

MCCMB 2025

Сборник тезисов

12-й Московской
конференции по
вычислительной
молекулярной биологии
(MCCMB)



Новый алгоритм поиска несовершенных палиндромов в ДНК

Г. А. Хазиев, О. А. Зверков, Л. И. Рубанов, А. В. Селиверстов,
В. А. Любецкий

*Институт проблем передачи информации имени А.А. Харкевича
Российской академии наук, 127051, Москва, Большой Каретный
переулок, д.19 стр.1, georgy.haziev@yandex.ru*

Нуклеотидная последовательность называется палиндромом, если она совпадает с комплементарной последовательностью, читаемой в обратном направлении (reverse complement). Последовательность, отличающаяся от (совершенного) палиндрома, называется несовершенным палиндромом. Несовершенные палиндромы, близкие к совершенным, могут быть сайтами кооперативного связывания с ДНК транскрипционных факторов, образующих гомодимеры (Datta and Rister, 2022), а также образовывать шпильки на РНК.

Редакционное расстояние или расстояние Левенштейна между двумя последовательностями вычисляется за квадратичное время (Янковский, 2024).

Нами предложен новый алгоритм, который получает на вход нуклеотидную последовательность x и выдаёт список длин префиксов w , для которых редакционное расстояние $\text{dist}(x, wc(w))$ между исходной последовательностью x и оптимальным палиндромом с таким префиксом w достигает минимума. Каждый префикс w определяет разбиение $x=wz$. Также алгоритм выдаёт минимальное значение редакционного расстояния. При этом надо выбрать палиндром из ансамбля палиндромов, определяемых входной последовательностью. Префиксы палиндромов из этого ансамбля совпадают с префиксами в x . При этом последовательность x может не содержать искомый палиндром целиком. В таком случае она содержит лишь несовершенный палиндром. Но в x содержится половина от искомого палиндрома. Оптимальные разбиения и редакционное расстояние вычисляются за квадратичное время. Вычислительная сложность алгоритма ниже, чем в случае полного перебора вариантов границы инвертированного участка. Для сравнения последовательностей

разной длины удобно использовать функцию $\text{imp}(x)$, значение которой равно отношению длины оптимального палиндрома к длине исходной последовательности. Значение этой функции говорит, насколько последовательность отличается от совершенного палиндрома, и равна нулю для совершенных палиндромов.

Алгоритм реализован на языке Python. Программа доступна по адресу <http://lab6.iitp.ru/~pali>, где также приведен пример использования. Использование библиотеки Numba позволяет значительно уменьшить время работы программы.

Эмпирическая оценка медианы значений $\text{imp}(x)$ для случайных нуклеотидных последовательностей из нескольких сотен нуклеотидов близка к значению 0.26. Но эта медиана немонотонно зависит от длины и заметно выше для коротких последовательностей.

Также разработан второй алгоритм для выделения из последовательности x более короткого несовершенного палиндрома y с меньшим значением $\text{imp}(y) < \text{imp}(x)$, получаемого удалением из исходной последовательности x префикса или суффикса, если это возможно. Эта операция называется триммингом. Основной идеей наших алгоритмов является проверка, верно ли, что одна из оптимальных длин префикса w последовательности x сильно отличается от половины длины x . Если так, то программа усекает x на разницу между длиной w и половиной длины x .

Кроме того, был разработан третий – альтернативный – алгоритм тримминга. Он выполняет заданное число последовательных усечений исходной последовательности x . Первое усечение производится на число нуклеотидов p , заданное как доля от длины x . После каждого усечения выполняется сравнение $\text{imp}(x)$ и $\text{imp}(x')$, где через x' обозначен результат усечения последовательности x на p' . Если выполнено неравенство $\text{imp}(x) > \text{imp}(x')$, то вычисляется следующее усечение уже для x' на p' . Иначе, на следующем шаге вычисляется усечение последовательности x на целую часть от числа $p'/2$ нуклеотидов. Эти алгоритмы также реализованы на языке Python. Программа доступна по адресу <http://lab6.iitp.ru/~trimmers>, где приведены краткое описание и пример использования.

