Математическая биология и биоинформатика
2024. Т. 19. № 2. С. 427–438. doi: 10.17537/2024.19.427

========== **БИОИНФОРМАТИКА** ==========

Выравнивание скрытого палиндрома

Зверков О.А.*[1], Селиверстов А.В.†[1], Шиловский Г.А.[1,2]

[1]*Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, Москва, Россия*
[2]*Московский государственный университет им. М.В. Ломоносова, Биологический факультет, Москва, Россия*

Аннотация. Цель работы состоит в обобщении известных алгоритмов для решения новых задач, возникающих в биоинформатике. Рассмотрены алгоритмы для оптимизации редакционного расстояния между последовательностями, из которых первая известна, а вторая вычисляется как скрытый палиндром произвольной длины. Существенно, что длина искомого палиндрома определяется в результате оптимизации. В первой задаче требуется выбрать палиндром из ансамбля палиндромов, определяемых второй входной последовательностью. При этом исходные последовательности могут не содержать искомый палиндром целиком. Но вторая последовательность содержит половину от искомого палиндрома. Первая входная последовательность используется для оптимизации. В другой задаче такой палиндром может быть частичным, когда лишь префикс комплементарен суффиксу. Такой частичный палиндром образует шпильку. Новые алгоритмы работают за квадратичное время, что быстрее, чем полный перебор допустимых палиндромов. Алгоритмы существенно используют линейную зависимость редакционного расстояния от длины сплошной делеции или вставки. С другой стороны, алгоритм для решения первой задачи позволяет вычислить сходство данной последовательности с каким-либо палиндромом. Однако в общем случае сравнение двух разных последовательностей не сводится к поиску палиндромов в каждой из них. Также обсуждается быстрый поиск субоптимальных решений. Созданы программные реализации рассмотренных алгоритмов. Они доступны по адресу http://lab6.iitp.ru/-/pali. Приведены некоторые примеры нуклеотидных последовательностей с вырожденными инвертированными повторами. В частности, рассмотрены инвертированные повторы в некодирующих областях ДНК пластид у цветковых растений и гены микроРНК. Также обсуждается возможное применение нашего метода для поиска консервативных вторичных структур РНК.

Ключевые слова: *выравнивание, палиндром, шпилька, редакционное расстояние, биоинформатика, вычислительная сложность.*

ВВЕДЕНИЕ

Вычисление редакционного расстояния (или расстояния Левенштейна [1]) между последовательностями и поиск наибольшей общей подпоследовательности удобны для поиска консервативных элементов генома. Близкие задачи комбинаторной оптимизации также рассматриваются без связи с генетикой [2]. Но мы сосредоточим внимание на инвертированных повторах или палиндромах, которые часто определяют структуру и

*zverkov@iitp.ru
†slvstv@iitp.ru

свойства некоторых биополимеров, включая ДНК и РНК.

В биоинформатике повторяющиеся участки включают сайты связывания транскрипционных факторов и ферментов рестрикции с ДНК [3]. При кооперативном связывании несколько сайтов расположены рядом, вообще говоря, на обеих комплементарных цепях ДНК. Поэтому такие сайты иногда образуют палиндром. С другой стороны, вырожденные инвертированные повторы могут кодировать шпильки на РНК [4, 5]. Такие шпильки также могут участвовать в процессинге РНК [6]. Прямые и инвертированные повторы на ДНК часто возникают в ходе эволюции и распространены как у животных [7, 8], так и у бактерий [9] и в пластидах растений [10, 11]. Поиск палиндромов в одной последовательности рассмотрен в работе [12].

Обычно для глобального выравнивания последовательностей и вычисления редакционного расстояния между ними применяют квадратичный по времени алгоритм Нидлмана – Вунша (Needleman–Wunsch) [13] и его варианты [14, 15]. Согласно общепринятой гипотезе [16], для любого $\varepsilon > 0$ не существует такого детерминированного алгоритма для поиска наибольшей общей подпоследовательности у двух последовательностей, длины которых равны m и n, что время работы ограничено сверху функцией $O((m+n)^{2-\varepsilon})$. В 1980 году Масек (W. J. Masek) и Патерсон (M. S. Paterson) [17] опубликовали алгоритм вычисления редакционного расстояния и поиска наибольшей общей подпоследовательности за время $O(mn/\log_2 n)$ при условии $n \geqslant m \geqslant \log_2 n$. Этот алгоритм использует разбиение исходных последовательностей на блоки равной длины по аналогии с работой [18]. Позже были предложены другие алгоритмы [19, 20]. В частности, в работе [19] предложен алгоритм с оценкой сложности, близкой к оценке для алгоритма Масека – Патерсона, для строк, сжатых методом Лемпеля – Зива (Lempel–Ziv). Известны квантовые алгоритмы, работающие за время, существенно меньшее квадратичного [21]. Также созданы параллельные алгоритмы для поиска наибольшей общей подпоследовательности [22, 23]. Особо рассматривались периодические последовательности [24, 25], последовательности с вырожденными прямыми повторами [26, 27] и последовательности, которые служат перестановками букв (большого) алфавита [28]. Перечисление всех наибольших общих подпоследовательностей рассмотрено в работе [29]. Поиск наибольших подстрок, то есть наибольших общих подпоследовательностей, элементы которых идут подряд, в частности, с данным числом несовпадений рассмотрен в работах [30, 31].

