РАЗРАБОТКА ПРОГРАММНОГО ПАКЕТА ДЛЯ ПОИСКА СМЕЖНЫХ КОНТИГОВ ПРИ СБОРКЕ ГЕНОМОВ DE NOVO

Н.Р. Мозоль, Л.Н. Валентович

Белорусский государственный университет, Минск, Беларусь
E-mail: nazarmozol@mail.ru

В данном исследовании представлена разработка программного обеспечения, направленного на автоматизацию процесса поиска и анализа смежных контигов в геномных сборках. Сборка геномов de novo метод конструирования геномов из большого количества (коротких или длинных) фрагментов ДНК без априорного знания о правильной последовательности или порядке этих фрагментов. Разработанный программный пакет реализован на языке C++, обладает высокой производительностью и эффективностью, работая быстрее по сравнению с аналогичными решениями. Это обеспечивает оперативное выполнение задач по поиску и анализу смежных контигов в геномных сборках, что значительно экономит время и ресурсы исследователей.

Ключевые слова: сборка генома; контиги; de novo сборка; биоинформатика.

Введение. Задачей сборки генома является восстановление последовательности ДНК (ее длина составляет от миллионов до миллиардов нуклеотидов у разных живых существ) на основании информации, полученной в результате секвенирования. [1].

De novo сборка не требует наличия референсного генома. Учитывая, что большинство организмов на данный момент еще не отсеквенированы, de novo сборка таких организмов может быть использована как первый этап в их изучении. Полезно собирать геном de novo даже при наличии референсного генома, так как при этом можно обнаружить участки, последовательности которых отсутствуют в предшествующей геномной сборке. Однако сборка de novo является алгоритмически сложным и вычислительно затратным процессом. Также данный подход отличается высокой чувствительностью к ошибкам.

Целью работы является разработка алгоритма и программы для поиска смежных контигов при сборке геномов de novo.

Возможности пакета. Разработанный пакет обнаруживает соседние контиги путём сопоставления их концов. Рассчитывается коэффициент LQ, отражающий степень связи между контигами:

$$LQ = \frac{n_a}{n} \cdot 100, \quad (1)$$

где n_a – это количество смежных концов контигов, а n – это общее количество концов контигов (то есть количество контигов = 2).

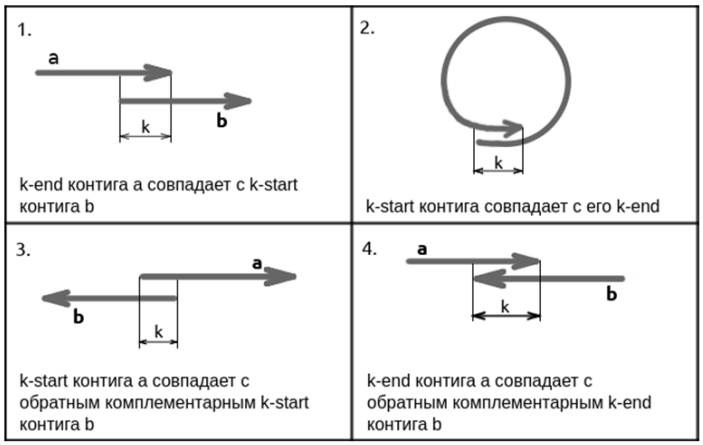

Рис. 1. Варианты смежных контигов

Ожидаемая длина генома вычисляется следующим образом:

$$L_g = \sum_{i=1}^{N_{cont}} L_i^{cont} m_i - \sum_{ovl \in O} L^{ovl}, \quad (2)$$

где L_g – это ожидаемая длина генома, N_{cont} – это количество контигов, L_i^{cont} – это длина i-го контига, O – это набор перекрытий между началами/концами контигов, L^{ovl} – это длина перекрытия ovl, m_i – это множественность i-го контига.

$$m_i = \frac{cov_i}{cov_1}, \quad (3)$$

где cov_i– это покрытие i-го контига.

