ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 58 **2022** **Вып. 2**

УДК 621.391 : 519.724.2

© 2022 г. **М. Ковачевич**

О МАКСИМАЛЬНОМ ЧИСЛЕ РАЗЛИЧИМЫХ СТРОК ПОД ДЕЙСТВИЕМ КОРОТКИХ ТАНДЕМНЫХ ДУПЛИКАЦИЙ[1]

Множество всех q-ичных строк, не содержащих повторяющихся подстрок длины $\leqslant 3$ (т.е. не содержащих подстрок вида aa, $abab$ и $abcabc$), образует код, исправляющий произвольное количество мутаций типа тандемных дупликаций длины $\leqslant 3$. Иными словами, любые две такие строки различимы в том смысле, что из них не могут образоваться одинаковые строки под действием некоторого количества тандемных дупликаций длины $\leqslant 3$. Показано, что этот код асимптотически оптимален по скорости, т.е. представляет собой максимальное множество различимых строк с точностью до субэкспоненциального множителя. Этот результат дает решение задачи о пропускной способности с нулевой ошибкой для последнего оставшегося случая каналов с тандемными дупликациями, удовлетворяющих свойству "единственности корневых строк".

Ключевые слова: тандемная дупликация, повторение тандема, ошибка дупликации, ошибка повторения, вклейка, хранение информации на основе ДНК, исправление ошибок, пропускная способность с нулевой ошибкой, код с ограничением, строка без повторений.

DOI: 10.31857/S0555292322020028, **EDN:** DZAPUX

§ 1. Введение

Тандемные дупликации – это ошибки типа "вклейки", естественным образом появляющиеся как мутации ДНК-строк и являющиеся поэтому потенциальным источником дефектов, возникающих в системах хранения информации на основе ДНК в живых организмах [1]. В то время как задача исправления тандемных дупликаций фиксированной и известной длины ℓ хорошо изучена, причем как при конечном [2, 3], так и при бесконечном [1, 4] числе ошибок, в отношении (по-видимому, гораздо более практически значимой) задачи исправления дупликаций переменной длины известно гораздо меньше. В частности, оптимальные коды, исправляющие все конфигурации дупликаций длины $\leqslant \ell$, были найдены только в специальных случаях $\ell = 1$ и $\ell = 2$ [1, теорема 32].

Нашим основным результатом является доказательство того факта, что аналогичная конструкция кодов, исправляющих неограниченное число тандемных дупликаций длины $\leqslant 3$ [1, теорема 27], также является асимптотически оптимальной по скорости. Этот результат дает решение задачи о пропускной способности с нулевой ошибкой для каналов с тандемными дупликациями во всех случаях, когда корневые строки (относительно дупликаций) любых строк единственны [1, теорема 40]. Однако при бóльших значениях ℓ корневые строки относительно дупликаций длины $\leqslant \ell$ перестают быть единственными [1, теорема 40], и следовательно, для решения за-

[1] Работа выполнена при поддержке исследовательско-инновационной программы Horizon 2020 Европейского Союза (номер гранта 856967), а также Секретариата по высшему образованию и научным исследованиям автономной провинции Воеводина, Сербия (номер проекта 142-451-2686/2021).

дачи о пропускной способности с нулевой ошибкой и смежных с ней задач в таких моделях понадобятся другие конструкции и верхние границы[2].

Помимо теоретико-информационных и кодовых вопросов упомянутого типа в литературе рассматривались и некоторые другие задачи, связанные с моделями с тандемными дупликациями переменной длины (см., например, [6–8]).

1.1. Описание модели. Для q-ичного алфавита будем использовать обозначение $\mathcal{A}_q := \{0, 1, \ldots, q-1\}$, а множество всех строк (или слов) над алфавитом $\mathcal{A}_q$ обозначим через $\mathcal{A}_q^* := \bigcup_{n=0}^{\infty} \mathcal{A}_q^n$. Длина строки $\boldsymbol{x} = x_1 \ldots x_n \in \mathcal{A}_q^n$ обозначается через $|\boldsymbol{x}| = n$. Строка, полученная конкатенацией двух строк $\boldsymbol{u}$ и $\boldsymbol{v}$, записывается в виде $\boldsymbol{uv}$. Строка $\boldsymbol{v}$ называется подстрокой (или фрагментом) строки $\boldsymbol{x}$, если существуют (возможно, пустые) строки $\boldsymbol{u}$ и $\boldsymbol{w}$, такие что $\boldsymbol{x} = \boldsymbol{uvw}$.

Пусть ℓ – фиксированное натуральное число. В *канале с тандемными* $(\leqslant \ell)$*-дупликациями* передаваемая по каналу строка $\boldsymbol{x}$ подвергается последовательному воздействию некоторого числа тандемных дупликаций длины $\leqslant \ell$ каждая, где тандемная дупликация длины k – это вставка точной копии подстроки длины k рядом с исходной подстрокой (неважно, вставляется ли она слева или справа от исходной, поскольку обе операции приводят к одной и той же новой строке). Предполагается, что число возникающих дупликаций заранее не известно ни передатчику, ни приемнику и может принимать любое значение из множества натуральных чисел $\{0, 1, 2, \ldots\}$. Более точно, канал описывается следующим образом:

- Вход: $\boldsymbol{x} \equiv \boldsymbol{x}^{(0)}$;
- Из множества $\{0, 1, 2, \ldots\}$ выбирается число t (количество дупликаций);
- Для $i = 1, \ldots, t$ повторяются следующие действия:
  - Из множества $\{1, \ldots, |\boldsymbol{x}^{(i-1)}|\}$ произвольным образом выбирается позиция дупликации j в строке $\boldsymbol{x}^{(i-1)}$;
  - Из множества $\{1, \ldots, \min\{j, \ell\}\}$ произвольным образом выбирается длина дупликации k;
  - Строка $\boldsymbol{x}^{(i)}$ получается вставкой копии подстроки $x_{j-k+1}^{(i-1)} \ldots x_j^{(i-1)}$ рядом с исходной подстрокой в $\boldsymbol{x}^{(i-1)}$, т.е.

$$\boldsymbol{x}^{(i)} = x_1^{(i-1)} \ldots \overline{x_{j-k+1}^{(i-1)} \ldots x_j^{(i-1)}}\, \underline{x_{j-k+1}^{(i-1)} \ldots x_j^{(i-1)}}\, x_{j+1}^{(i-1)} \ldots x_{|\boldsymbol{x}^{(i-1)}|}^{(i-1)},$$

    где исходная дуплицируемая подстрока выделена чертой сверху, а вставленная копия – чертой снизу;
- Выход: $\boldsymbol{y} \equiv \boldsymbol{x}^{(t)}$.

