Trie benchmark results

File	Size in MB	Number of strings
bible.txt	3.841	30,105
world192.txt	1.826	32,650
dickens	9.345	157,038
mozilla	1.387	21,579
mr	0.043	1,045
nci	10.594	178,289
ooffice	0.020	599
osdb	0.032	890
reymont	6.103	6,058
samba	10.274	247,372
sao	0.942	8,282
webster	30.598	491,661
xml	2.375	26,657
x-ray	3.199	23,172
sources	104.845	2,507,311
pitches	6.090	25
proteins	651.355	2,265,632
dna	383.493	1,864
english	988.750	16,709,728
dblp.xml	164.885	2,949,908
coreutils	10.746	230,982
einstein.de.txt	0.900	2,828
einstein.en.txt	1.958	5,298
kernel	5.983	154,025
world-leaders	6.144	5,276

Table 1: Statistics about the trie test data used for experiments. The files were generated as follows: first, all lines containing nullbytes or containing less than 10 characters were removed from the original file. Next, lines containing any other line as proper substring were removed. Resulting files with less than 1 MB or less than 1000 lines were filtered out. The lines of the resulting remaining files then were used as input strings for trie construction.

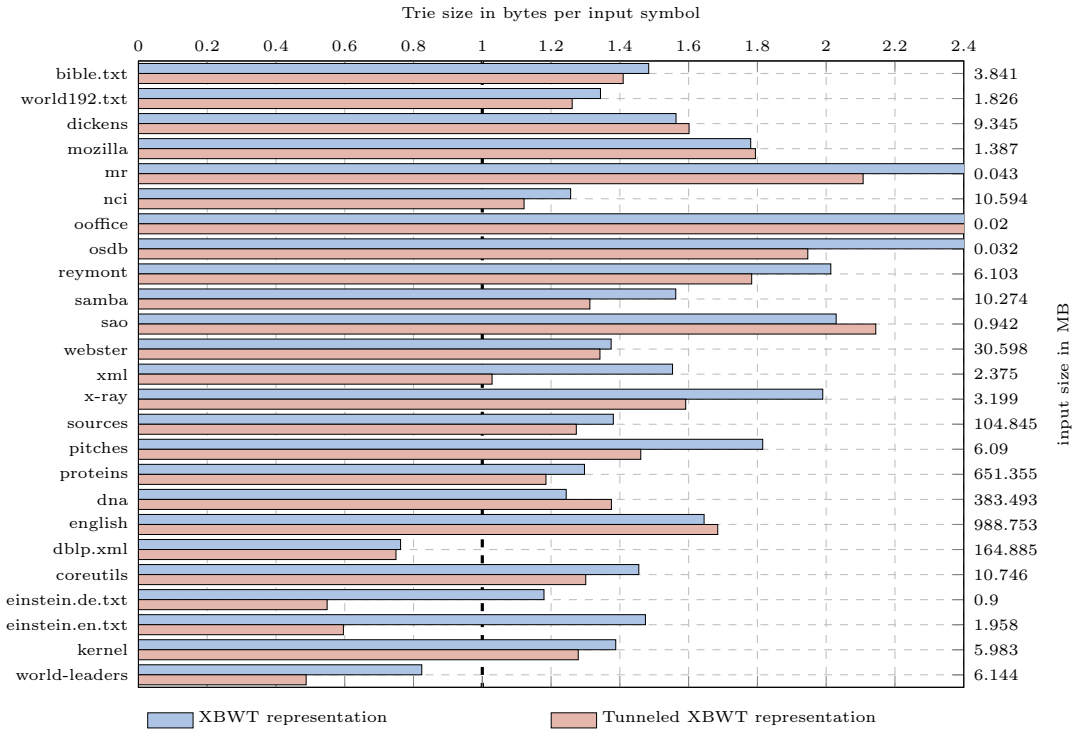

Figure 1: Size of trie representations measured relative to the size of the sum of lengths of all input strings.

