CPTS471 PA2 Report

Team

Miranda Stoddard

Ace Cassidy

System configuration

CPU: Intel Core i7-1065G7 Clock Rate: 1.30GHz × 8 Ram: 8 Gb DDR4

Construction performance

- dna-inputs/chr12.fas Time (seconds): 0.62
- dna-inputs/Slyco.fas Time (seconds): 0.04
- dna-inputs/Covid_USA-CA4.fasta Time (seconds): 0.01
- dna-inputs/colorblind_human_gene.fasta Time (seconds): 0.0
- dna-inputs/colorblind_mouse_gene.fasta Time (seconds): 0.00
- dna-inputs/Covid_Australia.fasta Time (seconds): 0.00
- dna-inputs/Covid_Brazil.fasta Time (seconds): 0.00
- dna-inputs/Covid_India.fasta Time (seconds): 0.00
- dna-inputs/Covid_Wuhan.fasta Time (seconds): 0.00
- dna-inputs/Human-BRCA2-cds.fasta Time (seconds): 0.00
- english-inputs/s1.fas Time (seconds): 0.00
- english-inputs/s2.fas Time (seconds): 0.00

Justification

Given the data points available (most times are zero due to running on fast hardware with low resolution timing) it is hard to create a trend line, but as we implemented the McCreight algorithm using suffix link insertion we expect linear time performance. The time to insert one basepair is 0.000000335 for the smallest one which is non-zero, then it's 0.000000257 for the middle one, and 0.000000575 for the largest one. I think the variation is because we don't have very good time specificity since it's large and then smaller and then large again. Based on this the time appears to be linear because it's not consistently growing.

Implementation constant

For each new character there’s the possibility of one internal node and one leaf node being created so the worst case implementation constant is 312.

BWT Index

Format of BWT output files will be “*/BWT.out” where * is the name of the input file

Exact matching repeat

•english-inputs/s1.fas = ANA

•english-inputs/s1.fas = ISSI

•colorblind_human_gene.fasta.BWT.out AAAATTAAAAATTAAAAA

•colorblind_mouse_gene.fasta.BWT.out TTTTTTTTTTTTT

•Covid_Australia.fasta.BWT.out AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

•Covid_Brazil.fasta.BWT.out ATCAGACAACTACTATT

•Covid_India.fasta.BWT.out ATCAGACAACTACTATT

•Covid_USA-CA4.fasta.BWT.out ATCAGACAACTACTATT

•Covid_Wuhan.fasta.BWT.out AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

