Project 3: DNA and LinkStrand Analysis

Brian Jordan (bjj17), Avishek Khan (ak410)

Non Linked List Hypothesis

- StringBuilderStrand runtime is O(bS) so it should be O(b) when N is constant
- StringStrand runtime is $O(b^2S)$ so it should be $O(b^2)$ when N is constant

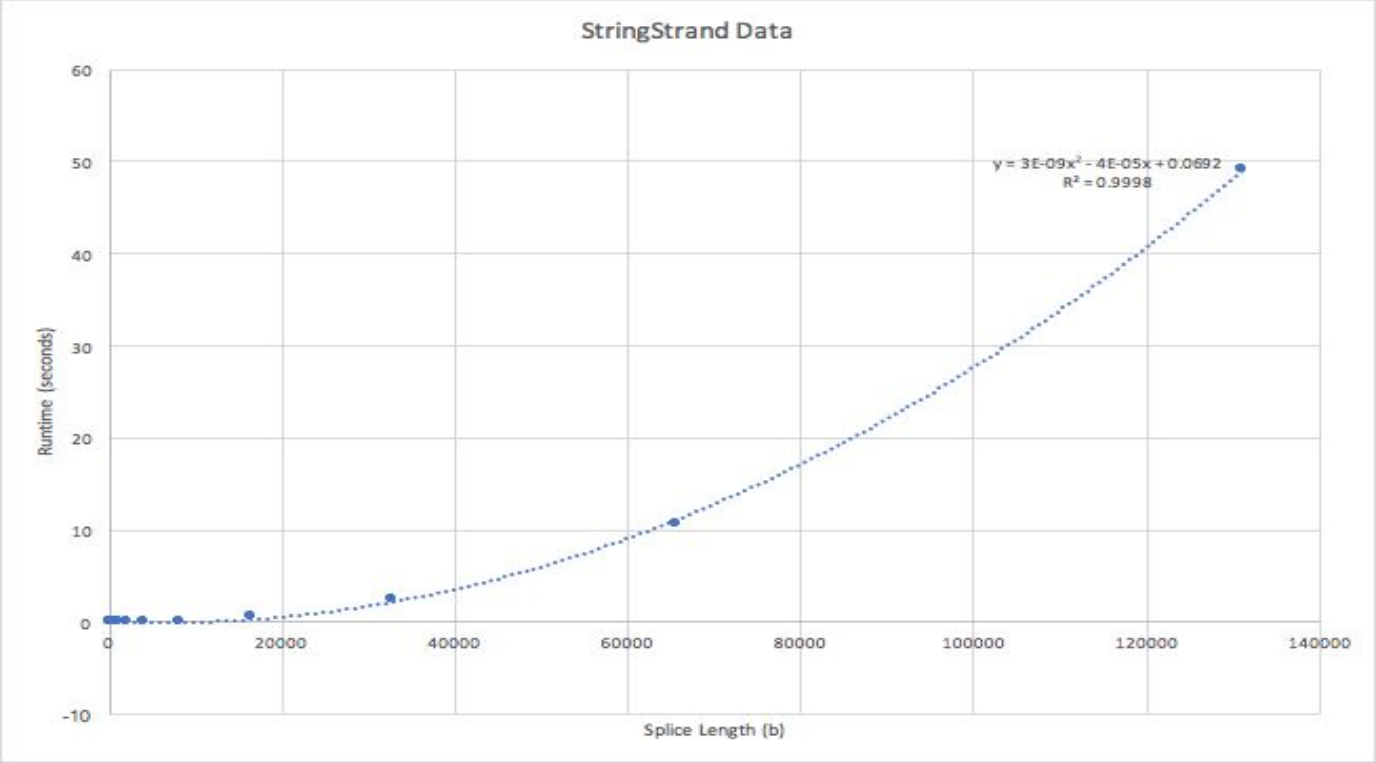
String Strand Data:

