**DNA Analysis**

**Yuanjie Jin & Junfei Liu**

1) Using SimpleStrand to create the recombinant strand is an O(N) operation where N is the size of the resulting recombinant strand. As the length of splicee grows, the code takes longer to execute.

Reasoning: In SimpleStrand, when it replaces the enzyme-recognized sequence with the splicee, the method of cutAndSplice calls append function. As in StringBuilder, such append method will copy the string to append and build on the original stored string. Therefore, the time this method takes correlates with the total length of all the strings to append, which is the length of the whole resulting recombinant strand.

Data: Besides ecoli.dat and ecolimed.dat, we also generated new data files: ecolimedx8, which contains 8 copies of ecolimed; ecolismall, which is a small truncation of ecolimed; and ecolimini, which is an even smaller truncation of ecolimed. The specific output of running DNABenchMark on each data file is listed in Table 1 on next page. The performance of SimpleStrand is shown in **Fig 1** below.

1.1) For each data file, as the length of splicee doubles, its length of resulting recombinant strand also increases. As is shown in Fig 1, as the lengths of recombinant sequences increase, the code will take longer to execute. And the correlation between them is linear.

1.2) Comparing between different data files, although they have different append numbers, as their recombinant lengths get close, the time lengths they take to run the code are also similar.

1.3) Therefore, creating recombinant strand using SimpleStrand is an O(N) operation.

**Table 1.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SimpleStrand generated recombinant of ecoli.dat** | | | **dna length = 4,639,221** | **cutting at enzyme gaattc** |
| Class | splicee | recomb | time |  |
| SimpleStrand: | 256 | 4,800,471 | 0.056 | # append calls = 1290 |
| SimpleStrand: | 512 | 4,965,591 | 0.054 | # append calls = 1290 |
| SimpleStrand: | 1,024 | 5,295,831 | 0.048 | # append calls = 1290 |
| SimpleStrand: | 2,048 | 5,956,311 | 0.047 | # append calls = 1290 |
| SimpleStrand: | 4,096 | 7,277,271 | 0.044 | # append calls = 1290 |
| SimpleStrand: | 8,192 | 9,919,191 | 0.081 | # append calls = 1290 |
| SimpleStrand: | 16,384 | 15,203,031 | 0.089 | # append calls = 1290 |
| **SimpleStrand generated recombinant of ecolimedx8.dat** | | | **dna length = 2,561,280** | **cutting at enzyme gaattc** |
| Class | splicee | recomb | time |  |
| SimpleStrand: | 256 | 2,651,280 | 0.026 | # append calls = 720 |
| SimpleStrand: | 512 | 2,743,440 | 0.027 | # append calls = 720 |
| SimpleStrand: | 1,024 | 2,927,760 | 0.026 | # append calls = 720 |
| SimpleStrand: | 2,048 | 3,296,400 | 0.027 | # append calls = 720 |
| SimpleStrand: | 4,096 | 4,033,680 | 0.024 | # append calls = 720 |
| SimpleStrand: | 8,192 | 5,508,240 | 0.029 | # append calls = 720 |
| SimpleStrand: | 16,384 | 8,457,360 | 0.069 | # append calls = 720 |
| SimpleStrand: | 32,768 | 14,355,600 | 0.117 | # append calls = 720 |
| **SimpleStrand generated recombinant of ecolimed.dat** | | | **dna length = 320,160** | **cutting at enzyme gaattc** |
| Class | splicee | recomb | time |  |
| SimpleStrand: | 256 | 331410 | 0.002 | # append calls = 90 |
| SimpleStrand: | 512 | 342930 | 0.002 | # append calls = 90 |
| SimpleStrand: | 1,024 | 365970 | 0.011 | # append calls = 90 |
| SimpleStrand: | 2,048 | 412050 | 0.002 | # append calls = 90 |
| SimpleStrand: | 4,096 | 504210 | 0.003 | # append calls = 90 |
| SimpleStrand: | 8,192 | 688530 | 0.003 | # append calls = 90 |
| SimpleStrand: | 16,384 | 1057170 | 0.004 | # append calls = 90 |
| SimpleStrand: | 32,768 | 1794450 | 0.004 | # append calls = 90 |
| SimpleStrand: | 65,536 | 3269010 | 0.01 | # append calls = 90 |
| SimpleStrand: | 131,072 | 6218130 | 0.03 | # append calls = 90 |
| SimpleStrand: | 262,144 | 12116370 | 0.084 | # append calls = 90 |
| SimpleStrand: | 524,288 | 23912850 | 0.132 | # append calls = 90 |
| **SimpleStrand generated recombinant of ecolismall.dat** | | | **dna length = 200,040** | **cutting at enzyme gaattc** |
| Class | splicee | recomb | time |  |
| SimpleStrand: | 256 | 206,290 | 0.002 | # append calls = 50 |
| SimpleStrand: | 512 | 212,690 | 0.001 | # append calls = 50 |
| SimpleStrand: | 1,024 | 225,490 | 0.002 | # append calls = 50 |
| SimpleStrand: | 2,048 | 251,090 | 0.001 | # append calls = 50 |
| SimpleStrand: | 4,096 | 302,290 | 0.002 | # append calls = 50 |
| SimpleStrand: | 8,192 | 404,690 | 0.002 | # append calls = 50 |
| SimpleStrand: | 16,384 | 609,490 | 0.011 | # append calls = 50 |
| SimpleStrand: | 32,768 | 1,019,090 | 0.002 | # append calls = 50 |
| SimpleStrand: | 65,536 | 1,838,290 | 0.005 | # append calls = 50 |
| SimpleStrand: | 131,072 | 3,476,690 | 0.009 | # append calls = 50 |
| SimpleStrand: | 262,144 | 6,753,490 | 0.019 | # append calls = 50 |
| SimpleStrand: | 524,288 | 13,307,090 | 0.078 | # append calls = 50 |
| SimpleStrand: | 1,048,576 | 26,414,290 | 0.143 | # append calls = 50 |
| **SimpleStrand generated recombinant of ecolimini.dat** | | | **dna length = 100,020** | **cutting at enzyme gaattc** |
| Class | splicee | recomb | time |  |
| SimpleStrand: | 256 | 101,520 | 0.001 | # append calls = 12 |
| SimpleStrand: | 512 | 103,056 | 0.001 | # append calls = 12 |
| SimpleStrand: | 1,024 | 106,128 | 0.001 | # append calls = 12 |
| SimpleStrand: | 2,048 | 112,272 | 0.001 | # append calls = 12 |
| SimpleStrand: | 4,096 | 124,560 | 0.001 | # append calls = 12 |
| SimpleStrand: | 8,192 | 149,136 | 0.007 | # append calls = 12 |
| SimpleStrand: | 16,384 | 198,288 | 0.001 | # append calls = 12 |
| SimpleStrand: | 32,768 | 296,592 | 0.002 | # append calls = 12 |
| SimpleStrand: | 65,536 | 493,200 | 0.001 | # append calls = 12 |
| SimpleStrand: | 131,072 | 886,416 | 0.001 | # append calls = 12 |
| SimpleStrand: | 262,144 | 1,672,848 | 0.011 | # append calls = 12 |
| SimpleStrand: | 524,288 | 3,245,712 | 0.008 | # append calls = 12 |
| SimpleStrand: | 1,048,576 | 6,391,440 | 0.02 | # append calls = 12 |
| SimpleStrand: | 2,097,152 | 12,682,896 | 0.076 | # append calls = 12 |

