, the respective estimate is tagged as either being a lower- or upper estimate of $\mathrm{K}_{m,w}(i,j)$ and assigned (imputed) the value of the median $\mathrm{K}_{m,w}(i,*)$ estimate. This has the following reason: If a mutation strongly decreases Kd (increases binding), all sequences carrying this mutation may be selected, and none- or too little amounts of sequence may remain non-selected. Thus, $\mathrm{K}_{m,w}(i,j)$ may not be accurately determined and we can assume that $\mathrm{K}_{m,w}(i,j)$ may in fact be lower than estimable.
Using the fields **y-axis from** and **y-axis to** the user can provide y-axis limits for the plot.
Two plots will appear on the screen:

1. **All mutational effects.** The upper plot shows the effects of all three mutations per sequence position $\log_2(\frac{Kd_m}{Kd_w})$ as candlesticks in different colors (A C G U). The black dots mark the median Kd estimates and the bars span the interquartile range. The vertical lines mark the range of the Kd estimates excluding outliers. The distribution of the Kd estimates is computed from re-sampling. Above the x-axis, the wild type nucleotide at this position is shown. The * denotes whether a mutation significantly alters the relative ($Kd_m/Kd_w \neq 1$), indicated by the distinct colors as defined above. The * above the letter indicates that the mutation m significantly increases the Kd value (decreasing binding), and below the nucleotide a significant decrease in Kd (improvement in function).

2. **Maximal mutational effects.** The lower plot highlights the mutation with the strongest effect (positive or negative) per position $\log_2(\frac{Kd_{m_{\max}}}{Kd_w})$. Again the * denotes whether the mutation with the strongest effect exerts a significant effect and the color identifies the specific mutation.

If the sequence is longer than the screen, it is possible to scroll through the sequence to investigate the region of interest.
The plots can also be saved (button **save images**) as separate files named *relKdWtMut* and *maxEffectOnKd* for the upper and lower plot respectively. If only a particular region should be plotted and saved, the user can define it in (**from position** and **to position**).
With **save results** all final results will be saved into tables to the subdirectory `pathtoresults/KdResults`, where the file *PositionWiseKdEstimates* contains all information per position for each possible mutation and the file *PositionWiseMaxKd* only contains the results for the mutation with the strongest effect (in either direction).

Output

Plots
All plots will be saved in the subdirectory `pathtoresults/plots`. They can be saved as eps- and/or pdf-file (checkbox). When saving the plots, the user is asked to give an (optional) suffix for the filename, in order to save several plots for different conditions.
Tables
Tabular result files of the Kd computation (last step after applying the quality criteria) are exported as csv-files (button **save results**) in the subdirectory `pathtoresults/KdResults`. Optionally, the user can give a suffix for the files, to generate different output files with different parameter settings. The file *PositionWiseKdEstimates* saves the following information (tab separated):

- position
- wild type base w (A = 1, C = 2, G = 3, U = 4)
- max. effect base $m_{\max}$ (A = 1, C = 2, G = 3, U = 4)
- for each nucleotide (mutation) m = A, C, G, U
  - median Kd_m/Kd_w
  - p-values for mutation $Kd_m/Kd_w \neq 1$
  - #resamplings for Kd_m/Kd_w
  - #lower estimates for Kd_m/Kd_w
  - #upper estimates for Kd_m/Kd_w
  - 5th percentile of Kd_m/Kd_w estimate
  - 95th percentile of Kd_m/Kd_w estimate

The file *PositionWiseMaxKd* contains the information for the maximal effecting mutation (tab separated):

- position
- wild type base w (A = 1, C = 2, G = 3, U = 4)
- max. effect base $m_{\max}$ (A = 1, C = 2, G = 3, U = 4)
- median $Kd_{m_{\max}}/Kd_w$
- p-value for $Kd_{m_{\max}}/Kd_w \neq 1$
- #re-samplings for mutation $Kd_{m_{\max}}/Kd_w$

- #lower estimates for $Kd_{m_{\max}}/Kd_w$
- #upper estimates for $Kd_{m_{\max}}/Kd_w$
- 5th percentile of $Kd_{m_{\max}}/Kd_w$ estimate
- 95th percentile of $Kd_{m_{\max}}/Kd_w$ estimate