Всюду далее будем считать, что $\ell = 3$.

Пример 1. Примером того, как канал действует на передаваемую строку $\boldsymbol{x} \in \mathcal{A}_3^8$, является следующий список строк, каждая строка в котором получена из предыдущей путем тандемной дупликации длины $\leqslant 3$:

$$\boldsymbol{x} = 0\,1\,1\,2\,0\,2\,1\,0, \tag{1.1a}$$

$$\boldsymbol{x}^{(1)} = \overline{0}\,\underline{0}\,1\,1\,2\,0\,2\,1\,0, \tag{1.1b}$$

$$\boldsymbol{x}^{(2)} = 0\,0\,1\,1\,\overline{2\,0\,2}\,\underline{2\,0\,2}\,1\,0, \tag{1.1c}$$

$$\boldsymbol{x}^{(3)} = 0\,0\,1\,1\,2\,0\,2\,2\,0\,\overline{2\,1}\,\underline{2\,1}\,0, \tag{1.1d}$$

$$\boldsymbol{x}^{(4)} = 0\,0\,\overline{1\,1}\,\underline{1\,1}\,2\,0\,2\,2\,0\,2\,1\,2\,1\,0. \tag{1.1e}$$

Здесь $t = 4$, и выходом канала является строка $\boldsymbol{y} = \boldsymbol{x}^{(4)}$.

[2] Первые конструкции кодов для таких моделей (при $\ell \in \{4, 5, \ldots\}$) были приведены в [5].

Будем говорить, что строка $\boldsymbol{y}$ является t-потомком строки $\boldsymbol{x}$, или что $\boldsymbol{x}$ является t-предком $\boldsymbol{y}$, если $\boldsymbol{y}$ можно получить из $\boldsymbol{x}$ путем последовательного применения t тандемных дупликаций длины $\leqslant 3$. Множество всех t-потомков строки $\boldsymbol{x}$ обозначим через $D^t(\boldsymbol{x})$. Отметим, что строка может принадлежать различным множествам $D^t(\boldsymbol{x})$ и $D^s(\boldsymbol{x})$, $s \neq t$, поскольку в модели разрешены дупликации различных длин, т.е. множество $D^t(\boldsymbol{x}) \cap D^s(\boldsymbol{x})$ не обязано быть пустым (например, 0 1 1 1 1 является как 1-потомком строки 0 1 1, полученным при одной дупликации длины 2, так и 2-потомком этой же строки 0 1 1, полученным двумя дупликациями длины 1 каждая). Множество всех потомков строки $\boldsymbol{x}$ обозначим через $D^*(\boldsymbol{x}) := \bigcup\limits_{t\geqslant 0} D^t(\boldsymbol{x})$, где $D^0(\boldsymbol{x}) := \{\boldsymbol{x}\}$. В этих обозначениях $D^*(\boldsymbol{x})$ для заданной входной строки $\boldsymbol{x}$ – множество всех возможных выходов канала с тандемными ($\leqslant 3$)-дупликациями.

1.2. Различимые строки и безошибочная передача. Две строки $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}_q^*$ называются неразличимыми (confusable) в заданном канале связи, если при их передаче по этому каналу на его выходе может появиться одна и та же строка; в противном случае они называются различимыми (non-confusable). В нашей терминологии $\boldsymbol{x}$ и $\boldsymbol{y}$ неразличимы, если у них имеется общий потомок, т.е. $D^*(\boldsymbol{x}) \cap D^*(\boldsymbol{y}) \neq \varnothing$. Множество строк $\mathcal{C} \subseteq \mathcal{A}_q^*$ называется кодом *с нулевой ошибкой* [9] для заданного канала, если любые два несовпадающих кодовых слова $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ различимы. Заметим, что код с нулевой ошибкой исправляет *все* конфигурации ошибок, реализуемые в канале, т.е. приемник может однозначно декодировать любую заданную строку на выходе канала, поскольку в $\mathcal{C}$ имеется лишь одно кодовое слово, способное породить эту строку. Код с нулевой ошибкой $\mathcal{C} \subseteq \mathcal{A}_q^n$ называется оптимальным, если не существует никакого другого кода с нулевой ошибкой $\mathcal{C}' \subseteq \mathcal{A}_q^n$, такого что $|\mathcal{C}'| > |\mathcal{C}|$.

Скорость кода $\mathcal{C} \subseteq \mathcal{A}_q^n$, выражаемая в битах на символ, определяется как $\dfrac{1}{n}\log_2|\mathcal{C}|$. Пропускная способность с нулевой ошибкой для канала с алфавитом $\mathcal{A}_q$ на входе определяется как $\limsup\limits_{n\to\infty}$ по всем скоростям оптимальных кодов с нулевой ошибкой в $\mathcal{A}_q^n$. Эта величина представляет собой максимальное число бит на символ, которые можно безошибочно передать по заданному каналу.

§ 2. Корневые и неприводимые строки относительно дупликаций

Последовательно применяя операцию дедупликации, т.е. удаления дуплицированных подстрок длины $\leqslant 3$, каждую строку $\boldsymbol{x}$ можно привести к ее *корневой* строке $R(\boldsymbol{x})$, не содержащей повторяющихся подряд подстрок длины $\leqslant 3$. Более того, как показано в [1, теорема 24], корневые строки единственны в том смысле, что независимо от порядка, в котором выполняются дедупликации, процесс гарантированно приведет к одной и той же строке. (Подчеркнем, что это "свойство единственности корневых строк" выполнено лишь для моделей с тандемными дупликациями длины (i) $=\ell$, (ii) $\leqslant 2$ или (iii) $\leqslant 3$. Оно не выполняется, например, в моделях с тандемными дупликациями длины $\leqslant \ell$ при $\ell \in \{4, 5, \ldots\}$; см. [1, теорема 40].)

