ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том ?? **20??** **Вып. на рецензировании**

УДК 621.391 : 519.218.5

© 20?? г. **Г.А. Хазиев, О.А. Зверков, С.А. Пирогов, А.В. Селиверстов, В.А. Любецкий**

МИНИМАЛЬНОЕ РАССТОЯНИЕ МЕЖДУ СЛУЧАЙНОЙ ПОСЛЕДОВАТЕЛЬНОСТЬЮ И ЕЁ СОВЕРШЕННЫМИ ПАЛИНДРОМАМИ [1]

Для произвольной последовательности x часто рассматривают нормированное длиной x минимальное расстояние Левенштейна $\mathrm{imp}(x)$ между x и переменным словом вида $w\mathrm{c}(w)$, где w – любой префикс в x, а $\mathrm{c}(\cdot)$ – перестановка букв в w в обратном порядке с заменой их на «двойственные» буквы. Например, взаимозаменой `A` с `T` и `G` с `C`. Свойства важной теоретически и в приложениях функции $\mathrm{imp}(x)$ остаются математически не изученными. Поэтому мы провели обширные компьютерные эксперименты, касающиеся поведения $\mathrm{imp}(x)$ на случайных последовательностях. В них наблюдается, что доля последовательностей x, близких к совершенным палиндромам $w\mathrm{c}(w)$, очень мала; медиана и среднее значений для $\mathrm{imp}(x)$ существенно зависят от вероятностей отдельных букв. У длинных последовательностей x эти характеристики слабо зависят от длины x, а для коротких, напротив, значимо зависят от длины. Но они весьма зависят от GC-состава; и среднеквадратичное отклонение мало зависит от него. Рассмотрены и другие характеристики. Результаты применимы, в частности, в задачах биоинформатики. Например, при поиске сайтов кооперативного связывания транскрипционных факторов и промоторов.

Ключевые слова: палиндром, случайная последовательность, GC-состав, биоинформатика.

DOI: `10.31857/S05552923??`, **EDN:** ??

§ 1. Введение

В работе рассматривается минимальное нормированное длиной случайной последовательности x, редакционное расстояние между x и последовательностями специального вида (говорят «словами»), у которых префиксы те же самые, что у x. Эти

[1] Работа выполнена в рамках государственного задания ИППИ РАН, утвержденного Минобрнауки России.

слова однозначно вычисляются по каждому префиксу w в x и имеют вид $w\mathrm{c}(w)$, где функция $(\cdot)$ определена ниже. Само расстояние обозначается $\mathrm{imp}(x)$ как функция от x.

В работе рассматриваются последовательности в четырёх буквенном алфавите. В биоинформатическом контексте буквы называются нуклеотидами и обозначаются `A`, `C`, `G`, `T`; для них определяются так называемые комплементарные пары $\{\mathtt{A},\mathtt{T}\}$ и $\{\mathtt{G},\mathtt{C}\}$. Точнее, определяется отношение комплементарности, которое записывается как $\mathrm{c}(\mathtt{A}) = \mathtt{T}$, $\mathrm{c}(\mathtt{T}) = \mathtt{A}$, $\mathrm{c}(\mathtt{C}) = \mathtt{G}$ и $\mathrm{c}(\mathtt{G}) = \mathtt{C}$; и для любого слова $w = y_1 \dots y_k$ (в роли префикса) $\mathrm{c}(w)$ определяется как $\mathrm{c}(w) = \mathrm{c}(y_k) \dots \mathrm{c}(y_1)$. Хотя результаты приводятся для нуклеотидных алфавита и комплементарности, предлагаемая методика пригодна для любых алфавита и отношения комплементарности на нём; нуклеотидный случай функции $\mathrm{imp}(x)$ выбран вследствие её широкой востребованности в биоинформатике. Слово (так говорят, чтобы отличать его от последовательности x) вида $w\mathrm{c}(w)$, для любого подслова w, называется совершенным палиндромом. Откуда следует, что $|w| = |\mathrm{c}(w)|$ и такое слово чётной длины; где $|\cdot|$ – длина слова или последовательности. В биоинформатике это определение выражают словами: a совпадает с комплементарной к ней последовательностью, читаемой в обратном направлении [1]. Несовершенным палиндромом назовём последовательность x близкую по расстоянию к совершенному палиндрому $w\mathrm{c}(w)$, где w – некоторый префикс в x. Здесь подразумевается достаточная близость и $w\mathrm{c}(w)$, но выбор границы сверху (говорят: порога) этой близости зависит от прикладной задачи. Ниже предложен вариант такого порога, но приведённые компьютерные эксперименты относятся к последовательностям без ограничения на порог. Поиск несовершенных палиндромов и выбор этого порога и весьма нетривиален и важен во многих прикладных задачах, включая биофизические. В частности, несовершенные палиндромы, близкие по последовательности к совершенным, могут служить сайтами кооперативного связывания с ДНК транскрипционных факторов, образующих гомодимеры [2]; такие сайты обычно несовершенные. Несовершенные палиндромы, близкие к совершенным, входят как подпоследовательность в состав многих некодирующих РНК [3]; они распространены в геномах эукариот [4, 5] и прокариот [6]. Вместо префиксов аналогичным образом рассматриваются суффиксы или даже слова, определённым образом расположенные в x. Расстояние между произвольной последовательностью x и соответствующим ей совершенным палиндромом $w\mathrm{c}(w)$ определяется как редакционное расстояние (расстояние Левенштейна) [7], которое затем нормируется длиной последовательности x. Результаты приведены для случая, когда расстояние Левенштейна не использует веса. Это расстояние между двумя последовательностями, как и их наибольшая общая подпоследовательность, вычисляются за квадратичное время от их длины. Известны алгоритмы вычисления редакционного расстояния за немного меньшее время, чем квадратичное [8, 9]; математическое ожидание длины наиболь-