•Human-BRCA2-cds.fasta.BWT.out AAGAGATACAGAAT

•Slyco.fas.BWT.out AATCAATGCAATTTAGGAGGAATCAATGCAATTTAGGAGGAATCAATG

•chr12.fas.BWT.out =
CTCATGTTTGCCGCTCTGATGGTGCGGAAAAAACTGCTCCATGAAGCAAACTGTCCGGGCAAATCCTTT
CACGCTCGGGAAGCTTTGTGAAAGCCCTTCTCTTTCAACCCATCTTTGCAACGAAAAAAAAAAAAAAAA
TAAAAAATAAAAAGACCAAATAGTAAATAGTAACTTACATACATTAGTAAATGGTACACTCTTACACACTAT
CATCCTCATCGTATATTATAATAGATATATACAATACATGTTTTTACCCGGATCATAGAATTCTTAAGACAAAT
AAAATTTATAGAGACTTGTTCAGTCTACTTCTCTCTAAACTAGGCCCCGGCTCCTGCCAGTACCCACTTAG
AAAGAAATAAAAAACAAATCAGACAACAAAGGCTTAATCTCAGCAGATCGTAACAACAAGGCTACTCTAC
TGCTTACAATACCCCGTTGTACATCTAAGTCGTATACAAATGATTTATCCCCACGCAAAATGACATTGCAAT
TCGCCAGCAAGCACCCAAGGCCTTTCCGCCAAGTGCACCGTTGCTAGCCTGCTATGGTTCAGCGACGCC
ACAAGGACGCCTTATTCGTATCCATCTATATTGTGTGGAGCAAAGAAATCACCGCGTTCTAGCATGGATTC
TGACTTAGAGGCGTTCAGCCATAATCCAGCGGATGGTAGCTTCGCGGCAATGCCTGATCAGACAGCCGC
AAAAACCAATTATCCGAATGAACTGTTCCTCTCGTACTAAGTTCAATTACTATTGCGGTAACATTCATCAGT
AGGGTAAAACTAACCTGTCTCACGACGGTCTAAACCCAGCTCACGTTCCCTATTAGTGGGTGAACAATCC
AACGCTTACCGAATTCTGCTTCGGTATGATAGGAAGAGCCGACATCGAAGAATCAAAAAGCAATGTCGCT
ATGAACGCTTGACTGCCACAAGCCAGTTATCCCTGTGGTAACTTTTCTGGCACCTCTAGCCTCAAATTCC
GAGGGACTAAAGGATCGATAGGCCACACTTTCATGGTTTGTATTCACACTGAAAATCAAAATCAAGGGGG
CTTTTACCCTTTTGTTCTACTGGAGATTTCTGTTCTCCATGAGCCCCCCTTAGGACATCTGCGTTATCGTTT
AACAGATGTGCCGCCCCAGCCAAACTCCCCACCTGACAATGTCTTCAACCCGGATCAGCCCCGAATGG
GACCTTGAATGCTAGAACGTGGAAAATGAATTCCAGCTCCGCTTCATTGAATAAGTAAAGAAACTATAAA
GGTAGTGGTATTTCACTGGCGCCGAAGCTCCCACTTATTCTACACCCTCTATGTCTCTTCACAATGTCAAA
CTAGAGTCAAGCTCAACAGGGTCTTCTTTCCCCGCTGATTCTGCCAAGCCCGTTCCCTTGGCTGTGGTTT
CGCTAGATAGTAGATAGGGACAGTGGGAATCTCGTTAATCCATTCATGCGCGTCACTAATTAGATGACGAG
GCATTTGGCTACCTTAAGAGAGTCATAGTTACTCCCGCCGTTTACCCGCGCTTGGTTGAATTTCTTCACTT
TGACATTCAGAGCACTGGGCAGAAATCACATTGCGTCAACATCACTTTCTGACCATCGCAATGCTATGTTT
TAATTAGACAGTCAGATTCCCCTTGTCCGTACCAGTTCTAAGTTGATCGTTAATTGTAGCAAGCGACGGTCT
ACAAGAGACCTACCAAGGCCGTCTACAACAAGGCACGCAAGTAGTCCGCCTAGCAGAGCAAGCCCCAC
CAAGCAGTCCACAAGCACGCCCGCTGCGTCTGACCAAGGCCCTCACTACCCGACCCTTAGAGCCAATCC
TTATCCCGAAGTTACGGATCTATTTTGCCGACTTCCCTTATCTACATTATTCTATCAACTAGAGGCTGTTCAC
CTTGGAGACCTGCTGCGGTTATCAGTACGACCTGGCATGAAAACTATTCCTTCCTGTGGATTTTCACGGG
CCGTCACAAGCGCACCGGAGCCAGCAAAGGTGCTGGCCTCTTCCAGCCATAAGACCCCATCTCCGGATA
AACCAATTCCGGGGTGATAAGCTGTTAAGAAGAAAAGATAACTCCTCCCAGGGCTCGCGCCGACGTCTC
CACATTCAGTTACGTTACCGTGAAGAATCCATATCCAGGTTCCGGAATCTTAACCGGATTCCCTTTCGATG
GTGGCCTGCATAAAATCAGGCCTTTGAAACGGAGCTTCCCCATCTCTTAGGATCGACTAACCCACGTCCA