Наш метод применён к последовательностям высококонсервативных элементов (ВКЭ) в некодирующих областях геномов из четырёх млекопитающих (*Homo sapiens*, *Mus musculus*, *Sus scrofa*, *Macaque fascicularis*). Эти ВКЭ были найдены методом, который служит модификацией ранее разработанного авторами алгоритма (Rubanov *et al.*, 2016). Поиск ВКЭ выполнен в синтетических блоках, расположенных у человека на хромосомах 1–3, в которых предварительно были замаскированы точкой все кодирующие последовательности (CDS) и все участки низкой сложности. К началу и к концу каждого блока добавлялся прилегающий участок генома длиной в 1 Мб.

В результате нами выделено много несовершенных палиндромов с малым значением $\text{imp}(x)$. В частности, многие из этих палиндромов расположены в 5'-некодирующих областях генов.

Опишем более подробно ВКЭ, найденные на начальном локусе первой хромосомы человека и в синтетических локусах у других рассмотренных видов. Для тех ВКЭ из этих локусов, в составе которых после тримминга найден участок с $\text{imp}(y)$ не выше 0.2, медиана разностей $\text{imp}(x) - \text{imp}(y)$ составила 0.049. Лишь для 20 последовательностей алгоритм не нашёл лучшей последовательности меньшей длины: $\text{imp}(x) = \text{imp}(y)$. Для одного из ВКЭ в результате тримминга был выделен участок с минимальным значением $\text{imp}(y) = 0.128$, хотя у исходной последовательности $\text{imp}(x) = 0.209$. Наибольшее уменьшение величины $\text{imp}(x)$ составляет 0.232, что лишь немного меньше медианы. В этом случае у исходной последовательности $\text{imp}(x) = 0.406$, а после тримминга $\text{imp}(y) = 0.174$.

Отметим, что сами последовательности ВКЭ обычно мало похожи на палиндромы, поскольку обычно обладают высоким значением $\text{imp}(x)$, близким к медиане. Однако предлагаемый алгоритм выполняет усечение префиксов и суффиксов, позволяя выделить несовершенные палиндромы с низким значением $\text{imp}(x)$, но имеющие сравнительно большую длину.

Выполнен поиск тех несовершенных палиндромов в составе ВКЭ, которые попадают в известные промоторные области. Работа выполнена в рамках государственного задания ИППИ РАН, утвержденного Минобрнауки России.

1. В.О. Янковский (2024) Группы изометрий формальных языков относительно обобщенных метрик Левенштейна, *Математические заметки*, **116**(2): 306–315.
2. R.R. Datta, J. Rister (2022) The power of the (imperfect) palindrome: sequence-specific roles of palindromic motifs in gene regulation, *Bioessays*, **44**(4): e2100191.
3. L.I. Rubanov, A.V. Seliverstov, O.A. Zverkov, V.A. Lyubetsky (2016) A method for identification of highly conserved elements and evolutionary analysis of superphylum Alveolata, *BMC Bioinformatics*, **17**: 385.

Гибридное определение родительского происхождения гаплотипов

Д. А. Хлебников^{1,2}, А. А. Миронов^{1,2}, Alexander Gimelbrant³

¹*Факультет биоинженерии и биоинформатики, Московский Государственный Университет им. М.В. Ломоносова, Москва, Россия*

²*Институт Общей Генетики им. Н.И. Вавилова Российской Академии Наук, Москва, Россия*
dkhlebn@gmail.com

³*Altius Institute for Biomedical Sciences, Seattle, WA*

Гаплотипирование неродственных индивидов базируется на моделировании частот гаплотипов в популяции. Одной из ключевых задач, возникающих в этом контексте, является реконструкция материнских и отцовских гаплотипов в геноме потомка. При наличии данных о тройке (ребёнок и оба родителя), а в некоторых случаях — о двойке (ребёнок и один из родителей), становится возможным определение происхождения гаплотипов каждой хромосомы ребёнка от матери или отца. Однако этот подход сопряжён с существенными финансовыми затратами на секвенирование всех участников трио.

Несмотря на постоянное снижение стоимости секвенирования, сохраняется актуальность разработки более