

УДК 004.912  
ББК 81.112  
К83

**Крошемор М., Лекрок Т., Риттер В.**

К83 Алгоритмы обработки текста: 125 задач с решениями / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2021. – 312 с.: ил.

**ISBN 978-5-97060-952-1**

Сопоставление строк – одна из самых старых тем в теории алгоритмов, но по-прежнему занимает важное место в информатике. За прошедшие 20 лет мы видели технологические прорывы в таких разных приложениях, как информационный поиск и сжатие информации. Эта книга, представляющая собой богатое собрание задач и упражнений по важнейшим вопросам алгоритмов обработки текстов и комбинаторных свойств слов, предлагает студентам и исследователям приятный и прямой путь к изучению и практическому освоению концепций повышенного уровня.

Задачи взяты из многочисленных научных публикаций – как уже ставших классическими, так и сравнительно новых. Начав с основ, авторы рассматривают все более сложные задачи по комбинаторным свойствам слов (включая слова Фибоначчи и Туэ–Морса), поиску строк в тексте (включая алгоритмы Кнута–Морриса–Пратта и Бойера–Мура), эффективным структурам данных для представления текстов (включая суффиксные деревья и суффиксные массивы) и сжатия текста (включая методы Хаффмана, Лемпеля–Зива и Барроуза–Уилера).

Издание будет полезно в качестве пособия для подготовки к олимпиадам по информатике.

УДК 004.912  
ББК 81.112

Copyright Original English language edition published by Cambridge University Press is part of the University of Cambridge. Russian language edition copyright © 2021 by DMK Press. All rights reserved.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

ISBN 978-1-108-83583-1 (англ.)

ISBN 978-5-97060-952-1 (рус.)

© Maxime Crochemore, Thierry Lecroq, Wojciech Rytter, 2021  
© Перевод, оформление, издание, ДМК Пресс, 2021

# Содержание

<b>От издательства</b> .....	9
<b>Предисловие</b> .....	10
<b>Глава 1. Первые понятия стрингологии</b> .....	12
Слова .....	13
Периодичность .....	14
Регулярные структуры .....	16
Упорядочение .....	17
Примечательные слова .....	18
Автоматы .....	20
Префиксные деревья .....	22
Суффиксные структуры .....	22
Суффиксный массив .....	23
Сжатие .....	24
Соглашения о псевдокоде алгоритмов .....	25
Примечания .....	26
<b>Глава 2. Комбинаторные задачи</b> .....	27
1. Стрингологическое доказательство малой теоремы Ферма .....	28
2. Простой случай проверки однозначности декодирования .....	29
3. Магические квадраты и слово Туэ–Морса .....	30
4. Последовательность Ольденбургера–Колакоски .....	31
5. Бесквадратная игра .....	34
6. Слова Фибоначчи и фибоначчиева система счисления .....	36
7. Игра Витхоффа и слово Фибоначчи .....	38
8. Различные периодические слова .....	39
9. Вариации на тему слова Туэ–Морса .....	41
10. Слова Туэ–Морса и суммы степеней .....	42
11. Сопряженные слова и ротации слов .....	43
12. Сопряженные палиндромы .....	45
13. Много слов с большим числом палиндромов .....	47
14. Короткое суперслово перестановок .....	48
15. Короткая суперпоследовательность перестановок .....	50
16. Слова Сколема .....	52
17. Слова Лэнгфорда .....	54
18. От слов Линдона к словам де Брёйна .....	56
<b>Глава 3. Сопоставление с образцом</b> .....	59
19. Таблица границ .....	60
20. Кратчайшие покрытия .....	62

21. Короткие границы.....	64
22. Таблица префиксов .....	65
23. От таблицы границ к максимальному суффиксу .....	67
24. Тест периодичности.....	69
25. Строгие границы .....	71
26. Задержка последовательного сопоставления строк .....	73
27. Разреженный автомат сопоставления.....	75
28. Сопоставление со строкой, эффективное с точки зрения числа сравнений.....	77
29. Таблица строгих границ слова Фибоначчи .....	79
30. Слова с подстановочными переменными.....	81
31. Образцы, сохраняющие порядок.....	83
32. Параметрическое сопоставление .....	85
33. Таблица хороших суффиксов .....	87
34. Худший случай в алгоритме Бойера–Мура .....	89
35. Алгоритм Turbo-ВМ .....	91
36. Сопоставление с образцом при наличии универсального символа .....	93
37. Циклическая эквивалентность.....	94
38. Простое вычисление максимального суффикса.....	96
39. Самомаксимальные слова .....	98
40. Максимальный суффикс и его период.....	100
41. Критическая позиция в слове .....	102
42. Периоды префиксов слов Линдона .....	105
43. Поиск слов Зимина .....	107
44. Поиск нерегулярных двумерных образцов .....	109