ACTGCTGTTGACGTGGAACCTTTCCCCACTTCAGTCTTCAAAGTTCTCATTTGAATATTTGCTACTACCACC
AAGATCTGCACTAGAGGCCGTTCGACCCGACCTTACGGTCTAGGCTTCGTCACTGACCTCCACGCCTGC
CTACTCGTCAGGGCATCATATCAACCCTGACGGTAGAGTATAGGTAACACGCTTGAGCGCCATCCATTTTC
AGGGCTAGTTCATTCGGCCGGTGAGTTGTTACACACTCCTTAGCGGATTCCGACTTCCATGGCCACCGTC
CGGCTGTCTAGATGAACTAACACCTTTTGTGGTGTCTGATGAGCGTGTATTCCGGCACCTTAACTCTACGT
TCGGTTCATCCCGCATCGCCAGTTCTGCTTACCAAAAATGGCCCACTAAAAGCTCTTCATTCAAATGTCCA
CGTTCAATTAAGTAACAAGGACTTCTTACATATTTAAAGTTTGAGAATAGGTCAAGGTCATTTCGACCCCG
GAACCTCTAATCATTCGCTTTACCTCATAAAACTGATACGAGCTTCTGCTATCCTGAGGGAAACTTCGGCA
GGAACCAGCTACTAGATGGTTCGATTAGTCTTTCGCCCCTATACCCAAATTCGACGATCGATTTGCACGTC
AGAACCGCTACGAGCCTCCACCAGAGTTTCCTCTGGCTTCACCCTATTCAGGCATAGTTCACCATCTTTCG
GGTCCCAACAGCTATGCTCTTACTCAAATCCATCCGAAGACATCAGGATCGGTCGATTGTGCACCTCTTG
CGAGGCCCCAACCTACGTTCACTTTCATTACGCGTATGGGTTTTACACCCAAACACTCGCATAGACGTTAG
ACTCCTTGGTCCGTGTTTCAAGACGGGCGGCATATAACCATTATGCCAGCATCCTTGACTTACGTCGCAGT
CCTCAGTCCCAGCTGGCAGTATTCCCACAGGCTATAATACTTACCGAGGCAAGCTACATTCCTATGGATTT
ATCCTGCCACCAAAACTGATGCTGGCCCAGTGAAATGCGAGATTCCCCTACCCACAAGGAGCAGAGGG
CACAAAACACCATGTCTGATCAAATGCCCTTCCCTTTCAACAATTTCACGTACTTTTTCACTCTCTTTTCAA
AGTTCTTTTCATCTTTCCATCACTGTACTTGTTCGCTATCGGTCTCTCGCCAATATTTAGCTTTAGATGGAAT
TTACCACCCACTTAGAGCTGCATTCCCAAACAACTCGACTCTTCGAAGGCACTTTACAAAGAACCGCACT
CCTCGCCACACGGGATTCTCACCCTCTATGACGTCCTGTTCCAAGGAACATAGACAAGGAACGGCCCCA
AAGTTGCCCTCTCCAAATTACAACTCGGGCACCGAAGGTACCAGATTTCAAATTTGAGCTTTTGCCGCTT
CACTCGCCGTTACTAAGGCAATCCCGGTTGGTTTCTTTTCCTCCGCTTATTGATATGCTTAAGTTCAGCGG
GTACTCCTACCTGATTTGAGGTCAAACTTTAAGAACATTGTTCGCCTAGACGCTCTCTTCTTATCGATAACG
TTCCAATACGCTCAGTATAAAAAAAGATTAGCCGCAGTTGGTAAAACCTAAAACGACCGTACTTGCATTAT
ACCTCAAGCACGCAGAGAAACCTCTCTTTGGAAAAAAAACATCCAATGAAAAGGCCAGCAATTTCAAGT
TAACTCCAAAGAGTATCACTCACTACCAAACAGAATGTTTGAGAAGGAAATGACGCTCAAACAGGCATG
CCCCCTGGAATACCAAGGGGCGCAATGTGCGTTCAAAGATTCGATGATTCACGGAATTCTGCAATTCACA