шей общей подпоследовательности двух случайных двоичных последовательностей рассмотрено в [10]. GC-состав нуклеотидной последовательности определяется как доля в ней букв `G` и `C`; можно аналогично рассматривать AT-состав. Напомним, что GC-состав одинаков для обеих цепей ДНК. Например, GC-состав всего генома человека 41%. Если в ДНК человека рассмотреть короткое окно, то в нём GC-состав меняется в широких пределах: в окне длиной несколько сотен нуклеотидов GC-состав меняется от 0% до 100%; а в окне длиной 100 тысяч нуклеотидов GC-состав меняется от 35% до 60%. Мы провели компьютерные эксперименты на основе описанного ниже метода, выясняющие свойства минимального расстояния между случайной последовательностью x и соответствующими ей совершенными палиндромами $w\mathrm{c}(w)$, которые частично приведены в разделе «Результаты».

§ 2. Метод

В [11] нами предложен алгоритм, который по последовательности x получает список длин префиксов w в x (от начала x), для которых нормированное на длину x редакционное расстояние dist между всей $x = wz$ и совершенным палиндромом $w\mathrm{c}(w)$ достигает минимума, вычисляемого по всем общим префиксам w в x и $w\mathrm{c}(w)$. Это расстояние, обозначается $\mathrm{imp}(x)$. Наш алгоритм выдаёт и функцию $\mathrm{imp}(x)$ – значение минимума для последовательности x, равное

$$\mathrm{imp}(x) = \frac{\min\{\mathrm{dist}(x, w\mathrm{c}(w))|x = wz\}}{|x|}$$

Обычно последовательности x и $w\mathrm{c}(w)$ разной длины. Отметим равенство

$$\mathrm{imp}(x) = imp(\mathrm{c}(x)).$$

Значение $\mathrm{imp}(x)$ говорит, насколько произвольная последовательность x отличается от ближайшего к ней совершенного палиндрома вида $w\mathrm{c}(w)$, где w пробегает все префиксы в x. Для x с чётной длиной выполняется $\mathrm{imp}(x) \leqslant \frac{1}{2}$; для x с нечётной длиной выполняется $\mathrm{imp}(x) \leqslant \frac{1}{2} + \frac{1}{2|x|}$, где $|x|$ – длина последовательности x. Например, для x из одной буквы $\mathrm{imp}(x) = 1$. В [12] получен алгоритм поиска (связной) строки u (в последовательности x), которая по редакционному расстоянию достаточно близка к какому-то совершенному палиндрому; и порог этой близости включён в квадратичную оценку сложности алгоритма. Эта и наша задачи важны в биоинформатике и в других приложениях; они различаются как глобальное, у нас, и локальное, в [12], выравнивания последовательностей. В прикладных задачах минимальное отклонение совершенного палиндрома $w\mathrm{c}(w)$ от всей последовательности x с общим префиксом w может быть велико, и величина отклонения имеет, в частности, биофизическое значение, как показано, например, в [2]. Совершенный палиндром $w\mathrm{c}(w)$, на котором достигается минимальное отклонение от x с произвольным общим префиксом w

в x, назовём оптимальным палиндромом для x. Нами доказано, что за квадратичное время от длины исходной последовательности x наш алгоритм строит указанный в начале этого раздела список длин префиксов, в то время как полный перебор всех префиксов в x с вычислением редакционного расстояния для каждого из них требует кубического времени. Компьютерная реализация на языке Python алгоритма, описанного в [11], доступна по адресу http://lab6.iitp.ru/-/pali вместе с примерами её использования.