2) Test SimpleStrand with multiple memory settings and report the power-of-two string that can be used without running out of memory with the input file ecoli.dat.

* 1. The Java runtime heap size is set using the command-line argument “-Xmx512M”. As is shown in the table 2 below, the longest recombinant strand constructed using the largest possible power-of-two splicee (131,072) is 89,176,791. And this process takes around 0.464 second.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 2. |  |  | dna length = 4,639,221 | cutting at enzyme gaattc |
| Class | splicee | recomb | time |  |
| SimpleStrand: | 256 | 4,800,471 | 0.056 | # append calls = 1290 |
| SimpleStrand: | 512 | 4,965,591 | 0.053 | # append calls = 1290 |
| SimpleStrand: | 1,024 | 5,295,831 | 0.047 | # append calls = 1290 |
| SimpleStrand: | 2,048 | 5,956,311 | 0.048 | # append calls = 1290 |
| SimpleStrand: | 4,096 | 7,277,271 | 0.043 | # append calls = 1290 |
| SimpleStrand: | 8,192 | 9,919,191 | 0.092 | # append calls = 1290 |
| SimpleStrand: | 16,384 | 15,203,031 | 0.097 | # append calls = 1290 |
| SimpleStrand: | 32,768 | 25,770,711 | 0.154 | # append calls = 1290 |
| SimpleStrand: | 65,536 | 46,906,071 | 0.239 | # append calls = 1290 |
| SimpleStrand: | 131,072 | 89,176,791 | 0.464 | # append calls = 1290 |