В этом контексте строка, не содержащая повторяющихся подстрок длины $\leqslant 3$, называется *неприводимой*. Другими словами, строка неприводима[3], если она не содержит подстрок вида $a\,a$, $a\,b\,a\,b$ и $a\,b\,c\,a\,b\,c$, где $a, b, c \in \mathcal{A}_q$. Через Irr_q обозначим множество всех неприводимых строк над $\mathcal{A}_q$, через $\mathrm{Irr}_q(n)$ – множество всех неприводимых строк длины n, а через $I_q(n)$ – мощность последнего, т.е. $I_q(n) := |\mathrm{Irr}_q(n)|$. Из вышеупомянутого "свойства единственности корневых строк" следует, что любые две различные неприводимые строки $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{Irr}_q$ различимы в канале с тандемными

[3] Неприводимые строки являются частным случаем строк с запрещенными конфигурациями, или строк с ограничениями [10], где множество запрещенных конфигураций имеет вид $\{a\,a, a\,b\,a\,b, a\,b\,c\,a\,b\,c\colon\ a, b, c \in \mathcal{A}_q\}$.

($\leqslant 3$)-дупликациями, т.е. $D^*(\boldsymbol{x}) \cap D^*(\boldsymbol{y}) = \varnothing$, и поэтому множество $\mathrm{Irr}_q(n)$ является кодом с нулевой ошибкой для этого канала [1, теорема 27].

Всюду далее будем предполагать, что $q \geqslant 3$, поскольку в случае двоичного алфавита рассматриваемая задача становится тривиальной. Например, над двоичным алфавитом существует лишь конечное число неприводимых строк, а именно $\mathrm{Irr}_2 = \{0, 1, 01, 10, 010, 101\}$, так что пропускная способность с нулевой ошибкой для канала с тандемными ($\leqslant 3$)-дупликациями над двоичным алфавитом равна нулю.

Для нас представляет интерес асимптотическое поведение величины $I_q(n)$ при $n \to \infty$, и в частности, экспонента ее скорости роста

$$\iota_q := \lim_{n\to\infty} \frac{1}{n} \log_2 I_q(n). \tag{2.1}$$

Величину ι_q можно охарактеризовать с помощью стандартных методов теории систем с ограничениями [10], например, как логарифм наибольшего собственного значения матрицы смежности направленного графа, представляющего собой диаграмму состояний системы, порождающей неприводимые строки. Здесь мы применим более простую характеризацию из [5, предложение 2], где было показано, что для $I_q(n)$ выполнено рекуррентное соотношение

$$I_q(n) = (q-2)I_q(n-1) + (q-3)I_q(n-2) + (q-2)I_q(n-3)$$

и поэтому

$$\iota_q = \log_2 r, \tag{2.2a}$$

где r – единственный положительный вещественный корень многочлена

$$x^3 - (q-2)x^2 - (q-3)x - (q-2),$$

т.е. r задается неявным условием

$$r^3 - (q-2)r^2 - (q-3)r - (q-2) = 0, \quad r > 0. \tag{2.2b}$$

В следующей лемме дается еще одна характеризация величины ι_q для троичного алфавита ($q = 3$), а также вытекающая из нее граница снизу на ι_q для бóльших алфавитов, которая будет использована в доказательстве нашего основного результата (теорема 1).

Лемма 1. *Для любых $q \geqslant 3$ и $\beta \in [0, 1]$ справедливо неравенство*

$$\iota_q \geqslant \frac{H(\beta)}{1+2\beta}, \tag{2.3}$$

где $H(\beta) := -\beta \log_2 \beta - (1-\beta)\log_2(1-\beta)$ – функция двоичной энтропии. Равенство в (2.3) достигается тогда и только тогда, когда $q = 3$ и $\beta = \bar{\beta}$, где $\bar{\beta}$ – единственное положительное решение уравнения $(1-x)^3 = x$.

Доказательство. Докажем соотношение

$$\iota_3 = \max_{0\leqslant\beta\leqslant 1} \frac{H(\beta)}{1+2\beta}, \tag{2.4}$$

из которого будет немедленно следовать утверждение леммы (так как ι_q монотонно возрастает по q). Приравнивая производную функции $\dfrac{H(\beta)}{1+2\beta}$ к нулю, убеждаемся, что эта функция достигает максимума в единственной положительной вещественной точке, удовлетворяющей уравнению $(1-x)^3 = x$, назовем ее $\bar{\beta}$. Тогда правую часть

равенства (2.4) можно представить в виде

$$\frac{H(\bar\beta)}{1+2\bar\beta} = \log_2\Big(\bar\beta^{\frac{-\bar\beta}{1+2\bar\beta}}(1-\bar\beta)^{\frac{-1+\bar\beta}{1+2\bar\beta}}\Big) = -\log_2(1-\bar\beta). \tag{2.5}$$

С другой стороны, мы знаем, что $\iota_3 = \log_2 r$, где r – единственный положительный вещественный корень уравнения $x^3 - x^2 - 1 = 0$ (см. (2.2)). Поэтому доказательство равенства в (2.4) равносильно доказательству того, что $-\log_2(1-\bar\beta) = \log_2 r$, т.е. что $(1-\bar\beta)^{-1}$ является решением уравнения $x^3 - x^2 - 1 = 0$. В этом можно убедиться непосредственно, подставляя $(1-\bar\beta)^{-1}$ вместо x и используя тот факт, что $(1-\bar\beta)^3 = \bar\beta$. ▲

§ 3. Различимость строк в канале с тандемными ($\leqslant 3$)-дупликациями

В этом параграфе изложим несколько фактов об эволюции строк под действием тандемных дупликаций длины $\leqslant 3$, основным из которых будет вывод верхней границы на максимальное число попарно различимых строк в заданном конусе потомков $D^*(\boldsymbol{x})$ (предложение 1). По поводу дальнейшего изучения комбинаторных и алгоритмических аспектов (не)различимости в ($\leqslant 2$)- и ($\leqslant 3$)-каналах с тандемными дупликациями отсылаем читателя к работе [11].

В следующей лемме утверждается, что множество попарно различимых строк, каждая из которых является 1-потомком данной строки $\boldsymbol{x}$, не может состоять из более чем двух элементов. В доказательстве также иллюстрируются условия, при которых две различимые строки могут быть получены применением различных мутаций к одной и той же строке $\boldsymbol{x}$ (см. уравнения (3.1) ниже).