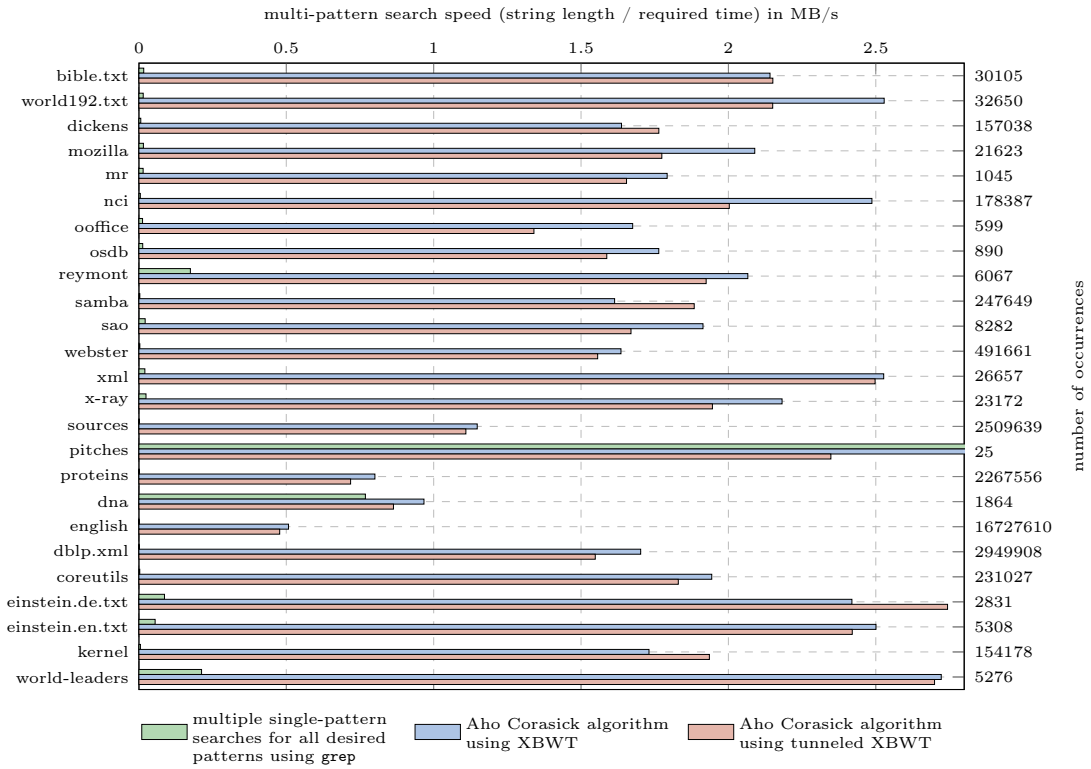

Figure 2: Speed of the multi-pattern search using the Aho Corasick algorithm compared to multiple single-pattern searches for all patterns using `grep`. The Aho Corasick algorithm uses a priori constructed tries of the test files and searches for all occurrences of the test file lines within the same test file. The `grep` speed is estimated by executing `grep` with the last pattern of each file and multiplying the required time by the number of overall patterns.

