# SpliceMIT – Splice Modelling Intronic Technology

MIT iGEM 2017 – By Qianchang (Dennis) Wang

## Overview

SpliceMIT is a tool to generate antisense oligonucleotides (ASO) for a given DNA/RNA sequence, and then analyze and output the most effective ASOs. This tool could be used for a variety of scenarios. For our test samples, SpliceMIT was used to generate and analyze the gRNA sequences (complementary to the region on the intron) in the CRISPR – dCas13a system to control alternative splicing by targeting mRNA sequences.

## Mechanism

To compute and obtain the most effective antisense oligonucleotides, we considered four factors:

1. GC content of the ASO
2. Off-target binding
3. RNA-binding proteins competition on the ASO binding site
4. Secondary structure of the ASO

The general process is:

1. Add additional features if needed (e.g. Add pre-gRNA secondary structure analysis part if the goal is to find gRNAs instead of ASOs)
2. Ask user inputs, including DNA/RNA sequence, preferred ASO length, etc.
3. Run the calculations that require online platform first.
4. Store all data locally.
5. Analyze all local data.
6. Output highest ranked ASOs.

Anywhere in this process, the user could choose his or her preferred scoring or rank mechanism, or ideally all of them and compare to each other. (For our sample test trials, we have done data analysis comparing all three rank/scoring

mechanisms. You could use the same file to analyze the results based on your own test samples.)

- **GC conten**t:

Measure the amount of Guanine and Cytosine bases on the ASO, then calculate the proportion of GC bases. Based on previous research results, the effective GC content is between 35% and 70%[1] [2] [3] . Then the GC content value would be either calculated into an arbitrary score or used to apply cutoffs.

- **Off-target binding:**

Search for potential off-target binding with human cDNA obtained from ensembl[4]. The original link is: ftp://ftp.ensembl.org/pub/release-89/fasta/homo_sapiens/cdna/.

The one we use is GRCh38 (Genome Reference Consortium Human genome build 38), released in May 7 2014. Our program searches for complementary sequences using a loop function. In the loop, the function skips a variable number of consecutive sequences each time based on the disparity of the previous pairing, therefore reduces a running time of about 30 minutes to now about 90 seconds. An alternative version of predicting off-target is to use Nucleotide BLAST, or blastn, on National Center for Biotechnology Information. Selenium (a Python Web-driver package) is used to guide the Chrome or Firefox explorer to automatically run each trial. As BLAST has a large database, the running time could be between 10 seconds to 3 minutes based on different sequences. The two programs are both incorporated into the program, and user can choose the one that they prefer to use under their specific circumstance. The two programs both return the number of predicted off-target bindings, and the mismatch value for each. The mismatch value is the number of differed bases between the complement of the segment on the intron sequence and the analyzed ASO sequence. The user can also set the mismatch threshold, which is set by default to 4. Any segments of the intron will not be considered a predicted off-target binding if the mismatch value exceeds the mismatch threshold. Then the data will be used to calculated into an arbitrary score, which will be explained later.

- **RNA-binding proteins competition:**

RNA-binding protein (RBP) competes with other types of proteins to bind at a region. The level of competition could affect the binding effectiveness of a certain ASO. In our model, we consider RBPs that either completely or partially cover the complementary sequences of the ASO on the given DNA/RNA strand. We utilize RBPmap[5], a tool that predict RBPs and their binding site on a given RNA sequence. We include all current RNA-binding proteins found naturally in Human and Mouse. That, in total, counts for 92 different RBPs. For our test samples, which focus on alternative splicing, the splicing factors such as the hnRNP family are also included in those 92 RBPs. A sample output would look like this:

**Protein: HNRNPA1(Hs/Mm)**

| Position | Motif | Occurrence | Z-score | P-value |
|---|---|---|---|---|
| 28 | rgnyag | gaucgaucguagcugcuagcuagcuagcuagcuagcuacguagcuagcuagcuagc | 4.036 | 2.72e-05 |
| 32 | rgnyag | gaucguagcugcuagcuagcuagcuagcuagcuacguagcuagcuagcuagcuagc | 4.055 | 2.51e-05 |
| 44 | rgnyag | uagcuagcuagcuagcuagcuacguagcuagcuagcuagcuagcuagcuagcuagc | 4.073 | 2.32e-05 |
| 48 | rgnyag | uagcuagcuagcuagcuacguagcuagcuagcuagcuagcuagcuagcuagcuagc | 4.073 | 2.32e-05 |
| 52 | rgnyag | uagcuagcuagcuacguagcuagcuagcuagcuagcuagcuagcuagcuacg | 4.073 | 2.32e-05 |
| 56 | rgnyag | uagcuagcuacguagcuagcuagcuagcuagcuagcuagcuagcuacgucgu | 4.055 | 2.51e-05 |
| 60 | rgnyag | uagcuacguagcuagcuagcuagcuagcuagcuagcuagcuacgucguca | 4.036 | 2.72e-05 |

**Protein: MBNL1(Hs/Mm)**

| Position | Motif | Occurrence | Z-score | P-value |
|---|---|---|---|---|
| 28 | gcuugc | gaucgaucguagcugcuagcuagcuagcuagcuagcuacguagcuagcuagcuagc | 3.039 | 1.19e-03 |
| 32 | gcuugc | gaucguagcugcuagcuagcuagcuagcuagcuacguagcuagcuagcuagcuagc | 3.065 | 1.09e-03 |
| 44 | gcuugc | uagcuagcuagcuagcuagcuacguagcuagcuagcuagcuagcuagcuagcuagc | 3.078 | 1.04e-03 |
| 48 | gcuugc | uagcuagcuagcuagcuacguagcuagcuagcuagcuagcuagcuagcuagcuagc | 3.078 | 1.04e-03 |
| 52 | gcuugc | uagcuagcuagcuacguagcuagcuagcuagcuagcuagcuagcuagcuacg | 3.065 | 1.09e-03 |
| 56 | gcuugc | uagcuagcuacguagcuagcuagcuagcuagcuagcuagcuagcuacgucgu | 3.065 | 1.09e-03 |
| 60 | gcuugc | uagcuacguagcuagcuagcuagcuagcuagcuagcuagcuacgucguca | 3.039 | 1.19e-03 |