Лемма 2. *Рассмотрим произвольную строку $\boldsymbol{x}$ и множество $D^1(\boldsymbol{x})$ ее 1-потомков, и пусть $\mathcal{C} \subseteq D^1(\boldsymbol{x})$ – код с нулевой ошибкой для канала с тандемными ($\leqslant 3$)-дупликациями. Тогда $|\mathcal{C}| \leqslant 2$.*

Доказательство. Рассмотрим $\boldsymbol{x}', \boldsymbol{x}'' \in D^1(\boldsymbol{x})$ и предположим, что мутации, переводящие $\boldsymbol{x}$ в $\boldsymbol{x}'$ и $\boldsymbol{x}''$, применяются к различным и непересекающимся подстрокам строки $\boldsymbol{x}$. Тогда $\boldsymbol{x}'$ и $\boldsymbol{x}''$ неразличимы, поскольку у них имеется общий потомок; действительно, достаточно применить к $\boldsymbol{x}'$ дупликацию, породившую $\boldsymbol{x}''$ из $\boldsymbol{x}$, и наоборот. Теперь предположим, что дупликации, переводящие $\boldsymbol{x}$ в $\boldsymbol{x}'$ и $\boldsymbol{x}''$, применяются к пересекающимся подстрокам $\boldsymbol{x}$. Оказывается, что во всех возможных случаях, *кроме одного*, то же самое рассуждение, что и для непересекающихся подстрок, показывает, что $\boldsymbol{x}'$ и $\boldsymbol{x}''$ неразличимы (мы продемонстрируем это для случаев, когда пересечение находится с правой стороны более длинной подстроки, так как остальные случаи симметричны им):

(i) Для случая пересекающихся подстрок длины 1 и 2 рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b}\,\underline{a\,b}\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,\overline{b}\,\underline{b}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,a\,b\,b\,\boldsymbol{v}$;

(ii) Для случая пересекающихся подстрок длины 2 и 2 рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,c\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b}\,\underline{a\,b}\,c\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,\overline{b\,c}\,\underline{b\,c}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,a\,b\,c\,b\,c\,\boldsymbol{v}$;

(iii) Для случая пересекающихся подстрок длины 2 и 3, имеющих пересечение длины 1, рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,c\,d\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b\,c}\,\underline{a\,b\,c}\,d\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,b\,\overline{c\,d}\,\underline{c\,d}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,c\,a\,b\,c\,d\,c\,d\,\boldsymbol{v}$;

(iv) Для случая пересекающихся подстрок длины 2 и 3, имеющих пересечение длины 2, рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,c\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b\,c}\,\underline{a\,b\,c}\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,\overline{b\,c}\,\underline{b\,c}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,c\,a\,b\,c\,b\,c\,\boldsymbol{v}$;

(v) Для случая пересекающихся подстрок длины 3 и 3, имеющих пересечение длины 1, рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\overline{a\,b\,c}d\,e\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,a\,b\,c\underline{a\,b\,c}d\,e\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,b\,c\,d\,e\underline{c\,d\,e}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,c\,a\,b\,c\,d\,e\,c\,d\,e\,\boldsymbol{v}$;

(vi) Для случая пересекающихся подстрок длины 3 и 3, имеющих пересечение длины 2, рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,c\,d\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b\,c}\underline{a\,b\,c}\,d\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,\overline{b\,c\,d}\underline{b\,c\,d}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,c\,a\,b\,c\,d\,b\,c\,d\,\boldsymbol{v}$;

(vii) Для случая пересекающихся подстрок длины 1 и 3 рассмотрим строку $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,c\,\boldsymbol{v}$ и заметим, что ее потомки $\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b\,c}\underline{a\,b\,c}\,\boldsymbol{v}$ и $\boldsymbol{x}'' = \boldsymbol{u}\,a\,b\,\overline{c}\,\underline{c}\,\boldsymbol{v}$ неразличимы, поскольку у них имеется общий потомок $\boldsymbol{u}\,a\,b\,c\,a\,b\,c\,c\,\boldsymbol{v}$.

Единственный случай, не вошедший в этот список, – это случай пересекающихся подстрок длины 1 и 3, когда пересечение возникает в середине более длинной подстроки. А именно, для $\boldsymbol{x} = \boldsymbol{u}\,a\,b\,c\,\boldsymbol{v}$, где символы $a, b, c \in \mathcal{A}_q$ различны, положим

$$\boldsymbol{x}' = \boldsymbol{u}\,\overline{a\,b\,c}\underline{a\,b\,c}\,\boldsymbol{v}, \tag{3.1a}$$

$$\boldsymbol{x}'' = \boldsymbol{u}\,a\,\overline{b}\,\underline{b}\,c\,\boldsymbol{v}. \tag{3.1b}$$

В этом случае мы не можем применить такое же рассуждение, как и выше, чтобы показать, что строки $\boldsymbol{x}'$ и $\boldsymbol{x}''$ неразличимы, и действительно, в общем случае это не обязательно верно. Например, если обе строки $\boldsymbol{u}$ и $\boldsymbol{v}$ – пустые, то $\boldsymbol{x}'$ и $\boldsymbol{x}''$ в (3.1) различимы, поскольку символ a не может оказаться после символа c в потомках строки $\boldsymbol{x}''$, в то время как *во всех* потомках слова $\boldsymbol{x}'$ символ a находится после c (аналогичный пример был приведен в [1]). Так происходит, поскольку фрагмент $a\,b\,c$, содержащийся в исходной строке $\boldsymbol{x}$, больше не возникнет в строке $\boldsymbol{x}''$, так как она "разбита" вставкой еще одной копии b. Наконец, в строке $\boldsymbol{x}''$ (соответственно, $\boldsymbol{x}'$) можно повторить дупликацию, породившую $\boldsymbol{x}'$ (соответственно, $\boldsymbol{x}''$) из $\boldsymbol{x}$, и таким образом показать, что $\boldsymbol{x}'$ и $\boldsymbol{x}''$ неразличимы, во всех случаях, кроме (3.1). Таким образом, код с нулевой ошибкой в $D^1(\boldsymbol{x})$ может содержать не более двух кодовых слов. ▲

Замечание 1. Не каждый случай вида (3.1) приводит к различимым потомкам. В качестве контрпримера предположим, что $\boldsymbol{u}$ – пустая строка, а $\boldsymbol{v} = a$, так что $\boldsymbol{x} = a\,b\,c\,a$, $\boldsymbol{x}' = \overline{a\,b\,c}\underline{a\,b\,c}\,a$ и $\boldsymbol{x}'' = a\,\overline{b}\,\underline{b}\,c\,a$. Тогда у $\boldsymbol{x}'$ и $\boldsymbol{x}''$ есть общий потомок $a\,b\,b\,c\,a\,b\,c\,a$, и поэтому они неразличимы. Однако для наших целей достаточно того факта, что (3.1) является *единственным* случаем, когда два потомка *могут быть* различимыми. В частности, это факт позволит нам вывести точную *верхнюю* границу на мощность оптимальных кодов с нулевой ошибкой.