Компьютерная генерация выборки независимых случайных последовательностей x с данной длиной n выполнялась по случайной величине, у которой вероятности букв `G` и `C` равны $\frac{gc}{2}$, а букв `A` и `T` равны $\frac{1-gc}{2}$, где gc – параметр, $0 \leqslant gc \leqslant 1$. С ростом объёма выборки среднее GC-составов отдельных последовательностей x из выборки стремится к gc, а среднеквадратичное отклонение GC-составов выборки стремится к $\sqrt{\frac{gc\cdot(1-gc)}{n}}$. В этом смысле параметр gc характеризует GC-состав случайной последовательности x с данной длиной n. Разработан скрипт, который генерирует независимые выборки одинакового заданного объёма из независимых случайных последовательностей x с заданной длиной n и данным gc. Ниже приводятся результаты, усреднённые по выборкам объёма сто тысяч при длине последовательностей $n = 1000$, и объёма миллион при длине последовательностей $n = 100$ или $n = 50$. Увеличение объёма и длины $n \geqslant 50$ приводят к качественно сходным результатам.

По выборке (переменная x пробегает её элементы) образуется числовая последовательность значений функции $\mathrm{imp}(x)$ или другой заданной функции $y = f(x)$, у которой одинаковым значениям приписывается кратность $f^1(y)$, т.е. доли вычисляются с учётом кратности, по соответствующему числу x-ов. В этом смысле ниже говорится о характеристиках функции $\mathrm{imp}(x)$ или другой функции $f(x)$.

§ 3. Результаты

На выборках объёма в миллион последовательностей для многих фиксированных длин в интервале 50-100-200 найдены эмпирические зависимости медианы $Me(n)$ и среднеквадратичного отклонения $\sigma(n)$ для функции $\mathrm{imp}(x)$ от длины $n = |x|$ последовательностей x. Типичный вид этих зависимостей показан на рис. 1– 2. Мы предполагаем, что $Me(n)$ при фиксированном GC-составе стремится к пилообразному графику с некоторой периодичностью и убывающей амплитудой; строгое описание предельного вида графика представляло бы значительный интерес.

При $gc = 0,5$ среднеквадратичное отклонение $\sigma(n)$ функции $\mathrm{imp}(x)$ хорошо аппроксимируется (с коэффициентом детерминации $R^2 = 0,999$) быстро убывающей функцией от длины $n = |x|$ последовательностей x:

$$\sigma_{50\%} = \frac{(n-0,6)^{0,35}}{2,58n}$$

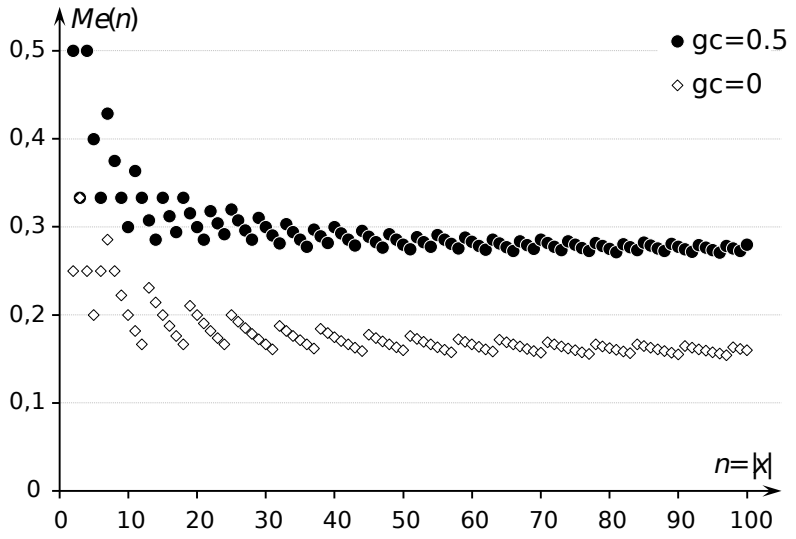

Рис. 1. Зависимость медианы $Me(n)$ (ордината) от данной длины $n = |x|$ (абсцисса) последовательностей x для функции $\mathrm{imp}(x)$ при значениях параметра $gc = 0,5$ и $gc = 0$.

При $gc = 0$ среднеквадратичное отклонение я $\sigma(n)$ аппроксимируется ($R^2 = 0,99$) другой быстро убывающей функцией от длины $n = |x|$ последовательностей x:

$$\sigma_{0\%} = \frac{(n - 0,1)^{0,36}}{2,65n}$$

Обе зависимости показаны на рис. 2а.

