

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Thesis statement . . . . .	3
1.2.1	Contributions . . . . .	4
1.3	Structure of the Thesis . . . . .	6
1.4	Software . . . . .	6
1.5	Notation . . . . .	7
<b>2</b>	<b>Basic Concepts</b>	<b>8</b>
2.1	Data Compression . . . . .	8
2.1.1	Entropy . . . . .	8
2.1.2	Encoding Sequences . . . . .	10
2.1.3	Direct Access to Variable-Length Codes . . . . .	12
2.2	Compact Data Structures . . . . .	13
2.2.1	Bit vectors . . . . .	13
2.2.2	Wavelet Trees . . . . .	15
2.2.3	Succinct Trees . . . . .	19
2.3	Hashing . . . . .	22
2.3.1	Hash Tables . . . . .	22
2.3.2	Rolling Hashing . . . . .	25
2.3.3	Bloom Filters . . . . .	26
2.3.4	Document Similarity . . . . .	27
<b>3</b>	<b>Indexing and Compressing Text</b>	<b>29</b>
3.1	Classical Indexes . . . . .	29
3.1.1	Suffix Array . . . . .	30
3.1.2	Suffix Tree . . . . .	31
3.2	Text Compression . . . . .	32
3.2.1	The Burrows-Wheeler Transform . . . . .	33
3.2.2	Grammars . . . . .	36
3.2.3	Other Compression Methods . . . . .	39
3.3	Self-Indexes . . . . .	40
3.3.1	FM-Index . . . . .	41
3.3.2	Bidirectional FM-Index . . . . .	42
3.3.3	The r-index . . . . .	43

3.3.4	The Grammar Index . . . . .	45
3.4	BWT Indexes for Labeled Directed Graphs . . . . .	50
3.4.1	Labeled Tries . . . . .	51
3.4.2	Directed Acyclic Graphs . . . . .	53
3.5	Algorithms for building the SA and the BWT . . . . .	55
3.5.1	Prefix-Free Parsing . . . . .	55
3.5.2	Induced Suffix Sorting . . . . .	57
<b>4</b>	<b>Computational Genomics</b>	<b>59</b>
4.1	DNA Sequences . . . . .	59
4.2	DNA Sequencing . . . . .	60
4.2.1	Sequencing File Format . . . . .	61
4.3	The de novo Assembly Problem . . . . .	62
4.3.1	The de Bruijn Framework . . . . .	64
4.3.2	The Overlap Graph Framework . . . . .	68
4.4	Reference Genomes . . . . .	69
4.5	Pangenomes . . . . .	70
<b>5</b>	<b>Grammar-Compressed Reads</b>	<b>73</b>
5.1	Motivation . . . . .	73
5.2	Definitions . . . . .	75
5.3	The LMSg Algorithm . . . . .	75
5.3.1	LMSg is for String Collections . . . . .	75
5.3.2	Simplifying the Grammar . . . . .	76
5.3.3	Analysis of LMSg . . . . .	77
5.3.4	Efficient Dictionary Construction . . . . .	78
5.4	Recompressing the Grammar . . . . .	79
5.5	Encoding the Grammar . . . . .	79
5.6	Experiments . . . . .	82
5.7	Results and Discussion . . . . .	83
<b>6</b>	<b>Computing the eBWT</b>	<b>85</b>
6.1	Encoding Information with Circular Strings . . . . .	85
6.2	Definitions . . . . .	87
6.3	Overview of infBWT . . . . .	87
6.4	Reconstructing the Alphabets . . . . .	88
6.4.1	Finding the Nonterminals in the Parse Tree . . . . .	88
6.4.2	Giving Ranks to the Labels . . . . .	89
6.4.3	Time Complexity for the Alphabet Reconstruction . . . . .	91
6.5	Computing the eBWT of the Compressed Text . . . . .	92
6.6	Inducing the eBWT . . . . .	93
6.7	Implicit Occurrences of the LMS Phrases . . . . .	99
6.8	Inducing the BWT in Run-Length Compressed Space . . . . .	100
6.8.1	Practical Considerations of nextBWT . . . . .	101
6.9	Experiments . . . . .	102
6.10	Results and Discussion . . . . .	103

<b>7</b>	<b>An Index for Navigating the Layout of Reads</b>	<b>105</b>
7.1	Definitions . . . . .	106
7.2	The Layout Query . . . . .	107
7.3	Computing Overlaps in a vo-dBG . . . . .	108
7.4	The Overlap Tree and rBOSS . . . . .	111
7.5	Simulating Bidirectionality . . . . .	113
7.6	Implementing the Layout Query . . . . .	115
7.7	The Layout Query and the BWT of the Reads . . . . .	117
7.8	Genome Assembly . . . . .	118
7.9	Experiments . . . . .	120
7.9.1	Space and Construction Time . . . . .	121
7.9.2	Time for the Primitives . . . . .	121
7.9.3	Genome Assembly . . . . .	123
<b>8</b>	<b>Succinct Colored de Bruijn Graphs</b>	<b>125</b>
8.1	Definitions . . . . .	126
8.2	Coloring a dBG of Reads . . . . .	127
8.2.1	Partial Coloring . . . . .	127
8.2.2	Unsafe Coloring . . . . .	127
8.2.3	Safe and Greedy Coloring . . . . .	128
8.2.4	Ambiguous Sequences . . . . .	131
8.3	Compressing the Colored dBG . . . . .	131
8.4	Reconstructing Unambiguous Sequences . . . . .	132
8.5	Assembling Contigs . . . . .	132
8.6	Experiments . . . . .	134
8.7	Results . . . . .	136
<b>9</b>	<b>Practical Locally Consistent Grammar</b>	<b>138</b>
9.1	Definitions . . . . .	140
9.2	A Grammar Self-Index based on LMS Parsing . . . . .	140
9.2.1	LMS parsing . . . . .	140
9.2.2	Computing the cuts during the pattern matching . . . . .	141
9.3	Experiments . . . . .	142
9.4	Results and Discussion . . . . .	143
9.5	Locally Consistent Grammars and Pangenomes . . . . .	145
<b>10</b>	<b>Conclusion and Further Work</b>	<b>147</b>
10.1	Summary of contributions . . . . .	147
10.2	Further Work . . . . .	149
	<b>Bibliography</b>	<b>151</b>