УДК 004.048: 631.46
https://doi.org/10.20914/2304-4691-2024-4-20-21

ПРИМЕНЕНИЕ АЛГОРИТМОВ КАРТИРОВАНИЯ ПРИ АНАЛИЗЕ МНОГОМЕРНЫХ ДАННЫХ СЕКВЕНИРОВАНИЯ

А.В. Линкина, Д.В. Шек

Воронежский институт высоких технологий, г. Воронеж, Россия

В настоящее время в результате необходимости обработки значительного объема многомерных массивов данных результатов секвенирования возникает острая необходимость интеграции методов машинного обучения и автоматизации при анализе полученных результатов. Это позволяет существенно ускорить процессы обработки информации и, что особенно важно, повысить точность исследований и устранить аномалии при сопоставлении последовательностей по сравнению с эталонным геномом. Поскольку это является фундаментальной задачей биоинформатики в большинстве современных case-средств, развитие программного инструментария и различных платформ для выполнения высокопроизводительного секвенирования является крайне актуальным. Различные методы картирования решают данные задачи по-разному, однако, исследователями отмечается, что при анализе ранее неопубликованных геномов усложняется задача их обратной «сборки», особенно при выполнении коротких чтений. На сегодняшний день имеется целый ряд программных решений, основанных на анализе результатов «длинных чтений», которые позволяют существенно сократить время и стоимость проведения полной последовательности геномов различных организмов (начиная от бактерий, заканчивая геномом человека). Среди таких платформ все большую популярность набирают решения от Oxford Nanopore Technologies, Pacific Biosciences, Illumina, Thermo Fisher Scientific и ряд других.

Однако, у отмеченных нами инструментов в отличие от секвенирования нового поколения (next-generation sequencing, NGS) при выполнении длинных чтений возрастает процент различного типа ошибок (по некоторым исследованиям принимающий значение до 15 % в сравнении с традиционным -1 %). Поскольку это связано с высокой зашумленностью, современные исследования в области картирования требуют разработки более совершенных алгоритмов выравнивания для эффективной работы при секвенировании de novo. Анализ существующих работ показывает, что наиболее эффективными алгоритмами выравнивания при этом являются IRA, LAMSA и Winnowmap2. Принцип их действия основан на разных механизмах. В LAMSA выравнивание основано на извлечении серии коротких фрагментов и нахождении приблизительного соответствия из каждого чтения на референсном геноме. Затем выполняется построение ациклического графа и выбирается основа (участок, «библиотека») для динамического программирования. Подход динамического программирования состоит в том, чтобы решить каждую подзадачу только один раз, сократив тем самым количество вычислений. Это особенно полезно в случаях, когда число повторяющихся подзадач экспоненциально велико, в частности при выполнении длинных чтений, например, при анализе экспрессии ген Возможность анализа экспрессии генов в биологических системах в биологических системах. Каждый из отобранных ранее участков содержит серии ко-линейных и / или не-колинейных значений. Затем, алгоритм классифицирует пробелы внутри «библиотек» на четыре категории (совпадение, дублирование, удаление и вставка) и заполняет пробелы, чтобы найти точки разрыва генома.

Алгоритм Winnowmap2 основан на специализации метода точного картирования длинных чтений на повторяющихся «библиотеках» в референсном геноме на основе чтений минимальной длины. В данном случае участки можно представить в виде подстрок, которые выравниваются на концах с эталоном с качеством отображения выше заданного пользователем порога. Winnowmap2 первоначально вычисляет MCAS из подмножества равноудаленных друг от друга стартовых позиций. При вычислении MCAS пространство поиска выравнивания

сокращается за счет использования алгоритма взвешенной минимизирующей выборки. Далее алгоритм извлекает якорные точки, которые колинеарны в каждом выравнивании. Наконец, для заполнения пробелов между парами последовательных якорных точек выполняется выравнивание по полосам.

Таким образом, нами проведен сравнительный анализ алгоритмов картирования при проведении секвенирования NGS, которые способствуют упрощению решения задач современной медицины, геномики, филогенетики и многом другим. Показано, что секвенирование неустойчиво к техническим ошибкам, которые могут влиять на научные выводы. Обосновано применение современных методов метаанализа многофакторных данных, которые представляют собой результаты длинных и коротких прочтений и сравнения на референсном геноме.

Работа выполнена при грантовой поддержке Федерального агентства по делам молодёжи (Росмолодёжь) Соглашение № 091-10-2023-069 от 23.05.2023 г. проект «Наука рядом».

Литература

1. Abnizova I., Leonard S., Skelly T., Brown A., Jackson D., Gourtovaia M., Qi G., Te Boekhorst R., Faruque N., Lewis K., Cox T. Analysis of context-dependent errors for illumina sequencing. J. Bioinform. Comput. Biol. 2012; 10(2):1241005

2. Dolled-Filhart M.P., Lee M., Jr., Ou-Yang C.W., Haraksingh R.R., Lin J.C. Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. Sci. World J. 2013:730210

3. te Boekhorst R., Naumenko F.M., Orlova N.G., Galieva E.R., Spitsina A.M., Chadaeva I.V., Orlov Y.L., Abnizova I.I. Computational problems of analysis of short next generation sequencing reads. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2016; 20(6):746–755. DOI 10.18699/VJ16.191