Collection	Characters	Documents	Avg. doc. len.	gzip-compr.	xz-compr.
ENWIKI-BIG	8,945,231,276	3,903,703	2,291.47	37.68	25.19
ENWIKI-SML	68,210,334	4,390	15,537.66	36.60	26.15
PROTEINS	58,959,815	143,244	411.60	52.24	11.31

Table 1: Statistics of the character based collections.

Identifier	sdsl type
GREEDY	`doc_list_index_greedy<>`
QPROBING	`doc_list_index_qprobing<>`
SADA	`doc_list_index_sada<>`

Table 2: Class definition of character indexes used in the experiment.

Collection	Index size in MiB (fraction of original collection)		
	GREEDY	QPROBING	SADA
ENWIKI-BIG	27,042.76 (3.17)	27,042.76 (3.17)	23,913.72 (2.80)
ENWIKI-SML	130.49 (2.01)	130.49 (2.01)	199.61 (3.07)
PROTEINS	161.67 (2.87)	161.67 (2.87)	147.92 (2.62)

Table 3: Size of character indexes.

Collection	Words	Documents	Avg. doc. len.	gzip-compr.	xz-compr.
ENWIKI-BIG-INT	1,690,724,944	3,903,703	433.11	63.13	50.66
ENWIKI-SML-INT	12,741,343	4,390	2,902.36	71.75	62.88

Table 4: Statistics of the word based collections.

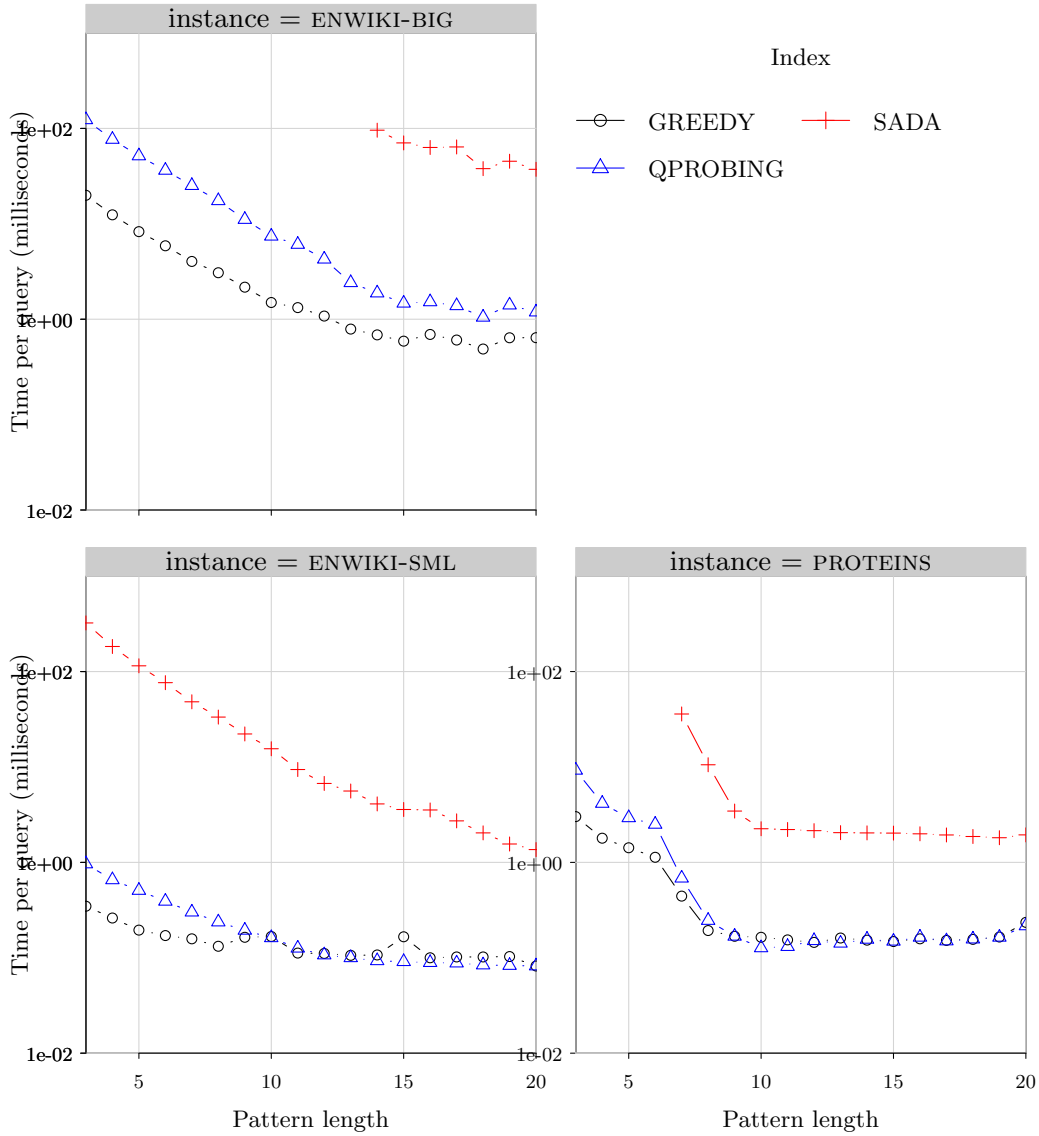

Figure 1: Average query time to find the top-10 documents (TFxIDF measure) for different pattern length using character based indexes. For each query length, 200 pattern were queried.