Если в заголовках последовательностей нет информации о покрытии, множественность рассчитывается следующим образом:

$$m_i = max\big(1.0, min(N_S, N_E)\big), \quad (4)$$

где N_S – это количество контигов, которые перекрываются с началом i-го контига, N_E – с концом [4].

Сравнение с предшествующими программными пакетами. Combinator-FQ, реализованный на языке Python, основан на анлогичном алгоритме. Для сравнения был выбран единый Fasta файл, длиной минимального перекрытия были выбраны 5 пар оснований, длина максимального – 50 пар оснований. Использованы Python 3.11.4, C++ 1.19.9 [2, 3].

Зависимость времени работы программ от количества прочтений имеет характер, показанный на рис. 2. В среднем, разработанный программный пакет оказался на 20% быстрее предшественника, что делает его более выгодным в использовании.

Заключение. В ходе исследования был разработан программный пакет, направленный на автоматизацию процесса поиска и анализа смежных

контигов в геномных сборках. Этот пакет позволяет обнаруживать соседние контиги путем сопоставления их концов, что является важным этапом в сборке геномов de novo.

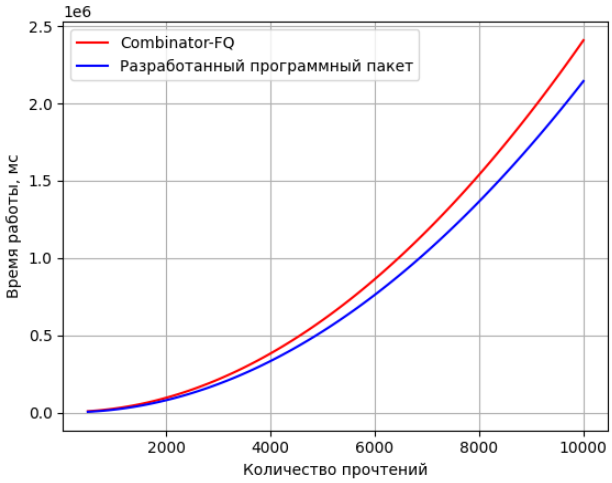

Рис. 2. График скорости выполнения программ

Помимо этого, программный пакет предоставляет разнообразную информацию о смежных контигах, включая коэффициент LQ, который отражает степень связи между контигами, а также ожидаемую длину генома. Эти данные позволяют получать дополнительные инсайты и информацию о структуре и организации генома. Программный пакет реализован на языке C++ и обладает высокой производительностью и эффективностью, что позволяет ему работать быстрее по сравнению с аналогичными решениями. Полученные результаты показывают, что разработанный программный пакет может экономить время и ресурсы исследователей при работе с геномными данными, что делает его ценным инструментом в области биоинформатики и геномики [4].

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. *А.В. Александров, С.В. Казаков, С.В. Мельников, А.А. Сергушичев.* Метод de novo сборки контигов геномных последовательностей на основе совместного применения графов Де Брюина и графов перекрытий [Электронный ресурс]. URL: https://is.ifmo.ru/works/2012/telematika/alexandrov_telematika.pdf (дата обращения: 30.03.2024).
2. Combinator-FQ [Электронный ресурс]. URL: https://github.com/masikol/combinator-FQ/tree/main (дата обращения: 30.03.2024).
3. Detector-Of-Adjacent-Contigs [Электронный ресурс]. URL: https://github.com/Nazarmmm/Detector-Of-Adjacent-Contigs (дата обращения: 30.03.2024).
4. *Сиколенко М. А., Сергеев Р. С., Валентович Л. Н.* Метод оценки полноты нуклеотидных данных для сборки геномных последовательностей на основе расчёта доли смежных контигов [Электронный ресурс]. URL: https://elib.bsu.by/handle/123456789/248660 (дата обращения: 30.03.2024).