МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ им. А. М. ГОРЬКОГО

А. М. Шур

КОМБИНАТОРИКА СЛОВ

Утверждено
редакционно-издательским советом университета
в качестве учебного пособия по курсу
«Комбинаторика слов»

Екатеринбург
Издательство Уральского университета
2003

УДК 519.1(075.8)
Ш967

Р е ц е н з е н т ы:

кафедра вычислительных методов и уравнений математической физики радиотехнического факультета Уральского государственного технического университета — УПИ (заведующий кафедрой доктор физико-математических наук, профессор П. С. Мартышко);

В. И. Трофимов, доктор физико-математических наук (Институт математики и механики УрО РАН)

Шур А. М.

Ш967 Комбинаторика слов: Учеб. пособие. – Екатеринбург: Изд-во Урал. ун-та, 2003. – 96 с.

ISBN 5-7996-0168-8

Пособие представляет собой первый учебник по комбинаторике слов на русском языке. Рассматриваются комбинаторные проблемы, связанные с понятиями «периодичность» и «избегаемость». Изложение опирается на базовый курс алгебры и дискретной математики.

Адресовано аспирантам и студентам, специализирующимся в дискретной математике, компьютерной математике, теоретической информатике.

УДК 519.1(075.8)

ISBN 5-7996-0168-8

ВВЕДЕНИЕ

Для чего нужно изучать слова?

Символьные последовательности, или *слова*, представляют собой наиболее популярный объект комбинаторных исследований в последние десятилетия. Откуда же поступает «заказ» на анализ свойств таких последовательностей? Можно, хотя и достаточно условно, выделить «три источника» комбинаторики слов.

Один источник — это математика, в первую очередь алгебра. Приведем пример: пусть S — полугруппа с порождающим множеством X. Тогда каждый ее элемент s есть произведение элементов из X: $s = x_1x_2\ldots x_n$, т. е. элемент полугруппы представлен словом. Каким должно быть слово, чтобы представлять элемент s? Каковы самые короткие слова, представляющие s? Представляют ли два заданных слова один и тот же элемент полугруппы (*проблема равенства слов*)? Подобных вопросов, ответы на которые очень важны в алгебраических исследованиях, немало.

Кроме алгебры словами интересуются и другие разделы математики. Упомянем лишь один из них — символическую динамику, занимающуюся «грубым» анализом поведения динамических систем. Пусть материальная точка перемещается в некоторой области, а мы измеряем ее координаты (дискретно, с определенным временны́м интервалом). Разобьем область перемещения точки на конечное число частей и каждой части поставим в соответствие некоторый символ. Тогда, заменив результат каждого измерения на символ соответствующей ему части разбиения, мы преобразуем таблицу результатов измерений в слово. Это слово называется *символической траекторией* точки, и его анализ дает богатую информацию о характере движения.

Второй источник — это компьютерные науки. Языки программирования и трансляторы к ним, кодирование информа-

ции, сжатие и восстановление данных, псевдослучайные последовательности, криптография — вот далеко не полный перечень направлений теоретической информатики, в которых необходим анализ последовательностей символов. На стыке алгебры и информатики появились теория формальных языков и теория автоматов, тесно связанные с изучением свойств слов. Поясняющие примеры для второго источника излишни.

В третий источник можно объединить все прочие науки (и даже шире — сферы, требующие использования математических методов) и привести три достаточно непохожих примера. Первый, очевидный, пример — лингвистика; эта наука, в частности, анализирует не только слова — последовательности букв, но и предложения — «слова» в алфавите частей речи. Второй, менее очевидный, но, безусловно, очень актуальный пример — молекулярная биология. Как известно из школьного курса биологии, генетическая информация «записана» в виде молекул ДНК, представляющих собой очень длинные (от тысяч до сотен миллионов элементов) цепочки-«слова» в «алфавите» из четырех «символов» — нуклеотидов. Расшифровка способов кодирования, хранения, восстановления, считывания и записи генетической информации представляет собой важнейшую (но не единственную!) комбинаторную задачу молекулярной биологии.

Наконец, пример задачи, успешно решенной методами комбинаторики слов, можно взять из шахмат. Последнее значительное изменение в правила шахмат было внесено после того, как в конце 1930-х гг. голландский математик, чемпион мира по шахматам Макс Эйве доказал, что существует бесконечная шахматная партия, в которой никакая циклическая последовательность позиций не повторяется. Тем самым было доказано, что правило, фиксирующее ничью в партии при «повторении ходов», недостаточно для избежания возможности бесконечной игры. С подачи Эйве было принято следующее дополнительное правило: ничья фиксируется, если в течение последних 50 ходов в партии не было ни одного взятия либо хода пешкой.

К сожалению, в рамках пособия невозможно охватить все разделы комбинаторики слов, разросшейся в последние десятилетия до самостоятельной математической дисциплины. Поэтому здесь будут рассмотрены два важнейших свойства слов — периодичность и избегаемость, а также некоторые известные комбинаторные проблемы, связанные с этими свойствами.

Надеемся, что читатель, осиливший эту книгу, получит представление о том, чем и как занимается комбинаторика слов. А читателю, который хочет получить более глубокие и систематические знания по этой дисциплине, рекомендуем ознакомиться с фундаментальными трудами [14] и [8].

Определения и обозначения

1. Отправной точкой в комбинаторике слов является понятие *алфавита* — «базового» множества (как правило, конечного, в некоторых случаях — счетного), элементы которого называются *символами* или *буквами*. Символы будем обозначать буквами a, b, c, d, x, y, z (с индексами), а алфавиты — буквами Σ, Δ и A. Конечная последовательность букв называется *словом*. Слова мы обозначаем буквами $P, Q, R, S, T, U, V, W, X, Y, Z$ (с индексами). Символ λ будет использоваться для специального слова — *пустого*, т. е. не содержащего букв. Буквы в непустом слове будем нумеровать числами от 1 до n; таким образом, на слово можно смотреть как на функцию из начального отрезка натурального ряда в алфавит. Порядковый номер буквы в слове называется *позицией*. Обозначение $W(i)$ указывает на i-ю букву слова W. Количество букв в слове W называется *длиной* этого слова и обозначается $|W|$. Множество букв, входящих в слово W, обозначается $alph(W)$.

2. Множество всех непустых слов над фиксированным алфавитом Σ обозначается через Σ^+; для множества $\Sigma^+ \cup \{\lambda\}$ используется обозначение Σ^*. Произвольное подмножество из Σ^+ или Σ^* называется *формальным языком* или просто *язы-*

ком. Для обозначения языков мы используем буквы L, Λ, а также различные двухбуквенные сокращения.

3. На множестве Σ^+ (или Σ^*) можно ввести бинарную операцию *конкатенации* (приписывания) слов: если $U = a_1 \dots a_n$, $V = b_1 \dots b_m$, то $UV = a_1 \dots a_n b_1 \dots b_m$. Часто конкатенацию называют *умножением* слов. Эта операция ассоциативна, следовательно, множество Σ^+ является полугруппой относительно конкатенации, а множество Σ^* — моноидом (с нейтральным элементом λ). Алгебраические системы Σ^+ и Σ^* носят названия *свободной полугруппы над алфавитом* Σ и *свободного моноида над алфавитом* Σ соответственно.

4. *Гомоморфизмом* свободных полугрупп Σ^+ и Δ^+ называется произвольное отображение $f : \Sigma^+ \to \Delta^+$, сохраняющее конкатенацию (т. е. $f(UV) = f(U)f(V)$ для произвольных слов $U, V \in \Sigma^+$). Таким образом, гомоморфизм однозначно определяется своим действием на буквы из Σ. В комбинаторике слов гомоморфизмы свободных полугрупп обычно называют *морфизмами.*

5. Комбинаторика слов изучает не только конечные, но и счетные последовательности символов. Мы будем рассматривать последовательности, индексированные натуральными числами *(ω-слова)*, неположительными целыми числами *(ω^*-слова)*, а также всеми целыми числами *(Z-слова)*. Бесконечные слова будем обозначать теми же буквами, что и обычные, либо выделяя шрифтом, либо используя индекс $_\infty$ (справа для ω-слов, слева для ω^*-слов). Множества бесконечных слов мы называем языками, как и множества конечных слов.

6. Слово V называется *подсловом* слова W (говорят также, что W *содержит* V), если $W = PVQ$ для некоторых слов P, Q. При этом мы говорим, что V является *префиксом* W, если $P = \lambda$, и *суффиксом* W, если $Q = \lambda$. Префикс (суффикс, подслово) называется *собственным*, если он не совпадает с самим словом. Бинарные отношения «быть префиксом», «быть суффиксом» и «быть подсловом» являются отношениями частичного порядка на произвольном множестве слов; они на-

зываются префиксным, суффиксным и подсловным порядком соответственно. Префикс, суффикс и подслово бесконечного слова определяются аналогично. Если V — подслово в W, начинающееся в i-й и заканчивающееся в j-й позиции, то используется запись $V = W(i \dots j)$.

7. Любое представление слова или бесконечного слова в виде произведения подслов мы называем *разбиением* этого слова. Подслова-«сомножители» будем называть *блоками* разбиения.

Структура изложения

Пособие состоит из четырех глав, каждая глава разделена на параграфы (нумерация параграфов сквозная). Предпочтительно читать пособие последовательно, параграф за параграфом; однако, выборочное чтение также допустимо, поскольку многие темы не содержат (или почти не содержат) ссылок на утверждения за пределами темы. Математические утверждения в тексте подразделяются на *теоремы* (утверждения, наиболее важные принципиально), их *следствия*, а также *предложения* (утверждения, имеющие некоторое самостоятельное значение), *леммы* (имеющие только вспомогательное значение) и *замечания* (очевидные или почти очевидные). Каждый вид утверждений нумеруется отдельно. Кроме того, текст содержит *примеры* и *упражнения*, также нумеруемые. Сложные упражнения отмечены звездочкой. Содержащиеся в тексте *примечания* не являются математическими утверждениями и не нумеруются.

Глава 1. ПРЕДВАРИТЕЛЬНЫЕ СВЕДЕНИЯ

§1. Коммутирующие слова. Сопряженные слова

Как уже упоминалось во введении, на множестве слов над данным алфавитом определена ассоциативная операция конкатенации (приписывания) слов, которую мы обычно будем называть *умножением* слов. Понятие *степени* слова вводится естественным образом по отношению к этой операции. Слово, не являющееся степенью никакого другого слова, называется *примитивным*. Если $W = Z^n$, где Z — примитивно и $n \geq 1$, то Z называется *корнем* W.

Предложение 1.1. *Два слова коммутируют тогда и только тогда, когда они имеют общий корень.*

Доказательство. Обратное утверждение очевидно: из того, что $U = Z^n$ и $V = Z^m$, следует $UV = VU = Z^{m+n}$. Докажем прямое утверждение. Предположим противное: пусть существует такая пара слов U, V, что $UV = VU$, но общих корней U и V не имеют. Среди всех пар слов с таким свойством выберем пару, для которой длина UV минимальна. Если $|U| = |V|$, то $U = V$, что противоречит нашему предположению. Пусть без ограничения общности $|U| < |V|$. Тогда $V = UV_1$. Поскольку $UUV_1 = UV_1U$, получаем $UV_1 = V_1U$. Но $|UV_1| < |UV|$, а следовательно, пара коммутирующих слов U, V_1 имеет общий корень Z. Но если $U = Z^n$, $V_1 = Z^m$, то $V = Z^{n+m}$, т. е. Z является корнем V. Противоречие. Доказательство предложения завершено. □

Примечание. Метод доказательства от противного, использованный выше, называется *методом минимального контрпримера* и будет неоднократно применяться в дальнейшем.

Слова U и V называются *сопряженными*, если найдутся такие P и Q, что $U = PQ$, $V = QP$.

Замечание 1.1. Слова U и V являются сопряженными тогда и только тогда, когда V получается циклической перестановкой букв в U.

Замечание 1.2. Отношение сопряженности является отношением эквивалентности на множестве всех слов над данным алфавитом.

Предложение 1.2. *Если $UW = WV$ для некоторых слов U, V и W, то слова U и V являются сопряженными. При этом если $U = PQ$, $V = QP$, то $W = (PQ)^n P$ для некоторого $n \geq 0$.*

Доказательство. Пусть $UW = WV$. Тогда для всякого n

$$U^n W = U^{n-1}(UW) = U^{n-1}WV = \ldots = WV^n.$$

Из равенства $U^n W = WV^n$ следует, что более короткое из слов U^n, W является префиксом более длинного. Выберем n такое, что $|U^n| < |W| \leq |U^{n+1}|$. Тогда существуют $P \neq \lambda$ и Q такие, что $W = U^n P$, $WQ = U^{n+1}$. Отсюда $U = PQ$, $W = (PQ)^n P$ и $V = QP$. Предложение доказано. □

§2. Слова Линдона. Каноническое разбиение слова

Во многих случаях удобно, чтобы множество всех слов над заданным конечным алфавитом было вполне упорядоченным. Наиболее удобным для этой цели является отношение порядка, определяемое следующим образом. Пусть конечный алфавит Σ линейно упорядочен отношением $\leq$. Продолжим это отношение на все множество Σ^+. Для этого положим $U \leq V$, если либо U — префикс V, либо $U = PxQ_1$, $V = PyQ_2$ для некоторых (возможно, пустых) слов P, Q_1, Q_2 и букв x, y таких, что $x < y$. Полученное отношение называется отношением *лексикографического порядка*.

Словом Линдона называется непустое примитивное слово, минимальное лексикографически в своем классе сопряженности.

Замечание 1.3. Все циклические перестановки слова Линдона (и вообще любого примитивного слова) различны. Это немедленно следует из предложения 1.1.

Таким образом, слова Линдона можно рассматривать как «канонические представители» классов сопряженности. Следующая теорема посвящена разложению произвольного слова в произведение таких канонических представителей. Мы приведем доказательство из книги [9].

Теорема 1.1 (Линдон, 1975). *Всякое слово допускает представление в виде произведения слов Линдона, взятых в невозрастающем порядке. Указанное представление единственно.*

Пример 1.1. Для слова $W = bbabaaba$ $(a < b)$ разложение будет таким:

$$W = \mathbf{b}\ \mathbf{b}\ \mathbf{ab}\ \mathbf{aab}\ \mathbf{a}.$$

Лемма 1.1. *Примитивное слово является словом Линдона тогда и только тогда, когда оно лексикографически меньше любого своего непустого собственного суффикса.*

Доказательство. Обратное утверждение доказывается просто: всякий суффикс слова W является префиксом слова, сопряженного с W. Поскольку префикс слова всегда лексикографически меньше самого слова, то слово W, будучи лексикографически меньше всех своих собственных суффиксов, оказывается меньшим всех сопряженных ему слов.

Прямое утверждение докажем от противного. Пусть W — слово Линдона, $W = UV$, $U \neq \lambda$ и $W > V$. Возможны два случая.

1) V не является префиксом слова W. Тогда $V = PxQ_1$, $W = PyQ_2$ и $x < y$. Тогда $VU = PxQ_1U < W$; но VU и W — сопряженные, а значит, W не является словом Линдона, противоречие.

2) $W = VZ$. Тогда $UV = VZ$ и к этим словам применимо предложение 1.2. Получаем $U = PQ$, $Z = QP$, $V = (PQ)^nP$, $W = (PQ)^{n+1}P$. Поскольку W — слово Линдона, имеем

$$W = (PQ)^{n+1}P = P(QP)^{n+1} < (QP)^{n+1}P,$$

откуда $(PQ)^{n+1} < (QP)^{n+1}$ и, следовательно,

$$P(PQ)^{n+1} < P(QP)^{n+1} = W,$$

т. е. W не может быть словом Линдона. Данное противоречие завершает доказательство леммы. □

Лемма 1.2. *Если U и W — слова Линдона и $U < W$, то UW также является словом Линдона и $U < UW < W$.*

Доказательство. Неравенство $U < UW$ очевидно, докажем, что $UW < W$. Поскольку $U < W$, возможны два случая.
1) $W = UQ$. Тогда $W < Q$ по лемме 1.1, откуда

$$UW = UUQ < UQ = W.$$

2) $U = PxQ_1$, $W = PyQ_2$, $x < y$. Тогда

$$UW = PxQ_1W < PyQ_2 = W.$$

Теперь покажем, что UW — слово Линдона. Пусть V — собственный суффикс UW. Если V является собственным суффиксом W, то по лемме 1.1 $W < V$. Если же $V = PW$, то P — суффикс U, т. е. $U < P$ по лемме 1.1, откуда $UW < PW = V$. В итоге UW меньше любого своего собственного суффикса, а значит, является словом Линдона по лемме 1.1. □

Доказательство теоремы Линдона. Поскольку все слова длины 1 являются словами Линдона, всякое слово может быть представлено как произведение слов Линдона. Пусть k — минимально возможное число сомножителей в разложении слова W в произведение слов Линдона, $U_1, \ldots, U_k$ — слова Линдона и $W = U_1 \ldots U_k$. Тогда если для некоторого i выполняется $U_i < U_{i+1}$, то по лемме 1.2 U_iU_{i+1} есть слово Линдона, что невозможно в силу выбора k. Значит, $U_i \geq U_{i+1}$ для всех i; тем самым доказано, что слово можно представить в виде произведения слов Линдона, взятых в невозрастающем порядке. Осталось доказать, что такое представление единственно.

Используем метод минимального контрпримера. Пусть W — минимальное по длине слово, допускающее не единственное

разложение в произведение слов Линдона, взятых в невозрастающем порядке, а $W = U_1 \dots U_k$ и $W = V_1 \dots V_m$ — два таких разложения W. Предположим, что $U_1 \neq V_1$ (без ограничения общности $|U_1| > |V_1|$). Тогда $U_1 = V_1 \dots V_p V'$, где V' — собственный префикс V_{p+1}. Пусть $V' \neq \lambda$. Тогда, пользуясь определениями лексикографического порядка, слов $V_1, \dots V_m$ и леммой 1.1, получаем

$$U_1 < V' < V_{p+1} \leq V_1 < U_1,$$

противоречие. Если же $V' = \lambda$, то аналогично получаем

$$U_1 < V_p \leq V_1 < U_1,$$

снова приходя к противоречию. Таким образом, мы показали, что $U_1 = V_1$. Но тогда слово $W' = U_2 \dots U_k = V_2 \dots V_m$ допускает два различных разложения, противоречие с минимальностью W. Теорема доказана. $\square$

Глава 2. ПЕРИОДИЧЕСКИЕ СЛОВА

Пусть W — произвольное слово. Натуральное число $p \leq |W|$ называется *периодом* слова W, если $W(i) = W(i+p)$ для всех $i = 1, \ldots, |W|-p$. Очевидно, что длина слова является его периодом; период слова будем называть *нетривиальным*, если он меньше длины этого слова. Минимальный период слова будем обозначать через $per(W)$.

Замечание 2.1. Если слово имеет нетривиальный период, то его можно представить в виде UVU, $U \neq \lambda$ (другими словами, некоторый его префикс равен некоторому его суффиксу).

Пример 2.1. Слово *abaabaa* имеет нетривиальные периоды 3 и 6. Слово *aaabaaa* имеет нетривиальные периоды 4, 5 и 6.

Упражнение 2.1. Доказать, что слова Линдона не имеют нетривиальных периодов.

§3. Свойство взаимодействия периодов

Как показывает следующая теорема, всякое достаточно длинное слово с двумя заданными периодами имеет третий, меньший период, равный наибольшему общему делителю заданных. Это свойство периодических слов мы называем *свойством взаимодействия периодов.*

Примечание. Далее мы везде предполагаем, что ни один из заданных периодов не делит другой (и, в частности, не равен 1), чтобы не рассматривать тривиальный случай.

Теорема 2.1 (Файн, Вильф, 1965). *Если слово U имеет периоды p и q и $|U| \geq p + q - НОД(p, q)$, то U имеет также период $НОД(p, q)$.*

Данная теорема была впервые сформулирована не для слов, а для периодических функций, однако оказалась необычайно полезным инструментом именно в комбинаторике слов. Мы приведем одно из комбинаторных доказательств, опирающееся на графы бинарных отношений.

Доказательство. Вначале докажем утверждение теоремы для взаимно простых периодов. Пусть без ограничения общности $p > q$. Для слова U с периодами p и q рассмотрим симметричное бинарное отношение $\rho = \{(i,j) |\ j \in \{i{+}q, i{-}q, i{+}p, i{-}p\}\}$ на множестве $\{1, \ldots, |U|\}$ (т. е. на множестве позиций слова U). Через ρ^* обозначим рефлексивно-транзитивное замыкание отношения ρ. Тогда, очевидно, $(i,j) \in \rho^*$ влечет $U(i) = U(j)$, а если $(i,j) \notin \rho^*$, то в позициях i и j в слове U могут находиться как одинаковые, так и разные буквы.

С другой стороны, условие $(i,j) \in \rho^*$ выполнено тогда и только тогда, когда вершины i и j находятся в одной компоненте связности графа отношения ρ. В итоге если граф отношения ρ для некоторого слова связен, то слово имеет период 1 (т. е. все буквы в нем одинаковы). Если же граф отношения ρ имеет r компонент связности, то количество различных букв в слове может варьироваться от 1 до r.

Пусть $|U| = p{+}q$. Тогда в графе отношения ρ все вершины будут иметь степень 2: вершинами графа являются натуральные числа от 1 до $p{+}q$, а для всякого $i \in [1, p{+}q]$ ровно два значения j соответствуют вершинам — либо $i{+}p$ и $i{+}q$, либо $i{+}q$ и $i{-}q$, либо $i{-}q$ и $i{-}p$ (см. пример 2.2 ниже).

Рассматриваемый граф содержит q ребер вида $(i, i{+}p)$ и p ребер вида $(i, i{+}q)$. Построим в нем наидлиннейшую цепь, начинающуюся с вершины 1 и ребра $(1, p{+}1)$. Поскольку все вершины графа имеют степень 2, то, попав в вершину первый раз, мы можем из нее выйти, но не можем затем попасть в нее вторично. Таким образом, наша цепь закончится снова в вершине 1; при этом на каждом шаге мы переходили из вершины i либо (1) в вершину $(i{+}p)$, либо (2) в вершину $(i{-}q)$. Пусть обход построенного цикла состоит из α шагов (1) и β шагов (2). Тогда, во-первых, величина $\alpha{+}\beta$ не превосходит числа ребер в графе, т. е. $p{+}q$. Во-вторых, $1 + \alpha p - \beta q = 1$, откуда $\alpha p = \beta q$. В силу взаимной простоты p и q число α должно делиться на q, а число β — на p. Отсюда следует, что число ребер в цикле в точности равняется числу ребер (и вершин!) в графе, т. е.

весь граф совпадает с построенным циклом. Это означает, что 1) граф связен, 2) он останется связным при удалении одной вершины. Следовательно, граф отношения ρ для слова длины $p+q-1$ связен и утверждение теоремы доказано для взаимно простых p и q.

Теперь рассмотрим общий случай. Пусть $НОД(p,q) = d$ и $|U| = p+q-d$. Рассмотрим слова $U_i = U(i)U(d+i)U(2d+i)\ldots$ для всех $i = 1,\ldots,d$. Каждое слово U_i имеет взаимно простые периоды p/d, q/d и длину $p/d+q/d-1$. Из доказанного выше следует, что U_i имеет период 1. Следовательно, слово U имеет период d по определению. □

Пример 2.2. Граф отношения ρ на слове длины 10 с периодами 3 и 7 изображен на рис. 1.