Приведенное выше наблюдение справедливо в общем случае, а не только для 1-потомков строки $\boldsymbol{x}$. А именно, если $\boldsymbol{x}', \boldsymbol{x}'' \in D^*(\boldsymbol{x})$ получены применением к $\boldsymbol{x}$ двух разных конфигураций дупликаций, в каждой из этих строк можно повторить/воспроизвести дупликации, примененные ко второй, и таким образом показать, что у них имеется общий потомок. Единственный случай, когда такой процесс повторения дупликаций *может* в какой-то момент стать невозможным, это ситуация, когда к $\boldsymbol{x}'$ применяется дупликация длины 3, а к соответствующему фрагменту в $\boldsymbol{x}''$ – дупликация длины 1 (см. (3.1)), так что повторить в $\boldsymbol{x}''$ соответствующую мутацию, примененную к $\boldsymbol{x}'$, невозможно. Так происходит из-за того, что всякий раз, когда к строке применяется дупликация длины 2 или 3, *все* фрагменты длины $\leqslant 3$ исходной строки сохраняются в получившейся строке (с несколькими дополнительными подстроками, появляющимися в том месте, куда вставилась копия). Единственный случай, в котором фрагмент длины 3 исходной строки исчезает (не появляется в получившейся строке), – это после дупликации длины 1, как показано в (3.1b). На основе этого наблюдения мы выведем верхнюю границу на мощность оп-

тимальных кодов с нулевой ошибкой в множестве всех t-потомков данной строки $\boldsymbol{x}$ при любом t (предложение 1 ниже).

Пример 2. Чтобы пояснить вышесказанное, представим пример, приведенный в (1.1), в несколько ином виде (фрагмент **1 2 0** выделен, и показаны дупликации слева (соответственно, справа) от этого фрагмента, при которых копии вставляются слева (соответственно, справа) от оригинала):

$$\boldsymbol{x} = 0\,1\,\mathbf{1\,2\,0}\,2\,1\,0, \tag{3.2a}$$

$$\boldsymbol{x}^{(1)} = \underline{0}\,\overline{0}\,1\,\mathbf{1\,2\,0}\,2\,1\,0, \tag{3.2b}$$

$$\boldsymbol{x}^{(2)} = 0\,0\,1\,\mathbf{1}\,\overline{\mathbf{2\,0}\,2}\,\underline{2\,0\,2}\,1\,0, \tag{3.2c}$$

$$\boldsymbol{x}^{(3)} = 0\,0\,1\,\mathbf{1\,2\,0}\,2\,2\,0\,\overline{2\,1}\,\underline{2\,1}\,0, \tag{3.2d}$$

$$\boldsymbol{x}^{(4)} = 0\,0\,\underline{1\,1}\,\overline{1\,\mathbf{1}}\,\mathbf{2\,0}\,2\,2\,0\,2\,1\,2\,1\,0. \tag{3.2e}$$

Пусть $\boldsymbol{z} = 0\,1\,\mathbf{1\,2}\,2\,\mathbf{0}\,2\,1\,0$. Заметим, что в $\boldsymbol{z}$ можно воспроизвести все дупликации подстрок $\boldsymbol{x}$, которые либо не пересекаются с сегментом **1 2 0**, либо пересекаются с ним лишь частично[4], как в примерах, приведенных в (3.2):

$$\boldsymbol{x} = 0\,1\,\mathbf{1\,2\,0}\,2\,1\,0, \tag{3.3a}$$

$$\boldsymbol{z} = 0\,1\,\mathbf{1}\,\overline{\mathbf{2}}\,\underline{2}\,\mathbf{0}\,2\,1\,0, \tag{3.3b}$$

$$\boldsymbol{z}^{(1)} = \underline{0}\,\overline{0}\,1\,\mathbf{1\,2}\,2\,\mathbf{0}\,2\,1\,0, \tag{3.3c}$$

$$\boldsymbol{z}^{(2)} = 0\,0\,1\,\mathbf{1\,2}\,\overline{2\,\mathbf{0}\,2}\,\underline{2\,0\,2}\,1\,0, \tag{3.3d}$$

$$\boldsymbol{z}^{(3)} = 0\,0\,1\,\mathbf{1\,2}\,2\,\mathbf{0}\,2\,2\,0\,\overline{2\,1}\,\underline{2\,1}\,0, \tag{3.3e}$$

$$\boldsymbol{z}^{(4)} = 0\,0\,\underline{1\,1}\,\overline{1\,\mathbf{1}}\,\mathbf{2}\,2\,\mathbf{0}\,2\,2\,0\,2\,1\,2\,1\,0. \tag{3.3f}$$

Поэтому любая пара строк из (3.2) и (3.3) неразличима; например, общим потомком $\boldsymbol{z}$ и $\boldsymbol{x}^{(3)}$ является $\boldsymbol{z}^{(3)}$. Единственная мутация, которую нельзя повторить в $\boldsymbol{z}$, – это дупликация всего фрагмента **1 2 0**, поскольку соответствующий фрагмент в $\boldsymbol{z}$ больше не существует (он был "разбит" вставленным символом 2). Например, если бы нужно было превратить $\boldsymbol{x}^{(2)}$ из (3.2c) в строку

$$\boldsymbol{y} = 0\,0\,1\,\overline{\mathbf{1\,2\,0}}\,\underline{1\,2\,0}\,2\,2\,0\,2\,1\,0 \tag{3.4}$$

вместо $\boldsymbol{x}^{(3)}$, было бы уже невозможно применить тот же процесс, что и в (3.3).

Прежде чем сформулировать предложение 1, являющееся основным результатом этого параграфа, докажем одну полезную лемму.

Лемма 3. *Зафиксируем натуральные числа b, t, n, такие что $b \leqslant t \leqslant n$. Пусть $\mathcal{U} \subseteq \{l_1, l_3, *\}^n$ – множество строк, удовлетворяющих следующим двум условиям:*

(1) *Каждая строка в $\boldsymbol{u} \in \mathcal{U}$ имеет ровно b символов l_3, $t-b$ символов l_1, и $n-t$ символов $*$;*

(2) *Для любых двух различных строк $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}$ найдется позиция $i \in \{1, \ldots, n\}$, в которой $\{u_i, v_i\} = \{l_1, l_3\}$ (т.е. такая, что либо $u_i = l_1$ и $v_i = l_3$, либо $u_i = l_3$ и $v_i = l_1$).*

Тогда $|\mathcal{U}| \leqslant 2^{tH(b/t)}$.