Для приложений в биоинформатике удобно считать, что алфавит состоит из четырёх букв $\{\mathtt{A}, \mathtt{C}, \mathtt{T}, \mathtt{G}\}$, которые связаны отношением комплементарности: A комплементарно к T, C комплементарно к G. Инверсией будем называть замену последовательности x на комплементарную, обозначаемую через $\mathrm{c}(x)$. Например, $\mathrm{c}(\mathtt{AACT}) = \mathtt{AGTT}$. Отметим, что комплементарная (reverse complement) последовательность получается одновременной перестановкой букв в обратном порядке и заменой букв на комплементарные буквы.

Можно рассматривать любой алфавит с заданной инволюцией, определяющей комплементарную букву. Например, алфавит из нуля и единицы, полагая $\mathrm{c}(0) = 1$ и $\mathrm{c}(1) = 0$. Для нуклеотидных последовательностей нуль и единица могут кодировать пуриновые и пиримидиновые основания соответственно.

Мы рассмотрим обобщение выравнивания, при котором на вход подаются две последовательности x и y, а требуется найти оптимальное разбиение последовательности y в виде конкатенации $y = wz$, при котором минимально редакционное расстояние между исходной последовательностью x и некоторой новой последовательностью $w\mathrm{c}(w)$, которая называется палиндромом. Это определение палиндрома отличается от обычного. В частности, если никакая буква алфавита не комплементарна себе, то последовательность нечётной длины не может быть палиндромом. Вместо палиндрома

$w\mathrm{c}(w)$ можно рассматривать частичный палиндром $y\mathrm{c}(w)$, в котором лишь префикс w комплементарен суффиксу $\mathrm{c}(w)$.

Очевидно, такие задачи могут быть решены перебором всех разбиений типа $y = wz$ за кубическое время. Однако модификация известного алгоритма позволяет получить ответ за квадратичное время. При этом также вычисляется значение редакционного расстояния между последовательностями. Поскольку длина искомого участка w произвольная, мы говорим о скрытых палиндромах. Более того, исходные последовательности могут не содержать искомый палиндром целиком. Но вторая последовательность содержит половину от искомого палиндрома.

РЕЗУЛЬТАТЫ

Постановка задач и теоретические результаты

Элементы последовательностей нумеруются с первого, а не с нулевого.

Задача 1. *Для двух последовательностей x и y надо найти оптимальное разбиение последовательности y в виде конкатенации двух $y = wz$, при котором минимально редакционное расстояние между x и палиндромом $w\mathrm{c}(w)$.*

Обозначим через m длину последовательности x, а через n длину последовательности y. Для вычисления редакционного расстояния между последовательностями x и y достаточно вычислить значение $f(m, n)$ функции f, заданной рекуррентными соотношениями

$$f(j,k) = \begin{cases} j, & j \geqslant 0, k = 0 \\ k, & j = 0, k \geqslant 0 \\ \min\{f(j,k-1)+1, f(j-1,k)+1, f(j-1,k-1)+s(j,k)\}, & j > 0, k > 0, \end{cases}$$

где значения $s(j,k) \in \{0,1\}$ зависят от элементов последовательностей x и y:

$$s(j,k) = \begin{cases} 0, & x_j = y_k \\ 1, & x_j \neq y_k \end{cases}$$

Значение $f(j,k)$ равно редакционному расстоянию между префиксами, длины которых равны j и k соответственно. Также можно вычислять редакционные расстояния для суффиксов, начиная с хвостов обеих последовательностей. Но теперь рассмотрим суффиксы для x и $\mathrm{c}(y)$. Обозначим через $g(j,k)$ функцию, заданную рекуррентными соотношениями

$$g(j,k) = \begin{cases} j, & j \geqslant 0, k = 0 \\ k, & j = 0, k \geqslant 0 \\ \min\{g(j,k-1)+1, g(j-1,k)+1, g(j-1,k-1)+r(j,k)\}, & j > 0, k > 0, \end{cases}$$

где значения $r(j,k) \in \{0,1\}$ зависят от элементов последовательностей x и y:

$$r(j,k) = \begin{cases} 0, & x_{m-j+1} = \mathrm{c}(y_k) \\ 1, & x_{m-j+1} \neq \mathrm{c}(y_k) \end{cases}$$

Здесь через $\mathrm{c}(y_k)$ обозначена буква, комплементарная букве y_k в исходной последовательности y. Редакционное расстояние между x и $\mathrm{c}(y)$ равно значению $g(m,n)$.

