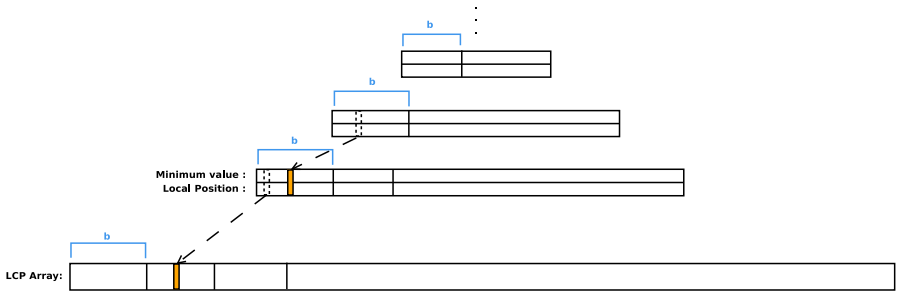


Figure 1: NPR original data structure

1 NPR CST-CN

In the original version (R. Canovas and G. Navarro. "Practical Compressed Suffix Trees". In Proc. SEA, 94–105, 2010.), the LCP array was divided into blocks of length b, and the minimum LCP within each block and its local position were stored. The values stored used $\frac{n}{b} \cdot \log n$ bits (the m array) and its positions $\frac{n}{L} \cdot \log b$ bits (the p array). Then a hierarchy of blocks was formed. On top of array m, a perfect b-ary tree was constructed where the leaves are the elements of the m and each internal node stores the minimum of the values stored in its children and its local position. The total space needed is $\frac{n}{b} \cdot (\log n + \log b) \cdot (1 + O(1/b))$ bits, so if $b = \omega(\log n)$, the space used is $o(n)$ bits. Figure 1 illustrates the NPR structure. In general the query time for NSV, PSV, and RMQ would depend of the value of b and the LCP structure used. In the worse case each operation would take $2 \cdot b \cdot t_{LCP} + O(1)$, where t_{LCP} is the time to extract an LCP value and the $O(1)$ indicates the time used to move in the b-ary tree.

The main problem of this structure is that the query time depends on the LCP structure used, being desirable to reduce the access to the LCP as much as possible without heavily increasing the space requirement. In order to achieve that, we propose to use smaller size blocks (sb) for the first two level, only storing for the first level the local position where the minimum value is within each block, and for upper levels store the minimum value and its position but using a bigger block size (b) for the other levels. The total space required is $\frac{n}{sb} \cdot \log sb \cdot (1 + \frac{1}{sb}) + \frac{n}{sb^2} \cdot \log n + \frac{n}{b * sb^2} \cdot (\log n + \log b) \cdot (1 + O(1/b))$ bits, and the query time per operation in the worse case is $4 \cdot sb \cdot t_{LCP} + O(1)$.

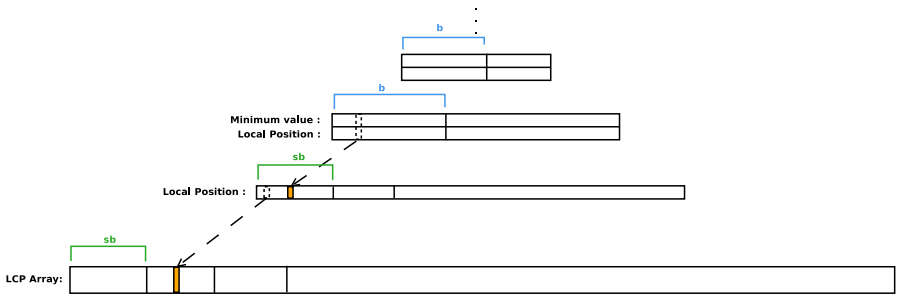


Figure 2: NPR new version

2 Experiments

Table 1 lists the test files used and their stats. For each of the compressed suffix tree (SADA, SCT3 and CN) used, it was required to select a compressed suffix array (CSA) representation, a longest common prefix (LCP) structure and a way to represent the topology of the tree (balanced parenthesis (BP) or NPR support).