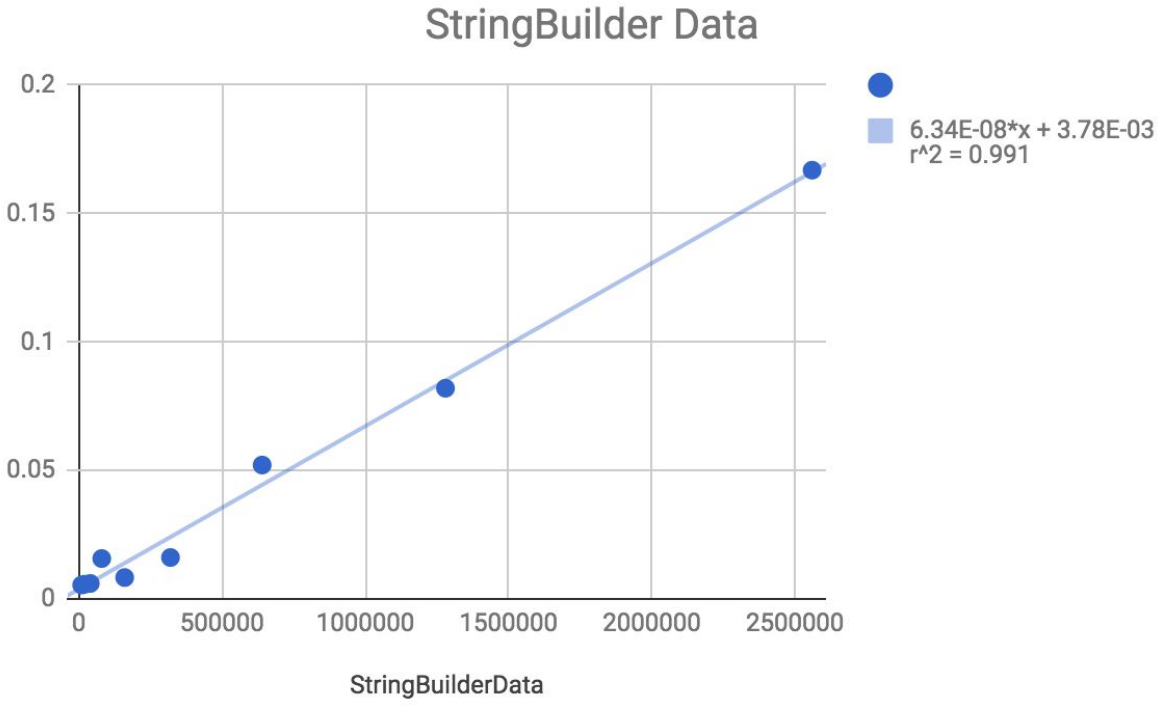
Number of Breaks	Runtime (seconds)
1	0.0000
2	0.0000
4	0.0000
8	0.0001
16	0.0001
32	0.0001
64	0.0001
128	0.0007
256	0.0010
512	0.0041
1024	0.0093
2048	0.0582
4096	0.1547
8192	0.1555
16384	0.6251
32768	2.4874
65536	10.6953
131072	49.1037

StringBuilder Data:

Length of Splice	Runtime (Seconds)
10000	0.0054
20000	0.0057
40000	0.006
80000	0.0157
160000	0.0083
320000	0.0161
640000	0.0521
1280000	0.082
2560000	0.1669

Graphs:


StringStrand Data
y = 3E-09x² - 4E-05x + 0.0692
R² = 0.9998
Runtime (seconds)
Splice Length (b)
60
50
40
30
20
10
0
-10
0
20000
40000
60000
80000
100000
120000
140000


StringBuilder Data
6.34E-08*x + 3.78E-03
r^2 = 0.991
0.2
0.15
0.1
0.05
0
0
500000
1000000
1500000
2000000
2500000
StringBuilderData

Explanation:

In the Analysis class that we created, the StringBuilderStrand and StringStrand classes' runtimes were tested by changing the value of b, the number of breaks in the DNA strand. For testing the StringStrand class, 18 different b values spanning from 0 to 163840 were tested. For testing the StringBuilderStrand class, 9 different b values spanning from 10,000 to 3,638,400 were tested. For the number of breaks in the dna strand being tested, myEnzyme, an arbitrary string, was appended to the created DNA StringBuilder object. From this, we performed five trials, each time replacing the string with myEnzyme with mySplice. These tests were averaged to get the runtime data for each b value. As a result, as the number of breaks increased, so did the length of the dna strand and runtime.

When looking at the data from the average of five trials, the StringStrand data point **matched** the hypothesis that the class would run at $O(b^2s)$ time as N was held constant, because the data points fit a quadratic regression with a correlation coefficient of 0.9998, indicating a strong correlation. StringBuilderStrand data had a linear regression with a correlation coefficient of 0.991, indicating a strong linear fit, meaning that it ran at O(bs) time. This also **matches** the hypothesis proposed.