Обозначим через h функцию $h(j,k) = f(j,k) + g(m-j,k)$. Значение $h(j,k)$ равно сумме двух редакционных расстояний. Первое между двумя префиксами для x и y, а второе между соответствующими суффиксами для x и $\mathrm{c}(y)$. При этом x совпадает с конкатенацией рассмотренных префикса и суффикса для x. Искомое редакционное расстояние между x и $w\mathrm{c}(w)$ равно минимальному значению h при значениях $0 \leqslant j \leqslant m$ и $0 \leqslant k \leqslant n$. При этом значение k, соответствующее этому минимуму, определяет разбиение $y = wz$. Это разбиение и редакционное расстояние вычисляются за время $O(nm)$. Здесь мы отождествляем время работы алгоритма с алгебраической сложностью. Это описание алгоритма для решения задачи 1 легко реализовать в виде программы. Реализация рассмотрена ниже.

Если последовательности x и y совпадают друг с другом, то минимальное расстояние в задаче 1 равно нулю для точного палиндрома и только для него.

А теперь сформулируем вторую задачу, в которой оптимизируется частичный палиндром. Искомая последовательность содержит исходную последовательность y.

Задача 2. *Для двух последовательностей x и y надо найти оптимальное разбиение последовательности y в виде конкатенации двух $y = wz$, при котором минимально редакционное расстояние между x и новой последовательностью $y\mathrm{c}(w)$, которая в общем случае будет частичным палиндромом.*

Будем использовать обозначения из решения задачи 1. Искомое редакционное расстояние между x и $y\mathrm{c}(w)$ равно минимуму суммы $f(j,n) + g(m-j,k)$ при $0 \leqslant j \leqslant m$ и $0 \leqslant k \leqslant n$. При этом значение k, соответствующее этому минимуму, определяет разбиение $y = wz$. Это разбиение и редакционное расстояние вычисляются за время $O(nm)$.

Перед обсуждением других методов рассмотрим примеры решения задач 1 и 2 для $m = 4$ и $n = 3$. В этом случае промежуточные вычисления легко проследить шаг за шагом. После этого рассмотрено обобщение алгоритма Масека – Патерсона. Далее рассмотрены детали программной реализации и другие примеры, имеющие биологический смысл.

Короткий пример

Рассмотрим две последовательности $x = \mathtt{ACGT}$ и $y = \mathtt{ACC}$. Здесь $m = 4$ и $n = 3$. Значение $f(j,k)$ равно редакционному расстоянию между префиксами в x и y, длины которых равны j и k соответственно.

	x	$-$	A	C	G	T
y	$f(j,k)$	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$
$-$	$k=0$	0	1	2	3	4
A	$k=1$	1	0	1	2	3
C	$k=2$	2	1	0	1	2
C	$k=3$	3	2	1	1	2

Значение $g(j,k)$ равно редакционному расстоянию между суффиксами в x и $\mathrm{c}(y)$, длины которых равны j и k соответственно. (Строки и столбцы переставлены.)

	x	$-$	A	C	G	T
c(y)	$g(j,k)$	$j=4$	$j=3$	$j=2$	$j=1$	$j=0$
G	$k=3$	2	1	1	2	3
G	$k=2$	2	1	0	1	2
T	$k=1$	3	2	1	0	1
$-$	$k=0$	4	3	2	1	0

В задаче 1 значения функции $h(j,k) = f(j,k) + g(4-j,k)$ такие:

$h(j,k)$	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$
$k=0$	4	4	4	4	4
$k=1$	4	2	2	2	4
$k=2$	4	2	0	2	4
$k=3$	5	3	2	3	5

Минимальное значение функции h достигается при $k = 2$. Поэтому оптимальное разбиение последовательности y такое: $w =$ AC и $z =$ C. Палиндром wc(w) совпадает с x, редакционное расстояние равно нулю.

Также легко вычислить значения сумм $h_2(j,k) = f(j,3) + g(4-j,k)$ для задачи 2.