Доказательство. Рассмотрим n бросков монеты, для которой вероятность выпадения орла равна b/t, и введем следующие события, индексированные строками из множества $\mathcal{U}$. Для $\boldsymbol{u} \in \mathcal{U}$ обозначим через $A_{\boldsymbol{u}}$ событие, состоящее в том, что на i-м

[4] Для $\boldsymbol{x}$ из (3.2a) подстроки длины $\leqslant 3$, частично пересекающиеся с подстрокой **1 2 0**, следующие: **1**, **2**, **0**, 1 **1**, **1 2**, **2 0**, **0** 2, 0 1 **1**, 1 **1 2**, **2 0** 2, **0** 2 1.

броске монеты выпал орел, если $u_i = l_3$, выпала решка, если $u_i = l_1$, и неважно, что выпало, если $u_i = *$, для всех $i = 1, \ldots, n$. Из условия (1) следует, что вероятность такого события равна

$$\Pr\{A_{\boldsymbol{u}}\} = \left(\frac{b}{t}\right)^b \left(1 - \frac{b}{t}\right)^{t-b} = 2^{-tH(b/t)}$$

для любой строки $\boldsymbol{u} \in \mathcal{U}$. При этом в силу условия (2) события $A_{\boldsymbol{u}}$ и $A_{\boldsymbol{v}}$ несовместны для любых $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}$, $\boldsymbol{u} \neq \boldsymbol{v}$. Отсюда $|\mathcal{U}| \Pr\{A_{\boldsymbol{u}}\} \leqslant 1$, что и требовалось доказать. ▲

Заметим, что приведенная верхняя граница на $|\mathcal{U}|$ не зависит от n (т.е. от числа символов $*$).

Предложение 1. *Рассмотрим строку $\boldsymbol{x} \in \mathcal{A}_q^n$, и пусть $\mathcal{C} \subseteq D^t(\boldsymbol{x})$ – код с нулевой ошибкой для канала с тандемными ($\leqslant 3$)-дупликациями, удовлетворяющий следующему условию: из t дупликаций, порождающих каждое кодовое слово $\boldsymbol{y} \in \mathcal{C}$ из слова $\boldsymbol{x}$, ровно b имеют длину 3. Тогда $|\mathcal{C}| \leqslant 2^{tH(b/t)}$.*

Доказательство. Как показано выше, если две строки $\boldsymbol{x}', \boldsymbol{x}'' \in D^t(\boldsymbol{x})$ различимы, то это с необходимостью означает, что к фрагменту $a\,b\,c$ в одном из предков строки $\boldsymbol{x}'$ была применена дупликация длины 3, а к среднему символу соответствующего фрагмента в некотором предке $\boldsymbol{x}''$ была применена дупликация длины 1, или наоборот. Другими словами, для любой пары кодовых слов кода с нулевой ошибкой в $D^t(\boldsymbol{x})$ найдется фрагмент, в котором они отличаются на мутацию длины 1/длины 3. Поэтому интересующий нас вопрос состоит в том, насколько велико может быть множество строк, любые две из которых отличаются на мутацию длины 1/длины 3 в некоторой позиции. (Неважно, что это за позиции и что происходит в промежутках между ними, единственное требование состоит в том, что любая пара кодовых слов в каком-то месте отличается на мутацию длины 1/длины 3, так как это единственная возможность для того, чтобы две строки могли стать различимыми.) Поэтому $|\mathcal{C}|$ можно ограничить сверху максимальной мощностью множества $\mathcal{U}$ строк над "алфавитом" {длина 1, длина 3, $*$}, удовлетворяющего следующим условиям: 1) любая строка в $\mathcal{U}$ содержит ровно b символов "длина 3" и $t - b$ символов "длина 1"; 2) любые две различные строки в $\mathcal{U}$ отличаются в некоторой позиции на символ "длина 1"/"длина 3". (Формальный символ $*$ служит для заполнения пустых позиций, которые могут возникать из-за того, что различные пары кодовых слов могут отличаться на мутацию длины 1/длины 3 в различных фрагментах; см. также пример 3 ниже.) Теперь требуемая граница получается применением леммы 3. ▲

Пример 3. Приведем пример множества строк $\mathcal{U}$ из вышеизложенного доказательства. Рассмотрим следующую строку:

$$\boldsymbol{x} = \overbrace{0\,\underbrace{1\,2}\,3}\,4\,\overbrace{5\,6\,\underbrace{7}\,8\,9}. \tag{3.5}$$

(Все символы в $\boldsymbol{x}$ обозначены по-разному для облегчения понимания, это никак не влияет на общность рассуждений.) Пусть применением тандемных дупликаций к четырем фрагментам в $\boldsymbol{x}$, отмеченных скобками сверху или снизу, получены следующие потомки в $D^3(\boldsymbol{x})$:

$$\boldsymbol{x}' = 0\,1\,2\,\underline{0\,1\,2}\,\underline{2}\,3\,4\,5\,6\,\underline{6}\,7\,8\,9, \tag{3.6a}$$

$$\boldsymbol{x}'' = 0\,1\,\underline{1}\,2\,3\,4\,5\,6\,\underline{6}\,7\,8\,9\,\underline{7\,8\,9}, \tag{3.6b}$$

$$\boldsymbol{x}''' = 0\,1\,\underline{1}\,2\,3\,4\,5\,6\,7\,\underline{5\,6\,7}\,8\,\underline{8}\,9, \tag{3.6c}$$

где подчеркнуты вставленные копии. Мутации, примененные к этим четырем фрагментам, можно описать с помощью следующих строк:

$$\boldsymbol{u}' = l_3\,l_1\,l_1\,*, \tag{3.7a}$$

$$\boldsymbol{u}'' = l_1 * l_1\, l_3, \tag{3.7b}$$
$$\boldsymbol{u}''' = l_1 * l_3\, l_1, \tag{3.7c}$$

где символ l_3 показывает, что к соответствующему фрагменту применена дупликация длины 3, символ l_1 показывает, что к среднему символу этого фрагмента применена дупликация длины 1, а $*$ означает, что к этому фрагменту ни одна из этих мутаций не применялась. Заметим, что для любой пары строк в (3.7) найдется координата, в которой одна из них равна l_1, а другая l_3.