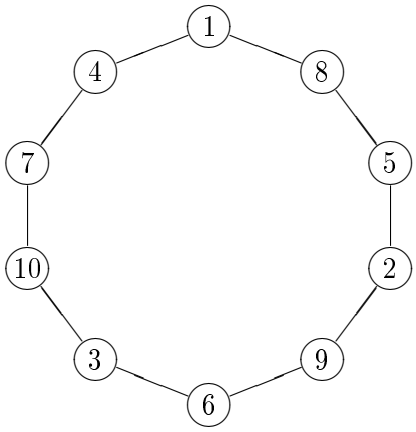


Рис. 1. Граф периодичности

Назовем *длиной взаимодействия* периодов p и q величину $L = L(p,q)$ такую, что слово длины L с периодами p и q необходимо имеет период $НОД(p,q)$, а слово длины $L-1$ может такого периода не иметь. Теорема Файна – Вильфа утверждает, что $L(p,q) \le p+q-НОД(p,q)$.

Предложение 2.1. $L(p,q) = p+q-НОД(p,q)$.

Доказательство. Как и в теореме, вначале докажем требуемое для взаимно простых p и q. Как показано выше, граф отношения ρ для слова длины $p+q$ является простым циклом. Так как $q \geq 2$, то вершины $(p+q)$ и $(p+q-1)$ не являются смежными по определению ρ. Следовательно, граф отношения ρ для слова длины $p+q-2$ получается из простого цикла удалением несмежных вершин, а значит, состоит из двух компонент связности, являющихся цепями. Таким образом, слово длины $p+q-2$ может не иметь периода 1.

Перейдем к общему случаю. Если слово U состоит менее чем из $p+q-d$ букв, где $d = НОД(p,q)$, то хотя бы одно из слов U_i (см. доказательство теоремы) имеет длину не более $p/d+q/d-2$, а значит, может не иметь периода 1; в этом случае само слово U не имеет периода d. Предложение доказано. $\square$

Примечание. Во всех оставшихся утверждениях текущего и следующего параграфов мы ограничимся рассмотрением случая взаимно простых периодов. Это позволит более компактно записывать формулы и при необходимости легко восстановить результат для общего случая при помощи слов U_i.

Следующее предложение показывает одно из направлений обобщения результата теоремы Файна – Вильфа.

Предложение 2.2. *Если слово U имеет взаимно простые периоды p и q и $|U| = p+q-r$, где $1 \leq r \leq q$, то U содержит не более r различных букв. Слово с указанными свойствами, имеющее r различных букв, единственно с точностью до переименования букв.*

Доказательство. Граф отношения ρ для произвольного слова длины $p+q-r$ получается из графа отношения ρ для слова длины $p+q$ (т. е. простого цикла) удалением r вершин с номерами $(p+q-r+1), \ldots, (p+q)$. Никакие две из этих вершин не являются в исходном графе смежными по определению ρ. Таким образом, полученный граф состоит из r компонент связности. Поскольку в позициях из одной компоненты связности

должны стоять одинаковые буквы, мы немедленно получаем требуемое утверждение. □

§4. Частичные слова. Взаимодействие периодов частичных слов

Как уже упоминалось во введении, обычное слово можно представлять себе как функцию

$$W : \{1, \ldots, n\} \to \Sigma.$$

Частичное слово получится, если произвести обобщение: всюду определенную функцию заменить на частичную. Таким образом, *частичным словом длины n над алфавитом Σ* называется частичная функция W из множества $\{1, \ldots, n\}$ во множество Σ. Область определения этой функции будем обозначать через $D(W)$.

Поскольку слово обычно рассматривается как конечная последовательность (цепочка) алфавитных символов, нужно определить частичное слово и на языке конечных последовательностей. А именно, частичным словом над алфавитом Σ будем называть произвольное слово над расширенным алфавитом $\Sigma \cup \{\diamond\}$. Дополнительный символ $\diamond$ будем называть *джокером*. Между определенными выше частичными функциями и словами над расширенным алфавитом существует очевидная биекция: джокеры в слове соответствуют позициям, на которых частичная функция не определена. Ниже, если не оговорено противное, под частичным словом мы будем понимать именно слово в расширенном алфавите, тем самым автоматически распространяя на частичные слова понятия длины, подслова, префикса, суффикса и т. п.

Частичные слова моделируют информацию, часть которой либо *не важна*, либо *недоступна*. В зависимости от того, какой из случаев имеет место, символ $\diamond$ несет различную смысловую нагрузку (в первом случае — это «здесь стоит буква, неважно какая», а во втором — «здесь стоит буква, неизвестно какая»).

Для первого случая примером может служить задача приблизительного поиска в тексте, для второго — задача восстановления информации, поврежденной при передаче или при порче носителя. Оба случая (хотя чаще — второй) встречаются в задачах молекулярной биологии и биоинформатики по расшифровке и анализу генетического материала, представляющего собой последовательности символов четырехэлементного алфавита.

Примечание. Джокер нельзя рассматривать как переменную: различные его вхождения в одно и то же слово могут «маскировать» различные буквы.

При исследовании частичных слов естественно начать с ответа на следующий обширный вопрос: какие комбинаторные свойства слов допускают обобщение на частичные слова? Основная цель данного параграфа — доказать, что свойство взаимодействия периодов допускает такое обобщение.

Понятие периода слова можно обобщить для частичных слов двумя способами. Пусть W — частичное слово. Натуральное число $p \leq |W|$ называется *локальным периодом* W, если $W(i) = W(i+p)$ для всех $i = 1, \ldots, |W| - p$ таких, что $i, i+p \in D(W)$. Натуральное число $p \leq |W|$ называется *периодом* W, если $W(i) = W(j)$ для всех $i, j \in D(W)$ таких, что $i \equiv j (\mathrm{mod} p)$. Оба этих понятия совпадают с обычным понятием периода для слов, но существенно различны для частичных слов: так, например, частичное слово $ab\diamond bc$ имеет локальный период 2, но не имеет периода 2.

Замечание 2.2. Частичное слово имеет период p тогда и только тогда, когда вместо входящих в него джокеров можно подставить буквы таким образом, что полученное слово будет иметь период p.

Свойство взаимодействия периодов для частичных слов сформулируем в точности так же, как и для слов (см. предыдущий параграф). Несложно заметить, что для локальных периодов свойство взаимодействия в общем случае не выполняется, если частичное слово содержит хотя бы два джокера. В са-

мом деле, рассмотрим сколь угодно длинное частичное слово с локальными периодами p и q и джокерами на позициях $(q{+}1)$ и $(p{+}1)$. Локальная периодичность не нарушится, если заменить букву в первой позиции на любую другую. Таким образом, можно получить частичное слово, в котором буквы на позициях 1 и $НОД(p,q){+}1$ различны и которое, следовательно, не имеет периода $НОД(p,q)$.

Тем не менее для периодов свойство взаимодействия выполняется. Справедлива следующая теорема.

Теорема 2.2. *Если частичное слово U имеет взаимно простые периоды p и q, $p > q$, содержит k джокеров и справедливо неравенство $|U| \geq qk + (p + q - 1)$, то U имеет период 1.*

Как и в доказательстве теоремы Файна – Вильфа, мы будем оперировать с графами бинарных отношений, определяемых периодами слова. Однако определение отношения ρ, сопоставляемого слову с двумя периодами, необходимо откорректировать для переноса на множество частичных слов.

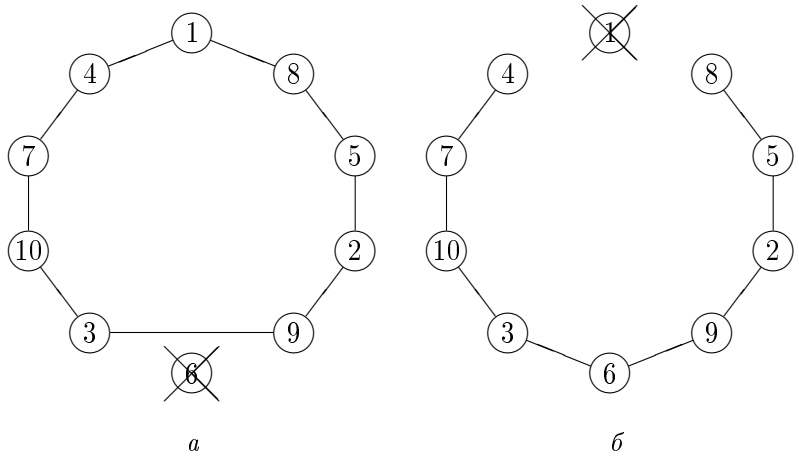
Итак, пусть частичное слово U имеет взаимно простые периоды p и q. Отношение ρ на множестве $D(U)$ определяется следующим образом: пара (i,j) принадлежит ρ тогда и только тогда, когда $i = j(\mathrm{mod}p)$ и всякое m, лежащее между i и j и равное им по модулю p, не принадлежит $D(U)$ (соответственно $i = j(\mathrm{mod}\, q)$ и всякое m, лежащее между i и j и равное им по модулю q, не принадлежит $D(U)$). Заметим, что, отношение ρ для слов, определенное в предыдущем параграфе, подходит под данное определение.

Замечание 2.3. Частичное слово U имеет период 1 тогда и только тогда, когда граф соответствующего ему отношения ρ связен.

Лемма 2.1. *Если частичное слово U длины $p{+}q$ имеет взаимно простые периоды p и q, $p > q$, содержит ровно один джокер в позициях $1, \ldots, q, p{+}1, \ldots, p{+}q$ и любое количество джокеров в позициях $q{+}1, \ldots, p$, то U имеет период 1.*

Доказательство. Как было показано при доказательстве тео-

ремы 2.1, граф отношения ρ для слова длины $p{+}q$ является простым циклом. Граф для частичного слова той же длины получается «изъятием» из этого цикла вершин, соответствующих позициям, в которых стоят джокеры. При этом результат «изъятия» вершины t, соединенной в графе с вершинами $(t{+}q)$ и $(t{-}q)$, будет таким, как в примере на рис. 2, *а*, а результат «изъятия» вершины t, соединенной в графе с вершинами $(t{+}q)$ и $(t{+}p)$ либо с вершинами $(t{-}q)$ и $(t{-}p)$ — таким, как в примере на рис. 2, *б*. Таким образом, если в слове длины $p{+}q$ разместить произвольное количество джокеров в позициях $q{+}1, \ldots, p$, то для получившегося частичного слова граф по-прежнему будет простым циклом. Добавление же джокера в одну из позиций $1, \ldots, q, p{+}1, \ldots, p{+}q$ приведет к разрыву цикла, но граф, тем не менее, останется связным. □



а *б*

Рис. 2. Добавление джокера в граф периодичности

Замечание 2.4. Для частичных слов длины $p{+}q{-}1$ лемма 1 неверна. Граф, соответствующий слову длины $p{+}q{-}1$, является цепью, и размещение джокера в любой из позиций $1, \ldots, q{-}1, p{+}1, \ldots, p{+}q{-}1$ приведет к разрыву этой цепи и, таким образом, к несвязности получившегося графа.

Будем называть джокер *существенным* для частичного слова длины $p+q$, если он расположен в его префиксе или суффиксе длины q.

Лемма 2.2 *Если частичное слово U со взаимно простыми периодами p и q, $p > q$, имеет подслово длины $p+q$, которое содержит не более одного существенного джокера, то U имеет период 1.*

Доказательство. Пусть указанное подслово расположено в U в позициях $t{+}1, \ldots, t{+}p{+}q$. По лемме 2.1 все буквы, стоящие на данных позициях, равны между собой; пусть они равны a. Достаточно доказать, что любая буква в U совпадает с a. Рассмотрим позицию s, занятую буквой и не принадлежащую выбранному подслову; пометим все позиции, равные s по модулю q. Тогда одна из позиций $t{+}1, \ldots, t{+}q$ и одна из позиций $t{+}p{+}1, \ldots, t{+}p{+}q$ окажутся помеченными. По условию леммы не более чем одна из них содержит джокер; следовательно, одна из них гарантированно содержит букву a. Поскольку все буквы в помеченных позициях равны, получаем $U(s) = a$. $\square$

Доказательство теоремы 2.2. Оценим максимальную длину частичного слова W, удовлетворяющего условиям теоремы и не имеющего периода 1. По лемме 2.2 W должно содержать не менее двух существенных джокеров в каждом подслове длины $p + q$. Поскольку в каждом таком подслове имеется $2q$ позиций для существенных джокеров, каждый джокер может являться существенным не более чем для $2q$ подслов. Далее, первый джокер в слове не может занимать последнюю позицию в подслове с двумя существенными джокерами; следовательно, он является существенным лишь для $2q - 1$ подслова. То же самое справедливо и для последнего джокера в слове.

Поскольку для каждого подслова длины $p + q$ необходимы два существенных джокера, количество таких подслов в W не превышает $(2qk - 2)/2$, т. е. $qk - 1$. С другой стороны, слово длины $|W|$ содержит $|W| - p - q + 1$ таких подслов. Итак,

$$|W| - p - q + 1 \leq qk - 1,$$

откуда $|W| \le qk + (p + q - 2)$. Поскольку в условии теоремы $|U| \ge qk + (p + q - 1)$, то U имеет период 1. Теорема доказана. □

Минимальное число $L = L(k, p, q)$ такое, что частичное слово длины L со взаимно простыми периодами p и q, $p > q$, и k джокерами обязательно имеет период 1, будем называть *длиной взаимодействия периодов p и q при наличии k джокеров*. Длину взаимодействия естественно считать функцией натурального аргумента k с параметрами p и q. Теорема 2.2 дает линейную оценку длины взаимодействия, показывая тем самым, что свойство взаимодействия периодов выполняется для всех частичных слов, количество джокеров в которых не превосходит некоторой фиксированной доли от длины слова. Рассмотрим, насколько точна оценка, приведенная в теореме 2.2. Предварительно заметим, что при $k = 0$ эта оценка совпадает с оценкой в теореме Файна – Вильфа, а следовательно, является точной.

Предложение 2.3. *Пусть $k \ge 1$, p и q взаимно просты, $p > q$. Тогда $L(k, p, q) = qk + (p + q - 1)$ в том и только в том случае, если $q = 2$ и $k = sp$ для некоторого натурального s.*

Доказательство. Пусть $q = 2$ и $k = sp$. Рассмотрим частичное слово W длины

$$|W| = qk + (p + q - 2) = (2s + 1)p,$$

состоящее из p букв, за которыми следуют p джокеров, снова p букв, затем p джокеров, и т. д., чередуя подслова из p букв и p джокеров. Потребуем, кроме того, чтобы все буквы, находящиеся в слове W на нечетных позициях, были равны a, а на четных — b.

$$W = \underbrace{\underbrace{aba\ldots}_{p}\underbrace{\diamond\diamond\ldots\diamond}_{p}\underbrace{aba\ldots}_{p}\ \ldots\ \underbrace{\diamond\diamond\ldots\diamond}_{p}\underbrace{aba\ldots}_{p}}_{(2s+1)\text{ блоков}}$$

Полученное частичное слово по построению имеет периоды p и 2 и содержит $sp = k$ джокеров, не имея при этом периода 1. С учетом теоремы 2.2 длина взаимодействия периодов p и 2 при наличии sp джокеров в точности равна указанной в условии предложения.

Доказательство предложения в обратную сторону является достаточно трудоемким, и мы вынуждены его опустить. Найти это доказательство можно в работе [22]. □

Приводимая ниже теорема 2.3 дает точную оценку длины взаимодействия для произвольных p и q и не слишком маленьких k. Доказательство п. 1 представляет собой **упражнение 2.2**, которое несложно выполнить, изучив пример, приведенный в доказательстве предыдущего предложения. Доказательство п. 2 многократно сложнее доказательства теоремы 2.2; оно полностью приведено в [30].

Теорема 2.3. *Пусть $k \geq 1$, p и q взаимно просты, $p > q$. Тогда*

1) $L(k, p, 2) = (2\lfloor \frac{k}{p} \rfloor + 1)p + (k \bmod p) + 1$;

2) если $q \geq 3$ и $k \geq \lfloor \frac{3p}{q} \rfloor + 3$, то

$$L(k, p, q) < \frac{pq}{p + q - 2} \cdot k + 4(q - 1)$$

и для любого $\varepsilon > 0$ и любого q найдутся такие p и k, что

$$L(k, p, q) > \frac{pq}{p + q - 2} \cdot k + 4(q - 1) - \varepsilon.$$

§5. Локальные периоды и критические разбиения

В этом параграфе нам будет удобно представлять слова изображенными на отрезке $[0, n]$ действительной прямой так, что i-я буква занимает отрезок $[i-1, i]$:

$$W = \underset{0}{\lfloor} \overset{a_1}{\underset{1}{\rfloor\lfloor}} \overset{a_2}{\underset{2}{\rfloor}} \quad \cdots \quad \overset{\lfloor a_n}{\underset{n-1}{\rfloor}} \overset{}{\underset{n}{\rfloor}}.$$

Число $p \in \mathbb{N}$ называется *локальным периодом слова* W *в точке* k, $k \in \mathbb{N}$, $k < |W|$, если слово $W(n_1 \ldots n_2)$ имеет период p, где $n_1 = \max\{k-p+1, 1\}$, $n_2 = \min\{k+p, |W|\}$. Другими словами, p есть локальный период W в точке k, если найдется слово Z длины p такое, что W и Z^2 «накладываются» друг на друга одним из четырех способов, указанных на рис. 3.

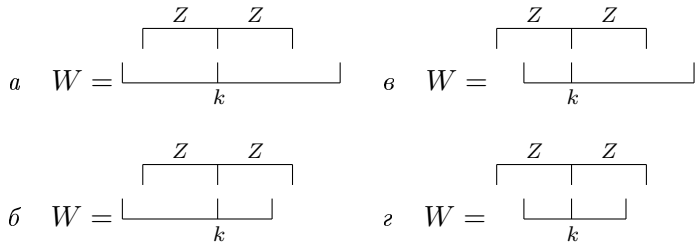


Рис. 3. Локальный период в точке k

Минимальный локальный период слова W в точке k обозначим через $per(W, k)$.

Пример 2.3. Пусть $W = aababaaa$. Тогда $per(W) = 6$, а вычисление локальных периодов дает

$per(W, 1) = 1$, $Z = a$,
$per(W, 2) = 5$, $Z = babaa$,
$per(W, 3) = 2$, $Z = ab$,
$per(W, 4) = 2$, $Z = ba$,
$per(W, 5) = 6$, $Z = aaabab$,
$per(W, 6) = 1$, $Z = a$,
$per(W, 7) = 1$, $Z = a$.

Очевидно, что всякий период слова является его локальным периодом в любой точке. Отсюда следует, что для всякого k $per(W, k) \leq per(W)$. Разбиение $W = UV$ называется *критическим*, если $per(W, |U|) = per(W)$. Таким образом, критическое разбиение отмечает точку, в которой локальная структура слова (с точки зрения периодических свойств) наиболее точно отражает его глобальную структуру. Для слова из примера 2.3 проведенный анализ обнаружил в точности одно критическое

разбиение: $W = \mathbf{aabab\ aaa}$. Следующая теорема гласит, что каждое слово обладает критическим разбиением, причем таким, что точка k расположена достаточно близко к началу слова (имеет место рис. 3, *в* или 3, *г*), а из доказательства теоремы следует эффективный алгоритм построения этого разбиения.

Теорема 2.4 (теорема о критическом разбиении, 1979). *Всякое слово W такое, что $|alph(W)| \geq 2$, $per(W) = p$, допускает критическое разбиение $W = UV$ с условием $|U| < p$.*

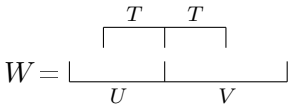
Обозначим $\Sigma = alph(W)$ и определим на Σ^+ два лексикографических порядка. Для этого рассмотрим произвольный линейный порядок на Σ; соответствующий ему лексикографический порядок на Σ^+ назовем *левым* и обозначим $\leq_l$. Лексикографический порядок, соответствующий обратному линейному порядку на Σ, назовем *правым* и обозначим $\leq_r$. Доказательство теоремы опирается на две леммы.

Лемма 2.3. *Если $U \leq_l V$ и $U \leq_r V$, то U — префикс V.*

Доказательство. Согласно определению лексикографического порядка, если U — не префикс V, то $U = PxQ_1$, $V = PyQ_2$, где $x, y \in \Sigma$, $x < y$. Но условие $x < y$ не может выполняться одновременно в двух взаимно обратных линейных порядках на Σ. Значит, остается лишь вариант, когда U является префиксом V. □

Лемма 2.4. *Пусть $W = UV$, где V — максимальный в отношении некоторого лексикографического порядка суффикс W. Тогда если слово T является одновременно суффиксом U и префиксом V, то $T = \lambda$.*

Доказательство. Указанное в условии взаимное расположение слов W, U, V, T изображено ниже:



Пусть $V = TV'$. Тогда V' — суффикс W, откуда $V \geq V'$, т. е. $TV' \geq V'$. С другой стороны, TV является суффиксом W, откуда $TV \leq V$ или $TTV' \leq TV'$. Последнее неравенство влечет $TV' \leq V'$. В итоге получаем $TV' = V'$, т. е. $T = \lambda$. $\square$

Доказательство теоремы 2.4. Пусть $W = UV$ и $W = U'V'$ — два разбиения W, причем V — максимальный суффикс W в отношении порядка $\leq_l$, а V' — максимальный суффикс W в отношении порядка $\leq_r$. Пусть также без ограничения общности $|U| \geq |U'|$. Докажем, что $W = UV$ является критическим разбиением.

Вначале докажем, что $U \neq \lambda$. Пусть $U = \lambda$. В этом случае $W = V = V'$. Предположим, что $W = xT$, где $x \in \Sigma$. Тогда T — суффикс W, откуда $T <_l V$, $T <_r V'$ по выбору V и V'. По лемме 2.3 T является префиксом W, т. е. $W = Ty$, $y \in \Sigma$. Но тогда

$$x = W(1) = T(1) = W(2) = T(2) = \ldots = W(|W|) = y,$$

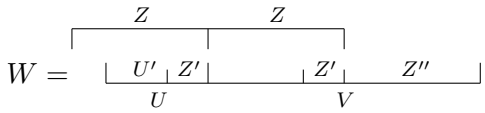
что противоречит условию теоремы $|alph(W)| \geq 2$. Данное противоречие доказывает, что слово U не пусто, следовательно, $W = UV$ является разбиением ($V \neq \lambda$, поскольку λ — минимальное слово в отношении любого лексикографического порядка).

Для завершения доказательства осталось показать, что всякий локальный период слова W в точке $|U|$ является периодом W. Тем самым будет доказано, что $per(W, |U|) = per(W)$ и рассматриваемое разбиение является критическим. Пусть p — локальный период W в точке $|U|$; рассмотрим случаи, изображенные на рис. 3, *а–г* ($k = |U|$; $|Z| = p$).

Случай «*а*» невозможен ввиду леммы 2.4: Z является непустым суффиксом U и префиксом V. В случае «*б*» V есть префикс Z, откуда $V < ZV$ в любом лексикографическом порядке, что противоречит предположению о том, что V — максимальный суффикс W в отношении $\leq_l$. Таким образом, случай «*б*» также невозможен. В случае «*г*», очевидно, p яв-

ляется периодом W, поскольку W есть подслово в Z^2. В оставшемся случае «*в*» нам необходимо доказать, что V имеет период p.

Поскольку $|U| \geq |U'|$, запишем $U = U'Z'$; тогда Z' является суффиксом Z, а $V' = Z'V$. Положим, кроме того, $V = ZZ''$:



По определению V имеем $V >_l Z''$, а по определению V' получаем $V' = Z'V >_r Z'Z''$, откуда $V >_r Z''$. Тогда по лемме 2.3 Z'' есть префикс V, т. е. $V = Z''T$ для некоторого T. Таким образом, $ZZ'' = Z''T$ и к этим словам можно применить предложение 1.2: существуют такие $P, Q \in \Sigma^+$, $n \geq 0$, что $Z = PQ$, $T = QP$, $Z'' = (PQ)^n P$. Тогда слово $V = ZZ''$ имеет период $|PQ| = |Z| = p$.