$h_2(j,k)$	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$
$k=0$	7	5	3	2	2
$k=1$	6	4	2	1	3
$k=2$	5	3	1	2	4
$k=3$	5	3	2	3	5

Минимальное значение суммы $f(j,3)+g(4-j,k)$ достигается дважды: при $k=1$ и $k=2$. Поэтому в задаче 2 оптимальное разбиение последовательности y либо $w =$ A и $z =$ CC, либо $w =$ AC и $z =$ C. Получаются три оптимальных выравнивания. В первом случае разбиения существует одно оптимальное выравнивание без делеций. Во втором случае разбиения существуют два оптимальных выравнивания:

```
AC–GT
ACCGT
```

и

```
A–CGT
ACCGT
```

Субоптимальные решения

Пусть длины m и n последовательностей x и y кратны числу $p = \lceil (\log_2 n)/\gamma \rceil$, где через $\gamma > 1$ обозначена некоторая константа, зависящая от числа букв в алфавите. В частности, $m = \Omega(\log_2 n)$. Выберем константу γ так, что вычислить таблицы значений $f(j,k)$ для всех пар блоков можно за линейное время $O(2^{\gamma p}) = O(n)$.

Поскольку блоки малого размера, некоторые из них повторяются много раз. Когда

известны таблицы значений $f(j,k)$ для всех пар блоков, вычисление большой таблицы значений $f(j,k)$ для исходных последовательностей легко сводится к добавлению блоков. При этом каждый раз достаточно знать числа на границе квадратной $p \times p$ таблицы, а числа внутри не будут использованы. Это позволяет уменьшить число шагов алгоритма динамического программирования [17].

Используя результат этих предварительных вычислений, за время $O(mn/\log_2 n)$ можно вычислить таблицу значений $h(ip, \ell p)$ с шагом p, а также минимальное из этих значений. Этот минимум определяет субоптимальное разбиение $y = wz$, для которого редакционное расстояние между x и $w\mathrm{c}(w)$ в задаче 1 близко к оптимальному. Также можно найти субоптимальное решение задачи 2.

В случае $p = o(m + n)$ предварительное вычисление таблиц значений $f(j,k)$ для всех пар блоков требует большого времени $2^{O(p)}$, которое осталось субэкспоненциальным. Однако если эти вычисления сделаны один раз, то потом выравнивание любых последовательностей можно выполнять за субквадратичное время, ограниченное функцией вида $O(mn/p)$. Более того, для хранения новых промежуточных данных достаточно памяти $O(mn/p^2)$. Также за это время вычислимы субоптимальные решения для задач 1 и 2. Однако в этом случае значения минимизирумого функционала определяется с большим шагом, следовательно, найденное субоптимальное решение может быть хуже оптимального по сравнению со случаем малого значения p.

Программная реализация

Для демонстрации применимости алгоритма мы используем программу на языке Python. Библиотека NumPy, используемая в данной реализации для представления матриц и выполнения базовых операций с ними, обеспечивает хороший баланс между скоростью разработки и производительностью.

При вычислении редакционного расстояния все замены дают одинаковый вклад в расстояние. Поэтому одновременная замена букв на комплементарные не меняет расстояние между последовательностями. С биологической точки зрения это соответствует симметрии между цепями ДНК. Поэтому значение функции $g(j,k)$, вычисляемое для последовательностей x и y, численно совпадает со значением функции $f(j,k)$ для последовательностей $\mathrm{c}(x)$ и y. Действительно, суффикс для x получается из префикса для $\mathrm{c}(x)$ при одновременной замене букв на комплементарные. Также суффикс для $\mathrm{c}(y)$ получается из префикса для y при одновременной замене букв на комплементарные.

Листинги доступны по адресу http://lab6.iitp.ru/-/pali. Далее приведены результаты работы на реальных примерах.

Инвертированные повторы в пластидах растений

Рассмотрим длинные инвертированные повторы в пластидах растений между сходящимися опероном *psbBT* и геном *psbN* у видов: *Arabidopsis thaliana*, *Aethionema cordifolium*, *Draba nemorosa*, *Barbarea verna*, *Arabis hirsuta*, *Capsella bursa-pastoris*, *Nasturtium officinale*, *Carica papaya*, *Citrus sinensis*, *Gossypium hirsutum*. Предполагается, что этот повтор участвует в терминации транскрипции [10]. В таблице 1 показаны множественное выравнивание и результаты решения задачи 1, когда последовательность x из *A. thaliana*, а y из очередного вида. Вставка нуклеотида у *A. hirsuta* привела к появлению трёх оптимальных палиндромов.

Для тех же последовательностей оптимальное решение задачи 2 соответствует пустому префиксу w, кроме *C. sinensis*, где этот префикс содержит лишь два нуклеотида. Это легко объяснить, поскольку исходные последовательности похожи друг на друга.