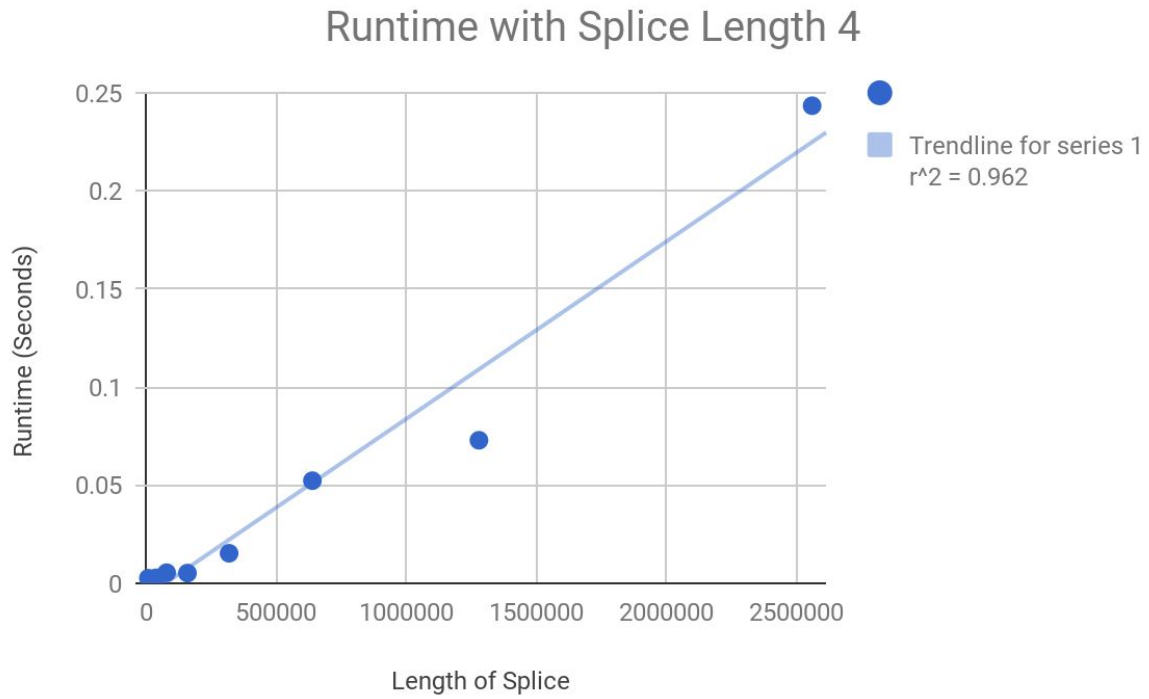
Linked List Hypothesis

- LinkStrand runtime is O(b) when N is constant regardless of the size of splice (s).

LinkStrand Data:

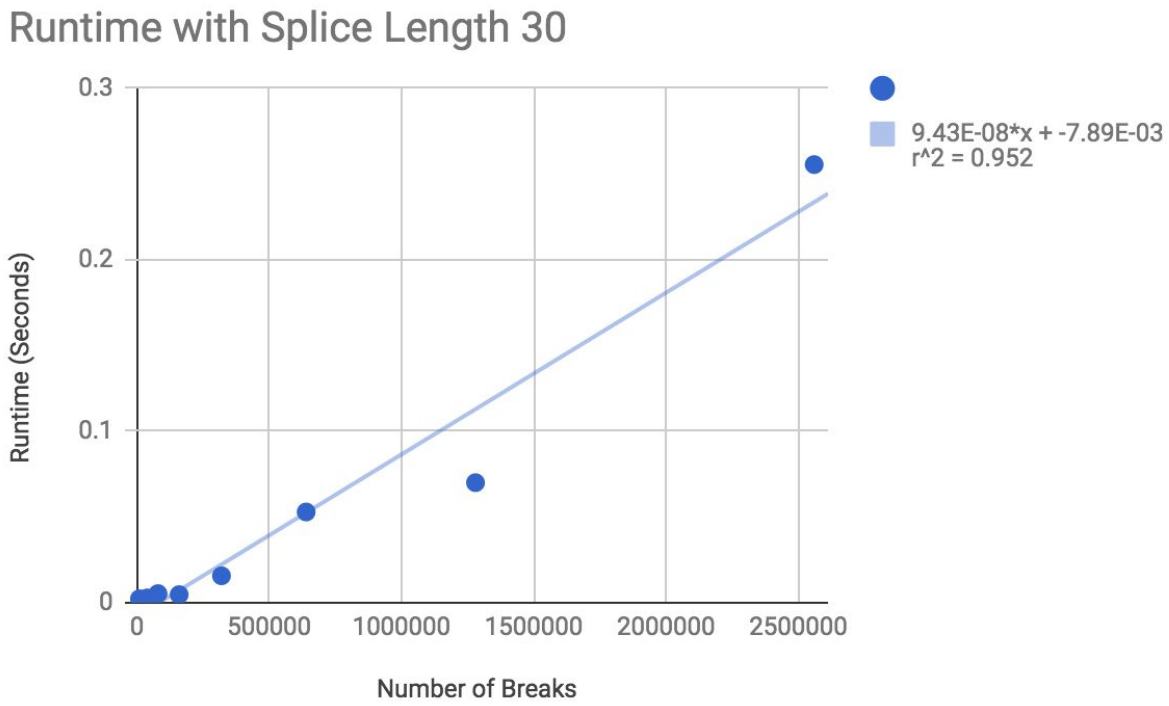
Splice: "bbbb", length 4:

10000	0.0026
20000	0.0019
40000	0.0028
80000	0.0054
160000	0.0052
320000	0.0153
640000	0.0523
1280000	0.0729
2560000	0.2434



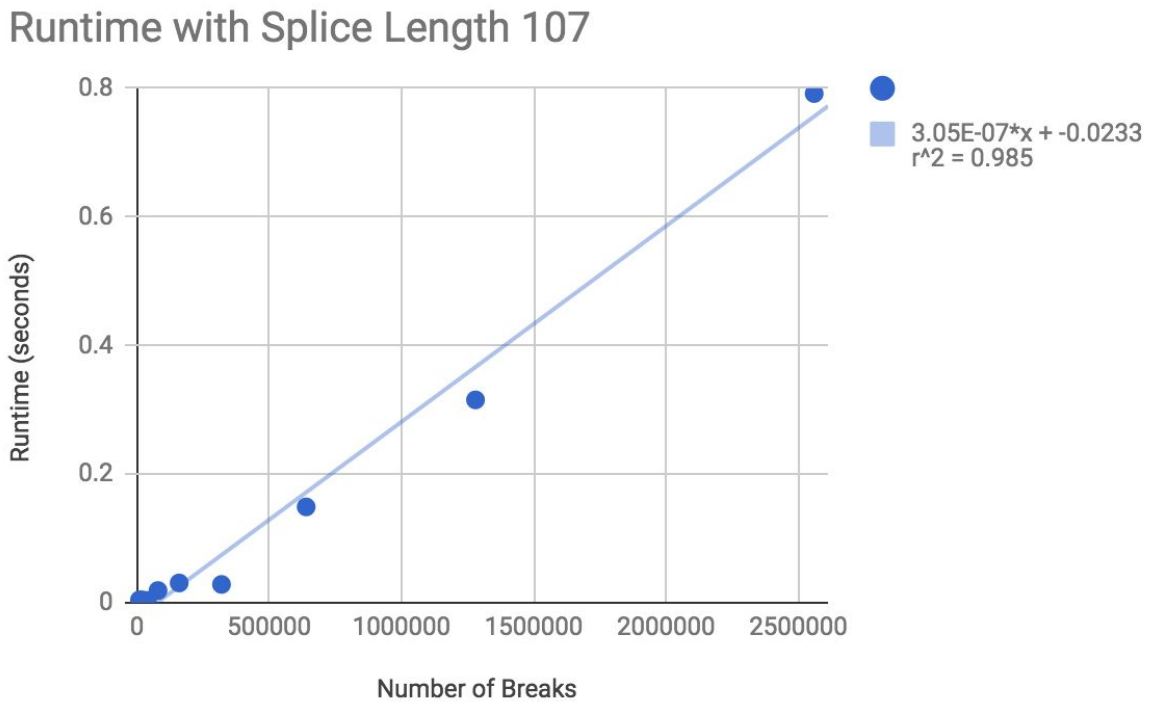
Splice: “”, length 30:

10000	0.0023
20000	0.0018
40000	0.0028
80000	0.0053
160000	0.0047
320000	0.0157
640000	0.0529
1280000	0.0700
2560000	0.2555



Splice: "asaskdfsajdflajsdlfkjalsdkfjlaksjdflaksjdflajsdflajsdlfkjalsdjflaskdfjlasjdfljalsdjkfasldfasdfaskjdflkasdjf", length: 30

10000	0.0044
20000	0.0044
40000	0.0035
80000	0.0189
160000	0.0307
320000	0.0284
640000	0.1490
1280000	0.3154
2560000	0.7918