§ 4. Пропускная способность с нулевой ошибкой для канала с тандемными ($\leqslant 3$)-дупликациями

Пусть $\mathcal{C}_q^\star(n) \subseteq \mathcal{A}_q^n$ – оптимальный код с нулевой ошибкой для канала с тандемными ($\leqslant 3$)-дупликациями. Для заданной неприводимой строки $\boldsymbol{x} \in \mathrm{Irr}_q$ положим

$$\mathcal{C}_q^\star(n;\boldsymbol{x}) := \mathcal{C}_q^\star(n) \cap D^*(\boldsymbol{x}).$$

Тогда $\mathcal{C}_q^\star(n;\boldsymbol{x})$ является оптимальным кодом с нулевой ошибкой на множестве всех потомков $\boldsymbol{x}$ длины n. Это так, поскольку в силу свойства единственности корневой строки для тандемных дупликаций длины $\leqslant 3$ любые два различных конуса потомков не пересекаются [1, следствие 26], и поэтому любые две строки, имеющие различные корневые строки, различимы. Иными словами, можно, не ограничивая общности, строить код по отдельности в конусе потомков для каждой возможной корневой/неприводимой строки. Этот факт явно сформулирован в следующей лемме; ее доказательство опущено, поскольку оно непосредственно вытекает из [1, следствие 26].

Лемма 4. *Оптимальный код длины n с нулевой ошибкой можно представить в виде несвязного объединения оптимальных кодов в каждом из конусов потомков*:

$$\mathcal{C}_q^\star(n) = \dot{\bigcup_{\boldsymbol{x}\in\mathrm{Irr}_q}} \mathcal{C}_q^\star(n;\boldsymbol{x}). \tag{4.1}$$

Следующее утверждение дает характеризацию экспоненты скорости роста мощности оптимальных кодов $\mathcal{C}_q^\star(n)$ или, эквивалентным образом, пропускной способности с нулевой ошибкой канала с тандемными ($\leqslant 3$)-дупликациями. В нем утверждается, что эта величина равна

$$\iota_q = \lim_{n\to\infty} \frac{1}{n} \log_2 I_q(n)$$

(см. (2.1) и (2.2)). Другими словами, пропускная способность с нулевой ошибкой достигается на кодах $\mathrm{Irr}_q(n)$, состоящих из неприводимых строк длины n.

Теорема 1. *Пропускная способность с нулевой ошибкой канала с тандемными ($\leqslant 3$)-дупликациями с алфавитом $\mathcal{A}_q$, $q \geqslant 3$, равна ι_q.*

Доказательство. Требуется показать, что

$$\lim_{n\to\infty} \frac{1}{n} \log_2 \bigl|\mathcal{C}_q^\star(n)\bigr| = \iota_q. \tag{4.2}$$

Так как $\mathrm{Irr}_q(n) \subseteq \mathcal{C}_q^\star(n)$, то мы знаем, что

$$\lim_{n\to\infty} \frac{1}{n} \log_2 |\mathcal{C}_q^\star(n)| \geqslant \lim_{n\to\infty} \frac{1}{n} \log_2 I_q(n) = \iota_q$$

(см. (2.1)), поэтому достаточно доказать противоположное неравенство

$$\lim_{n\to\infty}\frac{1}{n}\log_2|\mathcal{C}_q^\star(n)|\leqslant\iota_q.$$

Для этого мы упростим анализ, построив достаточно большой *подкод* $\mathcal{C}_q(n;m,t,b)\subseteq\subseteq\mathcal{C}_q^\star(n)$, имеющий ту же экспоненту скорости роста, что и оптимальный код $\mathcal{C}_q^\star(n)$, т.е.

$$\lim_{n\to\infty}\frac{1}{n}\log_2|\mathcal{C}_q^\star(n)|=\lim_{n\to\infty}\frac{1}{n}\log_2|\mathcal{C}_q(n;m,t,b)|,$$

при подходящем выборе параметров m,t,b.

Зафиксируем произвольную неприводимую строку $\boldsymbol{x}$ длины m, т.е. $\boldsymbol{x}\in\mathrm{Irr}_q(m)$, и пусть $\mathcal{C}_q(n;\boldsymbol{x},t,b)\subseteq\mathcal{C}_q^\star(n;\boldsymbol{x})$ – код, содержащий только те кодовые слова из $\mathcal{C}_q^\star(n;\boldsymbol{x})$, которые удовлетворяют следующим двум условиям: 1) каждое кодовое слово лежит в $D^t(\boldsymbol{x})$, т.е. является t-потомком строки $\boldsymbol{x}$; 2) из t дупликаций, переводящих $\boldsymbol{x}$ в заданный потомок/заданное кодовое слово, ровно b имеют длину 3. Определим такой подкод следующим образом:

$$\mathcal{C}_q(n;m,t,b):=\bigcup_{\boldsymbol{x}\in\mathrm{Irr}_q(m)}\mathcal{C}_q(n;\boldsymbol{x},t,b).\tag{4.3}$$

Из этой конструкции и леммы 4 следует, что

$$\mathcal{C}_q^\star(n)=\bigcup_{m,t,b}\mathcal{C}_q(n;m,t,b).\tag{4.4}$$

Теперь должно быть ясно, что величина $|\mathcal{C}_q(n;m,t,b)|$, максимизированная по всем возможным значениям m,t,b, имеет ту же самую экспоненту скорости роста, что и $|\mathcal{C}_q^\star(n)|$ (значения m,t,b выбираются для каждого n, т.е. оптимальные значения параметров m,t,b являются, вообще говоря, функциями от длины блока n). Это следует из равенства (4.4) и принципа Дирихле – мощность кода $\mathcal{C}_q^\star(n)$ растет экспоненциально быстро по длине блока n, а выбрать значения каждого из параметров m, t и b можно линейным количеством способов, поэтому по крайней мере для одного такого выбора код $\mathcal{C}_q(n;m,t,b)$ будет содержать экспоненциально много кодовых слов (с тем же показателем экспоненты). Таким образом, коды $\mathcal{C}_q(n;m,t,b)$ асимптотически оптимальны по скорости, т.е. на них достигается пропускная способность с нулевой ошибкой канала с тандемными ($\leqslant 3$)-дупликациями, когда параметры m,t,b выбираются правильным образом (так, чтобы максимизировать $|\mathcal{C}_q(n;m,t,b)|$).