In this work we used two representation of the CSA, the Sadakane's CSA and a version based on a wavelet tree of the BWT of the original text. In the case of the LCP array, we used the DAC implementation presented in sdsl-lite and the TREE2 representation (which only can be used by trees which stores the topology of the tree using balanced parenthesis). Table 2 shows the space used by each of these components depending on the parameters chosen.

Figures 3 to 10 presents the performance for each of the possible combinations of the components in their respective CST representation. For each SADA and SCT3 alternative we presented two points being the left most point the implementation using LCP-TREE2 and the right most point using LCP-DAC.

To test we followed the experimental protocols of (Gog's thesis). Given a node v, for measuring the time used by the operations $parent(v)$, $depth(v)$, $first\_child(v)$, $sibling(v)$, $node\_depth(v)$, and $child(v, c)$ (where c is a character obtained from a random positions in the text) we took 100000 random leaves of the CST and for each random leaf we add all nodes of the path from v to the root to the sampled nodes to test.

For the operation $Slink(v)$ we took 100000 random leaves of the CST and for each parent v of a random leaf we traverse its suffix link path to the

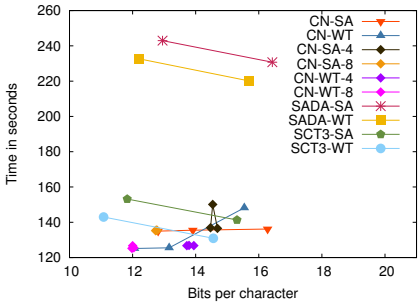
File Name	Size (Megabytes)	σ
`XML`	200.00	97
`English`	200.00	239
`Proteins`	200.00	27
`DNA`	200.00	16
`Sources`	200.00	230

Table 1: Text files used in the experiments.

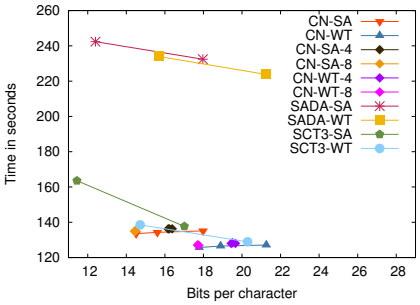
root, adding all the nodes visited to the sampled node to test. Finally for the $lca(v, w)$ operation we randomly choose 100000 pairs of leaf to query.

	Description	Size in bits per character				
		XML	English	Proteins	DNA	Sources
CSA-SA	Sadakane's CSA with sampling density $sa = 32$ and $isa = 64$	3.60	4.81	6.54	5.33	4.25
CSA-WT	CSA based on a Wavelet Tree of the BWT of the text (same sample as CSA-SA)	9.23	8.09	7.60	4.58	9.46
LCP-DAC	block size = 4	9.07	8.60	7.85	6.38	8.18
LCP-TREE2	density = 16	3.44	3.03	3.60	2.90	4.21
BP		4.20	4.56	4.51	4.72	4.58
NPR-SCT3		3.60	3.60	3.60	3.60	3.60
NPR-CN						
	block size = 8	4.57	4.57	4.57	4.57	4.57
	block size = 16	2.20	2.20	2.20	2.20	2.20
	block size = 32	1.10	1.10	1.10	1.10	1.10
NPR-CNR						
	block size = 8 and $sb = 4$	2.97	2.97	2.97	2.97	2.97
	block size = 16 and $sb = 4$	2.83	2.83	2.83	2.83	2.83
	block size = 32 and $sb = 4$	2.76	2.76	2.76	2.76	2.76
	block size = 8 and $sb = 8$	1.07	1.07	1.07	1.07	1.07
	block size = 16 and $sb = 8$	1.03	1.03	1.03	1.03	1.03
	block size = 32 and $sb = 8$	1.02	1.02	1.02	1.02	1.02