Далее, при $gc = 0,5$ на выборках объёма миллион последовательностей с теми же длинами найдена эмпирическая зависимость доли последовательностей x в зависимости от редакционного расстояния dist между x и оптимальным палиндромом $w\mathrm{c}(w)$. На рис. 3 показана зависимость числа N последовательностей x в зависимости от редакционного расстояния dist при $n = 100$. В частности, доля последовательностей, близких к таким совершенным палиндромам, очень мала: не нашлось ни одной последовательности на расстояниях меньше 16 и больше 37.

Колоколообразный вид этой зависимости сохраняется и для других длин и выборок.

Далее, вычислены эмпирические среднее, медиана и другие квантили для функции $\mathrm{imp}(x)$ и случайных нуклеотидных последовательностей в зависимости от длины и параметра gc. Например, для независимых случайных последовательностей x с длиной $n = |x| = 1000$, независимыми позициями и данным gc среднее $\mu(n, gc)$ и среднеквадратичное отклонение $\sigma(n, gc)$ для $\mathrm{imp}(x)$, приведены в табл.1. Для длин свыше 1000 среднее μ почти не зависит от длины n последовательности, рис. 4.

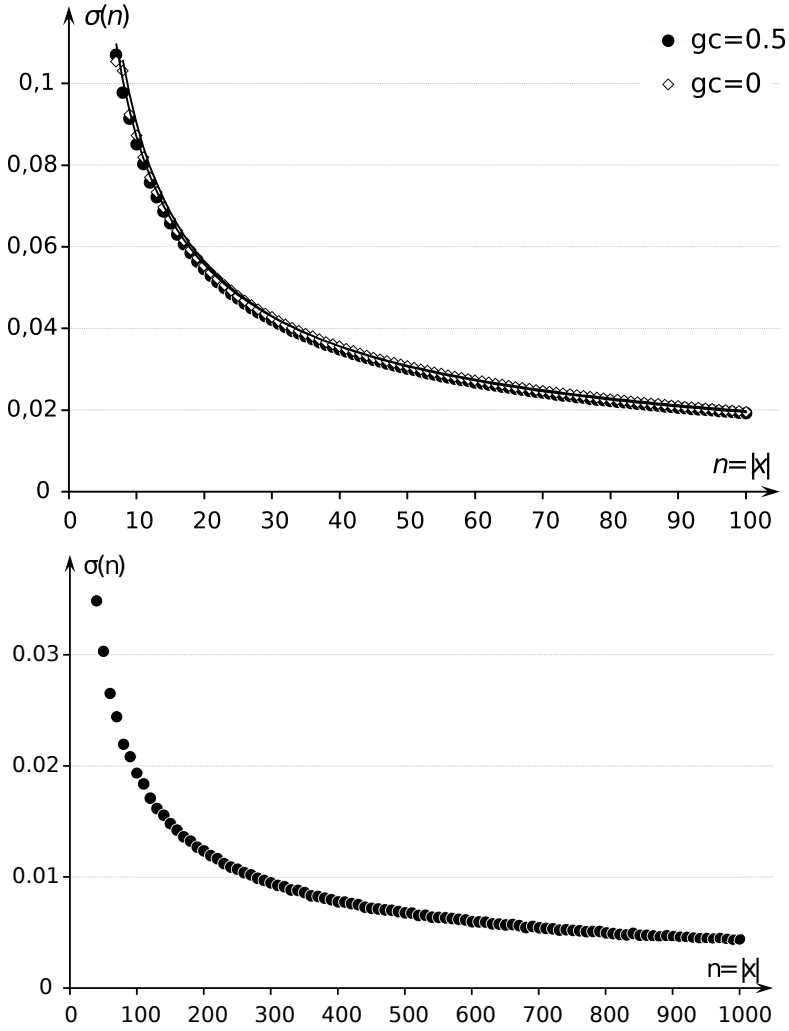

Рис. 2. (a) Зависимость среднеквадратичного отклонения $\sigma(n)$ (ордината) от длины $n = |x|$ (абсцисса) последовательностей x для функции $\mathrm{imp}(x)$ при $gc = 0,5$ и $gc = 0$. На рис. 2a показаны ещё соответствующие регрессии, их формулы приведены ниже. (b) Первая из этих зависимостей для большего интервала длин последовательностей.

Ниже "k%-квантиль равен q" означает, что k процентов выборки последовательностей x удовлетворяет неравенству $\mathrm{imp}(x) \leqslant q$ с наименьшим возможным q; а 50%-квантиль совпадает с медианой. Рассмотрены выборки случайных последовательностей x с длинами $n = |x| = 100$ и $n = |x| = 50$ при различных GC-составах; аналогичные соотношения наблюдаются и при $n = |x| = 1000$. В частности, из последних столбцов табл. 2-3 видно, что при любом GC-составе для почти всех последовательностей x значения $\mathrm{imp}(x)$ близки к медиане.