## **Глава 4. Эффективные структуры данных..... 110**

45. Списковый алгоритм для кратчайшего покрытия.....	111
46. Вычисление наибольших общих префиксов.....	112
47. От суффиксного массива к суффиксному дереву .....	114
48. Линейное суффиксное trie-дерево .....	117
49. Троичное префиксное дерево поиска .....	120
50. Наибольший общий фактор двух слов .....	122
51. Автомат подпоследовательностей .....	124
52. Проверка однозначности декодирования .....	126
53. Таблица LPF.....	128
54. Сортировка суффиксов слов Туэ–Морса.....	131
55. Простое построение суффиксного дерева.....	133
56. Сравнение суффиксов слова Фибоначчи.....	135
57. Устранимость двоичных слов .....	137
58. Устранимость множества слов .....	139
59. Минимальные уникальные факторы .....	141
60. Минимальные отсутствующие слова .....	143
61. Жадная суперстрока .....	146
62. Кратчайшая общая суперстрока коротких слов .....	149
63. Подсчет факторов по длине.....	151
64. Подсчет факторов, покрывающих позицию .....	153

65. Наибольшие факторы с одинаковой четностью.....	154
66. Установление свободы слова от квадратов с помощью словаря базовых факторов .....	155
67. Общие решения факторных уравнений.....	157
68. Поиск в бесконечном слове .....	159
69. Совершенные слова .....	161
70. Плотные двоичные слова .....	165
71. Факторный оракул .....	167

## **Глава 5. Регулярные структуры в словах..... 171**

72. Три квадрата префиксов .....	172
73. Точные границы количества вхождений степеней.....	174
74. Вычисление серий для алфавитов общего вида .....	176
75. Проверка перекрытий в двоичном слове .....	178
76. Игра, свободная от перекрытий.....	180
77. Заякоренные квадраты.....	182
78. Слова, почти свободные от квадратов .....	184
79. Двоичные слова с небольшим числом квадратов .....	186
80. Построение длинных свободных от квадратов слов .....	188
81. Проверка свободы морфизма от квадратов.....	190
82. Число квадратных факторов в помеченных деревьях.....	192
83. Подсчет квадратов в гребнях за линейное время .....	194
84. Кубические серии .....	196
85. Короткий квадрат и локальный период.....	198
86. Число серий .....	200
87. Вычислений серий над отсортированным алфавитом.....	203
88. Периодичность и факторная сложность .....	207
89. Периодичность морфических слов.....	208
90. Простые антистепени.....	210
91. Палиндромическая конкатенация палиндромов .....	211
92. Деревья палиндромов .....	212
93. Неустрашимые образцы.....	214

## **Глава 6. Сжатие текста..... 217**

94. Преобразование Барроуза–Уилера слов Туэ–Морса.....	218
95. Преобразование Барроуза–Уилера сбалансированных слов.....	219
96. Преобразование Барроуза–Уилера на месте.....	223
97. Факторизация Лемпеля–Зива.....	225
98. Декодирование Лемпеля–Зива–Уэлча .....	227
99. Стоимость кода Хаффмана .....	229
100. Кодирование Хаффмана с ограничением на длину.....	232
101. Динамическое кодирование Хаффмана .....	237
102. Кодирование длинами серий .....	239
103. Компактный факторный автомат.....	244
104. Сжатое сопоставление в слове Фибоначчи .....	247
105. Предсказание по частичному совпадению .....	249

106. Сжатие суффиксных массивов .....	251
107. Коэффициент сжатия жадных суперстрок .....	253

## **Глава 7. Разное .....**

108. Двоичные слова Паскаля .....	258
109. Самовоспроизводящиеся слова .....	260
110. Веса факторов .....	261
111. Разности вхождений букв .....	263
112. Факторизация с префиксами, свободными от границ .....	265
113. Тест примитивности для унарных расширений .....	267
114. Частично коммутативные алфавиты .....	269
115. Наибольшее ожерелье фиксированной плотности .....	270
116. Двоичные слова, эквивалентные по периодам .....	272
117. Динамическое генерирование слов де Брёйна .....	275
118. Рекурсивное генерирование слов де Брёйна .....	277
119. Уравнения в словах с заданными длинами переменных .....	279
120. Разнородные факторы над трехбуквенным алфавитом .....	281
121. Наибольшая возрастающая подпоследовательность .....	283
122. Неустранимые множества и слова Линдона .....	285
123. Синхронизация слов .....	287
124. Сейфооткрывающие слова .....	289
125. Суперслова укороченных перестановок .....	293

## **Литература .....**

## **Предметный указатель .....**