Теперь вычислим скорость построенных кодов. Согласно (4.3) и предложению 1 (в котором утверждается, что $|\mathcal{C}_q(n;\boldsymbol{x},t,b)|\leqslant 2^{tH(b/t)}$), мощность кода $\mathcal{C}_q(n;m,t,b)$ можно оценить сверху как

$$\bigl|\mathcal{C}_q(n;m,t,b)\bigr|\leqslant I_q(m)\cdot 2^{tH(b/t)},\tag{4.5}$$

а длину этого кода можно оценить снизу как

$$n\geqslant m+3b+(t-b)=m+t+2b\tag{4.6}$$

(начальная неприводимая строка имеет длину m, и ровно b дупликаций, порождающих ее потомков, имеют длину 3). Следовательно,

$$\frac{1}{n}\log_2\bigl|\mathcal{C}_q(n;m,t,b)\bigr|\leqslant\frac{\log_2 I_q(m)+tH(b/t)}{m+t+2b}.\tag{4.7}$$

Чтобы найти асимптотику этой величины при $n \to \infty$, нужно рассмотреть два случая, соответствующие разным выборам параметров m, t, b:

- $m = o(t)$. Пусть $\lim\limits_{t\to\infty} \dfrac{b}{t} = \beta \in [0,1]$. Тогда

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 \bigl|\mathcal{C}_q(n; m, t, b)\bigr| \leqslant \frac{H(\beta)}{1+2\beta} \leqslant \iota_q, \tag{4.8}$$

  где первое неравенство следует из (4.7), а второе совпадает с (2.3);
- $t = \mathcal{O}(m)$. Пусть $\liminf\limits_{m\to\infty} \dfrac{t}{m} = \tau \geqslant 0$ и $\lim\limits_{t\to\infty} \dfrac{b}{t} = \beta \in [0,1]$. Тогда

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 \bigl|\mathcal{C}_q(n; m, t, b)\bigr| \leqslant \frac{\iota_q + \tau H(\beta)}{1+\tau(1+2\beta)} \leqslant \iota_q. \tag{4.9}$$

Здесь снова первое неравенство вытекает из (4.7), а второе равносильно (2.3). Итак, все выборы значений параметров m, t, b дают асимптотическую скорость кодов $\mathcal{C}_q(n; m, t, b)$, не превосходящую ι_q. Поскольку все эти коды оптимальны по скорости, как указано во втором абзаце этого доказательства, равенство (4.2) тем самым доказано. ▲

§ 5. Заключение

Эволюция строк под действием тандемных дупликаций – интересная и нетривиальная задача, важная в нескольких областях исследований. В настоящей статье исследована различимость строк при тандемных дупликациях переменной длины – задача, инспирированная исправлением ошибок в каналах связи, в которых передаваемые сообщения подвергаются мутациям такого типа. А именно, для случая дупликаций длины $\leqslant 3$ получена верхняя граница на максимальную мощность множества попарно различимых строк, которая вместе с конструкцией из [1] определяет максимальную скорость, достижимую кодами, исправляющими *произвольное* число таких дефектов.

В случаях, когда корневые строки относительно дупликаций не единственны, например, для модели ($\leqslant \ell$)-тандемных дупликаций с параметром ℓ, большим чем 3, максимальные достижимые скорости остаются неизвестными. Благодаря свойству "неединственности корневых строк" анализ эволюции и различимости строк в таких моделях более сложен, и поэтому для решения задачи о пропускной способности с нулевой ошибкой и смежных с ней задач потребуются дальнейшие исследования и, возможно, другие методы.

СПИСОК ЛИТЕРАТУРЫ

1. *Jain S., Farnoud F., Schwartz M., Bruck J.* Duplication-Correcting Codes for Data Storage in the DNA of Living Organisms // IEEE Trans. Inform. Theory. 2017. V. 63. № 8. P. 4996–5010. https://doi.org/10.1109/TIT.2017.2688361
2. *Kovačević M., Tan V.Y.F.* Asymptotically Optimal Codes Correcting Fixed-Length Duplication Errors in DNA Storage Systems // IEEE Commun. Lett. 2018. V. 22. № 11. P. 2194–2197. https://doi.org/10.1109/LCOMM.2018.2868666
3. *Lenz A., Jünger N., Wachter-Zeh A.* Bounds and Constructions for Multi-Symbol Duplication Error Correcting Codes, https://arXiv.org/abs/1807.02874v3 [cs.IT], 2018.
4. *Kovačević M.* Zero-Error Capacity of Duplication Channels // IEEE Trans. Commun. 2019. V. 67. № 10. P. 6735–6742. https://doi.org/10.1109/TCOMM.2019.2931342
5. *Chee Y.M., Chrisnata J., Kiah H.M., Nguyen T.T.* Efficient Encoding/Decoding of GC-Balanced Codes Correcting Tandem Duplications // IEEE Trans. Inform. Theory. 2020. V. 66. № 8. P. 4892–4903. https://doi.org/10.1109/TIT.2020.2981069

6. *Farnoud F., Schwartz M., Bruck J.* The Capacity of String-Duplication Systems // IEEE Trans. Inform. Theory. 2016. V. 62. № 2. P. 811–824. https://doi.org/10.1109/TIT.2015.2505735
7. *Jain S., Farnoud F., Bruck J.* Capacity and Expressiveness of Genomic Tandem Duplication // IEEE Trans. Inform. Theory. 2017. V. 63. № 10. P. 6129–6138. https://doi.org/10.1109/TIT.2017.2728079
8. *Leupold P., Martín-Vide C., Mitrana V.* Uniformly Bounded Duplication Languages // Discrete Appl. Math. 2005. V. 146. № 3. P. 301–310. https://doi.org/10.1016/j.dam.2004.10.003
9. *Shannon C.E.* The Zero Error Capacity of a Noisy Channel // IRE Trans. Inform. Theory. 1956. V. 2. № 3. P. 8–19. https://doi.org/10.1109/TIT.1956.1056798
10. *Marcus B.H., Roth R.M., Siegel P.H.* An Introduction to Coding for Constrained Systems (unpublished manuscript), 5th ed., 2001. Available online at http://www.math.ubc.ca/~marcus/Handbook/.
11. *Chee Y.M., Chrisnata J., Kiah H.M., Nguyen T.T.* Deciding the Confusability of Words under Tandem Repeats in Linear Time // ACM Trans. Algorithms. 2019. V. 15. № 3. Art. 42 (22 pp.). https://doi.org/10.1145/3338514

Ковачевич Младен
Факультет технических наук,
Университет г. Нови-Сад, Сербия
kmladen@uns.ac.rs

Поступила в редакцию
04.10.2021
После доработки
20.04.2022
Принята к публикации
21.04.2022