* 1. When the size of the heap available to the Java runtime is doubled (-Xmx1024M), the next power-of-two strand (262,144) is now supported. The time it takes to construct the current longest recombinant strand (173,718,231) is 0.938 second, which is roughly the doubled time length of the previous 0.464 second.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 3 |  |  | dna length = 4,639,221 | cutting at enzyme gaattc |
| Class | splicee | recomb | time |  |
| SimpleStrand: | 256 | 4,800,471 | 0.056 | # append calls = 1290 |
| SimpleStrand: | 512 | 4,965,591 | 0.055 | # append calls = 1290 |
| SimpleStrand: | 1,024 | 5,295,831 | 0.051 | # append calls = 1290 |
| SimpleStrand: | 2,048 | 5,956,311 | 0.044 | # append calls = 1290 |
| SimpleStrand: | 4,096 | 7,277,271 | 0.044 | # append calls = 1290 |
| SimpleStrand: | 8,192 | 9,919,191 | 0.051 | # append calls = 1290 |
| SimpleStrand: | 16,384 | 15,203,031 | 0.101 | # append calls = 1290 |
| SimpleStrand: | 32,768 | 25,770,711 | 0.113 | # append calls = 1290 |
| SimpleStrand: | 65,536 | 46,906,071 | 0.236 | # append calls = 1290 |
| SimpleStrand: | 131,072 | 89,176,791 | 0.37 | # append calls = 1290 |
| SimpleStrand: | 262,144 | 173,718,231 | 0.938 | # append calls = 1290 |

2.3) Further thoughts as for mathematical prove: : the length of original DNA for recombination; the length of the largest possible splicee given a set memory; : the number of appends; : the length of the enzyme site to be replaced.

Let

Because

Thus, theoretically, the doubling of memory will always be able to support the doubling of splicee length.

Also, when the original size of the DNA to be spliced is significantly smaller than Memory and could be neglected,

This means memory doubling will only be able to support one round of splicee doubling.

However, it also needed to be pointed out that the memory is not allocated solo for the storage of resulting recombinant strand, as other factors may affect.

3) Using LinkStrand to create the recombinant strand is an O(B) operation where B is the number of breaks/splits created by the restriction enzyme.

Reasoning: In LinkStrand, repeated nodes are created in a way that they contain pointers to the exactly same splicee string. This representation avoids recopying the splicee string over and over again. Therefore, the append method takes constant time to generate a new node regardless of the base-pair length of the splicee, i.e. it is an O(1) operation. As a result, the overall running time should be O(B), where B is the number of breaks in the original strand.

Data: Besides ecoli.dat and ecolimed.dat, we also used other data files: ecolimedx8, which contains 8 copies of ecolimed; ecolimedx4, which contains 4 copies of ecolimed; and ecolimedx12, which contains 12 copies of ecolimed. The specific output of running DNABenchMark on each data file is listed in **Table 4** on next page. The performance of LinkStrand is shown in **Fig 2** below.

3.1) As is shown in **Fig2**, for each data file, as their resulting recombinant strands increase, their running time lengths roughly remain, with only small fluctuations.

3.2) Files with more append numbers take more time to run the code, showing correlation between time and append numbers.

The append number is approximately twice the number of breaks, so here we use append number as a reference of B.

3.3) A recombinant length range from 21,533,400 to 26,152,080 is picked, and the time length vs. number of appends is plotted in **Fig 3.** The correlation is shown to be linear.

|  |  |  |  |
| --- | --- | --- | --- |
| Data File | recomb | time | append |
| ecolimed | 23,912,850 | 0.001 | 90 |
| ecolimedx4 | 24,872,520 | 0.006 | 360 |
| ecolimedx8 | 26,152,080 | 0.012 | 720 |
| ecolimedx12 | 21,533,400 | 0.018 | 1080 |
| ecoli | 25,770,711 | 0.021 | 1290 |