Итак, мы доказали, что рассматриваемое разбиение является критическим. Кроме того, мы доказали, что если p — локальный период W в точке $|U|$, то имеет место один из случаев *«в»*, *«г»*, откуда $|U| < p$. Тем самым теорема полностью доказана. □

Из теоремы о критическом разбиении в качестве следствия можно получить критерий периодичности Z-слов (напомним, что Z-словами называются бесконечные последовательности алфавитных символов, индексированные целыми числами). Определения периода и локального периода Z-слова достаточно очевидны, но мы все же воспроизведем их. Число $p \in \mathbb{N}$ называется *периодом* Z-слова $\mathbf{W}$, если $\mathbf{W}(i) = \mathbf{W}(i+p)$ для любого целого i. Z-слово называется *периодическим*, если оно имеет период. Число $p \in \mathbb{N}$ называется *локальным периодом* Z-слова $\mathbf{W}$ в точке $k \in \mathbb{Z}$, если слово $\mathbf{W}(k-p+1 \ldots k+p)$ имеет период p.

Следствие 2.1 (критерий периодичности Z-слов). *Z-слово $\mathbf{W}$ является периодическим тогда и только тогда, когда ми-*

нимальные локальные периоды слова $\mathbf{W}$ *существуют во всех точках и ограничены в совокупности некоторой константой* $N \in \mathbb{N}$.

Доказательство. В одну сторону доказательство очевидно: если $\mathbf{W}$ имеет период, то он является его локальным периодом в любой точке. Для доказательства в обратную сторону вначале заметим, что если все конечные подслова $\mathbf{W}$ имеют период p, то и само $\mathbf{W}$ имеет период p. Согласно теореме 2.4 период любого конечного подслова $\mathbf{W}$ не превосходит N.

Пусть U_1 — конечное подслово $\mathbf{W}$, $per(U_1) = p_1$. Если все конечные подслова из $\mathbf{W}$, содержащие U_1, имеют период p_1, то положим $p = p_1$. Если же в $\mathbf{W}$ есть подслово $U_2 = PU_1Q$, не имеющее периода p_1, то положим $p_2 = per(U_2)$ $(p_2 > p_1)$, повторим исследование для U_2 и т. д. Поскольку последовательность $\{p_i\}$ строго возрастает и ограничена сверху числом N, она содержит лишь конечное число членов. Таким образом, на некотором (k-м) шаге все конечные подслова из $\mathbf{W}$, содержащие U_k, будут иметь период p_k и мы положим $p = p_k$. Поскольку всякое конечное подслово из $\mathbf{W}$ содержится в некотором конечном подслове, содержащем U_k, то все конечные подслова из $\mathbf{W}$ имеют период p. Следствие доказано. $\square$

Глава 3. ИЗБЕГАЕМОСТЬ

Данная глава посвящена свойству слов не содержать подслов определенного вида.

§6. Слова Туэ – Морса и некоторые их свойства

В этом параграфе мы познакомимся с наиболее знаменитым множеством слов во всей комбинаторике. Впервые слова Туэ – Морса появились на сцене в 1851 году в работе по теории чисел малоизвестного французского математика Пруэ (см. [19]). Хотя никаких «слов» в этой работе не было, алгоритм их построения был описан. Мы начнем со свойства, доказанного в этой работе, — о существовании «сверхуравновешенного» разбиения натурального ряда на два класса.

Теорема 3.1. *Существует разбиение множества $\mathbb{N}$ на два класса, обладающее следующим свойством: для любого $n \in \mathbb{N}$, $k \in \{1, \ldots, n\}$ сумма k-х степеней элементов одного класса, принадлежащих множеству $\{1, \ldots, 2^{n+1}\}$, равна сумме k-х степеней элементов другого класса, принадлежащих тому же множеству.*

Доказательство. Элементы одного множества назовем «синими», а другого — «зелеными». Разбиение (или «раскраску») натурального ряда произведем по индукции. База индукции: единицу окрасим в синий цвет. Предположение индукции: пусть отрезок $[1, 2^n]$ раскрашен. На шаге индукции раскрасим отрезок $[2^n{+}1, 2^{n+1}]$ следующим образом: элемент $2^n{+}i$ получает цвет, отличный от цвета элемента i. В итоге отрезок $[2^n{+}1, 2^{n+1}]$ становится «негативом» отрезка $[1, 2^n]$.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	...
с	з	з	с	з	с	с	з	з	с	с	з	с	з	з	с	з	...

Доказательство того, что эта раскраска обладает требуемыми свойствами, также проведем по индукции.

База индукции: $n = 1$ (отрезок $[1, 4]$). $1 + 4 = 2 + 3$.

Предположение индукции: на отрезке $[1, 2^n]$ сумма k-х степеней синих элементов равна сумме k-х степеней зеленых для любого $k \leq n-1$.

Шаг индукции: на отрезке $[1, 2^{n+1}]$ сумма k-х степеней синих элементов равна сумме k-х степеней зеленых для любого $k \leq n$. Докажем это. Вначале рассмотрим случай $k < n$. Тогда по предположению индукции на отрезке $[1, 2^n]$ сумма k-х степеней синих элементов равна сумме k-х степеней зеленых. Сравним суммы k-х степеней синих и зеленых элементов на отрезке $[2^n{+}1, 2^{n+1}]$. Множество всех синих (всех зеленых) элементов на отрезке $[x, y]$ обозначим через $[x, y]_{\text{с}}$ (соответственно $[x, y]_{\text{з}}$).

$$\sum_{i \in [2^n+1, 2^{n+1}]_{\text{с}}} i^k = \sum_{i \in [1, 2^n]_{\text{з}}} (2^n + i)^k = \sum_{i \in [1, 2^n]_{\text{з}}} \sum_{m=0}^{k} C_k^m 2^{nm} i^{k-m} =$$

$$= \sum_{m=0}^{k} C_k^m \cdot \sum_{i \in [1, 2^n]_{\text{з}}} 2^{nm} i^{k-m} = \sum_{m=0}^{k} C_k^m \cdot \sum_{i \in [1, 2^n]_{\text{с}}} 2^{nm} i^{k-m} =$$

$$= \sum_{i \in [1, 2^n]_{\text{с}}} \sum_{m=0}^{k} C_k^m 2^{nm} i^{k-m} = \sum_{i \in [1, 2^n]_{\text{с}}} (2^n + i)^k = \sum_{i \in [2^n+1, 2^{n+1}]_{\text{з}}} i^k.$$

Итак, равенство k-х степеней для случая $k < n$ доказано. Разберем случай $k = n$. Используя преобразования, аналогичные приведенным выше, вычислим разность сумм n-х степеней синих и зеленых элементов на отрезке $[2^n{+}1, 2^{n+1}]$.

$$\sum_{i \in [2^n+1, 2^{n+1}]_{\text{с}}} i^n - \sum_{i \in [2^n+1, 2^{n+1}]_{\text{з}}} i^n = \sum_{i \in [1, 2^n]_{\text{з}}} (2^n + i)^n - \sum_{i \in [1, 2^n]_{\text{с}}} (2^n + i)^n =$$

$$= \sum_{m=0}^{n} C_n^m \cdot \sum_{i \in [1, 2^n]_{\text{з}}} 2^{nm} i^{n-m} - \sum_{m=0}^{n} C_n^m \cdot \sum_{i \in [1, 2^n]_{\text{с}}} 2^{nm} i^{n-m} =$$

$$= \sum_{i \in [1, 2^n]_{\text{з}}} i^n - \sum_{i \in [1, 2^n]_{\text{с}}} i^n.$$

(Последнее равенство вытекает из предположения индукции: все остальные слагаемые внутренних сумм равны.) Отсюда немедленно следует, что разность сумм n-х степеней синих и зеленых элементов на всем отрезке $[1, 2^{n+1}]$ равна нулю. Шаг индукции завершен. □

Отметим, что «раскраска» из предыдущей теоремы есть не что иное, как ω-слово в бинарном алфавите $\{\text{с}, \text{з}\}$. С этим ω-словом (и его префиксами длины 2^n для некоторого n) мы будем встречаться в этой и следующей главах многократно.

Хронологически мы подошли к работам норвежского математика начала прошлого века А. Туэ. Ему принадлежит ряд выдающихся математических открытий; однако не вполне удачное размещение публикаций (все — на норвежском языке, в национальных журналах и сборниках) привело к тому, что научный мир познакомился с его трудами спустя полвека, успев повторить его основные достижения, иногда и не по одному разу.

Туэ был первым математиком, изучавшим свойства символьных последовательностей, так что его с полным правом можно считать «дедушкой» современной комбинаторики слов. Работы Туэ относятся к тому времени, когда такие технические достижения, как радио и телеграф, потребовали нового языка для передачи сообщений, основанного на малом количестве алфавитных символов. На сцену впервые вышел бинарный алфавит: два хорошо различимых сигнала — точка и тире — легли в основу кодировки алфавитов естественных языков, известной как азбука Морзе. Туэ изучал выразительные возможности алфавитов из двух и трех символов и пришел к выводу, что они весьма велики.

Сформулируем и докажем самый известный результат Туэ.

Теорема 3.2 (Туэ, 1906). *Существует сколь угодно длинное слово в бинарном алфавите, не содержащее подслов X^3 для произвольного X.*

Доказательство этого результата конструктивно: мы по-

строим бесконечную серию слов, удовлетворяющих требуемому свойству. Такая серия будет содержать сколь угодно длинные слова, поскольку количество слов фиксированной длины над конечным алфавитом конечно.

Определим две последовательности слов в бинарном алфавите $A = \{a, b\}$ совместной индукцией:

$$U_0 = a\,, \qquad V_0 = b\,,$$
$$U_{n+1} = U_n V_n\,, \qquad V_{n+1} = V_n U_n\,.$$

Слова U_n, V_n и есть *слова Туэ – Морса*.

Упражнение 3.1. Доказать следующие простые свойства слов Туэ – Морса.
1. $|U_n| = |V_n| = 2^n$.
2. U_n получается из V_n переименованием букв.
3. U_n является префиксом U_{n+1}.
4. U_n определяет раскраску Пруэ отрезка $[1, 2^n]$ натурального ряда.

Лемма 3.1. *Пусть $\phi : A^+ \to A^+$ — морфизм, задаваемый равенствами $\phi(a) = ab$, $\phi(b) = ba$. Тогда $U_{n+1} = \phi(U_n)$, $V_{n+1} = \phi(V_n)$ и, как следствие, $U_n = \phi^n(a)$, $V_n = \phi^n(b)$ для всех натуральных n.*

Морфизм ϕ, определенный в лемме 3.1, называется морфизмом Туэ.

Лемма 3.2. *Слово U тогда и только тогда равно $\phi(V)$ для некоторого V, когда U можно разбить на блоки длины 2, каждый из которых равен ab или ba. (Иными словами, когда всякое подслово длины 2, начинающееся в U с нечетной позиции, есть ab или ba.)*

Упражнение 3.2. Доказать леммы 3.1 и 3.2.

Теперь мы можем доказать теорему 3.2, а именно, тот факт, что слова U_n не содержат подслов вида X^3. (Это свойство, естественно, будет справедливо и для V_n, в силу симметрии.)

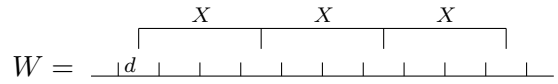
Доказательство. Воспользуемся методом минимального контрпримера. Пусть n — наименьшее натуральное число та-

кое, что U_n содержит подслово X^3 для некоторого X. Согласно леммам 3.1 и 3.2, слово U_n может быть разбито на блоки вида ab и ba. Рассмотрим все случаи расположения X^3 относительно границ этих блоков.

Случай 1. $|X| = 2k$.

1.1. X^3 начинается в U_n с нечетной позиции. Тогда $X = \phi(V)$ для некоторого V согласно лемме 3.2, откуда $X^3 = (\phi(V))^3$ и, поскольку $U_n = \phi(U_{n-1})$, V^3 является подсловом U_{n-1}, противоречие с минимальностью n.

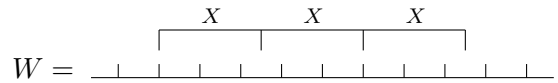
1.2. X^3 начинается в U_n с четной позиции.



Тогда $X(1) \neq X(2k)$ (они лежат в одном блоке) и $X(1) \neq d$ (они также лежат в одном блоке). Следовательно, $X(2k) = d$. Положим $Y = d\,X(1 \ldots 2k{-}1)$. Тогда Y^3 является подсловом в U_n и начинается в нем с нечетной позиции; применяя случай 1.1, получаем противоречие.

Случай 2. $|X| = 2k - 1$.

2.1. X^3 начинается в U_n с нечетной позиции.



Тогда $X(1) \neq X(2)$ (они находятся в одном блоке в первом вхождении X), $X(2) \neq X(3)$ (они находятся в одном блоке во втором вхождении X), $X(3) \neq X(4)$ (они находятся в одном блоке в первом вхождении X) и т. д. В итоге получаем $X(1) = X(3) = \ldots = X(2k{-}1)$. Но $X(1)$ и $X(2k{-}1)$ также находятся в одном блоке (на стыке первого и второго вхождения X). Противоречие.

Случай 2.2 (X^3 начинается в U_n с четной позиции) симметричен случаю 2.1. Доказательство теоремы завершено. $\square$

Итак, мы доказали, что слова Туэ – Морса не содержат подслов вида X^3 (т. е. *кубов*). На самом деле, можно практи-

чески дословно воспроизвести то же самое доказательство для получения более сильного результата.

Теорема 3.3 (Туэ, 1906). *Слова Туэ – Морса не содержат подслов вида X^2c, где $c = X(1)$, для произвольного $X \in A^+$.*

Упражнение 3.3*. Доказать теорему 3.3.

Примечание. Слова, которые не содержат подслов вида X^3, называются *бескубными*. Слова, которые не содержат подслов вида X^2c, где $c = X(1)$, называются *сильно бескубными*.

Ввиду теорем 3.2 и 3.3 естественно ожидать, что хотя в бинарном алфавите существует всего шесть слов, не содержащих подслов вида X^2 (*квадратов*), в алфавите из трех символов можно будет построить сколь угодно длинное бесквадратное слово. Это действительно так, но перед тем, как доказать очередную теорему Туэ, нам необходимо дать несколько определений.

Бесконечную последовательность слов $\{W_n\}$ будем называть *префиксной*, если для любого i слово W_i является собственным префиксом в W_{i+1}. Префиксная последовательность обладает следующими свойствами:

1) $\forall k \in \mathbb{N}\ \exists i \in \mathbb{N} :\ |W_i| \geq k$;

2) $\forall k \in \mathbb{N}\ \forall i, j \in \mathbb{N} :\ |W_i|, |W_j| \geq k \Rightarrow W_i(k) = W_j(k)$.

Эти свойства позволяют сопоставить префиксной последовательности $\{W_i\}$ ω-слово W_∞, называемое *пределом* $\{W_i\}$, по правилу $W_\infty(k) = W_i(k)$ для произвольного i с условием $|W_i| \geq k$.

Замечание 3.1. Последовательности $\{U_n\}$ и $\{V_n\}$ являются префиксными. Их пределы U_∞ и V_∞ мы будем называть ω-словами Туэ – Морса.

Примечание. Будем говорить, что ω-слово $\mathbf{W}$ порождено морфизмом f, если $\mathbf{W}$ есть предел префиксной последовательности $\{f^n(c)\}$ для некоторой буквы c.

Теорема 3.4 (Туэ, 1912). *Существует сколь угодно длинное бесквадратное слово в алфавите из трех символов.*

Доказательство. Построим ω-слово $\mathbf{T}$ над алфавитом $\{0,1,2\}$ по следующему правилу:

$$\mathbf{T}(i) = \begin{cases} 0, & U_\infty(i)U_\infty(i{+}1) = ab, \\ 1, & U_\infty(i)U_\infty(i{+}1) = ba, \\ 2, & U_\infty(i) = U_\infty(i{+}1). \end{cases}$$

Докажем, что $\mathbf{T}$ не содержит квадратов. Заметим вначале, что любые два соседних символа в $\mathbf{T}$ различны. Действительно, 00 и 11 не могут встретиться в $\mathbf{T}$ по определению, а наличие 22 означает, что U_∞ содержит aaa или bbb, что противоречит теореме 3.3.

Предположим, что $\mathbf{T}(i \ldots i{+}k{-}1) = \mathbf{T}(i{+}k \ldots i{+}2k{-}1)$. Тогда если $\mathbf{T}(j) \in \{0,1\}$, $i \le j \le i{+}k{-}1$, то $U_\infty(j) = U_\infty(j{+}k)$ и $U_\infty(j{+}1) = U_\infty(j{+}k{+}1)$. Поскольку $\mathbf{T}$ не содержит соседних двоек, получаем

$$U_\infty(i \ldots i{+}k{-}1) = U_\infty(i{+}k \ldots i{+}2k{-}1),$$

и при этом $U_\infty(i{+}k) = U_\infty(i{+}2k)$. Получили противоречие с теоремой 3.3. □

§7. Избегаемые экспоненты

Экспонентой слова называется отношение длины этого слова к его минимальному периоду. Таким образом, экспонента показывает, сколько раз укладывается в слове его самый длинный повторяющийся префикс и служит, тем самым, индексом «глобальной» периодичности слова. Обозначается экспонента слова U через $exp(U)$. Экспоненту можно трактовать как обобщение понятия степени на рациональные показатели (если слово X примитивно, то $exp(X^2) = 2$, $exp(X^3) = 3$, а экспонента слова X^2c, где $c = X(1)$, равна $2 + 1/|X|$).

Однако часто экспонента практически не дает информации о структуре слова. Так, например, $exp(abababab) = 1$; при этом основным элементом структуры данного слова является подслово $(ab)^4$ экспоненты 4. Информативным показателем

является *локальная экспонента* слова, обозначаемая $lexp(U)$ и определяемая равенством

$$lexp(U) = \max\{exp(V) |\ V \text{ — подслово } U\}.$$

Таким образом, локальная экспонента характеризует степень повторяемости фрагментов внутри слова. Отметим, что определенные в §6 бескубные и бесквадратные слова — это в точности слова, локальная экспонента которых меньше 3 (соответственно меньше 2), а сильно бескубные слова имеют локальную экспоненту, не превосходящую 2.

Будем говорить, что слово U *избегает* экспоненту β, если $lexp(U) < \beta$. Экспоненту β будем называть *избегаемой в алфавите* Σ, если существует бесконечно много слов над Σ, избегающих β. Если множество слов над Σ, избегающих β, конечно, β будем называть *неизбежной в алфавите* Σ. Поскольку конкретные значения символов, входящих в Σ, не влияют на свойство избегаемости, обычно мы будем говорить об n*-избегаемых* и n*-неизбежных* экспонентах, полагая $n = |\Sigma|$.

Фактически мы уже начали исследования выразительных возможностей алфавитов с точки зрения избегаемых экспонент. Результаты предыдущего параграфа можно переформулировать в виде следствия.

Следствие 3.1. *Всякая экспонента, большая 2, является 2-избегаемой. Экспонента 2 является 2-неизбежной и 3-избегаемой.*

Упражнение 3.4. Доказать, что всякая экспонента является 1-неизбежной.

Упражнение 3.5. Пусть экспонента β n-избегаема. Доказать, что
1) β является $(n{+}1)$-избегаемой;
2) любая экспонента, большая β, является n-избегаемой.

Определим функцию натурального аргумента $RT(n)$, называемую *границей повторяемости* (repetition threshold), следующим образом:

$$RT(n) = \inf\{\beta |\ \beta \ \ n\text{-избегаема}\}.$$

Задача оценки выразительных возможностей алфавитов с точки зрения экспонент сводится тем самым к вычислению функции $RT(n)$. Мы уже знаем, что $RT(2) = 2$, а $RT(1)$ не определена (можно положить $RT(1) = \infty$). Следующая гипотеза была сформулирована около тридцати лет назад.

Гипотеза (Дежан, 1972). Таблица значений функции $RT(n)$ выглядит следующим образом:

1	2	3	4	5	6	...	n	...
∞	2	7/4	7/5	5/4	6/5	...	$n/(n-1)$	...

Дежан доказала эту гипотезу для случая $n = 3$. Вначале она показала (перебором), что всякое слово над трехбуквенным алфавитом длины ≥ 39 содержит подслово экспоненты $\geq 7/4$, откуда $RT(3) \geq 7/4$. Далее, Дежан построила морфизм d со следующим свойством:

(i) если $lexp(W) \leq 7/4$, то $lexp(d(W)) \leq 7/4$.

Вот этот морфизм:

$d(a) = abcacbcabcbacbcacba$,
$d(b) = bcabacabcacbacabacb$,
$d(c) = cabcbabcabacbabcbac$.

Свойство (i) означает, что все слова $d^n(a)$ имеют экспоненту, не превышающую $7/4$; отсюда следует, что любая бóльшая экспонента 3-избегаема, а значит, $RT(3) = 7/4$.

Упражнение 3.6*. Доказать свойство (i).

Аналогичным образом в 1984 г. Пансье доказал гипотезу Дежан для четырехбуквенного алфавита. Заметим, что только случаи $n = 3$, $n = 4$ являются исключениями из общего правила гипотезы: $RT(n) = n/(n-1)$. Докажем, что это нижняя оценка для $RT(n)$.

Предложение 3.1 (Дежан, 1972). $RT(n) \geq n/(n-1)$.

Доказательство. Докажем, что экспонента $n/(n-1)$ является n-неизбежной. Возьмем какое-либо слово U длины n над n-буквенным алфавитом. Если в нем найдутся две одинаковые

буквы $U(i)$ и $U(j)$, то фрагмент $U(i \dots j)$ имеет период $(j-i)$, а значит,

$$lexp(U) \geq exp(U(i \dots j)) \geq \frac{j-i+1}{j-i} \geq \frac{n}{n-1}.$$

Таким образом, слово длины n, в котором хотя бы две буквы совпадают, имеет локальную экспоненту не меньшую, чем $n/(n-1)$. Рассмотрим теперь произвольное слово V длины $n+2$ над тем же алфавитом. Если у него есть подслово длины n с двумя совпадающими буквами, то $lexp(V) \geq n/n-1$, как доказано выше. Если же такого подслова нет, то, поскольку число различных букв равно n, получаем $V(1) = V(n+1)$ и $V(2) = V(n+2)$. Но тогда V имеет период n и

$$lexp(V) \geq exp(V) \geq (n+2)/n > n/(n-1)\,.$$

Итак, никакое слово длины $n+2$ в n-буквенном алфавите не избегает экспоненту $n/(n-1)$. Следовательно, эта экспонента n-неизбежна. □

В справедливости гипотезы Дежан в общем случае никто не сомневается, однако общего доказательства до сих пор нет. Красивая идея, предложенная Пансье, так и не получила своего воплощения в общем случае из-за значительных технических трудностей. Однако эта идея заслуживает того, чтобы привести ее здесь.

Чтобы доказать гипотезу Дежан для n-буквенного алфавита Σ, достаточно построить над Σ ω-слово, локальная экспонента которого равна $RT(n) = n/(n-1)$. Если такое ω-слово существует, то оно обладает следующим свойством: в любом его подслове длины $n-1$ все буквы различны, а в любом его подслове длины n либо все буквы различны, либо первая совпадает с последней. Это свойство немедленно следует из того, что экспонента любого подслова не превосходит $RT(n)$ по определению локальной экспоненты.

Всякому ω-слову $\mathbf{T}$ над Σ, обладающему этим свойством, можно сопоставить его *характеристическую последовательность* — бинарное ω-слово $\mathbf{H}$ по следующему правилу:

$\mathbf{H}(k) = 0$, если $\mathbf{T}(k) = \mathbf{T}(k{+}n{-}1)$,

$\mathbf{H}(k) = 1$, если $\mathbf{T}(k) \neq \mathbf{T}(k{+}n{-}1)$.