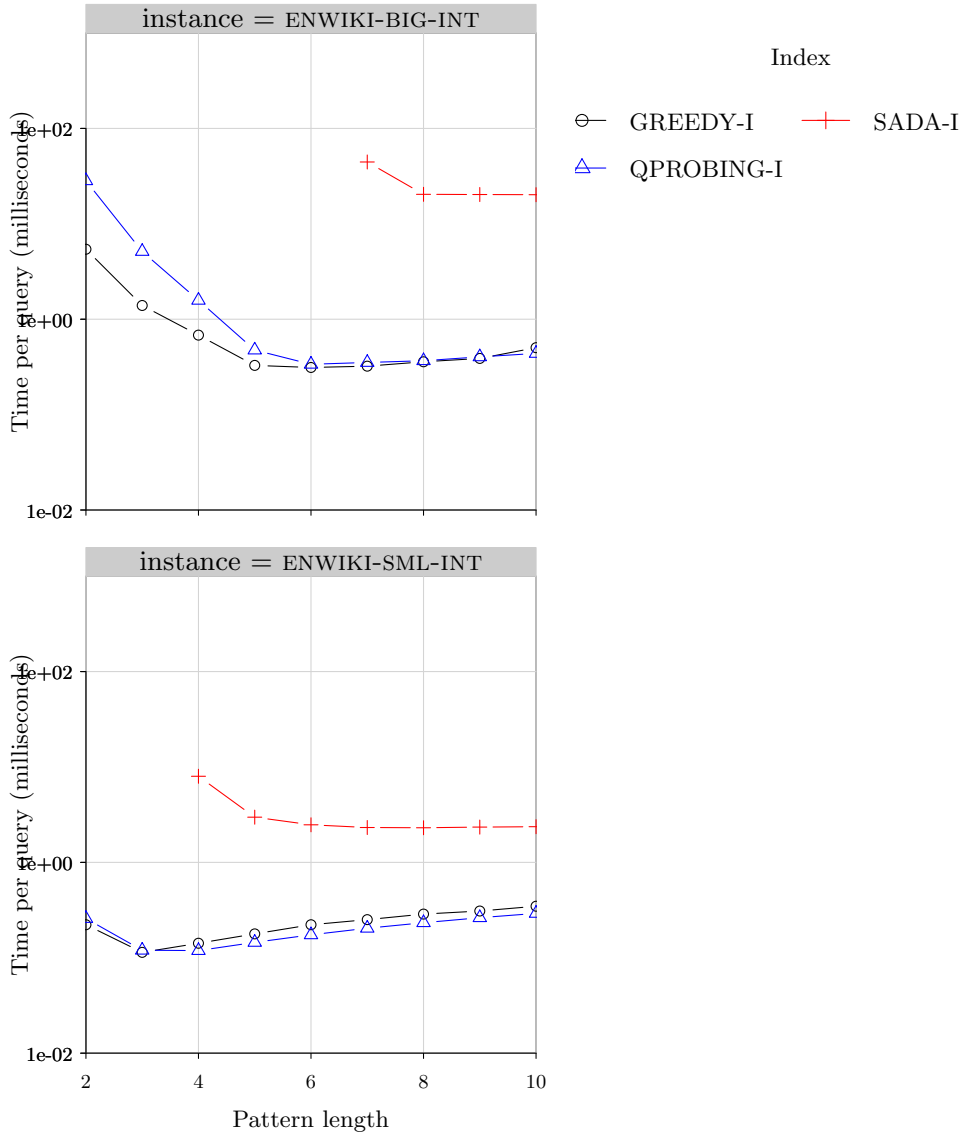

Figure 2: Average query time to find the top-10 documents (TFxIDF measure) for different pattern length using word bases indexes. For each query length, 200 pattern were queried.

Identifier	sdsl type
GREEDY-I	`doc_list_index_greedy<csa_wt<wt_int<rrr_vector<63>>, 1000000, 1000000>>`
QPROBING-I	`doc_list_index_qprobing<csa_wt<wt_int<rrr_vector<63>>, 1000000, 1000000>>`
SADA-I	`doc_list_index_sada<csa_wt<wt_int<rrr_vector<63>>, 30, 1000000>>`

Table 5: Class definition of word indexes used in the experiment.

Collection	Index size in MiB (fraction of original collection)		
	GREEDY-I	QPROBING-I	SADA-I
ENWIKI-BIG-INT	6,786.43 (1.46)	6,786.43 (1.46)	5,471.17 (1.18)
ENWIKI-SML-INT	38.05 (1.32)	38.05 (1.32)	45.29 (1.57)

Table 6: Size of word indexes.