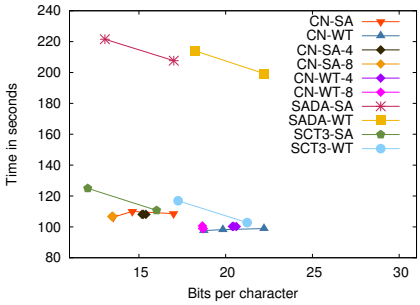
Table 2: Components



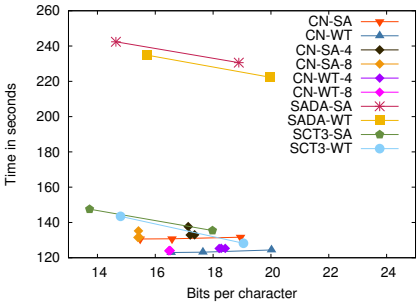
(a) DNA



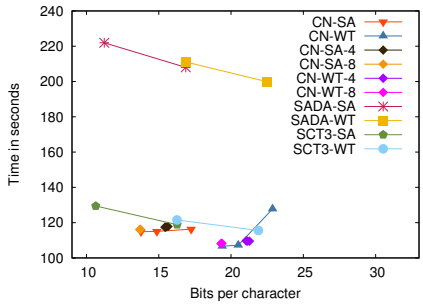
(b) English



(c) Sources

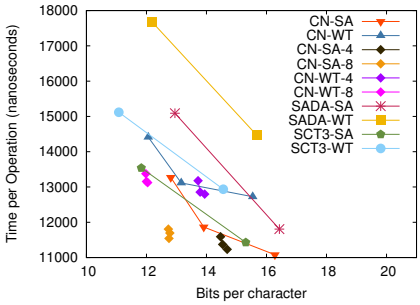


(d) Proteins

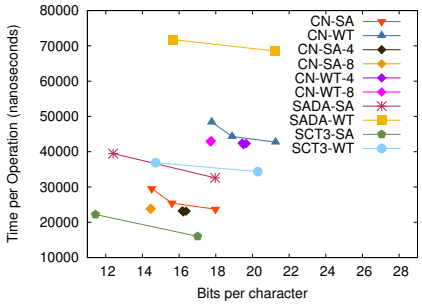


(e) XML

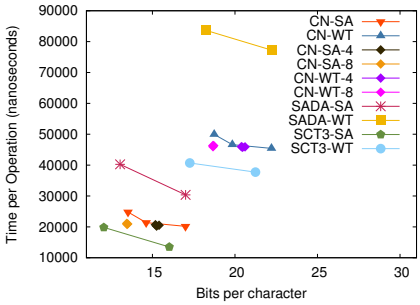
Figure 3: CST construction



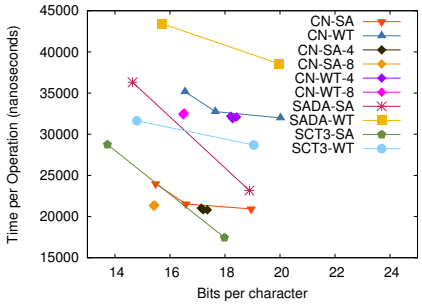
(a) DNA



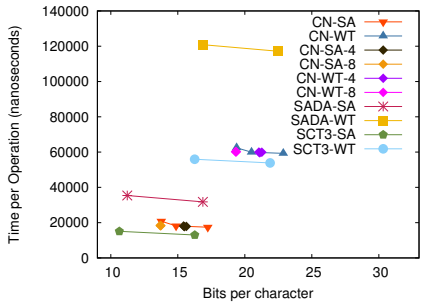
(b) English



(c) Sources

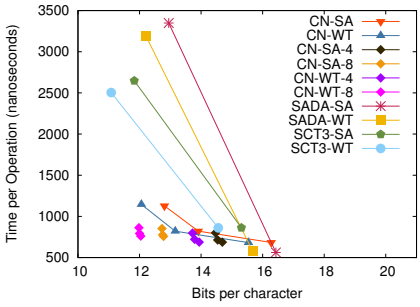


(d) Proteins

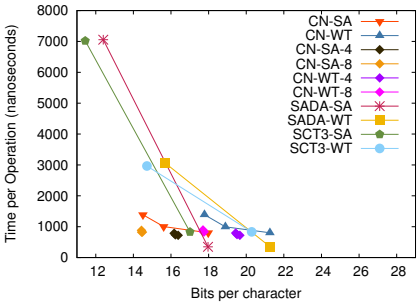