**Protein: MSI1(Hs/Mm)**

| Position | Motif | Occurrence | Z-score | P-value |
|---|---|---|---|---|
| 27 | uaguwrg | cgaucgaucguagcugcuagcuagcuagcuagcuagcuacguagcuagcuagcuagc | 4.545 | 2.75e-06 |
| 31 | uaguwrg | cgaucguagcugcuagcuagcuagcuagcuagcuacguagcuagcuagcuagc | 4.545 | 2.75e-06 |
| 35 | uaguwrg | cguagcugcuagcuagcuagcuagcuagcuacguagcuagcuagcuagcuagc | 3.803 | 7.15e-05 |
| 43 | uaguwrg | cuagcuagcuagcuagcuagcuacguagcuagcuagcuagcuagcuagcuagc | 4.561 | 2.55e-06 |
| 47 | uaguwrg | cuagcuagcuagcuagcuacguagcuagcuagcuagcuagcuagcuagcuagc | 4.561 | 2.55e-06 |
| 51 | uaguwrg | cuagcuagcuagcuacguagcuagcuagcuagcuagcuagcuagcuagcuacg | 4.561 | 2.55e-06 |
| 55 | uaguwrg | cuagcuagcuacguagcuagcuagcuagcuagcuagcuagcuagcuacgucgu | 4.561 | 2.55e-06 |
| 59 | uaguwrg | cuagcuacguagcuagcuagcuagcuagcuagcuagcuagcuacgucguca | 4.545 | 2.75e-06 |
| 63 | uaguwrg | cuacguagcuagcuagcuagcuagcuagcuagcuagcuagcuacgucguca | 4.545 | 2.75e-06 |

The outputs are in a user-friendly format as they are categorized by the RBP type. The “Position” column indicates the position from the 5’ end of the given DNA/RNA sequence. The “motif” column indicates the binding motif of a specific RBP. The “Occurrence” column visualizes the binding. To define a significant match, the match score of all the sites in the query sequence (input DNA/RNA sequence) are compared to the mean match score for the motif calculated for a

background of randomly chosen regulatory regions (composed of exonic and intronic regions around splice-sites and exons in UTR regions). The calculated Z-score is contained in the next column. The P-value is therefore the single-tailed probability value derived from these Z-scores.

- **ASO Secondary Structure:**

Some or a potential secondary structure could decrease the efficiency of binding or even inhibit the process completely. Any stem-loop or hairpin structure of the ASO could be devastating. Here we calculate the probability of obtaining a secondary structure for each generated ASO. SpliceMIT calculates the probabilities of a base binding to another on Nupack[6] web interface. Here is a sample output from Nupack for a random ASO sequence of 24 nucleotides:

```
% This is a data file for the plot containing Pair probabilities at 37.0 C.
%
% This file was generated by nupack.org on Sun Sep 17 13:55:06 2017 PST
%
% The data are in three columns.  Each row describes a base pair between
% base i and base j.  Column 1 contains i, column 2 contains j, and column
% 3 contains the probability that base i is paired with base j at equilibrium.
% If an entry of -1 appears in column 2, column 3 signifies the probability
% that base i is unpaired at equilibrium.
% Only values above a cutoff of 0.001 are reported.
%
% The sequence:
% AGAUCGUAGCUAGCAUCGAUCGUA
1       23      0.0039669
2       17      0.0050649
2       21      0.9900100
3       16      0.0050593
3       20      0.9909400
4       15      0.0049018
4       19      0.9919500
5       18      0.9928400
6       14      0.0226140
6       16      0.0024739
6       17      0.8817300
7       12      0.0013492
7       13      0.0380680
7       15      0.0492770
8       16      0.2004000
9       14      0.1707100
9       16      0.0080771
11      15      0.0031270
1       -1      0.9959400
2       -1      0.0047971
3       -1      0.0039539
4       -1      0.0030963
5       -1      0.0065959
6       -1      0.0930520
7       -1      0.9112500
8       -1      0.7995900
9       -1      0.8210700
10      -1      0.9998200
11      -1      0.9967500
12      -1      0.9982100
13      -1      0.9614000
14      -1      0.8066400
15      -1      0.9426600
16      -1      0.7835300
17      -1      0.1130400
18      -1      0.0070790
19      -1      0.0079757
20      -1      0.0090029
21      -1      0.0098969
22      -1      0.9998400
23      -1      0.9959300
24      -1      0.9999800
```

The first column and second column are both the position number of a base from 5’. The decimal figure in the third column represents the probability of a bond between the two bases indicated in the previous two columns. Our program uses the second part of the figure, which the second column of each row equals to -1. The decimal figure in those rows represents the probability of not having any bonds at the base position described by the first column. The probability of obtaining a non-secondary structure for a given ASO is therefore:
$P_n = P_1 \times P_2 \times \cdots \times P_{n_{base}}$ where $P_n$ represents probability of non-secondary

structure and $P_{nbase}$ represents the probability of non-binding at each base. Then this probability would be used to calculate arbitrary score, rank-product selection, or cutoff methods. Each will be explained later in the three sections.