Таблица 1. Выравнивание последовательностей, дистанционные расстояния $h_{\min}$ между оптимальным палиндромом $w\mathrm{c}(w)$ и последовательностью из *A. thaliana* и оптимальная длина $|w|$ префикса w из решения задачи 1, где x из *A. thaliana*, а y из очередного вида

Вид	**Последовательность** y	$h_{\min}$	$\lvert w\rvert$
A. thaliana	TTAACGTAATCAGCCTCCAAA–TATTTGGAGGCTGATTACGTTAA	0	22
A. cordifolium	TTGAAGTAATCAGCCTCCAAA–TATTTGGAGGCTGATTACTTCAA	4	22
D. nemorosa	TTGATGTAATCAGCCTCCAAA–TATTTGGAGGCTGATTACATCAA	4	22
B. verna	TTGACGTAATCAGCCTCCAAA–TATTTGGCGGCTGATTACGTCAA	2	22
A. hirsuta	TTGACGCAATCAGCCTCCAAAATATTTGGAGGCTGATTACGTCAA	6	21, 22, 23
C. bursa-pastoris	TTGACGTAATCAGCCTCCAAA–TATTAGGAGGCTGATTACGTCAA	2	22
N. officinale	TTGACGTAATCAGCCTCCAAA–TATTTGGAGGCTGATTACGTCAA	2	22
C. papaya	TTGAAGTAATCAGCCTCCCAA–TATTGGGAGGCTGATTACTTCAA	6	22
C. sinensis	TTGAAGTAATGGGCCTCCCAA–TATTGGGAGGCCCGTTACTTCCT	10	22
G. hirsutum	TTGAAGTAATGAGCCTCCCAA–TATTGGGAGGCTCATTACTTCAA	8	22

Сравнение генов микроРНК MIR195 у млекопитающих

Другим примером служит выравнивание генов MIR195 у человека и у голоносого вомбата *Vombatus ursinus*. Хотя сами микроРНК короткие, предшественниками микроРНК служат некодирующие РНК, образующие длинные шпильки. Выбранный пример интересен тем, что сверхэкспрессия MIR195 в гиппокампе крыс защищает от развития деменции, а ингибирование приводит к нарушению пространственной памяти [6, 32]. Транскрипты MIR195 разной длины: 87 у человека и 67 у вомбата. Идентификаторы генов MIR195 в базе данных Ensembl: у человека ENSG00000284112, у вомбата ENSVURG00010002446. Выравнивание показано в таблице 2.

Задача 1 решена в двух случаях. В первом случае входные последовательности $x = y$ совпадали с MIR195 человека. Их длина 87. Оптимальная длина префикса w равна 42, минимальное редакционное расстояние $h_{\min}$ между $w\mathrm{c}(w)$ и x составило 15. Во втором случае последовательностью x служил ген MIR195 человека, а последовательностью y ген MIR195 вомбата. Оптимальная длина префикса w равна 34, минимальное редакционное расстояние $h_{\min}$ между $w\mathrm{c}(w)$ и x составило 28. В каждом случае оптимальный палиндром отличается от исходной последовательности y. Но правые границы префикса w расположены близко друг от друга на выравниваниии и захватывают несколько нуклеотидов из концевой петли шпильки на РНК.

ОБСУЖДЕНИЕ

В задаче 1 надо выбрать палиндром из ансамбля палиндромов, определяемых второй входной последовательностью, которая обозначена через y. Префиксы палиндромов из этого ансамбля совпадают с префиксами в y. При этом исходные последовательности могут не содержать искомый палиндром целиком. Но в y содержится половина от искомого палиндрома. Первая входная последовательность x нужна для оптимизации. Оптимальный палиндром из ансамбля должен быть на минимальном расстоянии от x.

Таблица 2. Выравнивание MIR195 у человека *Homo sapiens* и вомбата *Vombatus ursinus*. По краям указаны номера нуклеотидов, отсчитываемые от начала последовательности

Вид	Начало		Конец
H. sapiens	1	AGCTTCCCTGGCTCTAGCAGCACAGAAATATTGGCACAGGGAAG	44
V. ursinus	1	CTGGCTTTAGCAGCACAGAAATATTGGCACCTGAGGG	37
H. sapiens	45	CGAGTC–TGCCAATATTGGCTGTGCTGCTCCAGGCAGGGTGGTG	87
V. ursinus	38	AAAGCCATGCCAGTATTGAGAGTGCTGCTC	67

Рассмотренное выравнивание соответствует инверсии участка на одной последовательности ДНК по сравнению с другой [33]. Если последовательность x содержит инвертированный повтор участка w, а последовательность y начинается с прямого повтора того же участка w, возможно, с небольшими изменениями, то в результате работы алгоритма соответствующий участок будет найден. Поскольку длина искомого участка w определяется алгоритмом оптимизации, алгоритм для задачи 1 ищет скрытый палиндром $w\mathrm{c}(w)$. Такие палиндромы могут быть сайтами кооперативного связывания транскрипционных факторов.

