

Contents

Introduction	1
Outline and Main Contributions	3
Publications	3
1 Basic Concepts	4
1.1 Empirical Entropy	4
1.2 Arrays	5
1.3 Rank/Select/Access (RSA)	6
1.3.1 Bitvectors	6
1.3.2 Wavelet Trees	7
1.4 Succinct Tree Representations	7
1.5 Range min-Max Tree	9
1.6 Suffix Array and Suffix Tree	11
1.7 Dictionary Compressors	11
1.8 Karp-Rabin Algorithm	13
2 Related Work	14
2.1 RSA-Structures using Grammar Compression	14
2.2 Compressed Suffix Arrays	15
2.3 Compressed Suffix Trees	16
2.3.1 The First Compressed Suffix Tree	16
2.3.2 Fully-Compressed Suffix Tree (FCST)	17
2.3.3 Faster Entropy-Bounded CST	17
2.4 Repetition-Aware Compressed Suffix Trees	18
2.4.1 The First Repetition-Aware CST	18
2.4.2 Grammar-Compressed Suffix Tree (GCST)	19
2.4.3 Other Approaches	19
2.5 Repetitive Datasets	19
2.6 Succinct Data Structure Library	20
3 Block Tree Improvements	22
3.1 Block Trees	22
3.1.1 Rank and Select	24
3.2 Previous Implementation of Block Trees	25
3.2.1 Definition	26
3.2.2 Construction	26

3.2.3	Compact Representation	27
3.3	Improvements	28
3.3.1	No Padding Block Tree	28
3.3.2	Compressed Compact Representation	29
3.3.3	Construction Algorithm	29
3.3.4	Elimination of First Levels	30
3.3.5	Pruning	30
3.3.6	New Sets of Rank/Select Fields	32
3.3.7	New Heuristics	32
3.4	Experimental Results	34
3.4.1	Setup and Datasets	34
3.4.2	Experiments and Structures	34
3.4.3	Comparisons and Analysis	35
4	Compression of Repetitive Differential Arrays	51
4.1	LCSA Adaptation	51
4.2	GCC Adaptation	52
4.3	Block Tree Adaptation	54
4.4	Experimental Results	54
4.4.1	Setup and Datasets	54
4.4.2	Experiments and Structures	55
4.4.3	Results and Analysis	55
5	Block Tree Compressed Topology	60
5.1	Block Tree Augmentation	60
5.1.1	Stored Fields	61
5.1.2	Fields Computed on the Fly	61
5.2	Description of Primitives	62
5.2.1	$leaf-rank(i)$ and $leaf-select(j)$	63
5.2.2	$fwd-search(i, d)$ and $bwd-search(i, d)$	63
5.2.3	$min-excess(i, j)$, $min-count(i, j)$ and $min-select(i, j, t)$	66
5.3	Further Optimizations on Primitives	69
5.3.1	r -ary rmM-tree on First Level	69
5.3.2	BT-CT Versions	70
5.4	Experimental Results	70
5.4.1	Setup and Datasets	70
5.4.2	Experiments and Structures	71
5.4.3	Results and Analysis	71
6	Our Repetition-Aware Compressed Suffix Trees	79
6.1	BT-CT Adaptation	79
6.1.1	Operation child	80
6.2	RLCSA Speed Up	80
6.3	Experimental Results	80
6.3.1	Setup and Datasets	80
6.3.2	Experiments and Structures	81
6.3.3	Results and Analysis	82

Conclusion and Future Work	96
Bibliography	103