## Rank Methods

**We have three methods to filter and produce as ranking of all potential ASO sequences: arbitrary score, sets of cutoff thresholds and rank-product of all factors in the model. The user can pick (a) rank method(s) based on the system constructed.**

- **Arbitrary Score**

An Arbitrary score is calculated respectively in each factor, and then the sum of those arbitrary score is used to rank the effectiveness and efficiency of the ASOs.

1. GC content of the ASO

   For GC content, there is simply a penalty score for a percentage outside the threshold range. This variable is named ‘GC_content_penalty_score’ and the value is set to 1000 in our model. According to the previous GC Content section, the effective content is between 35% and 70%. The distribution of GC content of one sample is shown in Figure 1:

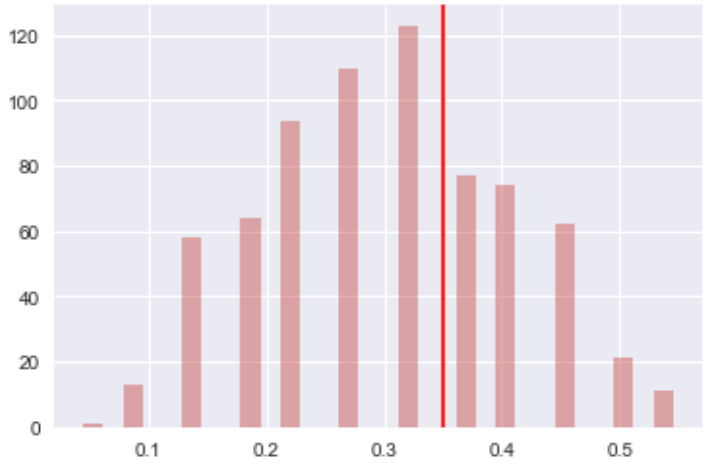



**Figure 1 – Distribution of GC Content.** This data is obtained based on our first test sample for the CRISPR system. In total, 708 gRNA sequences were analyzed. The red line represents the threshold of 35%

## 2. Off-target binding

The number of potential off-target binding and the mismatch number of each are already obtained. The penalty value for potential off-target bindings is set based on the number of mismatches. Those variables are named ‘mm0_score’ (when the target sequence is completely complementary to the ASO sequence), ‘mm1_score’ (when the target sequence has one base that is not complementary), etc. The model sets values up to 5 mismatches. It also has a ‘default_mm_score’ if the threshold of mismatches increases to over 5. Table 1 shows the values of those variables/penalty scores. The score distribution is shown in Figure 2. The majority have a value of 0 or less than 350. However, there is a cluster of outlier around 3000, which almost directly expels them from the list of effective ASO sequences.

| | |
|---|---|
| mm0_score | 10 |
| mm1_score | 20 |
| mm2_score | 200 |
| mm3_score | 300 |
| mm4_score | 450 |
| mm5_score | 1000 |
| default_mm_score | 10 |

**Table 1 - Penalty Scores for Off-Target Binding based on Mismatches**

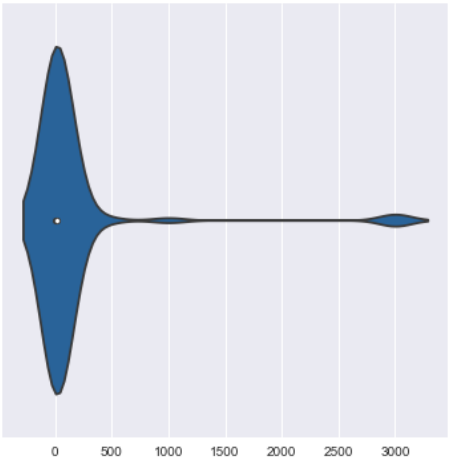



**Figure 2 – Violin Plot for Off-Target Scores.** The data used is from the test sample for out CRISPR-dCas13a system. 708 gRNA sequences were analyzed.

## 3. RNA-binding proteins competition on the ASO binding site

$$I_{RBP} = \sum_{i=1}^{N_{Base}} \sum_{n=1}^{N_{RBP}} a^{n-1} \cdot \frac{[RBP]}{K_{dRBP}} \cdot \frac{L_b}{L_m}$$

The algorithm for RBP interference:

while $a \in (0, 0.5]$.

a is the weight value;

k is the weight-power constant;

$K_d$ is the dissociation constant for the RBP;

$L_b$ is the length of binding;

$L_m$ is the length of RBP-binding motif;

[RBP] is the concentration of an RNA-binding protein (Default to N/A). $K_d$ of the RBPs were found through previous papers[7-22]. For the RBPs without previous dissociation constant records, the model uses the median value instead.

The first summation simply means to add RBP interference score on each base for RBPs whose first binding site is that base. The second summation intends to simplify a situation. On a segment of intron sequence, for example, several RNA-binding proteins (with different binding motif length) could bind in the region from base 5 to base 15. It creates more possibilities of combinations of RBPs, and the accumulation of those significantly increase the running time of the model. The summation of those RBPs' interference scores starting at the same base simplifies such situation.

$\frac{[RBP]}{K_{dRBP}} \cdot \frac{L_b}{L_m}$ is the score for an RNA-binding protein as $L_b/L_m$ considers the partial binding at both the 5' and 3' ends of the complementary sequence. For example, the complementary sequence is from base 300 to 330 from 5'. There could be RBPs binding from base 298 to 303, which does not completely bind in the "effective region". The value *a* assigns weighted values of multiple RBP scores on one base. The highest interference score at one base would be given a weight value of 100%. The second largest would be given a weight value of *a* and the third one would be given a weight value of *a*^2, so on and so forth. The weight value *a* gives decreasing score for each additional RNA-binding protein whose motif starts at one base.