Далее, наблюдается следующее. При $gc = 0,5$ и всех длинах $6 < n \leqslant 100$ последовательностей x медиана числовой последовательности, составленной из количеств

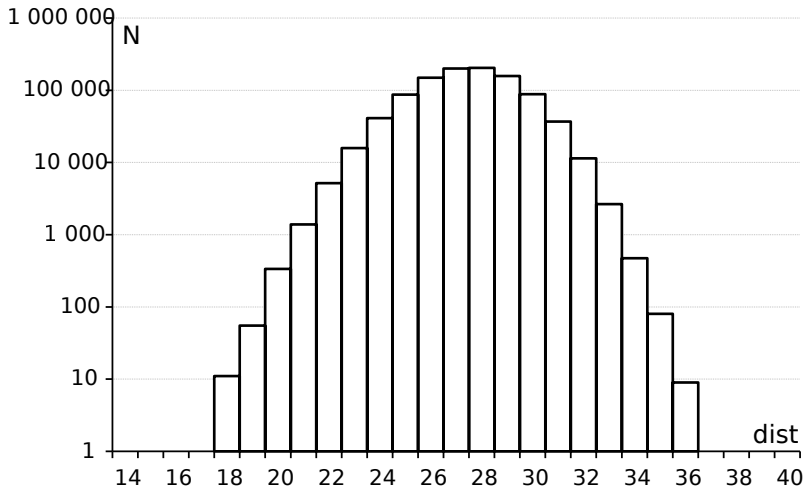

Рис. 3. Число N указано в логарифмическом масштабе. При $gc = 0,5$ показана зависимость числа N (ордината) последовательностей x (с длиной $n = 100$) в выборке объёма миллион от минимального редакционного расстояния dist для функции $\mathrm{imp}(x)$ (абсцисса) между x и оптимальным палиндромом $w\mathrm{c}(w)$.

Таблица 1. Среднее значение $\mu(n, gc)$ и среднеквадратичное отклонение $\sigma(n, gc)$ для функции $\mathrm{imp}(x)$ и длинных последовательностей с $n = |x| = 1000$ в зависимости от GC-состава.

Параметр gc	Среднее $\mu(1000, gc)$ для $\mathrm{imp}(x)$	Отклонение $\sigma(1000, gc)$ для $\mathrm{imp}(x)$
0,00	0,147	0,0044
0,05	0,172	0,0054
0,10	0,193	0,0056
0,15	0,211	0,0056
0,20	0,225	0,0054
0,25	0,237	0,0052
0,30	0,246	0,0049
0,35	0,253	0,0048
0,40	0,258	0,0046
0,45	0,261	0,0044
0,50	0,262	0,0044

оптимальных палиндромов в последовательности x по выборкам с данным n, равна двум; т.е. примерно в половине по x имеется $\geqslant 2$ оптимальных палиндромов. Мода этой последовательности равна 1. При любом $gc \in [0, 0, 5]$ и $n = 100$ 90%-й квантиль количества оптимальных палиндромов равен четырём, то есть примерно 10% последовательностей x соответствует $\geqslant 4$ по оптимальных палиндромов; а 99%-й квантиль равен 7. При этих gc среднее значение по выборкам этой числовой последовательности лежит в интервале $(2, 14, 2, 28)$. Среднее число оптимальных палиндромов при $n = |x| = 100$ и выборках объёма миллион слабо зависит от параметра gc, рис. 5.

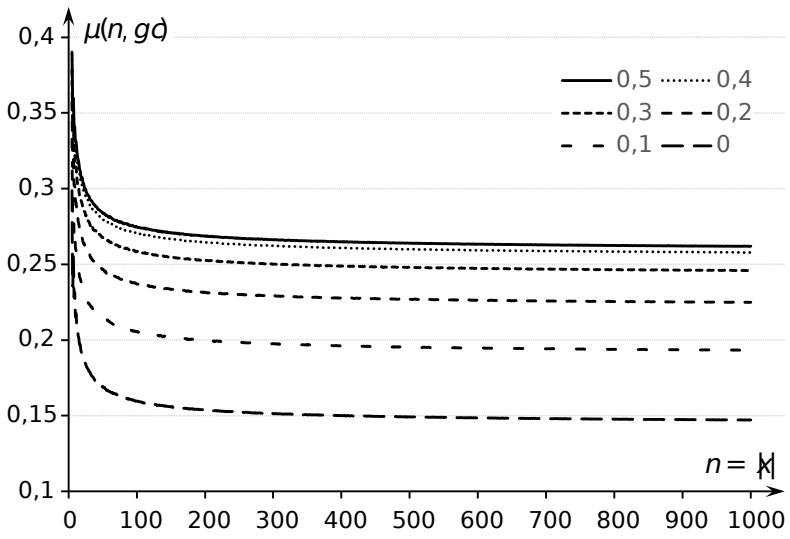

Рис. 4. Зависимости среднего значения $\mu(n, gc)$ для $\mathrm{imp}(x)$ от длины $n = |x|$ последовательностей x при GC-составах 20%, 30%, 40% и 50%.