Отметим, что всякая характеристическая последовательность определяет *единственное*, с точностью до переименования букв, ω-слово над Σ с рассматриваемым свойством. Тем самым свойства этого ω-слова над Σ, в том числе и его локальная экспонента, могут быть установлены на основании свойств характеристической последовательности. Таким образом, от ω-слов над произвольным алфавитом мы переходим к ω-словам над бинарным алфавитом, изучать которые неизмеримо проще. А доказательство гипотезы Дежан сводится к поиску морфизма над бинарным алфавитом такого, что порождаемое им ω-слово и будет характеристической последовательностью с требуемыми свойствами.

В заключение этого параграфа отметим любопытный факт: по определению величины $RT(n)$ ничего нельзя сказать о том, является ли сама экспонента $RT(n)$ n-избегаемой; однако если гипотеза Дежан справедлива, эта экспонента всегда n-неизбежна.

§8. Морфизмы, избегающие экспоненты

Скажем, что морфизм f избегает экспоненту β, если для любого U из условия «U избегает β» следует, что $f(U)$ избегает β. Ясно, что если над Σ существует нетривиальный (т. е. не отображающий Σ в себя) морфизм, избегающий экспоненту β, то β избегаема в Σ. Заметим, что доказательства избегаемости экспонент в предыдущем параграфе проводились именно по этому принципу. Морфизмы, избегающие заданную экспоненту, интересны и сами по себе; при этом, безусловно, наибольший интерес представляют морфизмы, избегающие любую экспоненту, превосходящую $RT(n)$ (про них можно сказать, что «они избегают все, что можно избежать»). Именно таким

является морфизм Туэ ϕ (см. §6), как показывают приводимые ниже предложения.

Предложение 3.2. *Пусть* $W \in A^+$ *и* $exp(W) > 1$. *Тогда*

$$exp(\phi(W)) = exp(W).$$

Доказательство. Пусть $p < |W|$ — минимальный период слова W. Тогда, очевидно, слово $\phi(W)$ имеет период $2p$; значит, $exp(\phi(W)) \geq exp(W)$. Предположим, что у $\phi(W)$ есть также период $t < 2p$. Пусть число t — четное. Для любого i $(2i{-}1)$-я буква в $\phi(W)$ совпадает с $(2i{+}t{-}1)$-й, а $(2i)$-я совпадает с $(2i{+}t)$-й. Но это означает, что $\phi(W(i)) = \phi(W(i{+}\frac{t}{2}))$, откуда $W(i) = W(i{+}\frac{t}{2})$; т. е. слово W имеет период $\frac{t}{2}$, меньший p; противоречие с минимальностью p. Следовательно, t — нечетное. Заметим, что равенство $t = 1$ невозможно по лемме 3.2, а значит, $t \geq 3$.

Представим слово $\phi(W)$ в виде произведения PQ, где $|P| = t$. Покажем, что слово Q не содержит подслов aa и bb. В самом деле, пусть Q содержит aa. Тогда на t позиций левее в слове $\phi(W)$ также находится подслово aa (по определению периода). Так как t — нечетно, то одно из этих подслов начинается в $\phi(W)$ с нечетной позиции, противоречие с леммой 3.2.

Итак, Q состоит из чередующихся букв. Таким образом, первая и $(2p{-}t{+}1)$-я буквы слова Q различны. Но эти две буквы суть $(t{+}1)$-я и $(2p{+}1)$-я буквы слова $\phi(W)$, которые по определению периода совпадают с первой буквой этого слова, а значит, и между собой. Полученное противоречие доказывает, что у слова $\phi(W)$ нет периодов, меньших $2p$, откуда следует требуемое утверждение. □

Предложение 3.3. *Пусть* $W \in A^+$ *и* $lexp(W) \geq 2$. *Тогда*

$$lexp(\phi(W)) = lexp(W).$$

Доказательство. Пусть $lexp(W) = \beta$. Из определений морфизма ϕ и локальной экспоненты очевидно следует неравенство $lexp(\phi(W)) \geq \beta$. Докажем, что при $\beta \geq 2$ справедливо и противоположное неравенство.

Предположим, что противоположное неравенство нарушается; выберем W такое, что $lexp(\phi(W)) > lexp(W) = \beta \geq 2$. Пусть V — максимальное по включению подслово в $\phi(W)$ такое, что $exp(V) > \beta$. Рассмотрим различные варианты расположения V в слове $\phi(W)$. Через V' при этом будем обозначать максимальное подслово слова V, начинающееся в $\phi(W)$ с нечетной позиции и заканчивающееся на четной (т. е. являющееся произведением блоков вида ab, ba).

Если $V = V'$, то слово $\phi^{-1}(V)$ существует и содержится в W. По предложению 3.2 экспонента последнего слова равна экспоненте V, противоречие с условием $lexp(W) = \beta$.

Пусть $V = V'c$, где c — буква, и $p = per(V)$. Поскольку $V' = \phi(Z)$ для некоторого подслова Z слова W (лемма 3.2), то из предложения 3.2 следует, что $exp(V') \leq \beta$. Отсюда $exp(V) > exp(V')$ и, значит, $p = per(V) = per(V')$, причем минимальный период слова V' четный по предложению 3.2. Таким образом, слово $Vc = V'cc$ не имеет периода p, так как слово V' не может содержать подслово cc, начинающееся с нечетной позиции (лемма 3.2). Тогда слово Vd, где d — буква, не совпадающая с c, имеет период p. Следовательно, $exp(Vd) > exp(V)$; осталось заметить, что $Vd \leq \phi(W)$. Противоречие с максимальностью V.

Случай $V = cV'$ $(c \in A)$ симметричен только что разобранному. В оставшемся случае $(V = cV'd)$ заметим, что пара условий $exp(V) > exp(V')$ и $exp(V) > 2$ обеспечивает совпадение минимальных периодов слов V и V'. Далее проводим то же рассуждение, что и в предыдущем абзаце. Тем самым доказательство предложения завершено. □

В гл. 4 мы еще вернемся к морфизму Туэ и докажем удивительный результат: морфизм Туэ — единственный (с точностью до возведения в степень и переименования букв) морфизм над бинарным алфавитом, избегающий все 2-избегаемые экспоненты.

А теперь перейдем к алфавитам больших размеров. Еще Туэ доказал удобный алгоритмический критерий для проверки

того, что заданный морфизм избегает экспоненту 2.

Теорема 3.5 (Туэ, 1912). *Пусть* $f : \Sigma^+ \to \Sigma^+$ *— морфизм, обладающий следующими свойствами:*
(1) если $|W| \le 3$ *и* $lexp(W) < 2$, *то* $lexp(f(W)) < 2$;
(2) если $a, b \in \Sigma$ *и* $f(a)$ *— подслово в* $f(b)$, *то* $a = b$.
Тогда f *избегает экспоненту 2.*

Для доказательства нам потребуется следующая лемма.

Лемма 3.3. *Пусть морфизм* f *удовлетворяет условиям (1),(2) теоремы 3.5,* $a \in \Sigma$, $W \in \Sigma^+$, $W = c_1 \dots c_n$, $f(c_i) = C_i$. *Тогда если* $f(W) = Xf(a)Y$, *то для некоторого* i *выполняется* $a = c_i$, $X = C_1 \dots C_{i-1}$, $Y = C_{i+1} \dots C_n$.

Доказательство. Рассмотрим все возможные расположения подслова $f(a)$ в слове $f(W) = C_1 \dots C_n$. Согласно условию (2), $f(a)$ не может быть собственным подсловом никакого блока C_i и не может содержать никакой блок C_i в качестве собственного подслова. Таким образом, $f(a)$ либо совпадает с некоторым блоком C_i, либо пересекает два соседних блока C_i и C_{i+1}. Показав, что второй случай невозможен, мы и докажем лемму.

Предположим, что $f(a)$ пересекает два соседних блока, как показано на рисунке. Тогда $C_i = PQ$, $f(a) = QR$, $C_{i+1} = RS$, где $P, Q, R, S \ne \lambda$.

$$f(W) = \underbrace{\overbrace{\ldots \mid P \quad Q \mid \quad R \quad S \mid \ldots}^{f(a)}}_{C_i \qquad\qquad C_{i+1}}$$

Рассмотрим слово $f(c_i a) = PQQR$. По условию (1) $lexp(c_i a) \ge 2$, т. е. $c_i = a$. Аналогично $a = c_{i+1}$ и, следовательно, $c_i = c_{i+1}$. Тогда получаем $|P| = |R|$ (поскольку $PQ = QR$), а значит, $P = R$ (поскольку $PQ = RS$). В итоге $PQ = QP$, откуда, согласно предложению 1.1, $PQ = Z^n$ для некоторого слова Z и некоторого $n > 1$. Таким образом, $lexp(f(c_i)) \ge 2$, противоречие с условием (1). Лемма доказана. □

Доказательство теоремы 3.5. Воспользуемся методом минимального контрпримера. Пусть W — кратчайшее бесквадрат-

ное слово над Σ такое, что $lexp(f(W)) \geq 2$. Тогда для некоторого непустого Y и подходящих (возможно, пустых) X и Z выполняется $f(W) = XYYZ$. Пусть $W = c_1 \dots c_n$ ($n \geq 4$ согласно условию (1)) и соответственно $f(W) = C_1 \dots C_n$. Заметим, что C_1 не может являться префиксом X: в этом случае мы могли бы отбросить первую букву в W и получить более короткий контрпример. Следовательно, X является собственным префиксом C_1; положим $C_1 = XC_1'$. Аналогично $C_n = C_n'Z$. Тогда

$$Y = C_1'C_2 \dots C_{i-1}C_i' = C_i''C_{i+1} \dots C_n',$$

где $C_i = C_i'C_i''$, причем $C_i'' \neq \lambda$. Заметим, что при выписанном нами представлении три блока — первый, i-й и последний — могут оказаться «раздробленными». Но поскольку $n \geq 4$, имеется блок, целиком расположенный внутри Y. Без ограничения общности пусть это блок C_2. Заметим, что это самое левое вхождение подслова $f(c_2)$ в Y (поскольку W — бесквадратное, $c_1 \neq c_2$; следовательно, $f(c_1c_2)$ не содержит других вхождений $f(c_2)$ по лемме 3.3). Рассмотрим самое левое вхождение подслова $f(c_2)$ во второй «экземпляр» Y. Нетрудно видеть, что это блок C_{i+1}. В самом деле, это некоторый блок C_j по лемме 3.3; если он расположен правее, чем C_{i+1}, то $C_i''C_{i+1}$ есть подслово в C_1'; поскольку $C_i'' \neq \lambda$, C_{i+1} является собственным подсловом C_1, противоречие с условием (2). Итак, мы получили, что $C_2 = C_{i+1}$ и $C_1' = C_i''$. Пользуясь условием (2), теперь легко получаем, что

$$\begin{cases} C_1' & = & C_i'', \\ C_2 & = & C_{i+1}, \\ C_3 & = & C_{i+2}, \\ & \dots & \\ C_{i-1} & = & C_{n-1}, \\ C_i' & = & C_n'. \end{cases} \tag{3.1}$$

Откуда, в частности, $n = 2i - 1$. Теперь рассмотрим слово $f(c_1c_ic_n) = XC_1'C_i'C_i''C_n'Z$. Из (3.1) следует, что между X и Z находится квадрат слова $C_1'C_i'$; согласно условию (1) левая

часть также содержит квадрат. Это означает, что $c_1 = c_i$ или $c_i = c_n$. Без ограничения общности предположим, что $c_1 = c_i$. Тогда из (3.1) и условия (2) мы получаем равенство $c_1 \dots c_{i-1} = c_i \dots c_{n-1}$. Противоречие с бесквадратностью W. Теорема доказана. □

§9. Отношение избегаемости

Сейчас мы дадим основное определение в теории избегаемости. Пусть Σ и Δ — два алфавита (не обязательно различных). Говорят, что слово $U \in \Sigma^+$ *избегает* слово $V \in \Delta^+$, если U не содержит подслов вида $h(V)$, где $h : \Sigma^+ \to \Delta^+$ — некоторый морфизм.

Замечание 3.2. $h(V) = h(V(1))\,h(V(2)) \dots h(V(k))$, где $k = |V|$. Таким образом, слово $h(V)$ можно воспринимать как результат применения к V подстановки фиксированных слов из Σ^+ вместо букв из Δ.

Основное определение согласуется с введенными ранее: так, например, слово U избегает целочисленную экспоненту k тогда и только тогда, когда U избегает слово x^k. Однако очевидно, что свойство избегать некоторое слово далеко не всегда можно выразить через свойство «избегать экспоненту», а свойство избегать экспоненту — через конечное множество свойств «избегать слово».

Упражнение 3.7. Доказать, что слово U является сильно бескубным (т. е. избегает любую экспоненту, большую двух) тогда и только тогда, когда U избегает слова x^3 и $xyxyx$.

Приводимая ниже лемма немедленно следует из основного определения.

Лемма 3.4. *Пусть слово U избегает слово V. Тогда*
(1) U избегает слово PVQ для любых P, Q;
(2) U избегает слово $h(V)$ для любого морфизма h;
(3) любое подслово U избегает V.

Упражнение 3.8. Доказать лемму 3.4.

На протяжении этой главы мы будем использовать еще несколько терминов. Слово вида $h(V)$, где h — морфизм, называют *гомоморфным образом* слова V. Кроме того, мы будем говорить, что слово U *изоморфно* слову V, если U можно получить из V переименованием букв (т. е. если U является образом V при изоморфизме). Слова из Δ^+ мы будем иногда называть *шаблонами*.

Основное определение естественным образом обобщается на множества слов, т. е. языки. Говорят, что язык $L \subseteq \Sigma^+$ *избегает* язык $\Lambda \subseteq \Delta^+$, если всякое слово $U \in L$ избегает все слова $V \in \Lambda$. Нетрудно заметить, что если язык L избегает язык Λ, то L можно дополнить до множества *всех* слов из Σ^+, избегающих все слова из Λ. В свою очередь, Λ можно дополнить до множества *всех* слов из Δ^+, избегаемых всеми словами из L. Данный параграф посвящен именно таким «максимальным» языкам.

Дадим несколько стандартных определений из универсальной алгебры. Пусть M — произвольное множество. Оператор f на множестве 2^M подмножеств из M называется *оператором замыкания*, если для любых $I, I_1, I_2 \in 2^M$ он удовлетворяет условиям

$$\begin{array}{ll} I \subseteq f(I) & \text{(возрастание)}, \\ (I_1 \subseteq I_2) \Longrightarrow (f(I_1) \subseteq f(I_2)) & \text{(монотонность)}, \\ f(f(I)) = f(I) & \text{(идемпотентность)}. \end{array}$$

Пусть на множестве 2^M задан оператор замыкания f. *Замкнутым множеством* называется подмножество I из 2^M, для которого $f(I) = I$. Множество замкнутых относительно данного оператора подмножеств из 2^M называется *системой замыканий* на множестве M.

Возьмем два произвольных алфавита Σ и Δ и рассмотрим бинарное отношение $\pi \subseteq \Sigma^+ \times \Delta^+$, состоящее из всех пар слов (U, V) таких, что U избегает V. Мы будем называть π отношением избегаемости. Отношению такого вида можно поставить в соответствие пару специальных функций, называемых *полярностями* отношения π. Одна из полярностей, обозначаемая $^{\rightarrow}$,

отображает множество языков над Σ во множество языков над Δ, а другая ($^{\leftarrow}$) — множество языков над Δ во множество языков над Σ. Для $L \subseteq \Sigma^+$, $\Lambda \subseteq \Delta^+$ положим по определению

$$L^{\rightarrow} = \{X \mid \forall U \in L \quad (U, X) \in \pi\},$$
$$\Lambda^{\leftarrow} = \{U \mid \forall X \in \Lambda \quad (U, X) \in \pi\}.$$

Языки $L^{\rightarrow}$ и $\Lambda^{\leftarrow}$ называются *полярами* языков L и Λ соответственно.

Упражнение 3.9. Доказать, что функции $^{\rightarrow\leftarrow}$ и $^{\leftarrow\rightarrow}$ являются операторами замыкания на множествах Σ^+ и Δ^+ соответственно, причем соответствующие этим операторам системы замыканий состоят в точности из поляров языков над Δ (соответственно над Σ).

Биекция между множествами замкнутых языков над Σ и Δ, определяемая полярностями, называется *соответствием Галуа* между этими множествами. Заметим, что множество всех слов из Σ^+, избегающих заданный язык $\Lambda \subseteq \Delta^+$, составляет замкнутое подмножество $\Lambda^{\leftarrow} \subseteq \Sigma^+$, а множество всех слов из Δ^+, избегаемых заданным языком $L \subseteq \Sigma^+$, составляет замкнутое подмножество $L^{\rightarrow} \subseteq \Delta^+$. Итак, данное соответствие Галуа содержит ответы на все вопросы типа «что избегает что?». Следующее предложение показывает, что замкнутые подмножества составляют хорошо известный в теории полугрупп класс подмножеств.

Подмножество I полугруппы S называется *идеалом*, если для любых $x \in I$, $a, b \in S$ элемент axb принадлежит I. (Это условие обычно записывают кратко в виде $SIS \subseteq I$.) Отметим, что идеалы представляют собой частный случай подполугрупп в S. Идеал I называется *вполне инвариантным*, если для любого $x \in I$ и любого гомоморфизма $h : S \rightarrow S$ элемент $h(x)$ принадлежит I.

Предложение 3.4.
(1) Язык $\Lambda \subseteq \Delta^+$ замкнут тогда и только тогда, когда он является вполне инвариантным идеалом полугруппы Δ^+.
(2) Язык $L \subseteq \Sigma^+$ замкнут тогда и только тогда, когда $\Sigma^+ \backslash L$ является вполне инвариантным идеалом полугруппы Σ^+.

Доказательство. (1) Как отмечалось выше, система замыканий на Δ^+ состоит из поляров языков над Σ. Таким образом, $\Lambda{=}L^{\rightarrow}$ для некоторого $L{\subseteq}\Sigma^+$. Пусть X — произвольное слово из Λ. Всякое слово из L избегает X и по лемме 3.4 избегает все его гомоморфные образы и все слова, его содержащие. По определению функции $^{\rightarrow}$ всякое слово из Δ^+, содержащее X или являющееся его гомоморфным образом, принадлежит Λ. Итак, Λ — вполне инвариантный идеал по определению.

(2) Так как система замыканий на Σ^+ состоит из поляров языков над Δ, имеем $L{=}\Lambda^{\leftarrow}$ для некоторого $\Lambda{\in}\Delta^+$. Пусть $U{\in}\Sigma^+{\setminus}L$. В этом случае

$$U = Ph(X)Q,$$

где $P,Q{\in}\Sigma^*$ и $h : \Delta^+{\to}\Sigma^+$ — гомоморфизм. Всякое слово, содержащее U, очевидно содержит $h(X)$ и, следовательно, принадлежит $\Sigma^+{\setminus}L$. Пусть f — произвольный гомоморфизм полугруппы Σ^+ в себя. Тогда слово $f(U)$ содержит $fh(X)$ и также принадлежит $\Sigma^+{\setminus}L$, поскольку fh является гомоморфизмом. Мы получили, что $\Sigma^+{\setminus}L$ — вполне инвариантный идеал по определению. □

§10. Избегаемые и неизбежные слова. Слова Зимина

Слово V называется *n-избегаемым*, если в алфавите из n символов найдется бесконечно много слов, избегающих V; в противном случае V называется *n-неизбежным*. *Избегаемым* называется слово, n-избегаемое для некоторого n, а *неизбежным* — слово, n-неизбежное при всех n.

Таким образом, все слова естественным образом разбиваются на классы: для каждого $n \geq 2$ можно рассмотреть класс n-избегаемых, но $(n{-}1)$-неизбежных слов; к указанным классам добавляется еще класс неизбежных слов. К настоящему времени, несмотря на усилия многих авторов, попрежнему неизвестно, существует ли общий алгоритм, позволяющий по произвольному слову определить, в какой класс

оно попадает. Более того, лишь про четыре класса известно, что они непусты: это класс 2-избегаемых слов (содержит x^3), класс 3-избегаемых 2-неизбежных слов (содержит x^2), класс 4-избегаемых 3-неизбежных слов (известный пример, приведенный в [1], представляет собой слово длины 14 над семибуквенным алфавитом) и класс неизбежных слов (содержит x). Тем не менее по произвольному слову можно определить, является ли оно избегаемым или неизбежным; в следующем параграфе мы укажем целых два алгоритмически проверяемых условия, эквивалентных неизбежности слова.

Одно из таких условий связано с «максимальными» неизбежными словами — словами Зимина, с которыми мы и познакомимся в этом параграфе. Начнем мы с простого предложения, «подсказавшего» конструкцию слов Зимина.

Предложение 3.5. *Слово xyx неизбежно.*

Доказательство. Рассмотрим произвольный конечный алфавит Σ и пусть $n = |\Sigma|$. Всякое слово $U \in \Sigma^+$ длины не меньшей, чем $2n+1$, содержит подслово вида cVc, где c — буква, а V — непустое слово. Но тогда $cVc = h(xyx)$ для морфизма h такого, что $h(x) = c$, $h(y) = V$. Таким образом, U не избегает xyx по определению. Это значит, что все слова над Σ, избегающие xyx, имеют длину, не превосходящую $2n$, а таких слов лишь конечное число. Итак, слово xyx n-неизбежно для произвольного n. □

Рассмотрим счетную серию слов над счетным алфавитом $\Delta = \{x_1, x_2, \ldots, x_n, \ldots\}$. (Подчеркнем, что каждое слово в серии является конечным, а значит, является словом над некоторым конечным подмножеством из Δ; вся же серия целиком может быть определена лишь над бесконечным алфавитом.)

$$\begin{aligned} Z_1 &= x_1, \\ Z_2 &= x_1x_2x_1, \\ &\ldots \\ Z_{n+1} &= Z_nx_{n+1}Z_n, \\ &\ldots \end{aligned}$$

Слова Z_n называются *словами Зимина.*

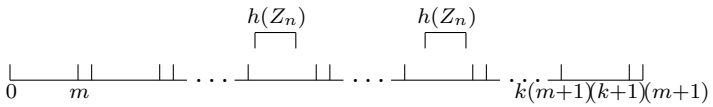
Предложение 3.6. *Слова Зимина неизбежны.*

Доказательство. Доказательство проведем по индукции. База индукции очевидна: любое слово над любым алфавитом является гомоморфным образом слова $Z_1 = x_1$. (Более того, в предложении 3.5 мы доказали, что слово Z_2 неизбежно.) Предположим, что слово Z_n неизбежно и докажем неизбежность Z_{n+1}.

Рассмотрим произвольный конечный алфавит Σ. По предположению индукции существует лишь конечное множество слов над Σ, избегающих Z_n. Значит, найдется такая константа m, что всякое слово над Σ, длина которого не меньше m, содержит подслово $h(Z_n)$ для некоторого морфизма h.

Теперь рассмотрим множество H всех морфизмов из $\{x_1, \ldots, x_n\}^+$ в Σ^+ таких, что образ любой буквы имеет длину, не превосходящую m. Это множество конечно, обозначим через k его мощность.

Возьмем произвольное слово над Σ длины $(k+1)(m+1)$ и разобьем его на блоки длины m, оставляя между соседними блоками промежуток в один символ.



Каждый блок содержит образ слова Z_n при некотором морфизме $h_i \in H$. Поскольку количество блоков превосходит k, хотя бы два блока содержат образы Z_n при одном и том же морфизме h. Таким образом, рассматриваемое слово содержит подслово $h(Z_n)Vh(Z_n)$, где $V \neq \lambda$. Определим морфизм $f : \{x_1, \ldots, x_n, x_{n+1}\}^+ \to \Sigma^+$ следующим образом: $f(x_i) = h(x_i)$ для всех $x = 1, \ldots, n$, $f(x_{n+1}) = V$. Тогда $f(Z_{n+1}) = h(Z_n)Vh(Z_n)$.