After this score is obtained. Another multiplication factor is assigned. The variable is named 'RBP_competition_weight', and set to $10^{-6}$. Figure 3 shows the distribution of weighted RBP interference scores on our test sample.

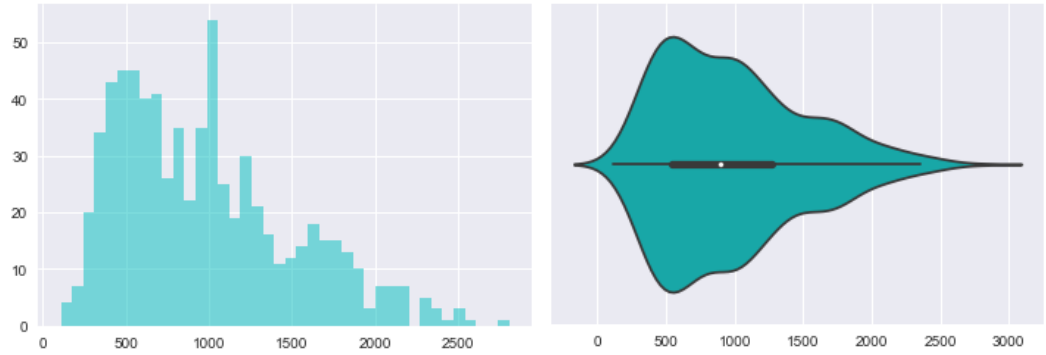



**Figure 3 – Histogram and Violin Plot of Weighted RBP Interference Score.** The data used is from the test sample for out CRISPR-dCas13a system. 708 gRNA sequences were analyzed. The median is around 890, while few outliers were presented.

## 4. Secondary structure of the ASO

The methods to obtain the secondary structure probabilities are already mentioned in the previous part of ASO Secondary Structure. Figure 4 shows the distribution of the raw probabilities. All values are extremely close to 0. The strong skewness will not show the fair comparison between the ASO sequences. Therefore, our model uses a weighted algorithm to modify those probabilities into a range between 0 and 1. The weighted algorithm is as following:

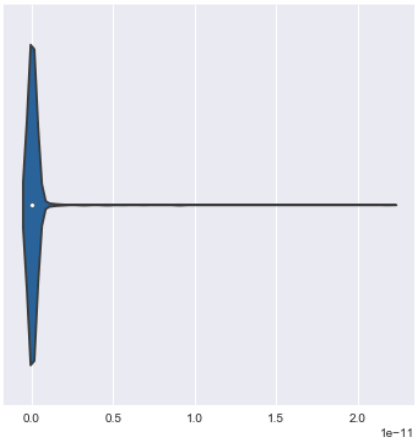
$$P = P_n^{k_w}$$

$$k_w = \frac{\log P_n}{\log P_{n_{max}} \cdot \log_{p_{target}} P_{n_{max}}}$$

$$P_{target} = 0.95$$

P is the weighted probability score
$K_w$ is the weight power
$P_{target}$ is the target maximum probability score
$P_{n\text{-}max}$ is the maximum value of all unweighted probabilities

$P_{target}$ could be set by the user, as it represents the desired maximum probability score in the range between 0 and 1. Then another multiplication factor is assigned. The variable is named as 'gRNA_2_structure_weight', and set to 3000 as default. Figure 5 shows the distribution of the final arbitrary scores of secondary structures in our test sample.



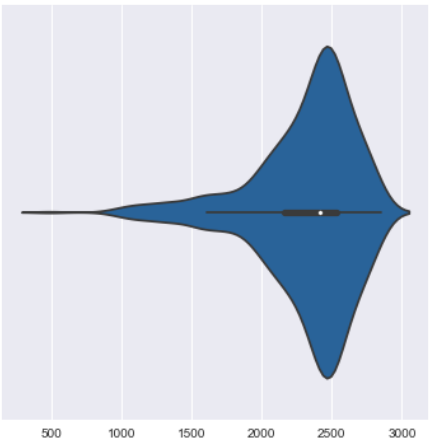




**Figure 4 (left) and 5 (right) – Distribution of Unweighted and Weighted Scores of Secondary Structure.** The data used is from the test sample for out CRISPR-dCas13a system. 708 gRNA sequences were analyzed. The left one is the distribution of raw probability scores, and all close to zero. The right one is the distribution of the weighted scores. The median is around 2440 and roughly symmetric.

The final arbitrary score is then calculated by adding the GC Content penalty score, Off-target binding score, RNA-binding proteins interference score, and subtracting the weighted secondary structure score from the sum. The efficiency of an ASO has a negative correlation with the final arbitrary score.

- **Cutoff**

The cut-off method preserves the basic scoring methods in the four sub-models. However, the weight and multiplication factors are removed. Instead, the model tests 10 sets of cut-off thresholds on 16 (currently completed) intron sequences. The data could be located on SpliceMIT's Github repository (https://github.com/MITiGEM2017/SpliceMIT) The selected introns are common ones in human and mouse, while their size also differs from 240 to 891 base pairs. Table 2 contains all 10 sets of cut-off thresholds. The model tested over 20 sets of thresholds. However, when thresholds of GC Content or Off-target Binding were adjusted alone while other thresholds stayed constant, the results did not change in almost all trials. Therefore, the thresholds the model provides here keep all GC Content thresholds at 0 and Off-target Binding thresholds at 20. The user could pick one set of thresholds or modify it based on the strictness.