Таблица 2. Квантили для функции $\mathrm{imp}(x)$ *и среднеквадратичное отклонение* $\sigma(n, gc)$ *для коротких последовательностей* x *с длиной* $n = |x| = 100$

Параметр gc	0,001%	0,01%	0,1%	1%	10%	50%	σ
0,00	0,08	0,09	0,10	0,12	0,14	0,16	0,020
0,05	0,09	0,11	0,12	0,14	0,16	0,18	0,022
0,10	0,11	0,12	0,14	0,15	0,18	0,21	0,022
0,15	0,13	0,14	0,15	0,17	0,19	0,22	0,022
0,20	0,14	0,15	0,17	0,19	0,21	0,24	0,022
0,25	0,15	0,17	0,18	0,20	0,22	0,25	0,021
0,30	0,17	0,18	0,19	0,21	0,23	0,26	0,021
0,35	0,17	0,18	0,20	0,22	0,24	0,27	0,020
0,40	0,18	0,19	0,21	0,22	0,25	0,27	0,020
0,45	0,19	0,19	0,21	0,23	0,25	0,27	0,019
0,50	0,18	0,20	0,21	0,23	0,25	0,28	0,019

Для всех последовательностей x небольшой длины вычислено редакционное расстояние

$$\mathrm{dist}(x) = |x|\mathrm{imp}(x)$$

между x и оптимальным палиндромом. Например, в таблице 4 для всех последовательностей x длины 15 в алфавите $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ (миллиард с небольшим) приведено их распределение в зависимости от значения $\mathrm{dist}(x)$.

Найдены средние значения для функции $\mathrm{imp}(x)$ по всем последовательностям фиксированной длины n, которое назовём «точным» значением. В частности, для длин $n = 15$ и $gc = 0,5$ и длин $n = 30$ и $gc = 0$ отличие этого точного среднего от

Таблица 3. Квантили и среднеквадратичное отклонение $\sigma(n, gc)$ для функции $\mathrm{imp}(x)$ *и ещё более коротких последовательностей x с длиной $n = |x| = 50$.*

Параметр gc	0,001%	0,01%	0,1%	1%	10%	50%	σ
0,00	0,04	0,06	0,08	0,1	0,14	0,16	0,031
0,05	0,06	0,08	0,1	0,12	0,16	0,2	0,033
0,10	0,08	0,1	0,12	0,14	0,18	0,22	0,034
0,15	0,08	0,1	0,12	0,16	0,18	0,24	0,034
0,20	0,1	0,12	0,14	0,16	0,2	0,24	0,033
0,25	0,12	0,14	0,16	0,18	0,22	0,26	0,032
0,30	0,12	0,14	0,16	0,2	0,22	0,26	0,032
0,35	0,12	0,16	0,18	0,2	0,24	0,28	0,031
0,40	0,14	0,16	0,18	0,2	0,24	0,28	0,031
0,45	0,14	0,16	0,18	0,22	0,24	0,28	0,030
0,50	0,14	0,16	0,18	0,22	0,24	0,28	0,030

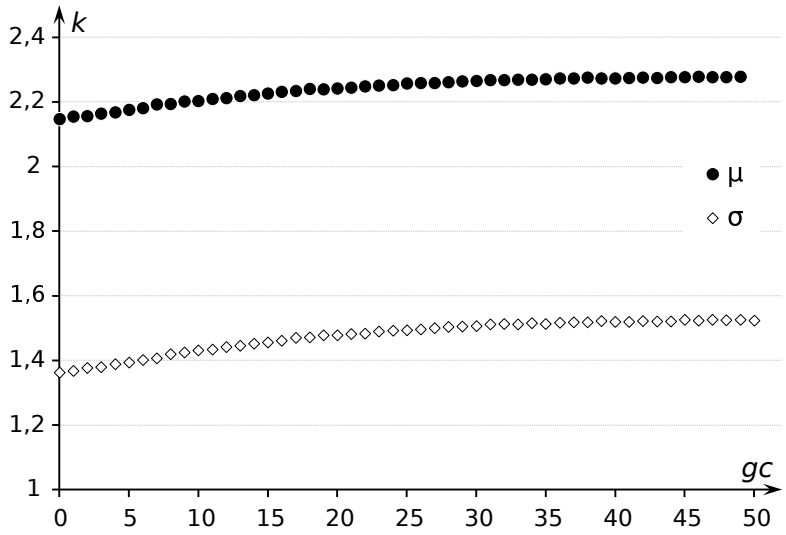

Рис. 5. Среднее число и стандартное отклонение числа оптимальных палиндромов при $n = |x| = 100$ и выборках объёма миллион в зависимости от параметра gc.