Решение задачи 1 в случае, когда исходные последовательности совпадают друг с другом, позволяет оценить, насколько эта последовательность близка к палиндрому. В частности, решение задачи 1 позволяет проверить, образует ли последовательность одну шпильку, возможно, с некоторыми несовпадениями нуклеотидов. Рассматривая две такие последовательности, можно оценить, насколько соответствующие шпильки близки между собой. Далее, рассматривая одновременно три и более последовательностей, можно оценить, насколько они в совокупности близки к одному палиндрому. В этом случае можно вычислять среднее редакционных расстояний $h_{\min}$, вычисляемых при решении задачи 1, между всеми парами последовательностей из рассматриваемого набора.

Частичный палиндром из задачи 2 также может соответствовать шпильке на кодируемой РНК. В этом случае два комплементарных плеча шпильки будут разделены петлей, а составляющие эту петлю нуклеотиды не обязаны быть комплементарными друг другу. Предсказание вторичной структуры РНК на основе термодинамических моделей имеет высокую вычислительную сложность, что требует создания других методов [4, 5]. Поэтому быстрые алгоритмы выравнивания палиндромов могут быть полезны для предсказания такой структуры. В частности, для предсказания новых микроРНК.

Обсуждаемые алгоритмы легко изменить для вычисления расстояний в обобщённой метрике Левенштейна, недавно рассмотренной В. О. Янковским [14]. В этом случае делеции и замены дают разный вклад в расстояние между последовательностями. Однако при поиске палиндромной структуры весовые коэффициенты могут отличаться от таковых, вычисленных на основании частот замен, вставок и делеций вне связи с сохранением палиндромной структуры. Несогласованные замены нуклеотидов нарушают палиндром. Кроме того, даже синхронные замены комплементарных нуклеотидов в составе шпильки РНК изменяют энергию связи. Поэтому применимость обобщённой метрики Левенштейна в этом случае требует уточнения. С другой стороны, наши алгоритмы существенно используют линейную зависимость редакционного расстояния от длины сплошной делеции или вставки. Поэтому не удалось обобщить наш метод, используя так называемый аффинный штраф за делеции.

ЗАКЛЮЧЕНИЕ

Созданы алгоритмы и программная реализация для поиска скрытого палиндрома, граница которого не задана *a priori*. Вычислительная сложность алгоритмов ниже, чем в случае полного перебора вариантов границы инвертированного участка. Примеры иллюстрируют применимость наших алгоритмов в биоинформатике, в частности, для поиска консервативных шпилек.

Исследование выполнено за счет гранта Российского научного фонда № 24-44-00099, https://rscf.ru/project/24-44-00099/.