| | GC Content | Off-target Binding | RBP Interference | Secondary Structure |
|---|---|---|---|---|
| 1 | 0 | 20 | $5*10^8$ | 0.01 |
| 2 | 0 | 20 | $5*10^8$ | 0.005 |
| 3 | 0 | 20 | $5*10^9$ | 0.005 |
| 4 | 0 | 20 | $5*10^8$ | 0.001 |
| 5 | 0 | 20 | $5*10^9$ | 0.0025 |
| 6 | 0 | 20 | $1*10^{10}$ | 0.001 |
| 7 | 0 | 20 | $1*10^9$ | 0.005 |
| 8 | 0 | 20 | $1*10^{10}$ | 0.0005 |
| 9 | 0 | 20 | $1*10^{10}$ | 0.00025 |
| 10 | 0 | 20 | $5*10^9$ | 0.00015 |

**Table 2 – Ten tested sets of thresholds for cut-off method.**

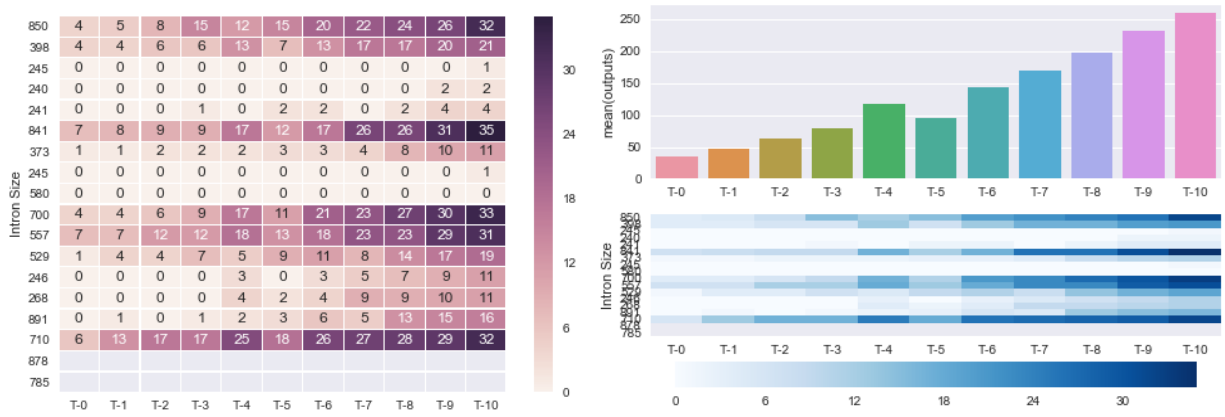



**Figure 6 (left) and 7 (right) – Heatmaps of Thresholds Outputs.** T-n represents the threshold number. For example, T-0 represents the thresholds used in the current model. T-1 represents the first set of thresholds in Table 2. The values in the columns in Figure 6 represent the number of outputs. The y-axis represents the size of the intron. Figure 7 combines the heatmap with the histogram. The y-axis represents the average number of outputs in each test sample.

- **Rank-Product**

The Rank-Product method also preserves the basic scoring methods in the four sub-models except the removal of the weight and multiplication factors. The method derives from a biologically motivated test for the detection of differentially expressed genes in replicated microarray experiments[23]. It is a simple non-parametric statistical method based on ranks of fold changes. The original calculation of rank product is as following.

Given n genes and k replicates, let $e_{g,j}$ be the fold change and $r_{g,i}$ the rank of gene g in the i-th replicate.

Compute the rank product via the geometric mean: $RP(g) = (\Pi_{i=1}^{k} r_{g,i})^{1/k}$

For our model, each ASO is ranked in the four scoring parts first. Then a rank product of that ASO is calculated. A final rank is therefore obtained based on the rank products.

- **Comparison of the Three Rank Methods**

The three rank methods were also analyzed. For the cut-off method, the analysis used the 10$^{th}$ set of thresholds as shown in Table 2. For the arbitrary score and the rank product methods, only the top 20 ranked ASOs remained in the list. Then the program found the identical ASO sequences between two lists of outputs. Therefore, three sets of number of overlapped sequences were produced as Table 3 shows. The three percentage histograms of overlapped sequences are shown in Figure 8, 9, 10.

| Intron size | ASO | Cutoff output | Cutoff vs RP | Cutoff vs weighted | RP vs. weighted |
|---|---|---|---|---|---|
| 850 | 708 | 32 | 9 | 3 | 4 |
| 398 | 301 | 21 | 7 | 17 | 0 |
| 245 | 143 | 1 | 0 | 0 | 0 |
| 240 | 183 | 2 | 2 | 0 | 3 |
| 241 | 143 | 4 | 4 | 0 | 0 |
| 841 | 582 | 35 | 0 | 12 | 1 |
| 373 | 258 | 11 | 0 | 5 | 3 |
| 245 | 141 | 1 | 1 | 0 | 0 |
| 580 | 386 | 0 | 0 | 0 | 0 |
| 700 | 544 | 33 | 5 | 12 | 0 |
| 557 | 379 | 31 | 0 | 19 | 3 |
| 529 | 406 | 19 | 6 | 5 | 0 |
| 246 | 169 | 11 | 1 | 9 | 3 |
| 268 | 167 | 11 | 3 | 8 | 0 |
| 891 | 725 | 16 | 6 | 3 | 0 |
| 710 | 550 | 32 | 4 | 14 | 0 |
| 878 | 627 | 79 | 0 | 18 | 4 |
| 91 | 45 | 2 | 2 | 0 | 10 |
| 91 | 56 | 8 | 8 | 3 | 3 |
| 45 | 17 | 7 | 7 | 4 | 9 |

**Table 3 – Analysis of Three Rank Methods.** The columns, from left to right, represent intron size, number of ASO sequences analyzed, number of ASO outputs from cut-off method, number of identical ASO sequences between cut-off and rank-product, number of identical ASO sequences between cut-off and arbitrary score, number of identical ASO sequences between arbitrary score and rank product.