Тем самым мы доказали, что всякое слово над Σ длины $(m+1)(k+1)$ содержит образ слова Z_{n+1} при некотором морфизме. Следовательно, слово Z_{n+1} $|\Sigma|$-неизбежно. Поскольку

алфавит Σ — произвольный, слово Z_{n+1} неизбежно. Шаг индукции доказан. $\square$

Следующее предложение содержит некоторые свойства слов Зимина, необходимые для дальнейшего изложения.

Предложение 3.7.
1) $|Z_n| = 2^n - 1$;
2) всякая буква, расположенная в слове Z_n *на нечетной позиции, равна* x_1, *а всякая буква, расположенная в* Z_n *на четной позиции, не равна* x_1;
3) $Z_{n+1} = h(Z_n)$, *где* h *— морфизм, определяемый равенствами* $h(x_1) = Z_2$, $h(x_i) = x_{i+1}$ *для всех* $i \geq 2$;
4) если в слове Z_n *стереть все вхождения буквы* x_1, *то получится слово, изоморфное* Z_{n-1};
5) если слово Z_n *является собственным подсловом слова* $V \in \{x_1, \ldots, x_n\}^+$, *то* $lexp(V) \geq 2$ *и* V *является 3-избегаемым.*

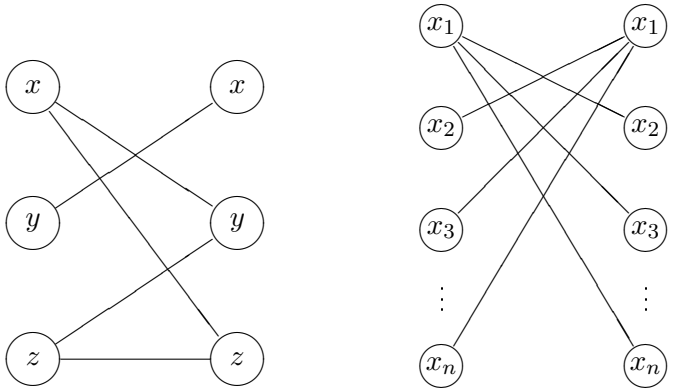
Упражнение 3.10. Доказать предложение 3.7.

§11. Связки и стирания. Теорема Бина – Эренфойхта – Макналти – Зимина

Пусть U — некоторое слово, $\Sigma = alph(U)$ — конечный алфавит. Пару множеств (B, C), где $B, C \subseteq \Sigma$, будем называть *связкой (относительно* U*)*, если для любого двухбуквенного подслова xy из U буква x принадлежит B тогда и только тогда, когда y принадлежит C.

Приведенное определение требует пояснений; проиллюстрируем его при помощи графов. Сопоставим слову U его *граф смежности*. Это двудольный граф, каждая доля которого проиндексирована множеством Σ. Вершина x из левой доли соединяется ребром с вершиной y из правой доли тогда и только тогда, когда в U есть подслово xy.

Пример 3.1. На рис. 4, *а* приведен граф смежности для слова $U = xyxzzyxz$, а на рис. 4, *б* — граф смежности для слова Зимина Z_n.



а *б*

Рис. 4. Графы смежности

Упражнение 3.11. Доказать, что пара (B, C) является связкой относительно U тогда и только тогда, когда для некоторого объединения компонент связности в графе смежности слова U множество B есть множество всех левых вершин, а C — множество всех правых вершин.

Упражнение 3.12. Найдите все связки для слов, графы смежности которых приведены на рис. 4.

Подмножество $A \subseteq \Sigma$ называется *свободным (относительно U)*, если $A \subseteq B \backslash C$ для некоторой связки (B, C). Таким образом, множество букв является свободным относительно U тогда и только тогда, когда его элементы не встречаются в U рядом.

Пример 3.2. Относительно слова $U = xyxzzyxz$ свободны множества $\{x\}$ и $\{y\}$; относительно слова Z_n свободны множество $\{x_1\}$ и любое подмножество из $\{x_2, \ldots, x_n\}$.

Назовем *стиранием множества* A следующее преобразование слова U: из последовательности букв слова U удаляются все элементы множества A. Полученная подпоследователь-

ность и является результатом стирания. Для этого преобразования будем использовать обозначение σ_A. Стирание множества A в слове U будем называть *свободным*, если A является свободным относительно U. Результат стирания множества A в слове U будем обозначать через U_A.

Пример 3.3. Множество $\{x_1\}$ свободно в Z_n; слово $(Z_n)_{\{x_1\}}$ изоморфно Z_{n-1}.

Последовательность $\sigma_{A_1}, \ldots, \sigma_{A_k}$ стираний в слове U будем называть последовательностью свободных стираний, если множество A_1 свободно в U, множество A_2 свободно в U_{A_1} и т. д.

Теперь мы можем сформулировать основную теорему.

Теорема 3.6 (Бин, Эренфойхт, Макналти, Зимин, 1979–1982). *Пусть U — произвольное слово, $n = |alph(U)|$. Тогда следующие условия эквивалентны:*
1) слово U неизбежно;
2) слово Зимина Z_n не избегает U;
3) существует последовательность свободных стираний в U, результат которой есть однобуквенное слово.

Мы приведем доказательство, предложенное М. В. Сапиром в [19]. Эквивалентность данных условий будем доказывать по циклу $1 \Rightarrow 3 \Rightarrow 2 \Rightarrow 1$. Импликация $2 \Rightarrow 1$ очевидна. В самом деле, если слово U — избегаемо, а Z_n содержит подслово $h(U)$ для некоторого морфизма h, то по лемме 3.4 избегаемо и слово Z_n, что противоречит предложению 3.6; значит, U неизбежно.

Две оставшиеся импликации нетривиальны. Мы начнем с более простой из них, импликации $3 \Rightarrow 2$. Проиллюстрировать эту импликацию достаточно легко. В результате последовательности свободных стираний у нас получилось однобуквенное слово x, изоморфное Z_1. Если на последнем шаге мы стерли букву y, то самое длинное слово, которое у нас могло быть перед последним шагом, есть yxy, изоморфное Z_2. Аналогично если перед этим мы стерли букву z, то у нас было «не больше» чем Z_3. Таким образом, исходное слово было «не больше» чем Z_n. Все же нужно строго доказать, что некоторый

гомоморфный образ каждого слова, встретившегося в последовательности стираний, сордержится в соответствующем Z_k.

$3 \Rightarrow 2$. Пусть U — произвольное слово, a — буква, не входящая в U. Нам потребуются следующие четыре слова, получаемые «насыщением» слова U при помощи буквы a:

$$\begin{aligned}
{}_0[U,a]_0 &= U(1)aU(2)a\ldots aU(|U|),\\
{}_1[U,a]_0 &= aU(1)aU(2)a\ldots aU(|U|),\\
{}_0[U,a]_1 &= U(1)aU(2)a\ldots aU(|U|)a,\\
{}_1[U,a]_1 &= aU(1)aU(2)a\ldots aU(|U|)a.
\end{aligned}$$

Заметим, что ${}_1[Z_k,a]_1$ изоморфно Z_{k+1}; более того, если V — подслово в Z_k, $\alpha,\beta\in\{0,1\}$, то ${}_\alpha[Z_k,a]_\beta$ изоморфно подслову в Z_{k+1}.

Пусть A — свободное подмножество относительно U, пара (B,C) — соответствующая связка и пусть h — гомоморфизм, определенный на $(alph(U)\backslash A)^+$. Определим гомоморфизм h^* на $(alph(U))^+$ следующим образом:

$$h^*(x)=\begin{cases} a, & x\in A,\\ {}_0[h(x),a]_0, & x\in C\backslash B,\\ {}_0[h(x),a]_1, & x\in B\cap C,\\ {}_1[h(x),a]_1, & x\in B\backslash(A\cup C),\\ {}_1[h(x),a]_0, & x\in alph(U)\backslash(B\cup C).\end{cases}$$

Докажем, что $h^*(U)={}_\alpha[h(U_A),a]_\beta$. Для этого нужно проверить, что на стыке h^*-образов двух букв из U всегда будет ровно одна буква a. По определению связки и свободного подмножества в слове U

– за буквой из B (т. е. из A, $B\cap C$ или $B\backslash(A\cup C)$) следует буква из C;

– за буквой не из B (т. е. из $C\backslash B$ или $alph(U)\backslash(B\cup C)$) следует буква не из C.

Требуемое условие немедленно следует из определения h^*.

Таким образом, если $h(U_A)$ содержится в Z_k, то $h^*(U)$ содержится в Z_{k+1}. Поскольку однобуквенное слово — результат последовательности свободных стираний в U — имеет гомоморфный образ, содержащийся в однобуквенном же сло-

ве Z_1, само слово U имеет гомоморфный образ в слове Z_m, где $(m-1)$ равно количеству стираний в последовательности; отсюда $m \leq |alph(U)| = n$. С учетом п. 3 предложения 3.7 рассматриваемая импликация доказана. □

$1 \Rightarrow 3$. Доказательство этой импликации требует нескольких вспомогательных утверждений. При этом соответствующие леммы удобнее формулировать и доказывать внутри общего доказательства.

Мы будем доказывать, что если не существует последовательности свободных стираний, сводящей U к однобуквенному слову, то U избегаемо. По слову U мы выберем подходящий алфавит A и построим морфизм γ над этим алфавитом такой, что для некоторого $a \in A$ все слова $\gamma^m(a)$ избегают U. Доказательство проведем методом минимального контрпримера.

Начнем с определения морфизма γ. Пусть r — натуральное число, а алфавит A состоит из r^2 символов a_{ij}, $1 \leq i, j \leq r$. Рассмотрим матрицу $r^2 \times r$, составленную из чисел $1 \ldots r$ следующим образом (рис. 5, *а*):

$$M = \begin{pmatrix} 1 & 1 & \ldots \\ \vdots & \vdots & \vdots \\ 1 & r & \ldots \\ 2 & 1 & \ldots \\ \vdots & \vdots & \vdots \\ 2 & r & \ldots \\ \vdots & \vdots & \vdots \\ r & 1 & \ldots \\ \vdots & \vdots & \vdots \\ r & r & \ldots \end{pmatrix} \qquad M_A = \begin{pmatrix} a_{11} & a_{12} & \ldots \\ \vdots & \vdots & \vdots \\ a_{11} & a_{r2} & \ldots \\ a_{21} & a_{12} & \ldots \\ \vdots & \vdots & \vdots \\ a_{21} & a_{r2} & \ldots \\ \vdots & \vdots & \vdots \\ a_{r1} & a_{12} & \ldots \\ \vdots & \vdots & \vdots \\ a_{r1} & a_{r2} & \ldots \end{pmatrix}$$

а *б*

Рис. 5. Построение морфизма γ

Каждый нечетный столбец матрицы M равен $1...1\ldots r...r$, а каждый четный — $1...r\ldots 1...r$. Заменим в M числа на буквы из A по следующему правилу: каждое число i в j-м столбце матрицы заменяется на a_{ij}. Строки полученной матрицы M_A (рис. 5, *б*), рассматриваемые как слова, и будут являться образами букв при отображении γ. Положим значение $\gamma(a_{ij})$ равным слову, записанному в строке M_A с номером $(i-1)r+j$. В дальнейшем изложении образы букв при γ будем называть *блоками*.

Приводимая ниже лемма следует из определения γ.

Лемма 3.5 *Морфизм γ обладает следующими свойствами.*
1. Длина любого блока равна r.
2. В каждом блоке все буквы различны.
3. На j-м месте в блоке стоит символ из множества $A_j = \{a_{1j}, \ldots, a_{rj}\}$.
4. Два различных блока не имеют одинаковых двухбуквенных подслов.

Упражнение 3.13. Доказать лемму 3.5.

Мы докажем, что при условии $r > 6n + 1$ (напомним, что $n = |alph(U)|$) все слова вида $\gamma^m(a_{11})$ избегают U. Предположим противное: существует m такое, что $\gamma^m(a_{11})$ содержит $h(U)$ для некоторого морфизма h. Рассмотрим минимальный контрпример: множество $alph(U)$ имеет минимально возможную для контрпримера мощность n и число m является минимальным для данного n.

Слова, являющиеся произведениями блоков, мы будем называть *целыми*. Пусть W — целое слово, V — подслово W. Тогда $V = P_1P_2P_3$, где либо P_1 является суффиксом блока (возможно, пустым), P_3 — префиксом блока (возможно, пустым), а P_2 есть произведение блоков, либо $P_1 = P_3 = \lambda$, а P_2 есть подслово блока, не являющееся ни суффиксом, ни префиксом.

Лемма 3.6. *Разбиение $V = P_1P_2P_3$ одинаково для всех вхождений V в W.*

Утверждение леммы 3.6 немедленно следует из свойств морфизма γ, доказанных в лемме 3.5.

В дальнейшем через W будем обозначать наименьшее целое подслово в $\gamma^m(a_{11})$, содержащее $h(U)$. Пусть $x \in alph(U)$. Тогда в соответствии с леммой 3.6 $h(x) = P_{x_1}P_{x_2}P_{x_3}$. Рассмотрим произвольный блок B из W. Ввиду минимальности W, B пересекается с U, а значит, может содержать некоторые слова вида P_{x_i} (заметим, что B может встречаться в W неоднократно). Согласно п. 2 леммы 3.5 вхождение каждого P_{x_i} в B единственно. Существует не более $3n$ различных слов P_{x_i}; таким образом, в B выбрано не более $3n$ различных подслов. Поскольку $|B| = r > 6n + 1$, то в B найдется двухбуквенное подслово, внутри которого нет границы никакого подслова P_{x_i} ($3n$ подслов имеют не более $6n$ различных границ).

Зафиксируем в B одно из двухбуквенных подслов с данным свойством и обозначим его через t_B. Отметим, что всякое слово P_{x_i} либо содержит t_B, либо не пересекается с ним. Кроме того, для разных блоков B_1 и B_2 слова t_{B_1} и t_{B_2} различны в силу п. 4 леммы 3.5. Таким образом, слово t_B является уникальной меткой блока B. Каждому слову t_B сопоставим новую букву y_B и рассмотрим слово W' над алфавитом $A \cup \{y_B |\ B$ — блок$\}$, полученное из W заменой каждого t_B на y_B:

$$W' = P_1 y_1 Q_1 \dots P_k y_k Q_k,$$

где P_i — префиксы блоков, Q_i — суффиксы блоков, и если $y_i = y_j$, то $P_i = P_j$ и $Q_i = Q_j$.

Определим морфизм g следующим образом: образ $g(x)$ буквы $x \in alph(U)$ есть слово $h(x)$, в котором каждое подслово t_B заменено на соответствующую букву y_B. Заметим, что $g(U)$ является подсловом в W'. В самом деле, всякое подслово t_B, находящееся внутри $h(U)$, целиком содержится в некотором слове $h(x)$; таким образом, переход от $h(U)$ к $g(U)$ есть в точности замена всех вхождений t_B на y_B.

Лемма 3.7. *Последовательность стираний $\sigma_{A_1}, \dots, \sigma_{A_r}$ в слове W' является последовательностью свободных стираний, ре-*

зультат которой есть слово $y_1 \dots y_k$. *Последовательность стираний* $\sigma_{A_1}, \dots, \sigma_{A_r}$ *в слове* $g(U)$ *является последовательностью свободных стираний, результат которой есть слово* $y_1 \dots y_k$, *кроме, быть может, первой и/или последней буквы.*

Доказательство этого утверждения почти очевидно: на каждом шаге, кроме последнего, за каждой буквой из стираемого множества A_j в текущем слове следует буква из A_{j+1} или буква y (это следует из п. 3 леммы 3.5). На последнем же шаге за буквой из стираемого множества A_r обязательно следует буква y (такая буква разделяет суффиксы двух любых соседних в W' блоков). Осталось заметить, что крайние буквы y_1 и y_k могут как входить, так и не входить в $g(U)$.

Свяжем теперь эту последовательность свободных стираний в $g(U)$ с последовательностью свободных стираний в самом слове U. Для этого потребуются еще две леммы.

Пусть V — произвольное слово, f — произвольный морфизм, определенный на $alph(V)$, и D — некоторое подмножество в $alph(f(V))$. Определим $D' \subseteq alph(V)$ следующим образом:

$$D' = \{x |\ alph(f(x)) \subseteq D\}.$$

На множестве $alph(V) \backslash D'$ определим морфизм f_D:

$$f_D(x) = (f(x))_D.$$

Напомним, что $V_{D'}$ есть результат стирания множества D' в слове V. Следующая лемма очевидна.

Лемма 3.8. $f_D(V_{D'}) = f(V)_D$.

Заметим также, что если в V найдутся две соседние буквы из множества D', то, поскольку f-образы этих букв состоят только из букв, принадлежащих D, в $f(V)$ найдутся две соседние буквы, принадлежащие D. Поскольку множество букв свободно в слове тогда и только тогда, когда никакие две буквы из этого множества не стоят в слове рядом, справедлива следующая лемма.

Лемма 3.9. *Если* D *свободно в* $f(V)$, *то* D' *свободно в* V.

Таким образом, последовательности $\sigma_{A_1}, \ldots, \sigma_{A_r}$ свободных стираний в $g(U)$ можно поставить в соответствие некоторую последовательность $\sigma_1, \ldots, \sigma_r$ свободных стираний в U. Результат этой последовательности стираний обозначим через $U_{A'}$. Заметим, что некоторые, а возможно и все, стирания в этой последовательности могут «стирать» пустое множество, так что возможна ситуация $U_{A'} = U$. Согласно лемме 3.8 получаем

$$g_A(U_{A'}) = (g(U))_A,$$

а по лемме 3.7 слово $g(U)_A$ есть $y_1 \ldots y_k$, быть может, без первой и/или последней буквы. Следовательно, к слову $g_A(U_{A'})$ можно применить морфизм α, отображающий каждую букву y_B в соответствующий блок B. В результате получится подслово в W (или само W). Таким образом, получаем, что слово $\alpha(g_A(U_{A'}))$ является подсловом в $\gamma^m(a_{11})$. По определению α, α-образ любого слова есть произведение блоков; таким образом, к нему можно применить отображение γ^{-1}. Следовательно, слово $\gamma^{-1}(\alpha(g_A(U_{A'})))$ является подсловом в $\gamma^{m-1}(a_{11})$ и одновременно образом слова $U_{A'}$ при морфизме $\gamma^{-1}\alpha g_A$.

Поскольку $|alph(U_{A'})| \leq |alph(U)| = n$ и $m-1 < m$, пара $(m-1, |alph(U_{A'})|)$ не дает контрпримеров по нашему предположению о минимальности пары (m, n). Поскольку $\gamma^{m-1}(a_{11})$ не избегает $U_{A'}$, существует последовательность свободных стираний в $U_{A'}$, результатом которой является однобуквенное слово. Соединив эту последовательность с последовательностью $\sigma_1 \ldots \sigma_r$, результатом которой является $U_{A'}$, получим последовательность свободных стираний в U, результатом которой является однобуквенное слово. Это противоречит исходному предположению. Тем самым импликация, а с ней и вся теорема, доказана. $\square$

§12. Бинарные 2-избегаемые слова

Несколько отступая от строгости изложения, можно сказать, что в двух предыдущих параграфах мы познакомились с наиболее «общеупотребительными» шаблонами — такими, без которых не обходится ни один достаточно длинный «текст» над любым конечным алфавитом. В данном параграфе, напротив, мы познакомимся с шаблонами, без которых вполне можно обойтись даже в минимальном нетривиальном алфавите. При этом мы ограничимся только бинарными шаблонами, т. е. содержащими не более двух различных букв. Полное описание множества 2-избегаемых бинарных слов дается в следующей теореме, обобщающей результаты, полученные несколькими авторами в 1989–1993 годах.

Теорема 3.7 (см. [6]). *Бинарное слово является 2-избегаемым тогда и только тогда, когда среди его гомоморфных образов нет подслов слов из множества*

$$M = \{xxyxx,\ xxyxy,\ xyxyy,\ xxyy,\ xyyx\}.$$

Таким образом, множество M состоит из всех максимальных (в смысле подсловного порядка) 2-неизбежных бинарных слов (шаблонов).

Доказательство этой теоремы состоит из простой и сложной части. Простая часть доказательства — проверить, что все шаблоны из M 2-неизбежны. Пример доказательства приведен на рис. 6 для шаблона $xyyx$. Будем строить префиксное дерево бинарных слов поярусно, начиная с корня (в силу симметрии можно предположить, что все слова начинаются с a). При построении удаляем узлы, помеченные словами, не избегающими W. Таким образом, в дереве остаются только узлы, помеченные словами, избегающими W; конечность дерева доказывает 2-неизбежность шаблона W.

Так, например, слово $abaaa$ не имеет в дереве потомков, поскольку слово $abaaaa$ содержит образ шаблона $xyyx$ при $x = y = a$, а слово $abaaab$ является образом этого шаблона при $x = ab$, $y = a$.

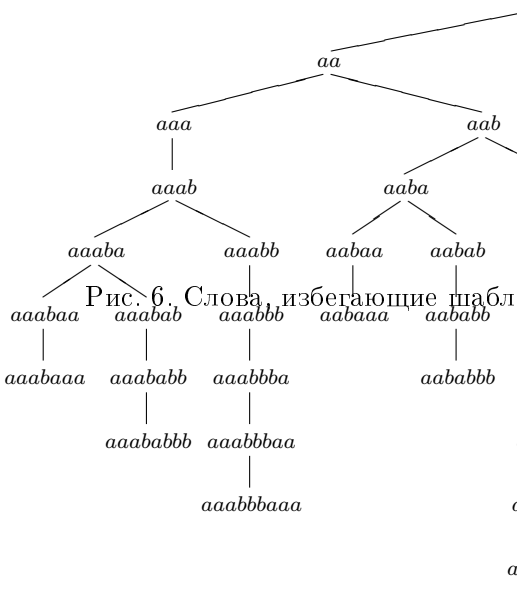


Рис. 6. Слова, избегающие шабл

Сложная часть доказательства состоит в проверке 2-избегаемости «минимальных» шаблонов, не входящих в M, т. е. таких, все собственные подслова которых являются подсловами слов из M. А именно, для каждого из шаблонов xxx, $xyxyx$, $xyxxy$, $xyyxy$, $xxyyx$, $xyyxx$, $xxyxyy$ нужно указать бесконечную последовательность слов, избегающих его. Как и в предыдущих параграфах, это делается при помощи морфизмов. Не приводя доказательств, укажем только морфизмы, на основе которых строятся бесконечные слова, избегающие требуемые шаблоны. Договоримся обозначать предел префиксной последовательности $\{f^n(a)\}$, где f — морфизм, a — буква, через $f^\infty(a)$. В построении бесконечных слов участвуют хорошо знакомый нам морфизм Туэ ϕ (см. §6) и еще четыре морфизма:

$$\begin{aligned} \mu:\ \{a,b,c\} &\to \{a,b,c\},\\ \mu(a) &= abc,\\ \mu(b) &= ac,\\ \mu(c) &= b; \end{aligned} \qquad \begin{aligned} \nu:\ \{a,b\} &\to \{a,b\},\\ \mu(a) &= aab,\\ \mu(b) &= bba; \end{aligned}$$

$$\begin{aligned} \pi:\ \{a,b,c\} &\to \{a,b\},\\ \pi(a) &= aa,\\ \pi(b) &= aba,\\ \pi(c) &= abbb; \end{aligned} \qquad \begin{aligned} \psi:\ \{a,b,c\} &\to \{a,b\},\\ \psi(a) &= aaa,\\ \psi(b) &= bbb,\\ \psi(c) &= ababab. \end{aligned}$$

Как известно из результатов §6, ω-слово Туэ – Морса U_∞, равное $\phi^\infty(a)$, избегает шаблоны xxx и $xyxyx$. Остальные шаблоны из списка избегаются следующим образом:

$\nu^\infty(a)$	избегает	$xxyxyy$;
$\psi(\mu_\infty(a))$	избегает	$xyxxy$;
$\pi(\mu_\infty(a))$	избегает	$xxyyx$.