(e) XML

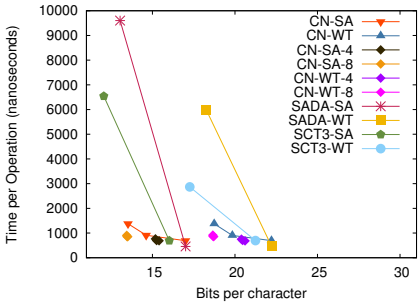
Figure 4: CST Child



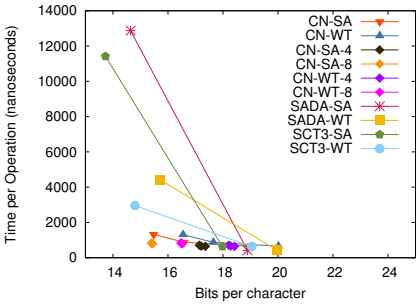
(a) DNA



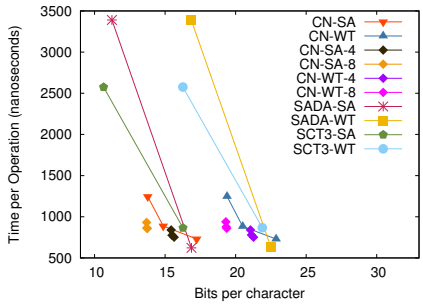
(b) English



(c) Sources

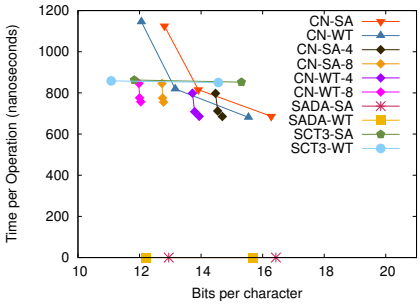


(d) Proteins

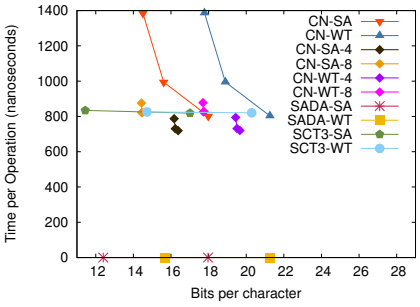


(e) XML

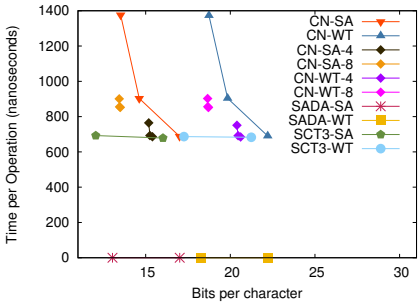
Figure 5: CST Depth



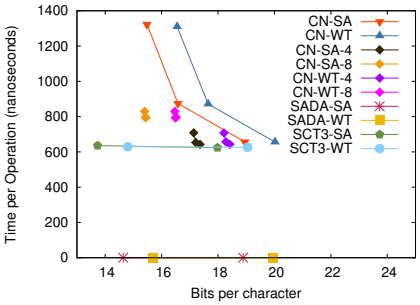
(a) DNA



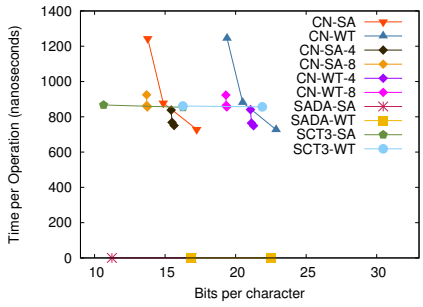
(b) English



(c) Sources

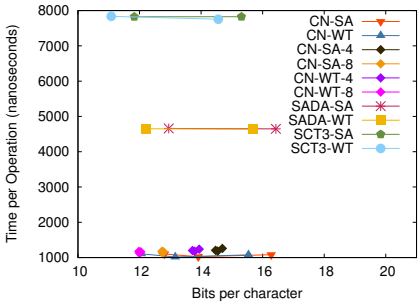


(d) Proteins

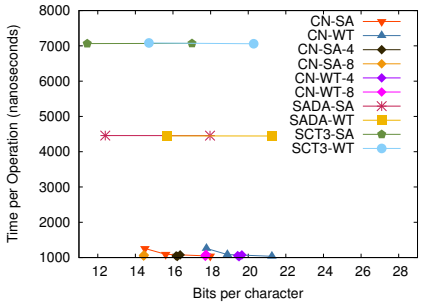