3.4) Comparing the same data file run by SimpleStrand and LinkStrand, the latter is more efficient both in time and memory. For example, SimpleStrand generates a 15,203,031 base pair recombinant from ecoli.dat in 0.089 second, while LinkStrand can generate a 10,825,939,671 base pair recombinant from ecoli.dat in 0.023 second.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 4** | | | | |
| **LinkStrand generated recombinant of ecolimed.dat dna length =320,160, cutting at enzyme gaattc** | | | | |
| Class | splicee | recomb | time |  |
| LinkStrand: | 256 | 331,410 | 0.002 | # append calls = 90 |
| LinkStrand: | 512 | 342,930 | 0.002 | # append calls = 90 |
| LinkStrand: | 1,024 | 365,970 | 0.001 | # append calls = 90 |
| LinkStrand: | 2,048 | 412,050 | 0.001 | # append calls = 90 |
| LinkStrand: | 4,096 | 504,210 | 0.002 | # append calls = 90 |
| LinkStrand: | 8,192 | 688,530 | 0.001 | # append calls = 90 |
| LinkStrand: | 16,384 | 1,057,170 | 0.002 | # append calls = 90 |
| LinkStrand: | 32,768 | 1,794,450 | 0.003 | # append calls = 90 |
| LinkStrand: | 65,536 | 3,269,010 | 0.001 | # append calls = 90 |
| LinkStrand: | 131,072 | 6,218,130 | 0.002 | # append calls = 90 |
| LinkStrand: | 262,144 | 12,116,370 | 0.002 | # append calls = 90 |
| LinkStrand: | 524,288 | 23,912,850 | 0.001 | # append calls = 90 |
| LinkStrand: | 1,048,576 | 47,505,810 | 0.002 | # append calls = 90 |
| LinkStrand: | 2,097,152 | 94,691,730 | 0.002 | # append calls = 90 |
| LinkStrand: | 4,194,304 | 189,063,570 | 0.001 | # append calls = 90 |
| LinkStrand: | 8,388,608 | 377,807,250 | 0.001 | # append calls = 90 |
| LinkStrand: | 16,777,216 | 755,294,610 | 0.001 | # append calls = 90 |
| **LinkStrand generated recombinant of ecolimedx4.dat dna length = 1,280,640 , cutting at enzyme gaattc** | | | | |
| Class | splicee | recomb | time |  |
| LinkStrand: | 256 | 1,325,640 | 0.006 | # append calls = 360 |
| LinkStrand: | 512 | 1,371,720 | 0.006 | # append calls = 360 |
| LinkStrand: | 1,024 | 1,463,880 | 0.006 | # append calls = 360 |
| LinkStrand: | 2,048 | 1,648,200 | 0.006 | # append calls = 360 |
| LinkStrand: | 4,096 | 2,016,840 | 0.008 | # append calls = 360 |
| LinkStrand: | 8,192 | 2,754,120 | 0.007 | # append calls = 360 |
| LinkStrand: | 16,384 | 4,228,680 | 0.006 | # append calls = 360 |
| LinkStrand: | 32,768 | 7,177,800 | 0.006 | # append calls = 360 |
| LinkStrand: | 65,536 | 13,076,040 | 0.006 | # append calls = 360 |
| LinkStrand: | 131,072 | 24,872,520 | 0.006 | # append calls = 360 |
| LinkStrand: | 262,144 | 48,465,480 | 0.006 | # append calls = 360 |
| LinkStrand: | 524,288 | 95,651,400 | 0.006 | # append calls = 360 |
| LinkStrand: | 1,048,576 | 190,023,240 | 0.005 | # append calls = 360 |
| LinkStrand: | 2,097,152 | 378,766,920 | 0.006 | # append calls = 360 |
| LinkStrand: | 4,194,304 | 756,254,280 | 0.006 | # append calls = 360 |
| LinkStrand: | 8,388,608 | 1,511,229,000 | 0.008 | # append calls = 360 |
| LinkStrand: | 16,777,216 | 3,021,178,440 | 0.006 | # append calls = 360 |
| **LinkStrand generated recombinant of ecolimedx8.dat; dna length = 2,561,280, cutting at enzyme gaattc** | | | | |
| Class | splicee | recomb | time |  |
| LinkStrand: | 256 | 2,651,280 | 0.012 | # append calls = 720 |
| LinkStrand: | 512 | 2,743,440 | 0.011 | # append calls = 720 |
| LinkStrand: | 1,024 | 2,927,760 | 0.012 | # append calls = 720 |
| LinkStrand: | 2,048 | 3,296,400 | 0.011 | # append calls = 720 |
| LinkStrand: | 4,096 | 4,033,680 | 0.013 | # append calls = 720 |