Шаблон $xyyxy$ изоморфен «перевернутому» шаблону $xyxxy$, и, следовательно, его избегает ω^*-слово, получаемое «переворачиванием» $\psi(\mu_\infty(a))$. Аналогично ω^*-слово, получаемое «переворачиванием» $\pi(\mu_\infty(a))$, избегает шаблон $xyyxx$. Список минимальных избегаемых шаблонов исчерпан.

Глава 4. КОМБИНАТОРНЫЕ ХАРАКТЕРИЗАЦИИ ФОРМАЛЬНЫХ ЯЗЫКОВ

В этой главе иллюстрируются некоторые методы комбинаторного анализа языков. Мы будем рассматривать только *факториальные* языки. Формальный язык называется факториальным, если вместе с любым словом он содержит все его подслова (в английском языке — factors; этим и объясняется термин). Таким образом, если язык задан свойством, то он будет факториальным тогда и только тогда, когда это свойство наследуется подсловами. Отметим два типа таких свойств. Первый тип составляют свойства «не содержать подслов определенного вида» (в частности, избегать некоторый набор слов). Пример: язык бинарных сильно бескубных слов OF. Ко второму типу относится свойство «являться подсловом» (в некотором фиксированном слове, конечном или бесконечном; в какомлибо из слов заданного множества). Пример: язык Туэ – Морса TM.

§13. Комбинаторная сложность

Комбинаторной сложностью факториального языка L называется функция $c_L : \mathbb{N} \to \mathbb{N}$ такая, что $c_L(n)$ равно количеству слов длины n в языке L. Комбинаторную сложность также определяют для конечного или бесконечного слова W. В этом случае $c_W(n)$ равно количеству различных подслов длины n в слове W (т. е. комбинаторная сложность слова есть комбинаторная сложность языка подслов этого слова).

Предложение 4.1. *Для любого факториального языка L либо функция $c_L(n)$ ограничена, либо $c_L(n) \geq n+1$ для всех n.*

Мы обойдем технические трудности и докажем предложение только в частном случае — для комбинаторной сложности бесконечного слова.

Доказательство. Пусть $\mathbf{T}$ — бесконечное слово (ω-слово для определенности), L — множество его конечных подслов. Слово $V \in L$ будем называть *продолжением* слова $U \in L$, если $V = Ud$ для некоторой буквы d. Очевидно, что если U является подсловом в $\mathbf{T}$, то найдется такая буква d, что Ud также является подсловом в $\mathbf{T}$; следовательно, всякое $U \in L$ имеет хотя бы одно продолжение. Поскольку продолжения различных слов различны, то слов длины $n{+}1$ в языке L не меньше, чем слов длины n. Таким образом, мы доказали, что функция $c_L(n)$ не убывает.

Пусть существует m, для которого выполняется равенство $c_L(m) = c_L(m{+}1) = C$. Тогда всякое подслово в $\mathbf{T}$ длины m имеет ровно одно продолжение. Отсюда следует, что всякое подслово в $\mathbf{T}$ длины $m{+}1$ имеет ровно одно продолжение (если бы такое слово имело более одного продолжения, то его суффикс длины m также имел бы более одного продолжения). Значит, $c_L(m{+}2) = C$ и по индукции $c_L(n) = C$ для всех $n > m$. Тогда функция $c_L(n)$ ограничена константой C.

Таким образом, если функция $c_L(n)$ неограничена, то она строго возрастает. Кроме того, в этом случае $c_L(1) \geq 2$ (если $c_L(1) = 1$, то комбинаторная сложность тождественно равна единице). Итак, $c_L(1) \geq 1{+}1$ и $c_L(n{+}1) > c_L(n)$ для всех n. Значит, $c_L(n) \geq n{+}1$. Предложение доказано. □

Напомним, что бесконечное слово называется периодическим, если оно имеет период. Будем называть ω-слово *финально периодическим*, если у него есть суффикс, являющийся периодическим ω-словом. Следующее предложение описывает класс ω-слов с ограниченной комбинаторной сложностью.

Предложение 4.2. *ω-слово имеет ограниченную комбинаторную сложность тогда и только тогда, когда оно является финально периодическим.*

Доказательство. Пусть ω-слово $\mathbf{T}$ имеет ограниченную комбинаторную сложность. Как было доказано в предыдущем

предложении, это означает, что все подслова в $\mathbf{T}$, длина которых не меньше некоторого фиксированного n, имеют единственное продолжение. Пусть P_1 есть префикс длины n в $\mathbf{T}$. Определим последовательность $\{P_k\}$ по следующему правилу: для любого i слово P_{i+1} есть суффикс длины n продолжения слова P_i. Поскольку количество слов длины n конечно, найдутся такие $i < j$, что $P_i = P_j$. Тогда прямо из определения следует, что $P_{i+1} = P_{j+1}$, $P_{i+2} = P_{j+2}$ и т. д. Обозначим $p = j-i$. Тогда i-я буква в $\mathbf{T}$ (это первая буква в P_i) совпадает с $(i+p)$-й (первая буква в P_j), $(i+1)$-я буква в $\mathbf{T}$ совпадает с $(i+p+1)$-й и т. д. Мы получили, что суффикс ω-слова $\mathbf{T}$, начинающийся с i-й позиции, имеет период p по определению, т. е. $\mathbf{T}$ является финально периодическим.

Обратно, пусть $\mathbf{T}$ — финально периодическое ω-слово, а его суффикс, начинающийся с i-й позиции, имеет период p. Тогда периодический суффикс содержит не более p различных подслов одной длины. Тем самым комбинаторная сложность ω-слова $\mathbf{T}$ ограничена константой $i-1+p$. □

Класс бесконечных слов, имеющих минимально возможную неограниченную комбинаторную сложность $c_W(n) = n+1$, известен в комбинаторике под именем *слов Штурма*. Мы приведем пример слова Штурма, порождаемого морфизмом очень простого вида. *Словом Фибоначчи* называется ω-слово, являющееся пределом префиксной последовательности $\{F_n\}$, где $F_n = \psi^n(a)$, а ψ — морфизм, определяемый равенствами

$$\psi(a) = ab, \qquad \psi(b) = a.$$

Заметим, что дли́ны слов F_n образуют в точности числовую последовательность Фибоначчи.

Предложение 4.3. *1)* $F_n = F_{n-1}F_{n-2}$*;*
2) $per(F_n) = |F_{n-1}|$*;*
3) слово Фибоначчи не является финально периодическим.

Упражнение 4.1*. Доказать предложение 4.3.

(«Звездочка» относится к п. 2; п. 1 тривиален, а п. 3 следует из п. 2.)

Из последнего пункта предложения 4.3 следует, что комбинаторная сложность слова Фибоначчи не меньше чем $n+1$. Докажем, что имеет место равенство.

Предложение 4.4. *Слово Фибоначчи имеет комбинаторную сложность, равную* $n+1$.

Пусть F — язык подслов слова Фибоначчи. Сделаем предварительные замечания.

Замечание 4.1. Слова bb и aaa не принадлежат F.

Замечание 4.2. Отображение ψ^{-1} можно применить к любому слову над $\{a, b\}$, начинающемуся с a и не содержащему подслов bb и aaa.

Упражнение 4.2. Доказать замечания 4.1 и 4.2.

Доказательство предложения 4.4 опирается на следующую лемму.

Лемма 4.1. *Не существует слова* X *такого, что* $aXa \in F$ *и* $bXb \in F$.

Доказательство. Воспользуемся методом минимального контрпримера. Пусть X — минимальное по длине слово такое, что $aXa, bXb \in F$. Тогда $X \neq \lambda$, $X \neq a$, первая и последняя буквы в X обе равны a (все условия следуют из замечания 4.1). Поскольку слово Фибоначчи начинается с a, то слово bXb не может быть его префиксом; следовательно, $abXb \in F$ ($bbXb \notin F$ по замечанию 4.1). Согласно замечаниям 4.1 и 4.2 к слову $abXb$ можно применить отображение ψ^{-1}. При этом условие $abXb \in F$ означает, что слово $abXb$ содержится в некотором F_n; тогда слово $\psi^{-1}(abXb)$ содержится в F_{n-1}, а значит, также принадлежит F. Положим $X = Ya$:

$$\psi^{-1}(abXb) = \psi^{-1}(abYab) = a\psi^{-1}(Y)a.$$

Итак, слово $a\psi^{-1}(Y)a$ принадлежит F. Теперь рассмотрим слово $aXa = aYaa$. Согласно замечаниям 4.1 и 4.2 слово $aXab$

принадлежит F и к нему применимо отображение ψ^{-1}:

$$\psi^{-1}(aXab) = \psi^{-1}(aYaab) = [\text{поскольку } Y(1) = a] = b\psi^{-1}(Y)ba.$$

Дублируя рассуждение, приведенное для bXb, получаем, что $b\psi^{-1}(Y)b \in F$. В результате слово $\psi^{-1}(Y)$ обладает тем же свойством, что и X. Поскольку $|\psi^{-1}(Y)| < |X|$, получаем противоречие с минимальностью X. □

Доказательство предложения 4.4. Докажем требуемое утверждение по индукции. База индукции: очевидно, $c_F(1) = 2$. Предположение индукции: пусть $c_F(k) = k{+}1$ для всех $k \leq n$. Шаг индукции: докажем, что $c_F(n{+}1) = n{+}2$. Слово $U \in F$ назовем *свободным*, если $Ua, Ub \in F$. Нетрудно видеть, что значение $c_F(k{+}1)$ равно $c_F(k)$ плюс количество свободных слов длины k (это количество не равно нулю, так как слово Фибоначчи не является финально периодическим). Таким образом, предположение индукции означает, что для каждого $k = 1, \ldots, n{-}1$ имеется ровно одно свободное слово длины k. Рассмотрим слова длины n. Для доказательства шага индукции необходимо показать, что существует ровно одно свободное слово длины n. Если слово длины n свободно, то его суффикс длины $n{-}1$ также свободен. Но такой суффикс — единственный по предположению индукции, обозначим его через X. Если в F имеется лишь одно слово длины n с суффиксом X, то оно, очевидно, и будет единственным свободным словом длины n. Если же оба слова aX и bX принадлежат F, то хотя бы одно из слов aXa, bXb не принадлежит F по лемме 4.1; следовательно, лишь одно из слов aX, bX является свободным. Шаг индукции доказан. □

Напомним, что запись $f(x) = O(g(x))$ означает, что существуют константы $0 < C_1 \leq C_2$ такие, что для всех x из области определения f выполнено двойное неравенство

$$C_1 \cdot g(x) \leq f(x) \leq C_2 \cdot g(x).$$

Следующая теорема из [8] дает исчерпывающее описание классов функций, представители которых могут быть комбинаторными сложностями ω-слов, порожденных морфизмами.

Теорема 4.1. *Комбинаторная сложность ω-слова, порожденного морфизмом, равна $O(1)$, либо $O(n)$, либо $O(n \log\log n)$, либо $O(n \log n)$, либо $O(n^2)$. Существует алгоритм, эффективно определяющий по виду морфизма, к какому классу принадлежат функции сложности ω-слов, порожденных этим морфизмом.*

Упражнение 4.3*. Доказать, что $c_{TM}(n) = O(n)$.

Существуют эффективные способы конструирования ω-слов, не использующие морфизмы. Содержательные примеры можно найти в [8]. Отметим здесь следующие факты. Во-первых, над любым алфавитом Σ существует ω-слово с комбинаторной сложностью $O(|\Sigma|^n)$. Например, такой сложностью будет обладать ω-слово $X_1X_2 \ldots X_n \ldots$, где X_i — i-е слово над Σ в каком-нибудь фиксированном лексикографическом порядке. Кроме того, с вероятностью 1 такой сложностью будет обладать случайная последовательность алфавитных символов (хотя это уже не относится к эффективным методам построения). Во-вторых, существуют нетривиальные комбинаторные приемы, позволяющие строить ω-слова сложности $O(n^k)$ для любого натурального k, сложности $O(\alpha^n)$ для любого рационального α, $1 < \alpha \leq |\Sigma|$, и даже промежуточной сложности, такой как, например, $O(\alpha^{\sqrt{n}})$ (наиболее новые результаты в этом направлении приведены в [7]).

В заключение параграфа отметим два результата о комбинаторной сложности факториальных языков, описываемых не бесконечными словами, а свойством избегаемости. Для языка CF бинарных бескубных слов комбинаторная сложность экспоненциальна, т. е. имеет вид $O(\alpha^n)$. Наилучшая известная оценка $\alpha \approx 1,455$ и один из общих методов получения таких оценок приведен в [25]. Что касается языка OF бинарных сильно бескубных слов, то его комбинаторная сложность полиномиальна, но ведет себя нерегулярно. Так, существуют константы

C_1, C_2 такие, что

$$C_1 \cdot n^{1,22} \leq c_{OF}(n) \leq C_2 \cdot n^{1,37},$$

но при этом предел

$$\lim_{n\to\infty} \frac{c_{OF}(n)}{n^\alpha}$$

не существует ни при каком α (см. [8]).

§14. Язык бинарных сильно бескубных слов

Важность языка OF всех бинарных сильно бескубных слов и интерес к нему исследователей неоднократно отмечались на протяжении гл. 3, поэтому здесь мы сразу приступим к формулировке и решению конкретных проблем. Представленный ниже список вопросов, разумеется, не полон, но достаточно представителен. Итак, попытаемся дать «портрет» языка OF.

1. Что избегает? Что представляет собой с точки зрения теории избегаемости?
Ответ. Избегает вполне инвариантный идеал, порожденный словами xxx и $xuxux$. Является замкнутым множеством, а именно — дополнением указанного идеала.
Ссылки. Упражнение 3.7, предложение 3.4.

2. Какова его комбинаторная сложность?
Ответ. Полиномиальна, но не может быть записана в виде $O(n^\alpha)$ ни для какого α; существуют константы C_1 и C_2 такие, что

$$C_1 n^{1,22} \leq c_{OF}(n) \leq C_2 n^{1,37}.$$

Ссылки. §13, [8].

3. Какими морфизмами сохраняется?
Ответ. Будет получен ниже, следствие 4.1.
Ссылки. [20, 21].

4. Какова внутренняя структура слов?
Ответ. Будет дано описание через слова из языка Туэ – Морса, теорема 4.2.
Ссылки. [21].

5. Что изменится, если от слов перейти к Z-словам?
Ответ. Языки OF и TM совпадут (следствие 4.2). Описание языка TM^Z дано в следующем параграфе.
Ссылки. [12, 29].

Большинство недостающих деталей «портрета» восполняют сформулированная ниже теорема и ее следствия.

Теорема 4.2. *Произвольное сильно бескубное слово W может быть представлено в виде $W = XYZ$, где*
1) $Y \in TM$;
2) $X, Z \in A^*$ и выполнены ограничения

$$\begin{aligned} |X| &\leq |W|/4, \\ |Z| &\leq |W|/4, \\ |X| + |Z| &\leq 3 \cdot |W|/10. \end{aligned}$$

В связи с формулировкой теоремы возникает естественный вопрос: можно ли что-нибудь сказать о структуре слов X и Z? Так как X и Z сильно бескубны, то они, очевидно, устроены так же, как и целое слово W. Заметим еще, что величины $|X|$ и $|Z|$ зависимы, как следует из последнего неравенства теоремы.

Доказательство теоремы проведем достаточно схематично, оставляя проверку ряда деталей в качестве упражнений. Доказательство основано на анализе результатов работы двух алгоритмов обработки сильно бескубных слов — простого (алгоритм А) и модифицированного (алгоритм A^M). Определение этих алгоритмов и обсуждение их свойств, в свою очередь, основаны на ряде вспомогательных утверждений, с которых мы и начнем.

В дальнейшем рассматриваются только слова над алфавитом $A = \{a, b\}$. Напомним, что два слова называются *изоморфными*, если они получаются друг из друга переименованием букв. Слово W будем называть *равномерным*, если оно допускает разбиение $W = cQ_1 \ldots Q_n d$, где $c, d \in \{a, b, \lambda\}$, а каждое Q_i равно ab или ba. Равномерные слова — это в точности слова, обладающие следующим свойством: к ним нужно добавить не более чем по одной букве в конец и/или в начало, чтобы

получилось слово из $\phi(A^+)$, где ϕ — морфизм Туэ – Морса. В свою очередь, ко всякому слову из $\phi(A^+)$ можно применить преобразование ϕ^{-1}, сокращающее длину слова вдвое и сохраняющее при этом многие важные свойства согласно предложениям 3.2, 3.3. В дальнейшем минимальное по длине слово из $\phi(A^+)$, содержащее данное равномерное слово W, будем обозначать через $\xi(W)$. Если таких слов два, как, например, для слова $W = aba$, договоримся брать то из них, для которого W является суффиксом (т. е. в данном примере $\xi(W) = baba$).

Замечание 4.3. Слово равномерно тогда и только тогда, когда его подслова вида c^2, где c — буква, начинаются либо всегда с четной, либо всегда с нечетной позиции.

Следующее утверждение дает критерий равномерности сильно бескубного слова.

Предложение 4.5. *Сильно бескубное слово W неравномерно тогда и только тогда, когда его префикс и/или суффикс изоморфен $aabaa$.*

Доказательство. Достаточность следует из определения равномерного слова; докажем необходимость. Выберем в W подслово c^2Vd^2 такое, что $c, d \in A$, $|V| = 2k-1$ и k — минимальное из возможных.

Случай 1: $k > 1$. Заметим, что V не содержит сегментов aa, bb в силу минимальности k. Следовательно, буквы a и b чередуются в V. Очевидно, что $V(1) \neq c$, $V(2k-1) \neq d$. В итоге W содержит подслово cVd, состоящее по крайней мере из пяти чередующихся букв. Получили противоречие с условием $W \in OF$. Случай 1 невозможен.

Случай 2: $k = 1$. Подслово c^2Vd^2 в этом случае изоморфно $aabaa$. Оно обязано являться префиксом или суффиксом W, так как в противном случае W содержит подслово, изоморфное aaa или $baabaab$ и не является сильно бескубным. Предложение доказано. □

Нам потребуются два вспомогательных утверждения, ха-

рактеризующих слова, квадраты которых входят в сильно бескубные слова и слова Туэ – Морса.

Предложение 4.6 *Пусть* $XX \in OF$. *Тогда* $|X| = 2^k$ *или* $|X| = 3 \cdot 2^k$ *для подходящего* $k \geq 0$.

Доказательство. Вначале покажем, что X не может иметь нечетную длину, большую 3. От противного: пусть $|X| = 2k+1$, $k \geq 2$. Так как слово X сильно бескубно, то X содержит подслово c^2, где c — буква. Следовательно, слово XX неравномерно, так как подслово c^2 начинается в нем как с четной, так и с нечетной позиции ввиду нечетности $|X|$. По предложению 4.5 получаем, что XX имеет префикс или суффикс $aabaa$ ($bbabb$). Пусть XX имеет префикс $aabaa$. Так как $|X| \geq 5$, то этот префикс является префиксом X, причем, очевидно, собственным. Тогда XX имеет подслово $aabaa$ «внутри»:

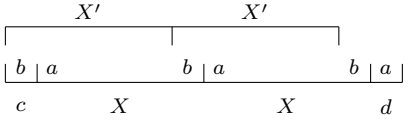
$$XX = \quad \underbrace{aabaa \qquad\qquad}\ \underbrace{aabaa \qquad\qquad}$$

Значит, $XX \notin OF$ (ср. доказательство предложения 4.5, случай 2), получили противоречие.

Теперь рассмотрим случай, когда число $|X|$ четно. Снова рассуждая от противного, предположим, что X — слово минимальной длины такое, что $|X| = m \cdot 2^k$, $m \geq 5$, $k \geq 1$ и $XX \in OF$. Заметим, что XX — равномерное слово (иначе по предложению 4.5 получим, как и в предыдущем абзаце, подслово $aabaa$ «внутри» XX). Таким образом, можно рассмотреть слово $\xi(XX)$. Возможны два случая.

Случай 1: $XX = \xi(XX)$. Так как $|X|$ — четное число, то получаем $X \in \phi(A^+)$, так как $XX \in \phi(A^+)$. Следовательно, сильно бескубное слово $\phi^{-1}(XX)$ является квадратом слова $\phi^{-1}(X)$, имеющего длину $m \cdot 2^{k-1}$, $m \geq 5$. Противоречие с минимальностью X.

Случай 2: $XX \neq \xi(XX)$. Так как $|XX|$ и $|\xi(XX)|$ — четные числа, то $\xi(XX) = cXXd$, где c и d — буквы. Всякий блок длины 2, начинающийся в $\xi(XX)$ с нечетной позиции, есть ab или ba. В предположении $X(1)=a$ имеем

$\xi(XX) =$ 

Покажем, что $X'X' \in OF$. Пусть это не так, тогда слово $X'X'$ имеет префикс $bYbYb$ ($Y \in A^*$), причем $YbYb = XX_1$, где X_1 — непустой префикс X (последнее условие есть следствие того факта, что слово $X'b$ содержится в XX). Если $|Y|$ четна, то $|Yb|$ нечетна и ≥ 5, так как $|X| \geq 10$; согласно доказанному выше, XX не может содержать $YbYb$, противоречие. Если же $|Y|$ нечетна, то XX имеет префикс $YbYbY(1)$ (по определению ξ), снова получили противоречие. Итак, $X'X' \in OF$, при этом $|X'| = |X|$ и $\xi(X'X') = X'X'$, т. е. второй случай мы свели к первому. Предложение доказано. □

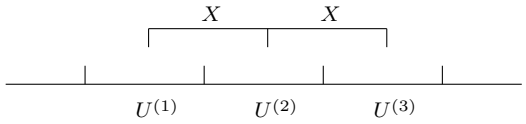
Упражнение 4.4. Доказать, что никакое слово XX не является префиксом (суффиксом) слова U_n.

Предложение 4.7. *Пусть $XX \in TM$. Если $|X| = 2^k$, то X изоморфно U_k. Если $|X| = 3 \cdot 2^k$, то X изоморфно $U_k V_k U_k$.*

Доказательство. Слово XX содержится в некотором U_n. Будем изображать их взаимное расположение на рисунках.

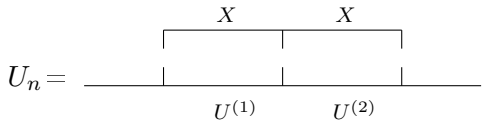
Пусть $|X| = 2^k$. Разобьем U_n на блоки длины 2^k. По определению U_n все они будут равны U_k или V_k.

Случай 1.

$U_n =$ 

Блоки $U^{(1)}$ и $U^{(2)}$ имеют одинаковые суффиксы и, следовательно, совпадают; блоки $U^{(2)}$ и $U^{(3)}$ имеют одинаковые префиксы и также совпадают. Случай 1 невозможен, так как U_n избегает x^3.

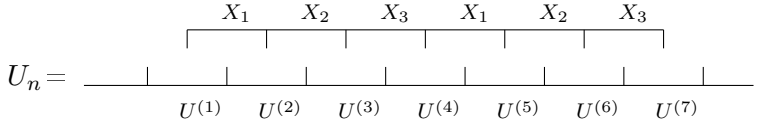
Случай 2.



Требуемое условие выполнено.

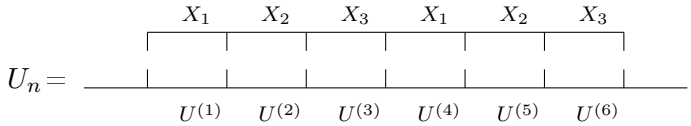
Пусть теперь $|X| = 3 \cdot 2^k$. Разобьем U_n и X на блоки длины 2^k.