СПИСОК ЛИТЕРАТУРЫ

1. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов. *Доклады АН СССР.* 1965. Т. 163. № 4. С. 845–848.
2. Леонтьев В.К. Восстановление циклических слов по фрагментам. *Проблемы передачи информации.* 2012. Т. 48. № 2. С. 121–126.
3. Тетуев Р.К., Назипова Н.Н. Статистическая модель предсказания сайтов связывания TALEN с ДНК на основе скользящего среднего. *Матем. биология и биоинформ.* 2023. Т. 18. № 2. С. 621–645. doi: 10.17537/2023.18.621
4. Fu L., Cao Y., Wu J., Peng Q., Nie Q., Xie X. UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Research.* 2022. V. 50. No. 3. P. e14. doi: 10.1093/nar/gkab1074
5. Chen C.C., Chan Y.M. REDfold: accurate RNA secondary structure prediction using residual encoder-decoder network. *BMC Bioinformatics.* 2023. V. 24. Article No. 122. P. 1–13. doi: 10.1186/s12859-023-05238-8
6. Гринкевич Л.Н. Роль микроРНК в обучении и долговременной памяти. *Вавиловский журнал генетики и селекции.* 2020. Т. 24. № 8. С. 885–896. doi: 10.18699/VJ20.687
7. Mikhailov K.V., Efeykin B.D., Panchin A.Y., Knorre D.A., Logacheva M.D., Penin A.A., Muntyan M.S., Nikitin M.A., Popova O.V., Zanegina O.N., Vyssokikh M.Y., Spiridonov S.E., Aleoshin V.V., Panchin Y.V. Coding palindromes in mitochondrial genes of Nematomorpha. *Nucleic Acids Research.* 2019. V. 47. No. 13. P. 6858–6870. doi: 10.1093/nar/gkz517
8. Nikolaeva O.V., Beregova A.M., Efeykin B.D., Miroliubova T.S., Zhuravlev A.Yu., Ivantsov A.Yu., Mikhailov K.V., Spiridonov S.E., Aleoshin V.V. Expression of hairpin-enriched mitochondrial DNA in two hairworm species (Nematomorpha). *International Journal of Molecular Sciences.* 2023. V. 24. No. 14. Article No. 11411. doi: 10.3390/ijms241411411
9. Мирошниченко Л.А., Арефьева Н.А., Джиоев Ю.П., Гусев В.Д., Борисенко А.Ю., Эрдынеев С.В., Букин Ю.С. Структура повторов в геномах сальмонелл. *Матем. биология и биоинформ.* 2023. Т. 18. № 2. С. 602–620. doi: 10.17537/2023.18.602
10. Lyubetsky V.A., Zverkov O.A., Rubanov L.I., Seliverstov A.V. Modeling RNA polymerase competition: the effect of σ-subunit knockout and heat shock on gene transcription level. *Biology Direct.* 2011. V. 6. No. 3. P. 1–16. doi: 10.1186/1745-6150-6-3
11. Зверков О.А., Русин Л.Ю., Селиверстов А.В., Любецкий В.А. Изучение вставок прямых повторов в микроэволюции митохондрий и пластид растений на основе кластеризации белков. *Вестник Московского университета. Серия 16. Биология.* 2013. № 1. С. 8–13.
12. Alzamel M., Hampson C., Iliopoulos C.S., Lim Z., Pissis S., Vlachakis D., Watts S. Maximal degenerate palindromes with gaps and mismatches. *Theoretical Computer Science.* 2023. V. 978. Article No. 114182. P. 1–16. doi: 10.1016/j.tcs.2023.114182
13. Needleman S.B., Wunsch Ch.D. A general method applicable to the search for similarities in

the amino acid sequence of two proteins. *Journal of Molecular Biology.* 1970. V. 48. No. 3. P. 443–453. doi: 10.1016/0022-2836(70)90057-4
14. Янковский В.О. Группы изометрий формальных языков относительно обобщенных метрик Левенштейна. *Математические заметки.* 2024. Т. 116. № 2. С. 306–315. doi: 10.4213/mzm13646
15. *Математические методы для анализа последовательностей ДНК.* Под ред. М.С. Уотермена. М.: Мир, 1999. 349 с.
16. Backurs A., Indyk P. Edit distance cannot be computed in strongly subquadratic time (unless SETH is false). *SIAM Journal on Computing.* 2018. V. 47. No. 3. P. 1087–1097. doi: 10.1137/15M1053128
17. Masek W.J., Paterson M.S. A faster algorithm computing string edit distances. *Journal of Computer and System Sciences.* 1980. V. 20. No. 1. P. 18–31. doi: 10.1016/0022-0000(80)90002-1
18. Арлазаров В.Л., Диниц Е.А., Кронрод М.А., Фараджев И.А. Об экономном построении транзитивного замыкания ориентированного графа. *Доклады АН СССР.* 1970. Т. 194. № 3. С. 487–488.
19. Crochemore M., Landau G.M., Ziv-Ukelson M. A subquadratic sequence alignment algorithm for unrestricted score matrices. *SIAM Journal on Computing.* 2003. V. 32. No. 6. P. 1654–1673. doi: 10.1137/S0097539702402007
20. Tiskin A. Semi-local longest common subsequences in subquadratic time. *Journal of Discrete Algorithms.* 2008. V. 6. No. 4. P. 570–581. doi: 10.1016/j.jda.2008.07.001
21. Akmal S., Jin C. Near-optimal quantum algorithms for string problems. *Algorithmica.* 2023. V. 85. P. 2260–2317. doi: 10.1007/s00453-022-01092-x
22. Тетуев Р.К., Пятков М.И., Панкратов А.Н. Параллельный алгоритм глобального выравнивания протяжённых аминокислотных и нуклеотидных последовательностей. *Матем. биология и биоинформ.* 2017. Т. 12. № 1. С. 137–150. doi: 10.17537/2017.12.137
23. Mishin N., Berezun D., Tiskin A. Efficient parallel algorithms for string comparison. *ICPP '21: Proceedings of the 50th International Conference on Parallel Processing.* 2021. No. 50. P. 1–10. doi: 10.1145/3472456.3472489
24. Tiskin A. Periodic string comparison. In: *Combinatorial Pattern Matching. CPM 2009.* Eds. Kucherov G., Ukkonen E. Springer, Berlin, Heidelberg, 2009. (Lecture Notes in Computer Science, vol. 5577). doi: 10.1007/978-3-642-02441-2
25. Золотов Б.А., Гаевой Н.С., Тискин А.В. Алгоритмы сравнения периодических строк. В: *IV Конференция математических центров России*: сборник тезисов. Санкт-Петербург, 2024. С. 143.
26. Sokol D., Benson G., Tojeira J. Tandem repeats over the edit distance. *Bioinformatics.* 2007. V. 23. No. 2. P. e30–e35. doi: 10.1093/bioinformatics/btl309
27. Sokol D., Tojeira J. Speeding up the detection of tandem repeats over the edit distance. *Theoretical Computer Science.* 2014. V. 525. P. 103–110. doi: 10.1016/j.tcs.2013.04.021
28. Tiskin A. Fast distance multiplication of unit-Monge matrices. *Algorithmica.* 2015. V. 71. P. 859–888. doi: 10.1007/s00453-013-9830-z
29. Conte A., Grossi R., Punzi G., Uno T. Enumeration of maximal common subsequences between two strings. *Algorithmica.* 2022. V. 84. P. 757–783. doi: 10.1007/s00453-021-00898-5
30. Kociumaka T., Radoszewski J., Starikovskaya T. Longest common substring with approximately k mismatches. *Algorithmica.* 2019. V. 81. No. 6. P. 2633–2652. doi: 10.1007/s00453-019-00548-x
31. Amir A., Charalampopoulos P., Pissis S.P., Radoszewski J. Dynamic and internal longest common substring. *Algorithmica.* 2020. V. 82. P. 3707–3743. doi:

10.1007/s00453-020-00744-0
32. Ai J., Sun L.-H., Che H., Zhang R., Zhang T.-Z., Wu W.-C., Su X.-L., Chen X., Yang G., Li K., Wang N., Ban T., Bao Y.-N., Guo F., Niu H.-F., Zhu Y.-L., Zhu X.-Y., Zhao S.-G., Yang B.-F. MicroRNA-195 protects against dementia induced by chronic brain hypoperfusion via its anti-amyloidogenic effect in rats. *Journal of Neuroscience.* 2013. V. 33. No. 9. P. 3989–4001. doi: 10.1523/JNEUROSCI.1997-12.2013
33. Chanin R.B., West P.T., Wirbel J., Gill M.O., Green G.Z.M., Park R.M., Enright N., Miklos A.M., Hickey A.S., Brooks E.F., Lum K.K., Cristea I.M., Bhatt A.S. Intragenic DNA inversions expand bacterial coding capacity. *Nature.* 2024. V. 634. P. 234–242. doi: 10.1038/s41586-024-07970-4

Рукопись поступила в редакцию 22.10.2024.
Переработанный вариант поступил 23.11.2024.
Дата опубликования 05.12.2024.

Mathematical Biology and Bioinformatics
doi: 10.17537/2024.19.427

========== **BIOINFORMATICS** ==========

Alignment of a Hidden Palindrome

Oleg Zverkov[1], Alexandr Seliverstov[1], Gregory Shilovsky[1,2]

[1]*Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia*
[2]*Lomonosov Moscow State University, Faculty of Biology, Moscow, Russia*

Abstract. The aim of the work is to generalize known algorithms to solve new problems arising in bioinformatics. We consider algorithms for optimizing the edit distance between sequences, the first of which is known and the second is a hidden palindrome of arbitrary length. It is important that the length of the desired palindrome is determined as a result of optimization. In the first task, it is necessary to select a palindrome from the ensemble of palindromes defined by the second input sequence. In this case, the original sequences may not contain the desired palindrome entirely. But the second sequence contains half of the desired palindrome. The first input sequence is used for optimization. In another task, such a palindrome may be partial, that is, only a prefix is complementary to a suffix. Such a partial palindrome forms a hairpin. The new algorithms run in quadratic time, which is faster than exhaustive search of admissible palindromes. The algorithms essentially exploit the linear dependence of the edit distance on the length of a continuous deletion or insertion. On the other hand, the algorithm for solving the first task allows us to calculate the similarity of a given sequence to any palindrome. However, in general, comparing two different sequences does not reduce to finding palindromes in each of them. Fast search for suboptimal solutions is also discussed. Software implementations of the considered algorithms are created. They are available at http://lab6.iitp.ru/-/pali. Some examples of nucleotide sequences with degenerate inverted repeats are given. In particular, we consider inverted repeats in noncoding regions of plastid DNA in flowering plants as well as microRNA genes. The possible application of our method to the search for conservative secondary structures of RNA is also discussed.

Key words: *alignment, palindrome, hairpin, edit distance, bioinformatics, computational complexity.*