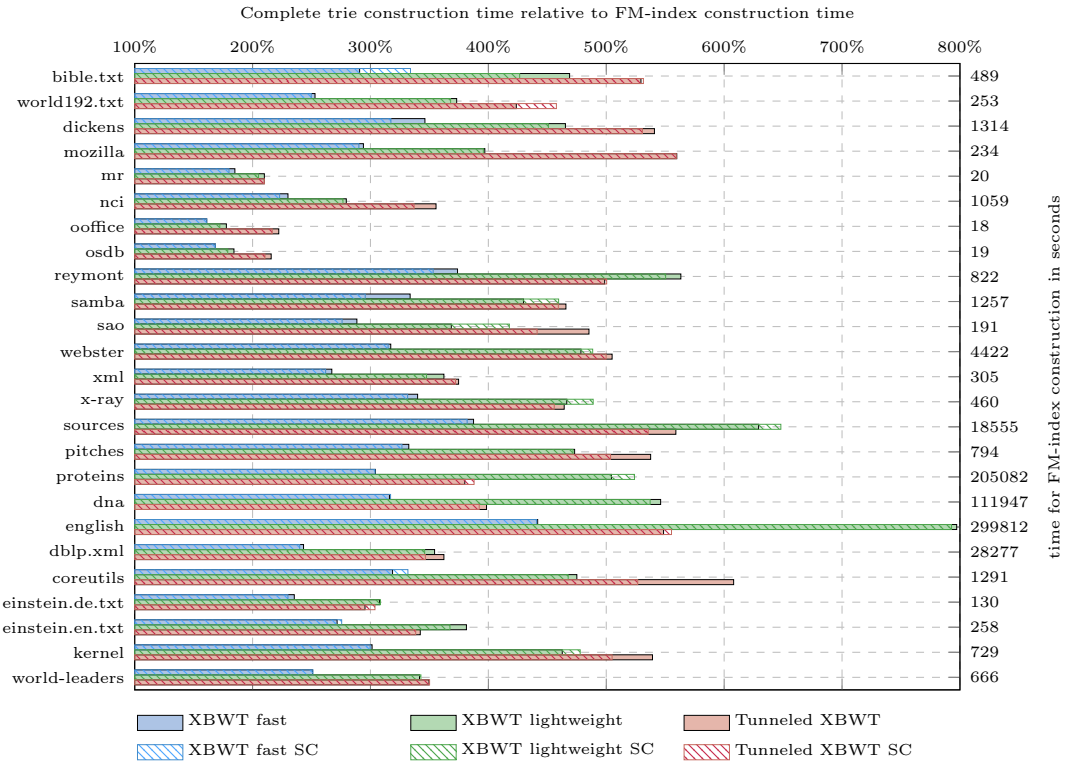

Figure 3: Trie construction timings of different algorithms relative to FM-index construction. Algorithm variants using a succinct counter instead of a normal counter are indicated by the same color and a north west line pattern.

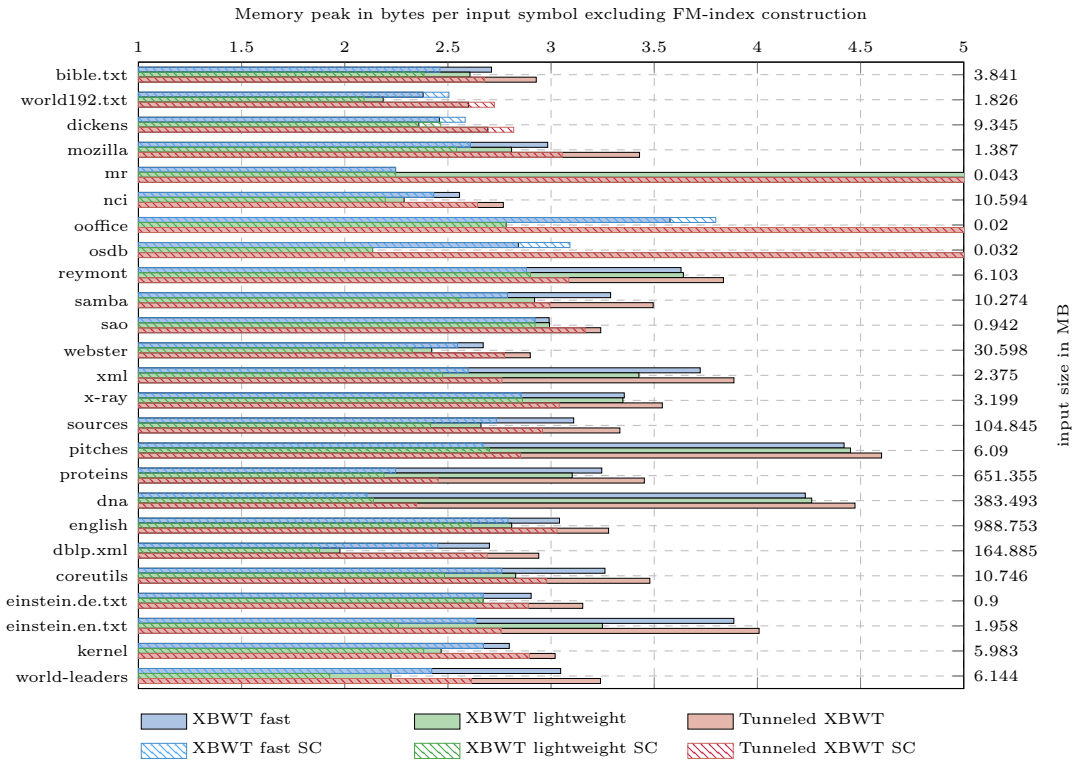

Figure 4: Memory peak during trie construction and excluding FM-index construction. The peak is measured in bytes per symbol of the input data. Algorithm variants using a succinct counter instead of a normal counter are indicated by the same color and a north west line pattern.