Случай 1.



Блоки $U^{(1)}$ и $U^{(4)}$, $U^{(2)}$ и $U^{(5)}$, $U^{(3)}$ и $U^{(6)}$, $U^{(4)}$ и $U^{(7)}$ совпадают. Так как U_n избегает $xyxyx$, случай 1 невозможен.

Случай 2.



В данном случае каждый из блоков X_i равен U_k или V_k; при этом либо $X_1 \neq X_2$, либо $X_2 \neq X_3$. Пусть $X_1 \neq X_2$. Тогда $X_3 = X_1$ либо $X_3 = X_2$. Если $X_3 = X_2$, то блок $U^{(6)}$ не заканчивает слово U_n согласно упражнению 4.4. При этом следующий блок не может равняться ни X_1, ни X_2, поскольку U_n сильно бескубно. Таким образом, $X_3 \neq X_2$, т. е. $X_3 = X_1$. Предположив $X_2 \neq X_3$, мы получим $X_3 = X_1$ симметричным рассуждением. Предложение доказано. □

Теперь определим основной алгоритм.

Алгоритм А. (Преобразует сильно бескубное слово W длины ≥ 2 в слово W_A.)
Шаг 0. Положить $W_0 = W$, $k = 0$.
Шаг 1. Если $W_k \in \{ab, ba\}$ *или* W_k *неравномерно, то положить* $W_A = W_k$ *и остановиться.*

Шаг 2. Положить $W_{k+1} = \phi^{-1}(\xi(W_k))$, $k = k+1$; *вернуться на шаг 1.*

Лемма 4.2. *Алгоритм A останавливается на любом исходном* $W \in OF$.

Упражнение 4.5. Доказать лемму 4.2.

Лемма 4.3. *Пусть* $W_k \notin OF$. *Тогда* $W_k = XXc$, *где* $c = X(1)$ *и* $XX \in OF$.

Доказательство. Предположим противное: $W_k = PXXcQ$, где $c = X(1)$ и без ограничения общности $Q \neq \lambda$. Тогда

$$\xi(W_{k-1}) = \phi(W_k) = \phi(P)\phi(X)\phi(X)cd\phi(Q),\ c \neq d.$$

Применяя к слову W_{k-1} оператор ξ, мы добавили не более чем по одной букве в начало и в конец; следовательно, можно записать $W_{k-1} = P'X'X'dQ'$, где X' — слово, получаемое из $\phi(X)$ перестановкой первой буквы в конец слова, $d = X'(1)$, а $Q' \neq \lambda$. Проведя данное рассуждение для каждого из k шагов работы алгоритма, получим, что и исходное слово $W = W_0$ не является сильно бескубным. Противоречие. □

Лемма 4.4. *Пусть* $W_k = PYQ$, $P, Q \in A^+$ *и* $|Y|/|W_k| = \alpha$. *Тогда найдутся такие* $P', Q' \in A^+$, *что* $W = P'\phi^k(Y)Q'$ *и* $|\phi^k(Y))|/|W| \geq \alpha$.

Доказательство. Выполняется

$$\xi(W_{k-1}) = \phi(W_k) = \phi(P)\phi(Y)\phi(Q),$$

где $|\phi(P)| \geq 2$ и $|\phi(Q)| \geq 2$. Следовательно, $W_{k-1} = P_1\phi(Y)Q_1$, где $P_1, Q_1 \in A^+$ и

$$\frac{|\phi(Y)|}{|W_{k-1}|} \geq \frac{|\phi(Y)|}{|\phi(W_k)|} = \frac{|Y|}{|W_k|} = \alpha.$$

Требуемое утверждение следует отсюда по индукции. □

Лемма 4.5. *Для слова W_A имеет место одно из следующих равенств (с точностью до изоморфизма).*

$1) W_A = ab$, $4) W_A = aabaa$,
$2) W_A = aaa$, $5) W_A = aabaabZ,\ Z \in A^*$,
$3) W_A = abbabba$, $6) W_A = Zabbabb,\ Z \in A^*$.

При этом $W_A = ab$ тогда и только тогда, когда $W \in TM$.

Упражнение 4.6. Доказать лемму 4.5.
Указание: использовать предложение 4.5 и лемму 4.3.

Требуемое в теореме разложение $W = XYZ$ будем искать в зависимости от вида слова W_A. Если W_A имеет вид 1) из леммы 4.5, то, согласно этой же лемме, можно взять $X = Z = \lambda$. Следующие две леммы определяют разложение W в случаях, когда W_A имеет вид 2) или 3).

Лемма 4.6. *Пусть $W_A = aaa$. Тогда $W = XYZ$, где $X, Y, Z \in TM$ и либо $X = \lambda$, $|Z| \leq |W|/8$, либо $Z = \lambda$, $|X| \leq |W|/8$.*

Доказательство. Пусть $W_A = W_t$. Используя лемму 4.3 и определение алгоритма А, получаем, что W_{t-1}=$ababa$, а для W_{t-2} справедливо одно из следующих равенств:

1) $W_{t-2} =$ $bbaabbaa$;
2) $W_{t-2} = abbaabbaa$;
3) $W_{t-2} =$ $bbaabbaab$.

Случай 1. Имеем $W = X'Y'Z'$, где X' является суффиксом $\phi^{t-2}(b)$, Z' является префиксом $\phi^{t-2}(a)$, $Y' = \phi^{t-2}(baabba)$ и оба слова $X'Y'$,$Y'Z'$ принадлежат TM. Пусть $|X'| \geq |Z'|$. Возьмем λ в качестве X; $X'Y'$ в качестве Y; Z' в качестве Z. Тогда требуемое условие выполнено. Случай $|Z'| \geq |X'|$ симметричен разобранному.

Случай 2. Имеем $W = X'Y'Z'$, где X' — суффикс $\phi^{t-2}(ab)$, Z' — префикс $\phi^{t-2}(a)$, $Y' = \phi^{t-2}(baabba)$, $X'Y' \in TM$ и $|X'| \geq 2^{t-2} \geq |Z'|$. Взяв соответственно λ,$X'Y'$ и Z' в качестве X,Y и Z, мы получаем требуемый результат.

Случай 3 симметричен случаю 2. □

Лемма 4.7. *Пусть* $W_A = abbabba$. *Тогда* $W{=}XYZ$, *где* $X, Y, Z \in TM$ *и либо* $X = \lambda$, $|Z| \leq |W|/4$, *либо* $Z = \lambda$, $|X| \leq |W|/4$.

Доказательство. Как и выше, пусть $W_A = W_t$. Пользуясь леммой 4.3 и определением алгоритма A, мы получаем, что одно из следующих равенств выполняется для слова W_{t-1}:

1) $W_{t-1} = \ bbabaabbabaa$;
2) $W_{t-1} = abbabaabbabaa$;
3) $W_{t-1} = \ bbabaabbabaab$.

Случай 1. Имеем $W = X'Y'Z'$, где X' — суффикс $\phi^{t-1}(bba)$, Z' — префикс $\phi^{t-1}(baa)$, $Y' = \phi^{t-1}(baabba)$ и оба слова $X'Y'$, $Y'Z'$ принадлежат TM. Пусть $|X'| \geq |Z'|$. Возьмем λ в качестве X; $X'Y'$ в качестве Y; Z' в качестве Z; требуемое условие выполнено. Случай $|Z'| \geq |X'|$ симметричен разобранному.

Случай 2. Имеем $W{=}X'Y'Z'$, где X' — суффикс $\phi^{t-1}(abba)$, Z' — префикс $\phi^{t-1}(baa)$, $Y' = \phi^{t-1}(baabba)$, $X'Y' \in TM$ и $|X'| \geq 3{\cdot}2^{t-1} \geq |Z'|$. Взяв соответственно λ,$X'Y'$ и Z' в качестве X,Y и Z, мы получим требуемый результат.

Случай 3 симметричен случаю 2. □

В случае когда слово W_A имеет вид 5) или 6) из леммы 4.5, мы не можем немедленно указать требуемое разбиение; вместо этого мы определим модификацию исходного алгоритма — алгоритм A^M, который можно применять и к словам вида $aabaaZ$, $Zaabaa$.

Алгоритм $\mathbf{A}^{\mathbf{M}}$. (Преобразует сильно бескубное слово W длины ≥ 2 в слово W_{A^M}.)
Шаг 0. Положить $W_0 = W$, $k = 0$.
Шаг 1. Если $W_k \in \{ab, ba, aaa, bbb, abbabba, baabaab\}$, *то положить* $W_{A^M} = W_k$ *и остановиться.*
Шаг 2. Если слово W_k *имеет префикс (суффикс)* $aabaa$ *или* $bbabb$, *то стереть первую и/или последнюю букву; полученное слово обозначить через* $\overline{W}_k$.

Шаг 3. Положить $W_{k+1} = \phi^{-1}(\xi(\overline{W}_k))$, $k = k+1$; *вернуться на шаг 1.*

Упражнение 4.7. Доказать леммы 4.2, 4.3 и 4.4, заменив алгоритм A на алгоритм A^M.

Итак, если слово W_A имеет вид 4), 5) или 6) из леммы 4.5, применим к слову W алгоритм A^M и попытаемся восстановить структуру W по виду слова W_{A^M}.

Из упражнения 4.7 и леммы 4.5 следует, что с точностью до изоморфизма слово W_{A^M} имеет один из следующих видов:

$W_{A^M} = ab$,
$W_{A^M} = aaa$,
$W_{A^M} = abbabba$.

Для поиска разбиения W в первом случае нам понадобятся два свойства слов Туэ – Морса, доказательство которых мы оставляем в качестве упражнения.

Лемма 4.8. *Пусть* $W \in TM$, $W = P_1P_2Z$, *где* P_i *изоморфно* U_k, *а* $|Z| \geq 2^k$. *Тогда префикс* Z *длины* 2^k *изоморфен* U_k.

Лемма 4.9. *Пусть слово* $W \in TM$ *имеет префикс* P_1P_2 *и суффикс* Q_1Q_2, *где* P_i *изоморфно* U_t, *а* Q_i *изоморфно* U_s, *причем* $t \geq s > 0$ *и* $|W| \geq 3 \cdot 2^t$. *Тогда* $|W|$ *делится на* 2^s.

Упражнение 4.8. Доказать леммы 4.8 и 4.9.

Вернемся к слову W_{A^M}.

Лемма 4.10 *Если* $W_{A^M} = ab$, *то* $W = XYZ$, *где*
1) $Y \in TM$;
2) $|X| < |W|/4$, $|Z| < |W|/4$, $|X| + |Z| \leq 3 \cdot |W|/10$.

Доказательство. Пусть $W_{A^M} = W_m$. Проанализируем работу алгоритма A^M. Легко видеть, что все слово W за исключением, возможно, некоторого префикса и некоторого суффикса, «потерявшихся» при удалении крайних букв в неравномерных словах W_t, содержится в $U_{m+1}{=}\phi^m(ab)$. Возьмем в качестве X упомянутый префикс, в качестве Z — упомянутый суффикс и в качестве Y — оставшееся подслово слова W. Первое утверждение леммы очевидно выполняется.

Пусть t' — наибольшее среди чисел t таких, что алгоритм A^M удаляет первую букву в слове W_t. Предполагая без ограничения общности, что в слове $W_{t'}$ была удалена буква a, мы получаем

$$W = XU_{t'}V_{t'}U_{t'}S\,,$$

где $X, S \in A^+$; и если $W_{t'} \neq aabaa$, то S содержит $U_{t'}$ в качестве префикса. Так как $|X| \leq |U_{t'}|$, мы имеем $|X| < |W|/4$. Симметричное рассуждение для конца слова W позволяет найти в W суффикс вида $TU_{s'}V_{s'}U_{s'}Z$ или $TV_{s'}U_{s'}V_{s'}Z$ и доказать требуемое утверждение для $|Z|$. Заметим, что слово $W_{s'}$ не может быть равным $aabaa$ или $bbabb$, поскольку в этом случае алгоритм удаляет первую букву в слове. Таким образом, слово T содержит в качестве суффикса блок $U_{s'}$ в первом варианте и блок $V_{s'}$ во втором.

Осталось доказать справедливость оценки для $|X|+|Z|$. Предположим, что мы получили, пользуясь вышеприведенными рассуждениями, префикс $XU_tV_tU_t$ и суффикс $U_sU_sV_sU_sZ$ или $V_sV_sU_sV_sZ$ для данного слова W. Если $|W| > 4 \cdot 2^t + |X|$, то W имеет префикс $XU_tV_tU_tU_t$ (см. выше замечание о префиксе слова S). Мы предполагаем без ограничения общности, что $t \geq s$. Если указанные нами префикс и суффикс не пересекаются в W, то, очевидно, выполняется неравенство $|X|+|Z| < |W|/4$. Пусть они пересекаются. Используя результаты лемм 4.8 и 4.9, а также сильную бескубность слова W, мы найдем для каждого s максимальное по длине пересечение.

1) $s = t$. Это равенство исключает случай $W_t = aabaa$ ($W_t = bbabb$).

$$W = \underbrace{XU_sV_s\overbrace{U_sU_s}^{\text{суффикс}}V_sU_sZ}_{\text{префикс}}$$

$$|X| \leq 2^s,\ |Z| \leq 2^s,\ |Y| = 6 \cdot 2^s;$$
$$|X| + |Z| \leq |W|/4.$$

2) $s = t - 1$.

$$W = \underbrace{XU_sV_sV_s\overbrace{U_sU_sV_sU_s}^{\text{суффикс}}Z}_{\text{префикс}}$$

$$|X| \le 2 \cdot 2^s,\ |Z| \le 2^s,\ |Y| = 7 \cdot 2^s;$$
$$|X| + |Z| \le 3 \cdot |W|/10.$$

Слово $W = bbabbaabaa$ ($X = bb$, $Z = a$) очевидно является кратчайшим, на котором достигается равенство

$$|X| + |Z| = 3 \cdot |W|/10\ .$$

3) $s \le t-2$. Рассматриваемые префикс и суффикс при этом условии пересекаются не более чем по трем блокам U_s (V_s), так как, с одной стороны, среди последних четырех таких блоков в блоке U_t или V_t есть в точности два блока U_s и два блока V_s; с другой стороны, первые четыре таких блока в рассматриваемых суффиксах равны либо $U_sU_sV_sU_s$, либо $V_sV_sU_sV_s$. Таким образом, максимальное возможное значение величины $(|X|+|Z|)/|W|$ в данном случае меньше, чем в предыдущем.

Доказательство леммы завершено. □

Лемма 4.11. *Если* $W_{A^M} = aaa$, *то* $W = XYZ$, *где* $Y \in TM$ *и* $|X| + |Z| \le |W|/4$.

Лемма 4.12. *Если* $W_{A^M} = abbabba$, *то* $W = XYZ$, *где* $Y \in TM$ *и* $|X| \le |W|/4$, $|Z| \le |W|/4$, $|X| + |Z| \le 7 \cdot |W|/24$.

Упражнение 4.9*. Доказать леммы 4.11 и 4.12, воспользовавшись доказательствами лемм 4.6 и 4.7.

Итак, в леммах 4.5–4.7, 4.10–4.12 мы исследовали все возможные случаи для результата применения алгоритма А к сильно бескубному слову W. Во всех случаях для W выполнено заключение теоремы 4.2. Таким образом, теорема 4.2 доказана.

Отметим два следствия доказанной теоремы. Первое из них является базовым результатом о сильно бескубных Z-словах, принадлежащим Готшалку и Хэдлунду (см. [12]).

Следствие 4.1. *Всякое конечное подслово сильно бескубного Z-слова принадлежит языку TM.*

Доказательство очевидно.

Кроме того, теорема 4.2 позволяет получить в качестве следствия результат Сиболда (см. [20]) о морфизмах, сохраняющих множество OF.

Следствие 4.2. *Множество морфизмов полугруппы A^+, сохраняющих язык OF, исчерпывается отображениями вида $\theta\phi^n$, где θ — изоморфизм полугруппы A^+, ϕ — морфизм Туэ – Морса, а $n \geq 0$.*

Доказательство. Пусть морфизм $f : A^+ \to A^+$ переводит множество OF в себя. Рассмотрим произвольное слово $U \in TM$. Существуют сколь угодно длинные слова P и Q такие, что $V = PUQ \in TM$. Выберем P и Q так, чтобы каждое из них содержало не меньше вхождений буквы a и буквы b, чем слово U. Имеем $f(V) \in OF$. Применив к слову $f(V)$ теорему 4.2, получаем, что центральная часть этого слова, а именно $f(U)$, заведомо лежит в TM. Таким образом, мы показали, что f сохраняет множество TM.

Так как $f(aa) = f(a)f(a) \in TM$, то по предложению 4.7 слово $f(a)$ изоморфно либо U_k, либо $U_kV_kU_k$ для подходящего k. То же самое справедливо для $f(b)$, $f(ab)$ и $f(ba)$, так как слова $f(bb), f(abab), f(baba)$ принадлежат TM. Легко проверить (перебором), что это возможно лишь когда одно из слов $f(a)$, $f(b)$ изоморфно U_k, а другое — V_k для одного и того же k. А это означает, что $f = \theta\phi^k$. Следствие доказано. □

§15. Z-язык Туэ – Морса TM^Z

Как было показано в предыдущем параграфе, $OF^Z = TM^Z$, т. е. Z-слово является сильно бескубным тогда и только тогда, когда всякое его конечное подслово содержится в некотором слове Туэ – Морса. Поскольку, как уже упоминалось, язык

Туэ – Морса имеет линейную комбинаторную сложность, естественно было бы ожидать, что язык TM^Z также «невелик» (например, имеет лишь счетную мощность). Однако, если вглядеться пристально, можно увидеть, что это не так. Вглядимся.

Бинарное Z-слово будем представлять себе как числовую прямую, на которой «расставлены» буквы алфавита (ср. §5): каждому отрезку $[i{-}1, i]$, где i — целое число, поставлена в соответствие i-я буква Z-слова. Определим элементарные операции над Z-словами: операция *автоморфизма* Aut (замена всех букв a на b и наоборот), операция *реверса* Rev (симметричное отражение Z-слова относительно точки 0) и набор операций *сдвига* $Shift_k$, k — целое (сдвиг всего Z-слова на k единиц вправо по числовой прямой). Два Z-слова будем называть *изоморфными*, если одно из них можно получить из другого с помощью элементарных операций. В этом параграфе мы перечислим все элементы множества TM^Z с точностью до изоморфизма и, в частности, убедимся, что их — континуум.

Пусть дано Z-слово $\mathbf{W}$. Разбиение $\mathbf{W}$ на блоки U_n и V_n (если такое существует) будем называть n-разбиением (тривиальное 0-разбиение существует для всякого Z-слова). Будем называть целочисленную точку x *граничной* для данного разбиения, если ближайшие к ней буквы $\mathbf{W}$ принадлежат разным блокам. *Порядком* точки x относительно $\mathbf{W}$ ($deg_{\mathbf{W}}(x)$) будем называть максимум из чисел n таких, что x является граничной для некоторого n-разбиения $\mathbf{W}$, либо ∞, если максимум не существует. Очевидно, что для произвольной точки x условие $deg_{\mathbf{W}}(x) \geq n$ выполнено тогда и только тогда, когда $deg_{\mathbf{W}}(x + k \cdot 2^n) \geq n$ для любого k.

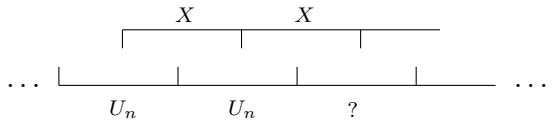
Предложение 4.8. *Для любого* $\mathbf{W} \in TM^Z$ *и любого натурального* n *существует единственное* n*-разбиение* $\mathbf{W}$.

Доказательство. Согласно предложению 4.5 всякое подслово из $\mathbf{W}$ равномерно, т. е. допускает разбиение на блоки $U_1 = ab$ и $V_1 = ba$; но тогда и само $\mathbf{W}$ можно разбить на такие блоки.

Рассмотрим Z-слово $\mathbf{W}'$, получаемое из $\mathbf{W}$ операцией ϕ^{-1}:

каждому блоку Q_i построенного разбиения $\mathbf{W}$ соответствует буква $\phi^{-1}(Q_i)$ в $\mathbf{W}'$. (Для однозначности определения потребуем, чтобы первой букве $\mathbf{W}'$ соответствовал блок, содержащий первую букву $\mathbf{W}$.) Очевидно, $\mathbf{W}' \in TM^Z$. По доказанному выше, существует 1-разбиение $\mathbf{W}'$, которое, в свою очередь, индуцирует 2-разбиение $\mathbf{W}$. Рассуждая по индукции, получаем, что существует n-разбиение $\mathbf{W}$ для всех n.

Предположим, что для некоторого n существуют два n-разбиения $\mathbf{W}$. Рассмотрим ω-подслово $\mathbf{W}$, начинающееся с граничной точки одного из разбиений двумя одинаковыми блоками:



Как видно из рисунка, следующий за U_nU_n блок выбранного разбиения начинается с той же буквы, что и U_n (т. е. равен ему). Противоречие с сильной бескубностью $\mathbf{W}$. Утверждение доказано. □

Следствие 4.3. *Пусть* $\mathbf{W} \in TM^Z$, x *— точка порядка* $\geq n$ *и* $k < 2^n$. *Тогда* $deg_{\mathbf{W}}(x+k) < n$.

Доказательство. Предположим, что $deg_{\mathbf{W}}(x+k) \geq n$. По определению порядка точки Z-слово $\mathbf{W}$ допускает два n-разбиения. Противоречие с предложением 4.8. □

Следствие 4.4. *Пусть* $\mathbf{W} \in TM^Z$. *Тогда из двух соседних точек порядка* $\geq n$ *одна имеет порядок* n, *а другая — порядок* $> n$.

Доказательство. Все точки, отстоящие от выбранных на расстояние, меньшее 2^n, имеют порядок, меньший n (следствие 4.3). Таким образом, ровно одна из выбранных точек будет являться граничной для $(n+1)$-разбиения $\mathbf{W}$. □

Следствие 4.5. Z-слово $\mathbf{W} \in TM^Z$ имеет не более одной точки порядка ∞.

Доказательство. С учетом следствия 4.4 — очевидно. □

Z-слово $\mathbf{W}$ будем называть Z-словом *бесконечного (конечного) порядка*, если $\mathbf{W}$ имеет (соответственно не имеет) точку бесконечного порядка.

Предложение 4.9. *Во множестве TM^Z существует в точности два неизоморфных Z-слова бесконечного порядка.*

Доказательство. Наличие в Z-слове точки порядка ∞ означает, что для любого $n \geq 0$ отрезок этого слова длины 2^n, примыкающий справа или слева к этой точке, равен U_n или V_n. Таким образом, справа от данной точки стоит, по определению, ω-слово U_∞ или V_∞, а слева — ω^*-слово ${}_\infty U$ или ${}_\infty V$. Перенесем точку порядка ∞ нашего Z-слова в 0 с помощью подходящей операции $Shift$ и заметим, что

$$Aut({}_\infty U U_\infty) = {}_\infty V V_\infty,$$
$$Aut({}_\infty U V_\infty) = Rev({}_\infty U V_\infty) = {}_\infty V U_\infty,$$

а из ${}_\infty U U_\infty$ нельзя получить ${}_\infty U V_\infty$ с помощью элементарных операций. Доказательство завершено. □

Построим пример сильно бескубного Z-слова конечного порядка при помощи так называемого «метода раскачивания».

Возьмем числовую прямую и начнем расставлять на ней буквы. Пусть $x_0 = \frac{1}{3}$, $W_0 = a = \mathbf{W}(1)$.

i-й шаг: если $W_{i-1} = U_{i-1}$, то дописываем к нему V_{i-1}; если $W_{i-1} = V_{i-1}$, то дописываем к нему U_{i-1}; дописывание производим с того конца, который ближе к точке x_0. Полученный блок U_i или V_i и будет являться словом W_i.