среднего по выборкам не превышает 0,2%. Аналогичная точная медиана совпадает с медианой по выборке, если медиана редакционного расстояния между x и оптимальным палиндромом целое число. Например, это условие нарушается при $gc = 0$, то есть для последовательностей в алфавите $\{\mathtt{A}, \mathtt{T}\}$ и длинах $n = 2$ или $n = 6$, когда медианы расстояния равны 0,5 и 1,5 соответственно.

§ 4. Заключение

Нами выполнены обширные компьютерные эксперименты, которые позволили предположить следующие утверждения. Доля случайных последовательностей x с длиной $n \geqslant 100$ и оценкой GC-состава параметром gc, которые близки к оптималь-

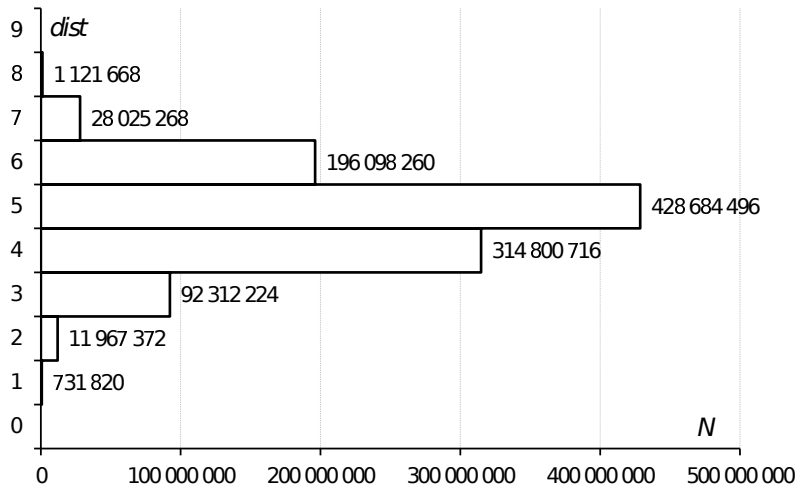

Рис. 6. Для всех последовательностей x длины 15 в алфавите $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ приведено распределение их числа в зависимости от значения $\mathrm{dist}(x)$.

ным палиндромам, очень мала. Распределение числа таких последовательностей x в зависимости от расстояния до оптимальных палиндромов имеет колоколообразную форму, рис. 3. Для почти всех последовательностей x значение $\mathrm{imp}(x)$ близко к среднему значению $\mu(n, gc)$ по всем x. Среднее значение $\mu(n)$ существенно зависит от параметра gc, достигая максимума при $gc = 0,5$, а симметрично расположенные минимумы этого среднего достигаются при $gc = 0$ и $gc = 1$ (с заменой в последовательностях x из выборки букв $\mathtt{A}$ на $\mathtt{G}$ и $\mathtt{T}$ на $\mathtt{C}$). Квантили и среднее значение для функции $\mathrm{imp}(x)$ зависят от длины последовательности x. Значения квантилей стабилизируются при длинах от 1000 букв. Для коротких последовательностей, длины которых типичны для сайтов связывания транскрипционных факторов с ДНК, среднее значение $\mu(n, gc)$ оказывается гораздо выше, рис. 4, чем его предельное значение. Мы предполагаем, что это справедливо для любого GC-состава, хотя скорости сходимости к предельному значению различные. Квантили числовой последовательности значений $\mathrm{imp}(x)$ для случайных последовательностей x указывают на зависимость квантилей от длины x и её GC-состава, что позволяет отсечь последовательности, неотличимые от типичного случая. Иными словами, позволяет по экспериментальным длинам и GC-составу выбрать верхнюю границу $\mathrm{imp}(x)$, при которой экспериментальные последовательности неслучайно близки к оптимальным палиндромам. В частности, при поиске несовершенных палиндромов в высококонсервативных элементах генома [13], которые часто имеют небольшую длину, а их GC-состав значительно отличается от среднего значения по геному. Здесь важно учитывать длину и GC-состав индивидуальных последовательностей в геноме, и трудно подобрать универсальный порог для отсечения типичных случаев. Тем не менее, проведённые компьютерные эксперименты предлагают в биоинформатическом контексте грубую оценку верхней границы при поиске несовершенных палиндромов