TTACGTATCGCATTTCGCTGCGTTCTTCATCGATGCGAGAACCAAGAGATCCGTTGTTGAAAGTTTTTAATA
TTTTAAAATTTCCAGTTACGAAAATTCTTGTTTTTGACAAAAATTTAATGAATAGATAAAATTGTTTGTGTTT
GTTACCTCTGGGCCCCGATTGCTCGAATGCCCAAAGAAAAAGTTGCAAAGATATGAAAACTCCACAGTG
TGTTGTATTGAAACGGTTTTAATTGTCCTATAACAAAAGCACAGAAATCTCTCACCGTTTGGAATAGCAAG
AAAGAAACTTACAAGCCTAGCAAGACCGCGCACTTAAGCGCAGGCCCGGCTGGACTCTCCATCTCTTGT
CTTCTTGCCCAGTAAAAGCTCTCATGCTCTTGCCAAAACAAAAAAATCCATTTTCAAAATTATTAAATTTCT
TTAATGATCCTTCCGCAGGTTCACCTACGGAAACCTTGTTACGACTTTTAGTTCCTCTAAATGACCAAGTTT
GTCCAAATTCTCCGCTCTGAGATGGAGTTGCCCCCTTCTCTAAGCAGATCCTGAGGCCTCACTAAGCCAT
TCAATCGGTACTAGCGACGGGCGGTGTGTACAAAGGGCAGGGACGTAATCAACGCAAGCTGATGACTT
GCGCTTACTAGGAATTCCTCGTTGAAGAGCAATAATTACAATGCTCTATCCCCAGCACGACGGAGTTTCA
CAAGATTACCAAGACCTCTCGGCCAAGGTTAGACTCGCTGGCTCCGTCAGTGTAGCGCGCGTGCGGCCC
AGAACGTCTAAGGGCATCACAGACCTGTTATTGCCTCAAACTTCCATCGGCTTGAAACCGATAGTCCCTC
TAAGAAGTGGATAACCAGCAAATGCTAGCACCACTATTTAGTAGGTTAAGGTCTCGTTCGTTATCGCAATT
AAGCAGACAAATCACTCCACCAACTAAGAACGGCCATGCACCACCACCCACAAAATCAAGAAAGAGCTC
TCAATCTGTCAATCCTTATTGTGTCTGGACCTGGTGAGTTTCCCCGTGTTGAGTCAAATTAAGCCGCAGGC
TCCACTCCTGGTGGTGCCCTTCCGTCAATTCCTTTAAGTTTCAGCCTTGCGACCATACTCCCCCCAGAACC
CAAAGACTTTGATTTCTCGTAAGGTGCCGAGTGGGTCATTAAAAAAACACCACCCGATCCCTAGTCGGCA
TAGTTTATGGTTAAGACTACGACGGTATCTGATCATCTTCGATCCCCTAACTTTCGTTCTTGATTAATGAAAA
CGTCCTTGGCAAATGCTTTCGCAGTAGTTAGTCTTCAATAAATCCAAGAATTTCACCTCTGACAATTGAATA
CTGATGCCCCCGACCGTCCCTATTAATCATTACGATGGTCCTAGAAACCAACAAAATAGAACCAAACGTC
CTATTCTATTATTCCATGCTAATATATTCGAGCAATACGCCTGCTTTGAACACTCTAATTTTTTCAAAGTAAA
AGTCCTGGTTCGCCAAGAGCCACAAGGACTCAAGGTTAGCCAGAAGGAAAGGCCCCGTTGGAAATCCA
GTACACGAAAAAATCGGACCGGCCAACCGGGCCCAAAGTTCAACTACGAGCTTTTTAACTGCAACAACT
TTAATATACGCTATTGGAGCTGGAATTACCGCGGCTGCTGGCACCAGACTTGCCCTCCAATTGTTCCTCG
TTAAGGTATTTACATTGTACTCATTCCAATTACAAGACCCGAATGGGCCCTGTATCGTTATTTATTGTCACTA
CCTCCCTGAATTAGGATTGGGTAATTTGCGCGCCTGCTGCCTTCCTTGGATGTGGTAGCCGTTTCTCAGG