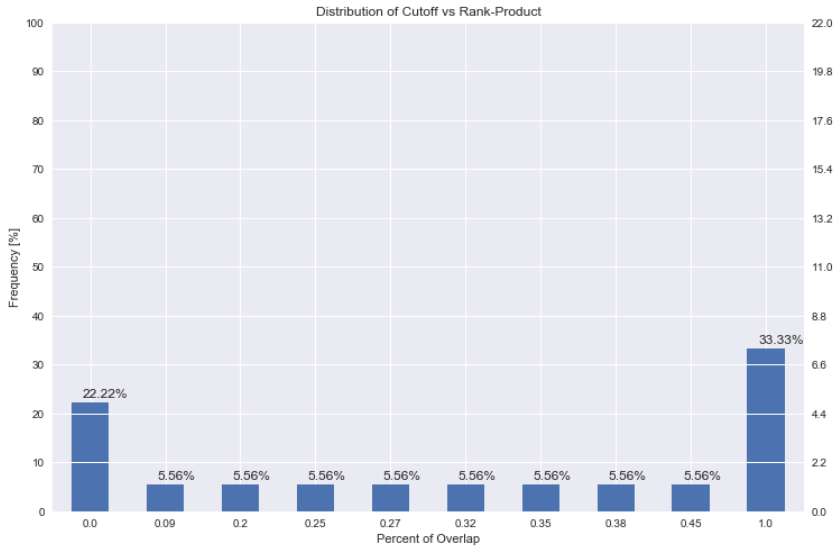



**Figure 8 – Histogram of Percentage of Overlapped ASO Sequences Between Cut-off Outputs and Rank Product Outputs.** The x-axis represents the percentage of overlapped ASO sequences in each test sample. The y-axis represents the frequency of a percentage value.

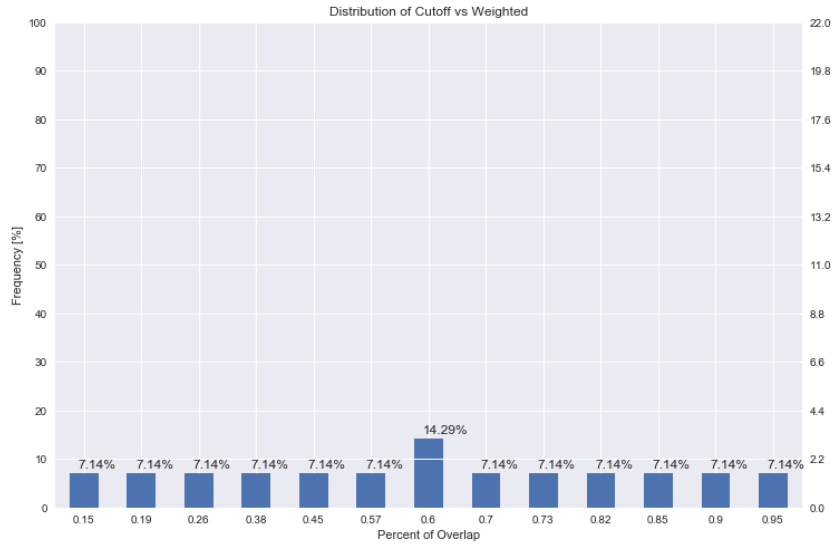



**Figure 9 – Histogram of Percentage of Overlapped ASO Sequences Between Cut-off Outputs and Arbitrary Score Outputs.** The x-axis represents the percentage of overlapped ASO sequences in each test sample. The y-axis represents the frequency of a percentage value.

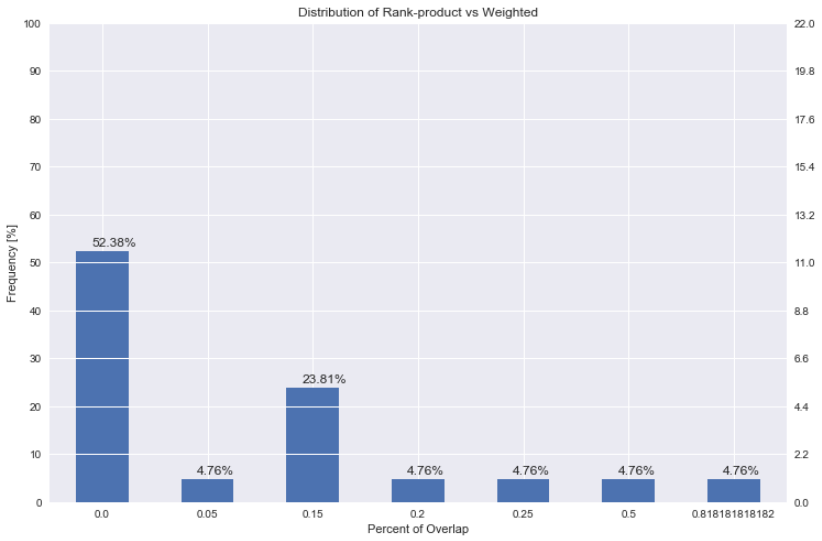



**Figure 10 – Histogram of Percentage of Overlapped ASO Sequences Between Rank Product Outputs and Arbitrary Score Outputs.** The x-axis represents the percentage of overlapped ASO sequences in each test sample. The y-axis represents the frequency of a percentage value.