(e) XML

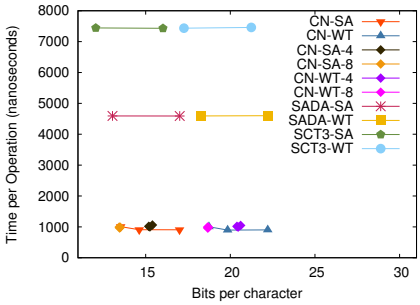
Figure 6: CST First Child



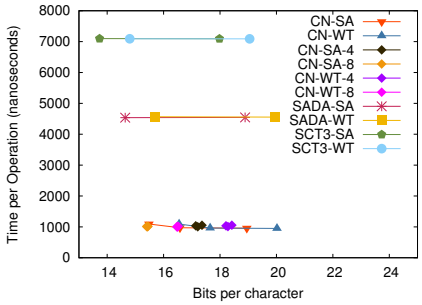
(a) DNA



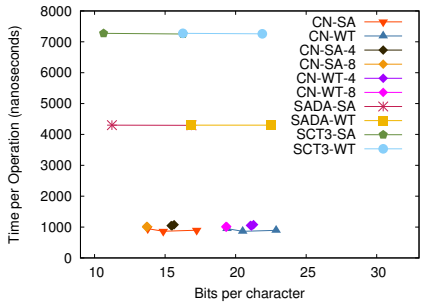
(b) English



(c) Sources

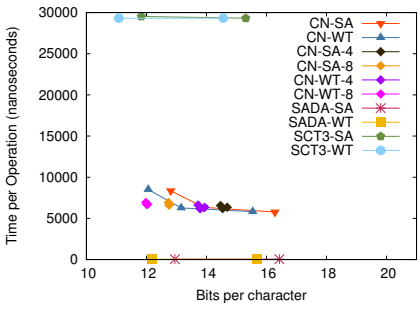


(d) Proteins

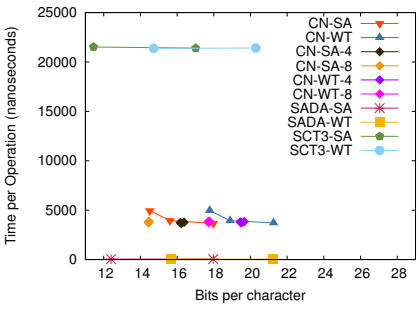


(e) XML

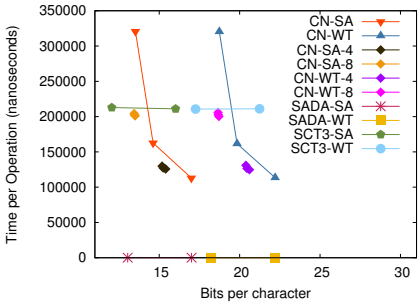
Figure 7: CST LCA



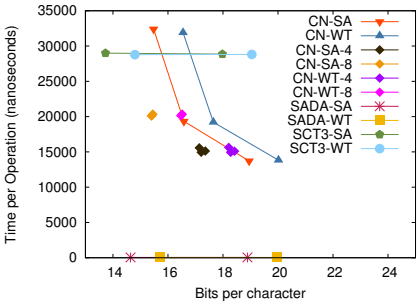
(a) DNA



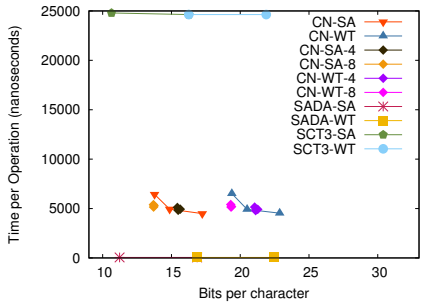
(b) English



(c) Sources

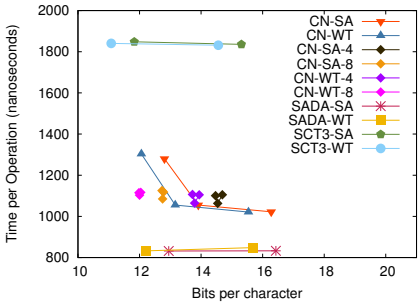


(d) Proteins