| LinkStrand: | 8,192 | 5,508,240 | 0.014 | # append calls = 720 |
| LinkStrand: | 16,384 | 8,457,360 | 0.013 | # append calls = 720 |
| LinkStrand: | 32,768 | 14,355,600 | 0.013 | # append calls = 720 |
| LinkStrand: | 65,536 | 26,152,080 | 0.012 | # append calls = 720 |
| LinkStrand: | 131,072 | 49,745,040 | 0.012 | # append calls = 720 |
| LinkStrand: | 262,144 | 96,930,960 | 0.012 | # append calls = 720 |
| LinkStrand: | 524,288 | 191,302,800 | 0.012 | # append calls = 720 |
| LinkStrand: | 1,048,576 | 380,046,480 | 0.012 | # append calls = 720 |
| LinkStrand: | 2,097,152 | 757,533,840 | 0.013 | # append calls = 720 |
| LinkStrand: | 4,194,304 | 1,512,508,560 | 0.012 | # append calls = 720 |
| LinkStrand: | 8,388,608 | 3,022,458,000 | 0.013 | # append calls = 720 |
| LinkStrand: | 16,777,216 | 6,042,356,880 | 0.012 | # append calls = 720 |
| **LinkStrand generated recombinant of ecolimedx12.dat; dna length = 3,841,920; cutting at enzyme gaattc** | | | | |
| Class | splicee | recomb | time |  |
| LinkStrand: | 256 | 3,976,920 | 0.016 | # append calls = 1080 |
| LinkStrand: | 512 | 4,115,160 | 0.016 | # append calls = 1080 |
| LinkStrand: | 1,024 | 4,391,640 | 0.017 | # append calls = 1080 |
| LinkStrand: | 2,048 | 4,944,600 | 0.017 | # append calls = 1080 |
| LinkStrand: | 4,096 | 6,050,520 | 0.02 | # append calls = 1080 |
| LinkStrand: | 8,192 | 8,262,360 | 0.022 | # append calls = 1080 |
| LinkStrand: | 16,384 | 12,686,040 | 0.019 | # append calls = 1080 |
| LinkStrand: | 32,768 | 21,533,400 | 0.018 | # append calls = 1080 |
| LinkStrand: | 65,536 | 39,228,120 | 0.019 | # append calls = 1080 |
| LinkStrand: | 131,072 | 74,617,560 | 0.018 | # append calls = 1080 |
| LinkStrand: | 262,144 | 145,396,440 | 0.018 | # append calls = 1080 |
| LinkStrand: | 524,288 | 286,954,200 | 0.017 | # append calls = 1080 |
| LinkStrand: | 1,048,576 | 570,069,720 | 0.018 | # append calls = 1080 |
| LinkStrand: | 2,097,152 | 1,136,300,760 | 0.017 | # append calls = 1080 |
| LinkStrand: | 4,194,304 | 2,268,762,840 | 0.018 | # append calls = 1080 |
| LinkStrand: | 8,388,608 | 4,533,687,000 | 0.017 | # append calls = 1080 |
| LinkStrand: | 16,777,216 | 9,063,535,320 | 0.018 | # append calls = 1080 |
| **LinkStrand generated recombinant of ecoli.dat; dna length = 4,639,221; cutting at enzyme gaattc** | | | | |
| Class | splicee | recomb | time |  |
| LinkStrand: | 256 | 4,800,471 | 0.019 | # append calls = 1290 |
| LinkStrand: | 512 | 4,965,591 | 0.019 | # append calls = 1290 |
| LinkStrand: | 1,024 | 5,295,831 | 0.021 | # append calls = 1290 |
| LinkStrand: | 2,048 | 5,956,311 | 0.019 | # append calls = 1290 |
| LinkStrand: | 4,096 | 7,277,271 | 0.022 | # append calls = 1290 |
| LinkStrand: | 8,192 | 9,919,191 | 0.022 | # append calls = 1290 |
| LinkStrand: | 16,384 | 15,203,031 | 0.021 | # append calls = 1290 |
| LinkStrand: | 32,768 | 25,770,711 | 0.021 | # append calls = 1290 |
| LinkStrand: | 65,536 | 46,906,071 | 0.02 | # append calls = 1290 |
| LinkStrand: | 131,072 | 89,176,791 | 0.021 | # append calls = 1290 |
| LinkStrand: | 262,144 | 173,718,231 | 0.02 | # append calls = 1290 |
| LinkStrand: | 524,288 | 342,801,111 | 0.025 | # append calls = 1290 |
| LinkStrand: | 1,048,576 | 680,966,871 | 0.02 | # append calls = 1290 |
| LinkStrand: | 2,097,152 | 1,357,298,391 | 0.02 | # append calls = 1290 |
| LinkStrand: | 4,194,304 | 2,709,961,431 | 0.022 | # append calls = 1290 |
| LinkStrand: | 8,388,608 | 5,415,287,511 | 0.02 | # append calls = 1290 |
| LinkStrand: | 16,777,216 | 10,825,939,671 | 0.023 | # append calls = 1290 |