

О выделении длинной подстроки, близкой к совершенному палиндрому

Хазиев Г.А.¹, Коковихина А.А.², Зверков О.А.¹, Шиловский Г.А.¹,
Селиверстов А.В.¹

¹ Институт проблем передачи информации им. А.А. Харкевича РАН, 127051,
г.Москва, Б. Каретный пер., д.19, стр.1

² ФГАОУ «Национальный исследовательский университет ИТМО» 197101, г.
Санкт-Петербург, Кронверкский проспект, д.49, лит. А.
khaziev@iitp.ru

Аннотация В разных областях часто встречаются совершенные палиндромы – последовательности, комплементарные сами себе. Рассматривается алгоритм `de_sharker`, выделяющий и удаляющий длинный некомплементарный участок в нуклеотидной последовательности. Данный подход позволяет найти последовательности, содержащие длинные подпоследовательности, близкие к палиндромам. Проверена работа алгоритма на наборе палиндромов, встречающихся в РНК.

Ключевые слова: Совершенный палиндром, шпильки, РНК, биоинформатика.

1 Введение

Последовательность называется совершенным палиндромом, если она комплементарна сама себе. В биоинформатике полезно оценить близость последовательности к совершенному палиндрому. Близкие к палиндрому последовательности, например, могут играть роль в экспрессии генов [2]. Ранее авторы разработали алгоритм `palindrome_self_alignment`, находящий разбиение строки x вида $x = wz$, при котором редакционное расстояние между x и конкатенацией $wc(w)$, где c – функция комплементарности, минимально [?]. Для оценки близости последовательности к палиндрому авторы предлагают использовать функцию

$$\text{imp}(x) = \frac{\min\{\text{dist}(x, wc(w)) \mid x = wz\}}{|x|}. \quad (1)$$

Чем ближе последовательность к совершенному палиндрому, тем ближе значение `imp` к нулю.

Часто последовательность может иметь высокое значение функции `imp`, однако содержать в себе достаточно длинную подстроку, близкую к совершенному палиндрому (то есть иметь низкое значение функции `imp`). Отчасти это может быть связано с наличием некомплементарных концов –

termini. Для отделения близкой к палиндром подстроки авторы предложили алгоритмы тримминга [?].

2 Алгоритм de_shapker

Алгоритм de_shapker анализирует строку x используя матрицу H , получающуюся в результате работы алгоритма `palindrome_self_alignment`. В данной матрице алгоритм итеративно ищет наименьшую по l_1 норме непрерывную квадратную подматрицу S_l . Размер подматрицы задаётся как параметр `window_size`. Затем, из индексов (u, v) верхнего левого элемента S_l в матрице H выбирается индекс i начала петли в палиндроме с помощью функции `strategy`. После выбора i , происходит удаление символов с i по $i + \text{window_size}$. Для получившейся в результате строки x' вычисляется значение функции `imp` и сравнивается с значением, полученным на предыдущей итерации (на первой итерации берётся значение для исходной строки x). Если значение `imp` уменьшилось, то предполагается, что обнаружен участок петли и начинается следующая итерация с увеличением `window_size` на значение `window_delta`. Иначе предполагается, что петля либо обнаружена полностью на предшествующей итерации, либо отсутствует в строке и возвращается строка, полученная на последней итерации. После выполнения всех итераций возвращается строка, полученная на последней итерации.

3 Палиндромы в РНК

Авторы проанализировали работу алгоритма `deshapker` на наборе данных miRBase [3], содержащем 38589 последовательностей. Из рассмотрения были исключены вырожденные последовательности. К оставшимся последовательностям предварительно были применены алгоритмы тримминга. Последовательность считалась близкой к палиндрому, если значение функции `imp` для неё не превосходило 0.2. После тримминга было обнаружено 29411 таких последовательностей. Среди последовательностей, которые не были отмечены как близкие к палиндрому был произведён запуск функции `deshapker` с разными функциями `strategy` и разным числом итераций. Параметры `window_size` и `window_delta` равны 2. Результаты поиска последовательностей приведены в таблице 1.

Таблица 1. Число найденных палиндромов в зависимости от выбора функции `strategy` и числа итераций.

Число итераций	$\min(u, v)$	$\max(u, v)$	$\lceil \text{mean}(u, v) \rceil$	$\lfloor \text{mean}(u, v) \rfloor$
3	2577	2287	2303	2577
5	2688	2363	2397	2682
10	2707	2376	2415	2702

При $\text{strategy}(u, v) = \min(u, v)$ и $\text{strategy}(u, v) = \lfloor \text{mean}(u, v) \rfloor$ было найдено больше последовательностей, близких к палиндрому. Также увеличение числа итераций с 5 до 10 давало меньший прирост, чем увеличение числа итераций с 3 до 5. Это может быть связано с малым количеством последовательностей, содержащих длинные петли.

Финансирование (Acknowledgements) Работа выполнена в рамках государственного задания ИППИ РАН, утвержденного Минобрнауки России.

Список литературы

1. Khaziev, G.A., Seliverstov A.V., Zverkov, O.A., Searching for an Imperfect Palindrome Computer algebra : 6th International Conference Materials, Moscow, 23–25 июня 2025 года. – Moscow: RUDN University, 62–65 (2025) <https://doi.org/10.22363/12585-2025-6-013>
2. Ganapathiraju, M.K., Subramanian, S., Chaparala, S. et al. A reference catalog of DNA palindromes in the human genome and their variations in 1000 Genomes. Hum Genome Var **7**(40), 1–12 (2020). <https://doi.org/10.1038/s41439-020-00127-5>
3. Kozomara, A., Birgaoanu, M., Griffiths-Jones, S. miRBase: from microRNA sequences to function. Nucleic Acids Res. **47**(D1) D155–D162 (2019). <https://doi.org/10.1093/nar/gky1141>