HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
MATEMAATTIS-LUONNONTIETEELLINEN TIEDEKUNTA
MATEMATISK-NATURVETENSKAPLIGA FAKULTETEN
FACULTY OF SCIENCE

# String Sorting in Python – Comparison of Several Algorithms

Onni Koskinen, Arturs Polis, and Lari Rasku

All algorithms were written from scratch, striving for idiomatic and easily understandable Python code over low-level or implementation-specific optimizations whenever possible. Empirical measurements on the performance of these algorithms were made.

Here we try to explore the relative performance of different algorithms and analyze the reasons behind strengths and weaknesses of the algorithms used.

## DATA SET

The timing test data consisted of the PROTEINS, DNA and ENGLISH datasets from the Pizza&Chili Corpus, in addition to a set of URLs from Ranjan Sinha's ref1 data ref2 for his original Burstsort paper.

A 100MB and a 200MB sample of each dataset was used. The ENGLISH datasets were not used as-is, but with each word split on its own line, in order to make the algorithms sort individual words and not entire lines. The statistics file documents some stringological properties of these datasets.

ref1 https://sites.google.com/site/ranjansinha/home
ref2 http://www.cs.mu.oz.au/ rsinha/resources/data/sort.data.zip

| Dataset | Number of strings | Sum of lengths | Max string length | alphabet size | Sum of LCP array |
|---|---|---|---|---|---|
| dna.100MB | 618 | 104856983 | 3732300 | 15 | 4501 |
| dna.200MB | 1114 | 209714087 | 3732300 | 15 | 8948 |
| proteins.100MB | 359505 | 104498096 | 36805 | 24 | 18853436 |
| proteins.200MB | 709116 | 209006085 | 36805 | 24 | 50076184 |
| urls.100MB | 3284368 | 101569109 | 372 | 114 | 94113004 |
| urls.200MB | 6576059 | 203139142 | 560 | 114 | 191545831 |
| words.100MB | 18502734 | 85200064 | 112 | 211 | 83643408 |
| words.200MB | 37003241 | 170395992 | 112 | 220 | 168115390 |

## ALGORITHMS

### MSD RADIX SORT

MSD Radix sort text MSD Radix sort textblock MSD Radix sort text MSD Radix sort text MSD Radix sort text MSD Radix sort text MSD Radix sort text block MSD Radix sort text

### QUICKSORT ALGORITHMS

Quicksort text Quicksort text Quicksort text Quicksort text Quicksort text Quicksort text Quicksort text Quicksort text Quicksort text Quicksort text Quicksort text Quicksort text Quicksort text Quicksort text Quicksort text

### BURST SORT

Burst sort text Burst sort text Burst sort text Burst sort text Burst sort text Burst sort text Burst sort text Burst sort text Burst sort text Burst sort text Burst sort text Burst sort text Burst sort text

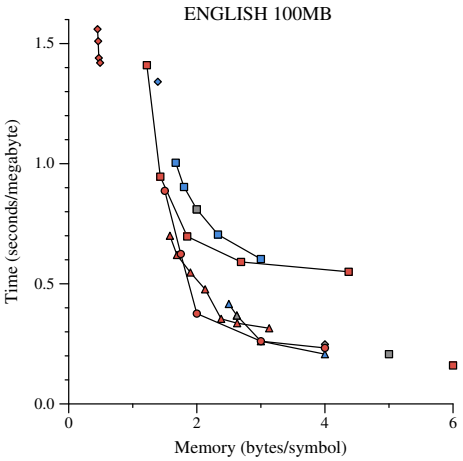
## EXPERIMENTAL RESULTS

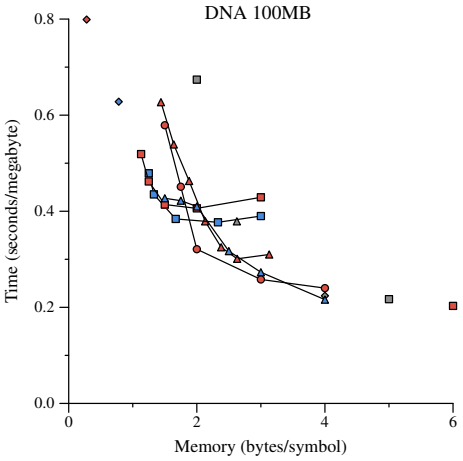
The graphs below show the time and space requirements of several algorithms on two texts. The algorithms are divided into three groups:

**New** algorithms based on reference point ranks, repetition shortcuts and wavelet trees

**Improved** implementations of wavelet trees and algorithms from [?]

**Prior** algorithms from [?, ?]




## REFERENCES

[1] U. Lauther and T. Lukovszki. Space efficient algorithms for the Burrows-Wheeler backtransformation. In *Proc. 13th Annual European Symposium on Algorithms*, volume 3669 of *LNCS*, pages 293–304. Springer, 2005.

[2] J. Seward. Space-time tradeoffs in the inverse B-W transform. In *Proc. IEEE Data Compression Conference*, pages 439–448. IEEE, 2001.

[3] U. Lauther and T. Lukovszki. Space efficient algorithms for the Burrows-Wheeler backtransformation. In *Proc. 13th Annual European Symposium on Algorithms*, volume 3669 of *LNCS*, pages 293–304. Springer, 2005.

[4] J. Seward. Space-time tradeoffs in the inverse B-W transform. In *Proc. IEEE Data Compression Conference*, pages 439–448. IEEE, 2001.

[5] U. Lauther and T. Lukovszki. Space efficient algorithms for the Burrows-Wheeler backtransformation. In *Proc. 13th Annual European Symposium on Algorithms*, volume 3669 of *LNCS*, pages 293–304. Springer, 2005.

[6] J. Seward. Space-time tradeoffs in the inverse B-W transform. In *Proc. IEEE Data Compression Conference*, pages 439–448. IEEE, 2001.