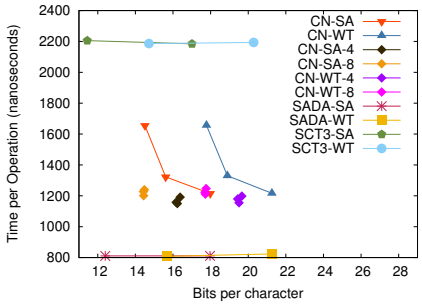


(e) XML

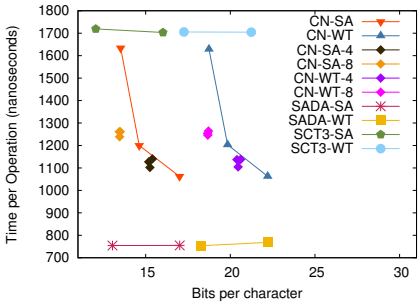
Figure 8: CST Node Depth



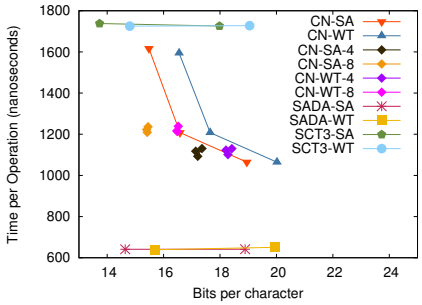
(a) DNA



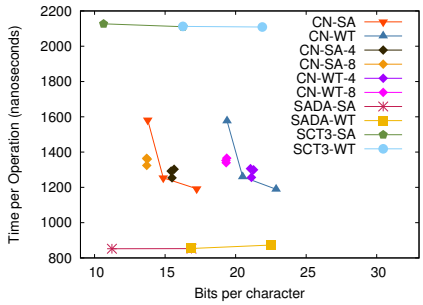
(b) English



(c) Sources

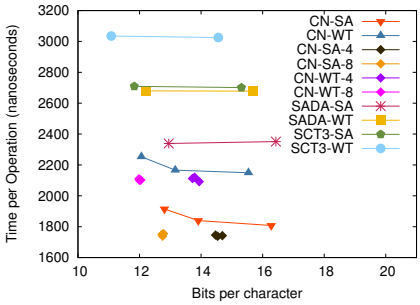


(d) Proteins

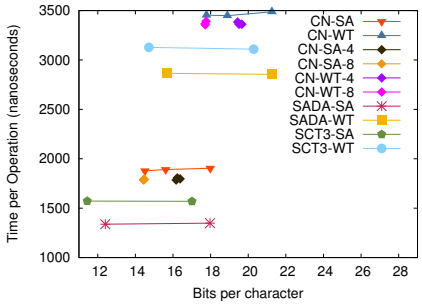


(e) XML

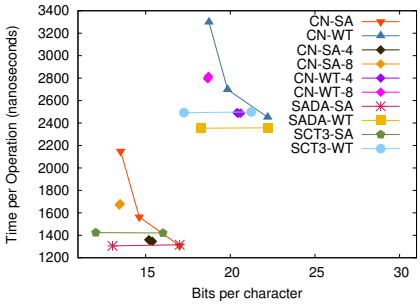
Figure 9: CST Sibling



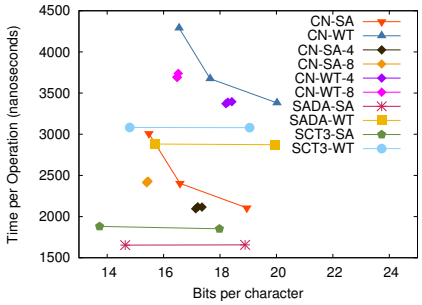
(a) DNA



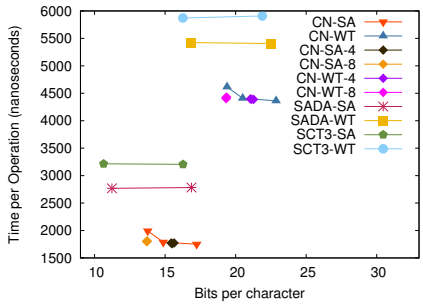
(b) English



(c) Sources



(d) Proteins



(e) XML

Figure 10: CST Suffix Link