Explanation:

In this data, the string was built in a similar way as the Non Linked Hypotheses Tests. The b coefficient was tested at 9 values ranging from 10,000 to 3,638,400. Additionally, the dna strand was built using a LinkStrand. Under this hypothesis, the runtime should have ran at O(b) time, regardless of the size of the splice. For this reason, we tested the size of the splice at three different sizes: at the values of 4, 30, and 107. In all three cases, after the data was plotted, the data fit a linear regression that a correlation coefficient of: 0.962, 0.952, and 0.985 respectively. Hence, the correlation was strong and **supports** the hypothesis that a LinkStrand would run in linear O(b) time since even as the size of the splice was changed, the regression remained linear.

Timing Results

Running DNABenchmark with ecoli.txt file
StringStrand Data:
dna length = 4,639,221
cutting at enzyme gaattc

Class	splicee	recomb	time	appends
StringStrand:	256	4,800,471	3.520	1290

```
StringStrand:          512          4,965,591          3.590  1290
StringStrand:        1,024          5,295,831          3.807  1290
StringStrand:        2,048          5,956,311          4.244  1290
StringStrand:        4,096          7,277,271          5.626  1290
StringStrand:        8,192          9,919,191          7.625  1290
StringStrand:       16,384         15,203,031         11.756 1290
StringStrand:       32,768         25,770,711         21.359 1290
StringStrand:       65,536         46,906,071         45.311 1290
StringStrand:      131,072         89,176,791        114.493 1290
StringStrand:      262,144        173,718,231        235.331 1290
StringStrand:      524,288        342,801,111        575.913 1290
Exception in thread "main" java.lang.OutOfMemoryError: Java heap space

StringBuilderStrand Data:
dna length = 4,639,221
cutting at enzyme gaattc
-----
Class                splicee            recomb  time   appends
-----
StringBuilderStrand:          256          4,800,471          0.038  1290
StringBuilderStrand:          512          4,965,591          0.031  1290
StringBuilderStrand:        1,024          5,295,831          0.054  1290
StringBuilderStrand:        2,048          5,956,311          0.024  1290
StringBuilderStrand:        4,096          7,277,271          0.029  1290
StringBuilderStrand:        8,192          9,919,191          0.030  1290
StringBuilderStrand:       16,384         15,203,031          0.064  1290
StringBuilderStrand:       32,768         25,770,711          0.080  1290
StringBuilderStrand:       65,536         46,906,071          0.147  1290
StringBuilderStrand:      131,072         89,176,791          0.242  1290
StringBuilderStrand:      262,144        173,718,231          0.489  1290
StringBuilderStrand:      524,288        342,801,111          0.681  1290
Exception in thread "main" java.lang.OutOfMemoryError: Java heap space

LinkStrand Data:
dna length = 4,639,221
cutting at enzyme gaattc
-----
Class                splicee            recomb     time    appends
-----
LinkStrand:           256          4,800,471  0.025   1290
LinkStrand:           512          4,965,591  0.020   1290
LinkStrand:         1,024          5,295,831  0.023   1290
LinkStrand:         2,048          5,956,311  0.020   1290
```

```
LinkStrand:           4,096         7,277,271 0.027 1290
LinkStrand:           8,192         9,919,191 0.029 1290
LinkStrand:          16,384        15,203,031 0.021 1290
LinkStrand:          32,768        25,770,711 0.021 1290
LinkStrand:          65,536        46,906,071 0.027 1290
LinkStrand:         131,072        89,176,791 0.024 1290
LinkStrand:         262,144       173,718,231       0.024 1290
LinkStrand:         524,288       342,801,111       0.023 1290
LinkStrand:       1,048,576       680,966,871       0.028 1290
LinkStrand:       2,097,152     1,357,298,391       0.030 1290
LinkStrand:       4,194,304     2,709,961,431       0.031 1290
LinkStrand:       8,388,608     5,415,287,511       0.024 1290
LinkStrand:      16,777,216    10,825,939,671       0.024 1290
LinkStrand:      33,554,432    21,647,243,991       0.025 1290
LinkStrand:      67,108,864    43,289,852,631       0.029 1290
LinkStrand:     134,217,728        86,575,069,911       0.021 1290
LinkStrand:     268,435,456       173,145,504,471       0.021 1290
LinkStrand:     536,870,912       346,286,373,591       0.020 1290
Exception in thread "main" java.lang.OutOfMemoryError: Java heap space
```