In Figure 8, there are 1/3 of the test samples, whose cut-off outputs and rank product outputs are identical. However, the rest all falls into the range below 50%, and there are 22.22% of the test samples have zero overlapped outputs. In Figure 9, 64.3% of the test samples had over 50% of the outputs as overlapped ASO sequences. Figure 8, however, shows that over 90% of test samples did not share over 50% of the outputs as overlapped ASO sequences. Also, over 50% of the test samples do not have any overlapped ASO sequences from rank-product outputs and arbitrary score outputs. While the rank product method seems to be distant from the arbitrary score method, the cut-off method matches fairly well with the other two rank methods. Figure 11 and 12 also demonstrates this idea. Figure 11 shows the cut-off methods correlates relatively strong with the other two methods. Figure 12 reflects a direct comparison between the methods, showing the poor match between rank product outputs and arbitrary score outputs while the cut-off outputs match well with the other two.

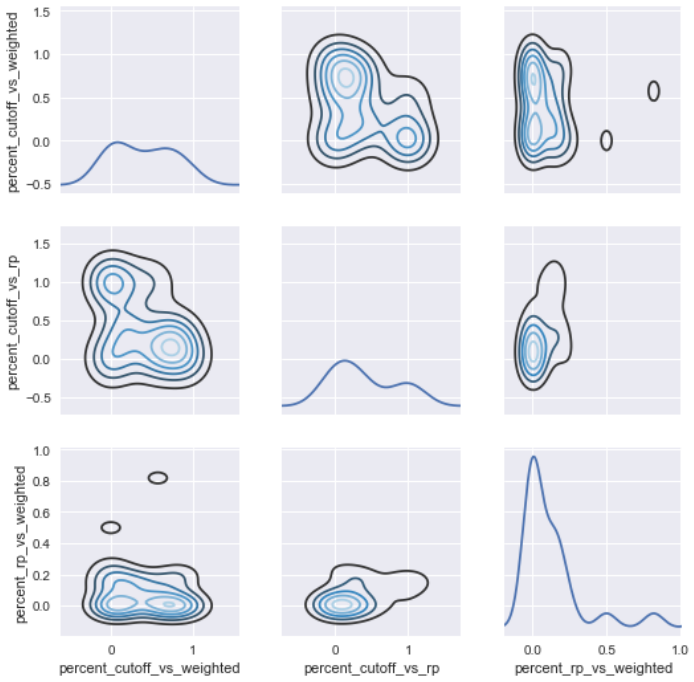



**Figure 11 – Pairplot of Overlapped ASO Seqeuences Ratios.** The pairplot shows the correlation between the distributions of overlapped ASO sequences percentages shown in Figure 8, 9 and 10. “RP” represents the rank product method while “weighted” represents the arbitrary score method.

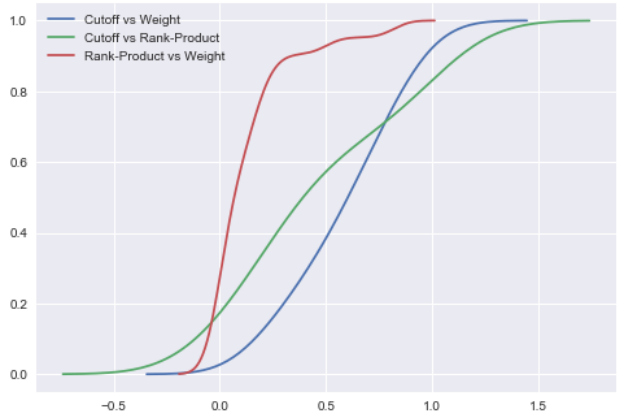



**Figure 12 – Accumulative Distribution of the Overlapped ASO Sequences Percentages.** The percentages shown in Figure 8, 9 and 10 are visualized as three accumulated curves. The curves fall slightly out of the range of 0 and 1 due to automatic predictions.

## About Test Sets – Our Application

- **Additional Factor – Pre-gRNA Secondary Structure**

For our test sets for SpliceMIT, we add one additional factor to the program – pre-gRNA secondary structure. The pre-gRNA structure is made up of two direct repeats and the gRNA sequence in between. Since we use pre-gRNA in the lab, the addition of this factor better simulates the situation. The default direct repeats sequence in our model is CCACCCCAATATCGAAGGGGACTAAAAC. The direct repeats are stem-loops, and we would prefer pre-gRNAs that would unlikely to have any binding between (a) bases on the gRNA region and (a) bases on either of the direct repeats. We also calculate the probabilities of binding on each individual base on Nupack web interface and then store the data locally. The algorithm for the probability score is similar to the one for crRNA structure.

$$P_n = P_1 \times P_2 \times \cdots \times P_{n_{base}}$$

This probability score is the product of all probabilities of binding of each individual gRNA base, therefore represents the probability of obtaining a desired pre-gRNA without any secondary structure in the gRNA region.

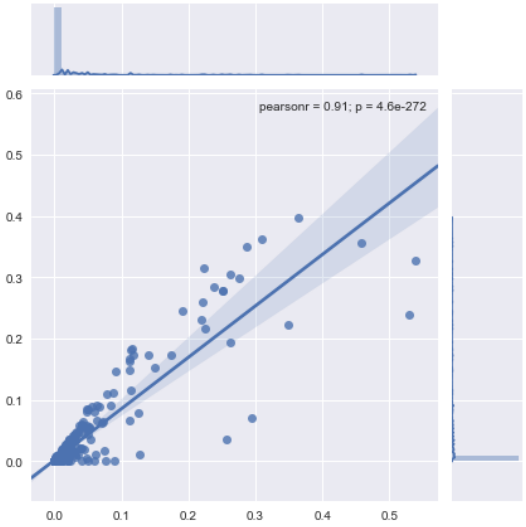



**Figure 13 – Correlation of Unweighted Probabilities of gRNA and pre-gRNA Sequences.** The x-axis and y-axis each represents the probabilities of gRNA and pre-gRNA in one test sample of 708 analyzed sequences.