x в форме неравенства $0 < \mathrm{imp}(x) < Me(|x|)$. Многие авторы искали алгоритмы для выделения короткого совершенного или почти совершенного палиндрома в длинной последовательности x. Мы предполагаем, что алгоритм низкой полиномиальной сложности для решения такой задачи возможен лишь при ограничениях на $\mathrm{imp}(x)$, которые превращают x практически в совершенный палиндром. Полученные выше оценки также могут найти применение, например, в теории кодирования в связи с q-ичными кодами, устойчивыми относительно появления вставок определённого вида. Коды, включая ДНК-коды, устойчивые к появлению коротких тандемных повторов, рассмотрены в [14]. Аналогичными вставками являются инвертированные повторы, которые образуют совершенный палиндром, когда инвертированный участок встраивается без зазора, или несовершенный палиндром, когда вставка отделена некоторым словом.

СПИСОК ЛИТЕРАТУРЫ

1. T. Mieno, M. Funakoshi, Y. Nakashima, S. Inenaga, H. Bannai, and M. Takeda. Computing maximal palindromes in non-standard matching models. *Information and Computation*, 304(105283):1–20, 2025.
2. R.R. Datta and J. Rister. The power of the (imperfect) palindrome: sequence-specific roles of palindromic motifs in gene regulation. *Bioessays*, 44(4):1–22, 2022.
3. Н.Н. Назипова. Разнообразие некодирующих РНК в геномах эукариот. *Математическая биология и биоинформатика*, 16(2):256–298, 2021.
4. K.V. Mikhailov, B.D. Efeykin, A.Y. Panchin, D.A. Knorre, M.D. Logacheva, A.A. Penin, M.S. Muntyan, M.A. Nikitin, O.V. Popova, O.N. Zanegina, M.Y. Vyssokikh, S.E. Spiridonov, V.V. Aleoshin, and Y.V. Panchin. Coding palindromes in mitochondrial genes of nematomorpha. *Nucleic Acids Research*, 47(13):6858–6870, 2019.
5. O.V. Nikolaeva, A.M. Beregova, B.D. Efeykin, T.S. Miroliubova, A.Yu. Zhuravlev, A.Yu. Ivantsov, K.V. Mikhailov, S.E. Spiridonov, and V.V. Aleoshin. Expression of hairpin-enriched mitochondrial dna in two hairworm species (nematomorpha). *International Journal of Molecular Sciences*, 24(14):1–17, 2023.
6. Л.А. Мирошниченко, Н.А. Арефьева, Ю.П. Джиоев, В.Д. Гусев, А.Ю. Борисенко, С.В. Эрдынеев, and Ю.С. Букин. Структура повторов в геномах сальмонелл. *Математическая биология и биоинформатика*, 18(2):602–620, 2023.
7. В.И. Левенштейн. Двоичные коды с исправлением выпадений, вставок и замещений символов. *Доклады АН СССР*, 163(4):845–848, 1965. Перевод: Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8), 707–710.
8. A. Tiskin. Semi-local longest common subsequences in subquadratic time. *Journal of Discrete Algorithms*, 6(4):570–581, 2008.
9. S. Grabowski. New tabulation and sparse dynamic programming based techniques for sequence similarity problems. *Discrete Applied Mathematics*, 212:96–103, 2016.

10. A. Tiskin. The chvátal–sankoff problem as a problem of symbolic dynamics. *Zapiski Nauchnykh Seminarov POMI*, 528:214–237, 2023.

11. О. Зверков, А. Селиверстов, and Г. Шиловский. Выравнивание скрытого палиндрома. *Математическая биология и биоинформатика*, 19(2):427–438, 2024.

12. M. Alzamel, C. Hampson, C.S. Iliopoulos, Z. Lim, S. Pissis, D. Vlachakis, and S. Watts. Maximal degenerate palindromes with gaps and mismatches. *Theoretical Computer Science*, 978(114182):1–16, 2023.

13. L.I. Rubanov, A.V. Seliverstov, O.A. Zverkov, and V.A. Lyubetsky. A method for identification of highly conserved elements and evolutionary analysis of superphylum alveolata. *BMC Bioinformatics*, 17(385):1–16, 2016.

14. М. Ковачевич. О максимальном числе различимых строк под действием коротких тандемных дупликаций. *Проблемы передачи информации*, 58(2):12–23, 2022.

Хазиев Георгий Андреевич
Зверков Олег Анатольевич
Пирогов Сергей Анатольевич
Селиверстов Александр Владиславович
Любецкий Василий Александрович
Институт проблем передачи информации
им. А.А. Харкевича Российской академии наук, Москва
khaziev@iitp.ru
zverkov@iitp.ru
pirogov@iitp.ru
slvstv@iitp.ru
lyubetsk@iitp.ru

Поступила в редакцию
26.06.2023
После доработки
26.09.2023
Принята к публикации
20.10.2023