Untitled

O. Denas

November 20, 2016

Report

Testing space usage on subsets of the `proteins.50MB` dataset from Pizza&Chilli. All tests were done on a prefix of the dataset of length |t| + |s|, for |t| = 1000 and various values of |s|.

Space usage

All figures are in bytes.

TODO:

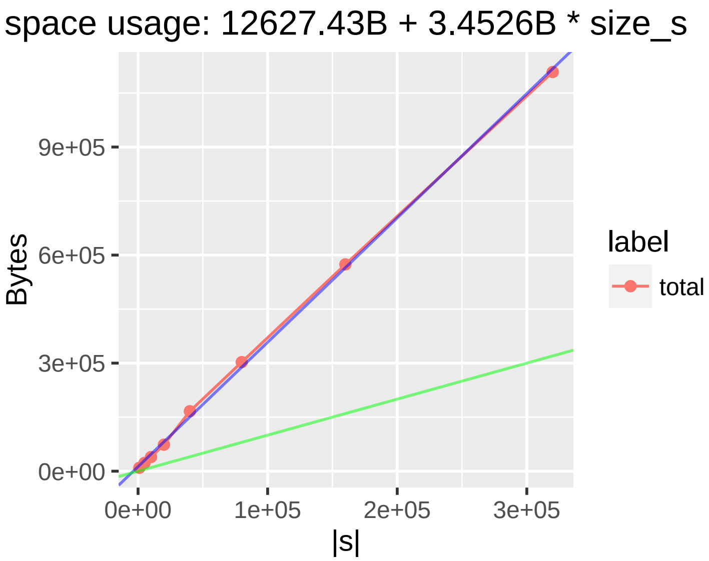
- the string `s` is not needed once the bwt is built
- the `bwt` is not needed once its wavelet tree is built

The algorithm allocates space for

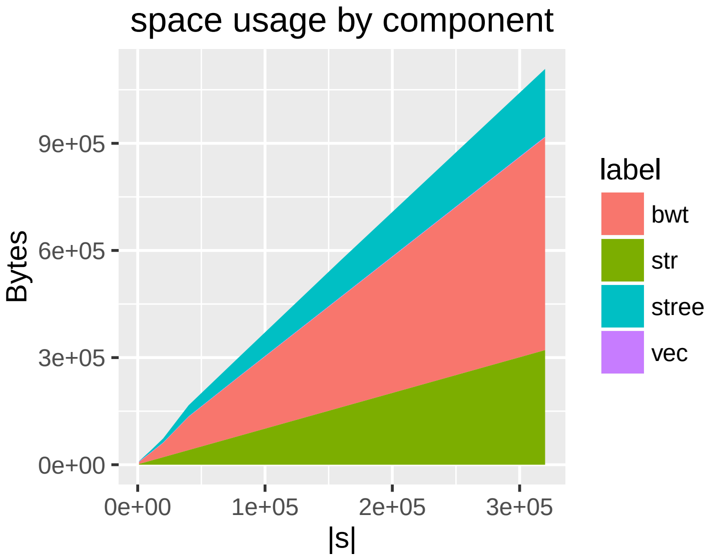
- **bwt**, the BWT index.
  - `bwt_wtree` the wavelet tree,
  - `alphabet, C` arrays of fixed size, `bwt` the actual BWT .
- **str**, the input strings `t` and `s`, `t` is fixed for all experiments `|t| =` 1000
- **stree**, the suffix tree topology
  - `stree_bp` balanced parenthesis bit vector,
  - `stree_bp_supp` navigational support for the bp vector,
  - rank and select support data structures for `stree_bp_supp`.
- **vec**, the following vectors
  - `runs, ms` fixed for all the experiments. `|runs| =` 158 and `|ms| =` 348
  - `runs rank/select` support, `ms select` support.

Data structures, **bwt** and **stree** data structures are built for `s` and its reverse, but space for `s` is released before the space for its reverse is used.

Total space usage. The blue (green) line is the linear fit (identity line).

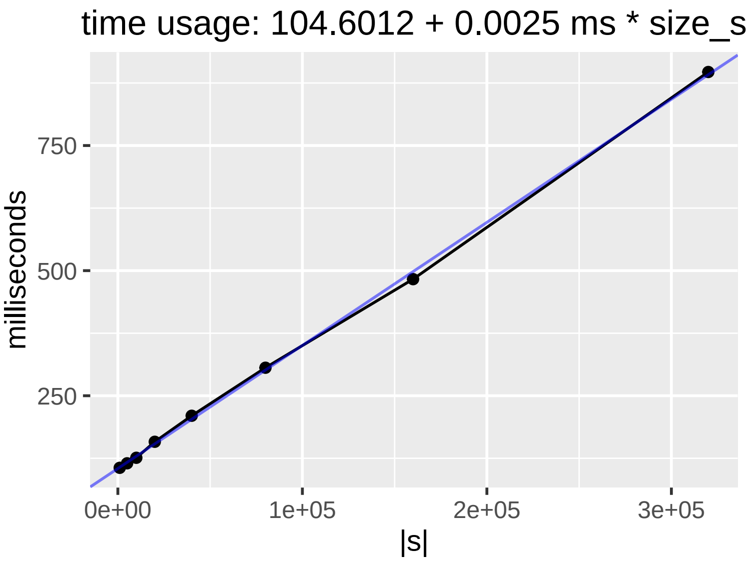


Total space usage by component



Time usage

Figures are in milliseconds. The blue line shows the linear fit.


time usage: 104.6012 + 0.0025 ms * size_s
milliseconds
750
500
250
0e+00
1e+05
2e+05
3e+05
|s|