Таким образом,

$W_1 = \mathbf{W}(0)\mathbf{W}(1) = ba$,
$W_2 = \mathbf{W}(0 \ldots 3) = baab$,
$W_3 = \mathbf{W}(-4 \ldots 3) = abbabaab$ и т. д.

Предельным результатом описанного процесса будет Z-слово $\mathbf{W}$, каждое подслово которого лежит в некотором W_i, т. е. принадлежит TM. Значит, мы действительно построили элемент TM^Z.

Точка, являющаяся серединой отрезка, занимаемого W_i $(i > 0)$, имеет по построению порядок $n = i - 1$. Докажем это по индукции. Точка $x = 0$, являющаяся серединой отрезка, занятого W_1, имеет порядок 0, так как отрезок длины 2, находящийся справа от нее, равен aa. Шаг индукции: пусть x,y — середины отрезков, занимаемых W_i и W_{i+1} соответственно. Тогда $|x-y| = 2^{i-1}$, т. е. x и y суть соседние точки порядка $\geq 2^{i-1}$ (по предположению индукции). По следствию 4.4 $deg_{\mathbf{W}}(y) \geq i$. Так как при построении W_{i+2} рядом с y оказываются два одинаковых блока (это два средних блока в слове, например, $U_{i+2} = U_iV_iV_iU_i$), то окончательно $deg_{\mathbf{W}}(y) = i$.

Середина отрезка, занимаемого W_i, находится на расстоянии $\frac{2^n}{3}$ от точки x_0, а значит, является ближайшей к x_0 точкой порядка $\geq n$ по следствию 4.3. В итоге имеем

$$\forall k\ \forall x\ (deg_{\mathbf{W}}(x) \geq k) \Longrightarrow (|x - x_0| \geq \tfrac{2^k}{3}).$$

Следовательно, Z-слово $\mathbf{W}$ не имеет точки порядка ∞.

Произвольному Z-слову $\mathbf{U} \in TM^Z$ поставим в соответствие две последовательности $\{y_n\}_0^\infty$ и $\{r_n\}_0^\infty$ по следующим правилам. Ближайшая к $x_0 = \frac{1}{3}$ точка порядка $\geq n$ относительно $\mathbf{U}$ есть y_n, а $r_n = |x_0 - y_n|$. Для всякого n реализуется одна из двух возможностей.

1) $deg_{\mathbf{U}}(y_n) \geq n + 1$. Тогда $y_{n+1} = y_n$ и соответственно $r_{n+1} = r_n$.

2) $deg_{\mathbf{U}}(y_n) = n$. Тогда по следствию 4.4 точки, отстоящие от y_n на расстояние 2^n, имеют порядок $\geq n+1$. Таким образом, одна из этих точек есть y_{n+1}; т. е. $|y_{n+1} - y_n| = 2^n$. Поскольку $r_k < 2^{k-1}$ для любого k (неравенство строгое, так как r_k — не целое число), то $r_{n+1} = 2^n - r_n$.

Таким образом, $\{r_n\}$ однозначно кодируется бинарной последовательностью $\{h_n\}_0^\infty$ по правилу

$$r_{n+1} = 2^n - r_n \quad \Longrightarrow \quad h_n = 1,$$
$$r_{n+1} = r_n \quad \Longrightarrow \quad h_n = 0$$

(можно записать $h_n = \operatorname{sgn}|r_{n+1} - r_n|$). Будем называть $\{h_n\}$ характеристической последовательностью слова $\mathbf{U}$. На языке характеристических последовательностей мы и сформулируем основную теорему данного параграфа.

Теорема 4.3. *(1) Всякая бинарная последовательность с бесконечным числом единиц является характеристической для некоторого сильно бескубного Z-слова конечного порядка.*

(2) Два сильно бескубных Z-слова конечного порядка изоморфны тогда и только тогда, когда их характеристические последовательности почти совпадают (т. е. различаются в конечном числе точек).

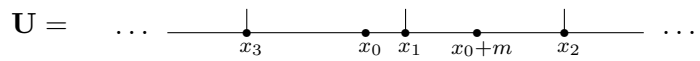
Следствие 4.6. *Множество TM^Z содержит континуум попарно неизоморфных элементов.*

Доказательство следствия 4.6. Рассмотрим множество всех бинарных последовательностей (их континуум) и отношение эквивалентности на нем: две последовательности эквивалентны тогда и только тогда, когда существует номер N, начиная с которого они совпадают. Последовательности можно рассматривать как ω-слова; каждое ω-слово, совпадающее с данным с N-й буквы, однозначно определяется своим префиксом длины $(N-1)$; таким образом, класс эквивалентности данного слова биективно отображается на A^*, т. е. счетен. Следовательно, классов эквивалентности — континуум. Согласно п. (2) теоремы 4.3 требуемое утверждение доказано. □

Доказательство теоремы 4.3. Нам будет удобнее доказать вначале п. (2). Пусть $\mathbf{U} \neq \mathbf{V} \in TM^Z$. Доказательство естественным образом разбивается на несколько этапов. Сначала докажем необходимость.

Этап 1. Если $\mathbf{V}{=}Aut(\mathbf{U})$, то порядки всех точек относительно $\mathbf{U}$ и $\mathbf{V}$ совпадают, т. е. совпадают и характеристические последовательности этих Z-слов.

Этап 2. Покажем, что если $\mathbf{V}{=}Shift_m(\mathbf{U})$, то существует номер N, начиная с которого последовательности $\{h_n^{\mathbf{U}}\}$ и $\{h_n^{\mathbf{V}}\}$ совпадают. Пусть $k = \max\{deg_{\mathbf{U}}(x) \mid x{\in}[x_0, x_0{+}m]\}$. Рассмотрим точку x_1 порядка k из отрезка $[x_0, x_0{+}m]$, а также точки x_2, x_3 — ближайшие к x_1 точки порядка $\geq k$. Согласно следствию 4.4 одна из них (пусть это x_2) имеет порядок $(k{+}1)$, а другая — порядок $\geq(k{+}2)$. В итоге картина такова (с точностью до переименования x_2 и x_3):

$\mathbf{U} =$ 

Видно, что x_3 является ближайшей точкой порядка $\geq k{+}2$ как к точке x_0, так и к точке $x_0{+}m$. Пусть $deg(x_3) = j$. Тогда ближайшей к x_0 и $x_0{+}m$ точкой порядка $(j{+}1)$ будет точка $x_3{+}2^j$. Рассуждая по индукции, получаем, что $y_n^{\mathbf{U}} = y_n^{\mathbf{V}} + m$ при всех $n \geq k{+}2$. Следовательно,

$$|y_{n+1}^{\mathbf{U}} - y_n^{\mathbf{U}}| = |y_{n+1}^{\mathbf{V}} - y_n^{\mathbf{V}}|,$$

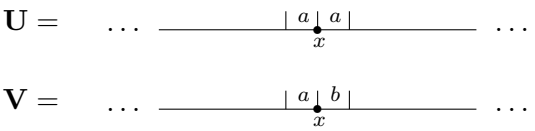
откуда $h_n^{\mathbf{U}} = h_n^{\mathbf{V}}$. Таким образом, можно взять $N = k + 2$.

Этап 3. На этапе 2 мы фактически показали, что характеристическая последовательность Z-слова и аналогичная последовательность, построенная исходя из произвольной точки x_0', совпадают, начиная с некоторого элемента (при этом нам не было важно, что $|x_0{-}x_0'|$ — целое число). Рассмотрим Z-слово $Rev(\mathbf{U})$. Очевидно, что, взяв за исходную для построения точку $x_0' = -\frac{1}{3}$, мы получим последовательность, совпадающую с $\{h_n^{\mathbf{U}}\}$. Следовательно, характеристические последовательности Z-слов $\mathbf{U}$ и $Rev(\mathbf{U})$ совпадают, начиная с некоторого номера N. Отметим, что для всех $n \geq \mathbb{N}$ выражения $y_n^{\mathbf{U}} - x_0$ и $y_n^{Rev(\mathbf{U})} - x_0$ имеют противоположные знаки.

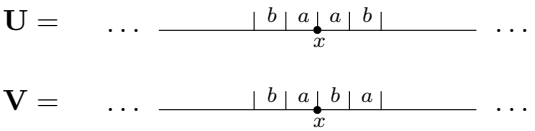
Итак, мы проверили требуемые условия для всех элементарных операций, т. е. доказательство необходимости закончено. Докажем достаточность.

Этап 4. Предположим, что $\{h_n^{\mathbf{U}}\} = \{h_n^{\mathbf{V}}\}$, и покажем, что $\mathbf{V} = Aut(\mathbf{U})$.

В самом деле, из равенства характеристических последовательностей следует, что порядки всех точек относительно $\mathbf{U}$ и $\mathbf{V}$ также равны. Если условие $\mathbf{V} = Aut(\mathbf{U})$ не выполняется, то найдется такая точка x, что в $\mathbf{U}$ и $\mathbf{V}$ либо совпадают буквы с номером x и не совпадают буквы с номером $(x{+}1)$, либо наоборот. Без ограничения общности можно считать, что имеет место следующая картина:

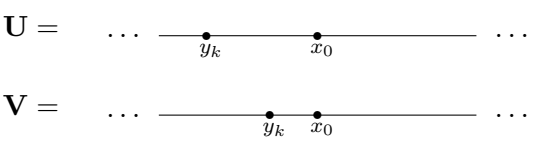


По определению $deg_{\mathbf{U}}(x{-}1) = 0$, откуда $deg_{\mathbf{U}}(x) \geq 1$ по следствию 4.4. Это означает, что картина такова:



Теперь по определению получаем $deg_{\mathbf{V}}(x{-}2) = 1$, т. е. $deg_{\mathbf{V}}(x) \geq 2$ по следствию 4.4. Рассуждая по индукции, мы получим, что для любого k порядок точки $(x{-}2^k)$ равен k; отсюда по следствию 4.4 порядок x строго больше k. Таким образом, x — точка порядка ∞, противоречие с условием.

Этап 5. Пусть теперь у Z-слов $\mathbf{U}$ и $\mathbf{V}$ совпадают характеристические последовательности, начиная с k-го члена. Рассмотрим вначале случай, когда знаки разностей $y_k^{\mathbf{U}} - x_0$ и $y_k^{\mathbf{V}} - x_0$ совпадают:



Так как $h_n^{\mathbf{U}} = h_n^{\mathbf{V}}$ при любом $n \geq k$, то разности $y_n^{\mathbf{U}} - x_0$ и $y_n^{\mathbf{V}} - x_0$ также будут одного знака; следовательно, равенство

$$|y_{n+1}^{\mathbf{U}} - y_n^{\mathbf{U}}| = |y_{n+1}^{\mathbf{V}} - y_n^{\mathbf{V}}|$$

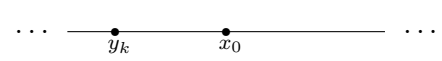
остается верным и при снятии модулей. Пусть $m = y_k^{\mathbf{U}} - y_k^{\mathbf{V}}$. Для всех точек x таких, что $deg_{\mathbf{U}}(x) \geq k$, выполняется

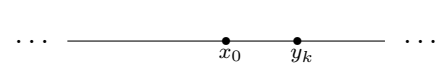
$$deg_{\mathbf{U}}(x) = deg_{Shift_m(\mathbf{V})}(x).$$

Согласно следствиям 4.3 и 4.4 порядок $(k-1)$ относительно двух данных слов имеют в точности все точки вида $x + 2^{k-1}$, где x — точка порядка $\geq k$. Таким образом, по индукции получаем, что порядки всех точек относительно $\mathbf{U}$ и $Shift_m(\mathbf{V})$ совпадают. Из доказанного на этапе 4 следует, что

$$\mathbf{U} \in \{Shift_m(\mathbf{V}), Aut(Shift_m(\mathbf{V}))\}.$$

Этап 6. Осталось рассмотреть случай, когда знаки разностей $y_k^{\mathbf{U}} - x_0$ и $y_k^{\mathbf{V}} - x_0$ противоположны:

$\mathbf{U} =$ 

$\mathbf{V} =$ 

Как отмечалось выше (этап 3), характеристические последовательности Z-слов $\mathbf{V}$ и $Rev(\mathbf{V})$ совпадают с некоторого номера N, причем для всех $n > N$ выражения $y_n^{\mathbf{V}} - x_0$ и $y_n^{Rev(\mathbf{V})} - x_0$ имеют противоположные знаки. Таким образом, согласно доказанному на предыдущем этапе Z-слово U совпадает либо с подходящим сдвигом $Rev(\mathbf{V})$, либо с автоморфизмом такого сдвига.

Доказательство п. (2) теоремы завершено. Перейдем к п. (1). Определим универсальный метод построения сильно бескубных Z-слов — «обобщенный метод раскачивания».

Пусть $d \in \{0, 1\}$, а $\{\gamma_i\}_1^\infty$ — произвольная последовательность натуральных чисел. Положим $x_0 = \frac{2-d}{3}$, $W_0 = \mathbf{W}(1) = a$

и определим индуктивно i-й шаг. Пусть слово W_{i-1} равнялось U_m или V_m. На i-м шаге к этому слову дописывается (с той стороны, которая ближе к точке x_0) $2^{\gamma_i}-1$ штук слов U_m и V_m так, чтобы полученное слово W_i равнялось $U_{m+\gamma_i}$ или $V_{m+\gamma_i}$ (это можно осуществить единственным способом).

Заметим, что всякое W_k равняется U_{Γ_k} или V_{Γ_k}, где

$$\Gamma_k = \sum_{i=1}^{k} \gamma_i .$$

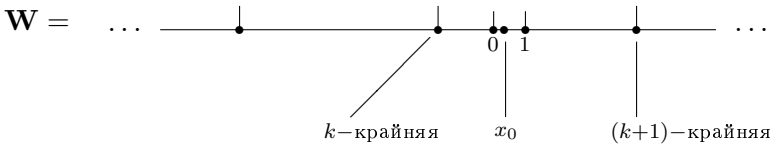
Какую бы последовательность $\{\gamma_i\}$ мы ни выбрали, данный метод приводит к построению Z-слова из TM^Z (всякое подслово полученного Z-слова лежит в некотором $W_k \in TM$); при этом последовательность, состоящая из единиц, приводит к построению Z-слова по обычному методу раскачивания.

Пусть $\{h_n\}_0^\infty$ — бинарная последовательность с бесконечным числом единиц. Выберем d и $\{\gamma_n\}_1^\infty$ для построения Z-слова обобщенным методом раскачивания. Пусть $d = h_0$, γ_1 есть номер первой единицы в $\{h_n\}_1^\infty$, а для $k > 1$ число γ_k есть разность номеров k-й и $(k-1)$-й единиц в $\{h_n\}_1^\infty$ (т. е. в $\{h_n\}_1^\infty$ единицы стоят в точности на всех позициях с номерами, равными частичным суммам последовательности $\{\gamma_n\}$).

Пусть обобщенный метод строит по d и $\{\gamma_n\}$ Z-слово $\mathbf{W}$. Проверим, что последовательность $\{h_n\}$ — характеристическая для $\mathbf{W}$.

Точку x назовем k-крайней, если при построении $\mathbf{W}$ она являлась концом отрезка, занимаемого W_k, но попадала внутрь отрезка, занимаемого W_{k+1}. Покажем, что k-крайняя точка имеет порядок Γ_k. В самом деле, слово W_k равно U_{Γ_k} или V_{Γ_k}; при дальнейшем построении добавляются сегменты, гарантированно являющиеся произведением блоков U_{Γ_k} и/или V_{Γ_k}. Таким образом, порядок k-крайней точки не меньше Γ_k по определению. Далее, легко проверить, что другим концом отрезка, занимаемого W_k, является $(k+1)$-крайняя точка. Ее порядок строго больше Γ_k, так как $\gamma_{k+1}>0$. Но расстояние между концами этого отрезка равно 2^{Γ_k}, откуда по следствию 4.4 порядок k-крайней точки в точности равен Γ_k.

Теперь построим характеристическую последовательность для $\mathbf{W}$. Заметим, что k-крайняя точка ($k \geq 1$) является ближайшей к $x_0 = \frac{1}{3}$ (и вообще к любой точке интервала $(0,1)$) точкой порядка $\geq \Gamma_k$. Более того, ближайшая к x_0 точка порядка $\geq \Gamma_k{+}1$ ($k \geq 0$) отстоит от k-крайней точки на расстоянии 2^{Γ_k} и, следовательно, совпадает с $(k{+}1)$-крайней точкой:



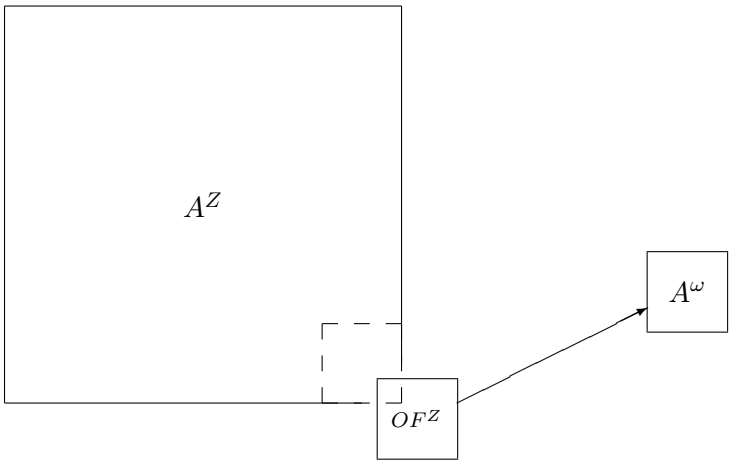
Таким образом, $y_1, \ldots, y_{\Gamma_1}$ совпадают с 1-крайней точкой и в общем случае $y_{\Gamma_k+1}, \ldots, y_{\Gamma_{k+1}}$ совпадают с $(k{+}1)$-крайней точкой. Точка y_0 совпадает, как легко убедиться, с 0-крайней точкой при $d = 1$ и с 1-крайней в противном случае. Отсюда получаем, что $h_0^{\mathbf{W}} = 1$ тогда и только тогда, когда y_0 — 0-крайняя точка, т. е. $d = 1$. Пусть $i > 0$. Имеем $h_i^{\mathbf{W}} = 1$ тогда и только тогда, когда $i = \Gamma_k$ для некоторого k. В результате $\{h_n^{\mathbf{W}}\} = \{h_n\}$, что и требовалось доказать.

Доказательство п. (1) и всей теоремы завершено. □

Замечание 4.4. Из доказательства теоремы следует, что всякое сильно бескубное Z-слово конечного порядка (с точностью до автоморфизма) можно построить обобщенным методом раскачивания с помощью выбранной бинарной последовательности с бесконечным числом единиц.

Замечание 4.5. Бинарные последовательности из «почти нулевого» класса (т. е. с конечным числом единиц) являются характеристическими для Z-слов бесконечного порядка. Таким образом, этому «вырожденному» классу последовательностей можно поставить в соответствие два вырожденных класса Z-слов бесконечного порядка. Этим завершена характеризация сильно бескубных Z-слов на языке бинарных последовательностей.

Если теперь заметить, что бинарные последовательности суть то же самое, что ω-слова в двухбуквенном алфавите, то мы получим впечатляющую разницу в масштабах между множествами Z-слов и ω-слов:



Этой иллюстрацией мы и закончим курс комбинаторики слов.

Список литературы

[1] *Baker K. A., McNulty G. F., Taylor W.* Growth problems for avoidable words // Theor. Comput. Sci. 1989. Vol. 69. P. 319–345.

[2] *Bean D. R., Ehrenfeucht A., McNulty G.* Avoidable patterns in strings of symbols // Pacific J. Math. 1979. Vol. 85. P. 261–294.

[3] *Berstel J., Boasson L.* Partial words and a theorem of Fine and Wilf // Theor. Comput. Sci. 1999. Vol. 218. P. 135–141.

[4] *Brandenburg F.-J.* Uniformly growing k-th power free homomorphisms // Theor. Comput. Sci. 1983. Vol. 23. P. 69–82.

[5] *Cassaigne J.* Unavoidable binary patterns // Acta Inf. 1993. Vol. 30. P. 385–395.

[6] *Cassaigne J.* Motifs evitables and regularites dans les mots: These de Doctorat. Univ. Paris 6, 1994.

[7] *Cassaigne J.* Constructing infinite words of intermediate complexity // Theor. Comput. Sci. (To appear).

[8] *Choffrut C., Karhumäki J.* Combinatorics on words // Handbook of formal languages. / Eds. Rosenberg G., Salomaa A. Berlin, 1997. Vol. 1, ch. 6.

[9] *De Luca A., Varricchio S.* Finiteness and regularity in semigroups and formal languages. Berlin, 1999.

[10] *Fine N. J., Wilf H. S.* Uniqueness theorem for periodic functions // Proc. Amer. Math. Soc. 1965. Vol. 16. P. 109–114.

[11] *Goralčik P., Vaniček T.* Binary patterns in binary words // Int. J. Alg. and Comp. 1991. Vol. 1. P. 387–392.

[12] *Gottschalk W. H., Hedlund G. A.* A characterization of the Morse minimal set // Proc. Amer. Math. Soc. 1964. Vol. 15. P. 70–74.

[13] *Lothaire M.* Combinatorics on words. Addison-Wesley, 1983.

[14] *Lothaire M.* Algebraic combinatorics on words. Cambridge Univ. Press, 2002.

[15] *Morse M., Hedlund G. A.* Unending chess, symbolic dynamics and a problem in semigroups // Duke Math. J. 1944. Vol. 11. P. 1–7.

[16] *Perrin D., Pin J.-E.* Semigroups and automata on infinite words // Semigroups, Formal languages and Groups. / Ed. Fountain J. Kluwer Acad. Publ., 1995. P. 49–72.

[17] *Pin J.-E.* Varieties of Formal Languages. London, 1986.

[18] *Restivo A., Salemi S.* Overlap-free words on two symbols // Lect. Notes Comp. Sci. 1984. Vol. 192. P. 196–206.

[19] *Sapir M. V.* Combinatorics on words with applications. LITP report. 1995. Vol. 32.

[20] *Séébold P.* Overlap-free sequences // Lect. Notes Comp. Sci. 1984. Vol. 192. P. 207–215.

[21] *Shur A. M.* Overlap-free words and Thue-Morse sequences // Int. J. Alg. and Comp., 1996. Vol. 6. P. 353–367.

[22] *Shur A. M., Konovalova Yu. V.* On the periods of partial words // Lect. Notes Comp. Sci. 2001. Vol. 2136. P. 657–665.

[23] *Sukhanov E. V., Shur A. M.* Galois connection in avoidability theory // Semigroups and their Applications, including Semigroup Rings. Berlin, 1998. P. 397–400.

[24] *Thue A.* Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen // Skr. Vid. Kristiania I. Mat. Naturv. Klasse I. 1912. P. 1–67.

[25] *Volkov L. M.* Number of words without forbidden factors // Publications Mathematicae. Debrecen, Hungary, 2001. Vol. 8. P. 47–54.

[26] *Зимин А. И.* Блокирующие множества термов // Мат. сб. 1982. Т. 119. С. 363–375.

[27] *Лаллеман Ж.* Полугруппы и комбинаторные приложения. М., 1985.

[28] *Саломаа А.* Жемчужины теории формальных языков. М., 1986.

[29] *Шур А. М.* Структура множества бескубных Z-слов в двухбуквенном алфавите // Изв. РАН. Сер. матем. 2000. Т. 64, №4. С. 201–224.

[30] *Шур А. М., Гамзова Ю. В.* Частичные слова и свойство взаимодействия периодов // Изв. РАН. Сер. матем. (В печати).