CTCCCTCTCCGGAATCGAACCCTTATTCCCCGTTACCCGTTGAAACCATGGTAGGCCACTATCCTACCATC
GAAAGTTGATAGGGCAGAAATTTGAATGAACCATCGCCAGCACAAGGCCATGCGATTCGAAAAGTTATTA
TGAATCATCAAAGAGTCCGAAGACATTGATTTTTTATCTAATAAATACATCTCTTCCAAAGGGTCGAGATTT
TAAGCATGTATTAGCTCTAGAATTACCACAGTTATACCATGTAGTAAAGGAACTATCAAATAAACGATAACT
GATTTAATGAGCCATTCGCAGTTTCACTGTATAAATTGCTTATACTTAGACATGCATGGCTTAATCTTTGAG
ACAAGCATATGACTACTGGCAGGATCAACCAGATAACTATCTTAAAAGAAGAAGCAACAAGCAGTAAAAA
AGAAAGAAACCGAAATCTCTTTTTTTTTTTCCCACCTATTCCCTCTTGCTAGAAGATACTTATTGAGTTTGG
AAACAGCTGAAATTCCAGAAAAATTGCTTTTTCAGGTCTCTCTGCTGCCGGAAATGCTCTCTGTTCAAAA
AGCTTTTACACTCTTGACCAGCGCACTCCGTCACCATACCATAGCACTCTTTGAGTTTCCTCTAATCAGGTT
CCACCAAACAGATACCCCGGTGTTTCACGGAATGGTACGTTTGATATCGCTGATTTGAGAGGAGGTTACA
CTTGAAGAATCACAGTCTTGCGACCGGCTATTCAACAAGGCATTCCCCCAAGTTTGAATTCTTTGAAATA
GATTGCTATTAGCTAGTAATCCACCAAATCCTTCGCTGCTCACCAATGGAATCGCAAGATGCCCACGATG
AGACTGTTCAGGTTAAACGCAAAAGAAACACACTCTGGGAATTTCTTCCCAAATTGTATCTCTCAATACGC
ATCAACCCATGTCAATTAAACACGCTGTATAGAGACTAGGCAGATCTGACGATCACCTAGCGACTCTCTCC
ACCGTTTGACGAGGCCATTTACAAAAACATAACGAACGACAAGCCTACTCGAATTCGTTTCCAAACTCTT
TTCGAACTTGTCTTCAACTGCTTTCGCATGAAGTACCTCCCAACTACTTTTCCTCACACTTGTACTCCATGA
CTAAACCCCCCCTCCCATTACAAACTAAAATCTTACTTTTATTTTCTTTTGCCCTCTCTGTCGCTCTGCCTTA
ACTACGTATTTCTCGCCGAGAAAAACTTCAATTTAAGCTATTCTCCAAAAATCTTAGCGTATATTTTTTTTCC
AAAGTGACAGGTGCCCCGGGTAACCCAGTTCCTCACTATTTTTTACTGCGGAAGCGGAAGCGGAAAATA
CGGAAACGCGCGGGAACATACAAAACATACAAAATATACCTTTCTCACACAAGAAATATATGCTACTTGC
AAAATATCATACCAAAAAACTTTTCACAACCGAAACCAAAACCAACGGATATCATACATTACACTACCACC
ATTCAAACTTTACTACTATCCTCCCTTCAGTTTCCCTTTTTCTGCCTTTTTCGGTGACGGAAATACGCTTCA
GAGACCCTAAAGGGAAATCCATGCCATAACAGGAAAGTAACATCCCAATGCGGACTATACCACCCCACC
ACACTCCTACCAATAACGGTAACTATTCTATGTTTTCTTACTCCTATGTCTATTCATCTTTCATCTGACTACCT
AATACTATGCAAAAATGTAAAATCATCACACAAAACATAAACAATCAAAATCAGCCATTTCCGCACCTTTTC
CTCTGTCCACTTTCAACCGTCCCTCCAAATGTAAAATGGCCTATCGGAATACATTTTCTACATCCTAACTAC
TATAAAACAACCTTTAGACTTACGTTTGCTACTCTCATGGTCTCAATACTGCCGCCGACATTCTGTCCCACA
TACTAAATCTCTTCCCGTCATTATCGCCCGCATCCGGTGCCGTAAATGCAAAACAAATACCATCTATGTCTT
CCACACCATCATTTTACTATGCCTGCCACCATCCATTTGTCTTTTGCACCATATCTTCATAACCTGTCACCTT
GAAACTACCTCTGCATGCCACCTACCGACCAACTTTCATGTTCTGTTTCGACCTACCTCTTGTAAATGACA
AATCACCTTTTTCATCGTATGCACCTTATTCTCCACATCACAATGCACTATTGCTTTTGCTTTTTCACCTGTC
ATATCCTATTGCTATTAGATGAAATATAATAAAAATTGTCCTCCACCCATAACACCTCTCACTCCCACCTACT
GAACATGTCTGGACCCTGCCCTCATATCACCTGCGTTTCCGTTAAACTATCGGTTGCGGCCATATCTACCA
GAAAGCACCGTTTCCCGTCCGATCAACTGTAGTTAAGCTGGTAAGAGCCTGACCGAGTAGTGTAGTGGG
TGACCATACGCGAAACTCAGGTGCTGCA