

Contents

Foreword xi

Preface xiii

Acknowledgments xix

About the sketches xxi

I ANALYSIS 1

Chapter 1 Probabilistic Models 3

1.1 Probabilistic models on words 4

1.2 Probabilistic tools 8

1.3 Generating functions and analytic tools 13

1.4 Special functions 16

1.4.1 Euler’s gamma function 16

1.4.2 Riemann’s zeta function 18

1.5 Exercises 19

Bibliographical notes 21

Chapter 2 Exact String Matching 23

2.1 Formulation of the problem 23

2.2 Language representation 25

2.3 Generating functions 28

2.4 Moments 31

2.5 Limit laws 33

2.5.1 Pattern count probability for small values of r 34

2.5.2 Central limit laws 35

2.5.3 Large deviations 37

2.5.4 Poisson laws 40

viii	Contents
2.6	Waiting times 43
2.7	Exercises 44
	Bibliographical notes 45
Chapter 3	Constrained Exact String Matching 47
3.1	Enumeration of (d, k) sequences 48
3.1.1	Languages and generating functions 51
3.2	Moments 54
3.3	The probability count 57
3.4	Central limit law 60
3.5	Large deviations 61
3.6	Some extensions 70
3.7	Application: significant signals in neural data 70
3.8	Exercises 73
	Bibliographical notes 73
Chapter 4	Generalized String Matching 75
4.1	String matching over a reduced set 76
4.2	Generalized string matching via automata 82
4.3	Generalized string matching via a language approach 93
4.3.1	Symbolic inclusion–exclusion principle 94
4.3.2	Multivariate generating function 96
4.3.3	Generating function of a cluster 98
4.3.4	Moments and covariance 103
4.4	Exercises 106
	Bibliographical notes 107
Chapter 5	Subsequence String Matching 109
5.1	Problem formulation 110
5.2	Mean and variance analysis 112
5.3	Autocorrelation polynomial revisited 118
5.4	Central limit laws 118
5.5	Limit laws for fully constrained pattern 122
5.6	Generalized subsequence problem 123
5.7	Exercises 129
	Bibliographical notes 132

II APPLICATIONS 133

Chapter 6 Algorithms and Data Structures 135

6.1 Tries 136

6.2 Suffix trees 140

6.3 Lempel–Ziv’77 scheme 143

6.4 Digital search tree 144

6.5 Parsing trees and Lempel–Ziv’78 algorithm 147

Bibliographical notes 153

Chapter 7 Digital Trees 155

7.1 Digital tree shape parameters 156

7.2 Moments 159

7.2.1 Average path length in a trie by Rice’s method 159

7.2.2 Average size of a trie 168

7.2.3 Average depth in a DST by Rice’s method 169

7.2.4 Multiplicity parameter by Mellin transform 176

7.2.5 Increasing domains 183

7.3 Limiting distributions 185

7.3.1 Depth in a trie 185

7.3.2 Depth in a digital search tree 189

7.3.3 Asymptotic distribution of the multiplicity parameter 193

7.4 Average profile of tries 197

7.5 Exercises 210

Bibliographical notes 216

Chapter 8 Suffix Trees and Lempel–Ziv’77 219

8.1 Random tries resemble suffix trees 222

8.1.1 Proof of Theorem 8.1.2 226

8.1.2 Suffix trees and finite suffix trees are equivalent 239

8.2 Size of suffix tree 243

8.3 Lempel–Ziv’77 248

8.4 Exercises 265

Bibliographical notes 266

Chapter 9 Lempel–Ziv’78 Compression Algorithm 267

9.1 Description of the algorithm 269

9.2 Number of phrases and redundancy of LZ’78 271

9.3 From Lempel–Ziv to digital search tree 279

x	Contents
9.3.1	Moments 282
9.3.2	Distributional analysis 285
9.3.3	Large deviation results 287
9.4	Proofs of Theorems 9.2.1 and 9.2.2 290
9.4.1	Large deviations: proof of Theorem 9.2.2 290
9.4.2	Central limit theorem: Proof of Theorem 9.2.1 292
9.4.3	Some technical results 294
9.4.4	Proof of Theorem 9.4.2 300
9.5	Exercises 307
	Bibliographical notes 309
Chapter 10	String Complexity 311
10.1	Introduction to string complexity 312
10.1.1	String self-complexity 313
10.1.2	Joint string complexity 313
10.2	Analysis of string self-complexity 314
10.3	Analysis of the joint complexity 315
10.3.1	Independent joint complexity 315
10.3.2	Key property 316
10.3.3	Recurrence and generating functions 317
10.3.4	Double depoissonization 318
10.4	Average joint complexity for identical sources 322
10.5	Average joint complexity for nonidentical sources 323
10.5.1	The kernel and its properties 323
10.5.2	Main results 326
10.5.3	Proof of (10.18) for one symmetric source 329
10.5.4	Finishing the proof of Theorem 10.5.6 338
10.6	Joint complexity via suffix trees 340
10.6.1	Joint complexity of two suffix trees for nonidentical sources 341
10.6.2	Joint complexity of two suffix trees for identical sources . 343
10.7	Conclusion and applications 343
10.8	Exercises 344
	Bibliographical notes 346
Bibliography	 347
Index	 363