In our sample test, the probabilities of secondary structure of gRNA sequence and pre-gRNA sequences are compared. The comparison is shown in Figure 13. The correlation efficient is 0.91, with an extremely low p-value. There is a strong positive relation between the two.

- **Other Changes**

As MIT 2017 iGEM Team uses the CRISPR system to target RNA sequences, a tracer RNA sequence is added while analyzing the secondary structure. The default tracer RNA sequence is set as "CCACCCCAAUAUCGAAGGGGACUAAAAC".

Bibliography:


1. Tessa G. Montague, José M. Cruz, James A. Gagnon, George M. Church, Eivind Valen; CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. Nucleic Acids Res 2014; 42 (W1): W401-W407. doi: 10.1093/nar/gku410
2. Wang, Tim et al. "Genetic Screens in Human Cells Using the CRISPR/Cas9 System." Science (New York, N.Y.) 343.6166 (2014): 80–84. PMC. Web. 5 July 2017.
3. Tsai, Shengdar Q. et al. "GUIDE-Seq Enables Genome-Wide Profiling of off-Target Cleavage by CRISPR-Cas Nucleases." Nature biotechnology 33.2 (2015): 187–197. PMC. Web. 5 July 2017.
4. ftp://ftp.ensembl.org/pub/release-89/fasta/homo_sapiens/cdna/
5. Inbal Paz, Idit Kosti, Manuel Ares, Jr, Melissa Cline, Yael Mandel-Gutfreund; RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res 2014; 42 (W1): W361-W367. doi: 10.1093/nar/gku406
6. J. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, N. A. Pierce. NUPACK: analysis and design of nucleic acid systems. J Comput Chem, 32:170–173, 2011.
7. Shlyakhtenko, Luda S. et al. "Atomic Force Microscopy Studies of APOBEC3G Oligomerization and Dynamics." Journal of structural biology 184.2 (2013): 10.1016/j.jsb.2013.09.008. PMC. Web. 3 July 2017.
8. Hollingworth, David et al. "KH Domains with Impaired Nucleic Acid Binding as a Tool for Functional Analysis." Nucleic Acids Research 40.14 (2012): 6873 – 6886.
9. Edwards, John M. et al. "Structural Insights into the Targeting of mRNA GU-Rich Elements by the Three RRMs of CELF1." Nucleic Acids Research 41.14 (2013): 7153 – 7166.
10. Choudhury, R et al. "The splicing activator DAZAP1 integrates splicing control into MEK/Erk-regulated cell proliferation and migration." Nature Communications 5, Article number: 3078 (2014)

11. Jiang, Z et al. "Molecular Cloning and Characterization of a Candidate Human Growth-Related and Time-Keeping Constitutive Cell Surface Hydroquinone (NADH) Oxidase." Biochemistry, 2008, 47 (52), pp 14028–14038

12. Dittmar, Kimberly A. et al. "Genome-Wide Determination of a Broad ESRP-Regulated Posttranscriptional Network by High-Throughput Sequencing." Molecular and Cellular Biology 32.8 (2012): 1468–1482.

13. Stefanovic, Snezana et al. "Fragile X Mental Retardation Protein Interactions with a G Quadruplex Structure in the 3′-Untranslated Region of NR2B mRNA." Molecular bioSystems 11.12 (2015): 3222–3230.

14. Niu C, Zhang J, Gao F, Yang L, Jia M, Zhu H, et al. (2012) FUS-NLS/Transportin 1 Complex Structure Provides Insights into the Nuclear Targeting Mechanism of FUS and the Implications in ALS. PLoS ONE 7(10): e47056. https://doi.org/10.1371/journal.pone.0047056

15. Ascano, Manuel et al. "FMR1 Targets Distinct mRNA Sequence Elements to Regulate Protein Expression." Nature 492.7429 (2012): 382–386.

16. Vognsen T, Møller IR, Kristensen O (2013) Crystal Structures of the Human G3BP1 NTF2-Like Domain Visualize FxFG Nup Repeat Specificity. PLoS ONE 8(12): e80947. https://doi.org/10.1371/journal.pone.0080947

17. Yang, Xiufeng et al. "The Dataset for protein–RNA Binding Affinity." Protein Science : A Publication of the Protein Society 22.12 (2013): 1808–1811.

18. Hafner, M et al. "Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP." Cell Press.Volume 141, Issue 1, 2 April 2010, Pages 129-141

19. Horisawa, K et al. "3′-Untranslated region of doublecortin mRNA is a binding target of the Musashi1 RNA-binding protein." Science Direct. Volume 583, Issue 14, 21 July 2009, Pages 2429–2434
20. Lee, Seung Hwan et al. "Poly(A) RNA and Paip2 Act as Allosteric Regulators of poly(A)-Binding Protein." Nucleic Acids Research 42.4 (2014): 2697–2707.
21. Gamarnik, Andrea V., and Raul Andino. "Interactions of Viral Protein 3CD and Poly(rC) Binding Protein with the 5′ Untranslated Region of the Poliovirus Genome." Journal of Virology 74.5 (2000): 2219–2226.
22. Clerte, Caroline, and Kathleen B. Hall. "Characterization of Multimeric Complexes Formed by the Human PTB1 Protein on RNA." RNA12.3 (2006): 457–475.
23. Rainer Breitling, Patrick Armengaud, Anna Amtmann, Pawel H. "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments." FEBS Press. 11 August 2004