МЕТОД СБОРКИ КОНТИГОВ ГЕНОМНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ИЗ ПАРНЫХ ЧТЕНИЙ С ОШИБКАМИ ВСТАВКИ И УДАЛЕНИЯ НА ОСНОВЕ СОВМЕСТНОГО ПРИМЕНЕНИЯ ГРАФОВ ДЕ БРЁЙНА И ГРАФОВ ПЕРЕКРЫТИЙ[1]

Александров А.В.

студент кафедры компьютерных технологий НИУ ИТМО,
alant@rain.ifmo.ru

Казаков С.В.

студент кафедры компьютерных технологий НИУ ИТМО,
svkazakov@rain.ifmo.ru

Сергушичев А.А.

студент кафедры компьютерных технологий НИУ ИТМО,
alserg@rain.ifmo.ru

Научный руководитель:

Царев Ф.Н.

к. т. н., ассистент кафедры программной инженерии
и верификации программ НИУ ИТМО

Аннотация: Сборка гыштп генома является одной из важных задач современной биологии и медицины. Большинство из современных программ для сборки, ориентированы на работу с чтениями, в которых содержатся ошибки замены. В данной работе предлагается алгоритм, использующий небольшой объем оперативной памяти и позволяющий использовать чтения с ошибками вставки и удаления.

Введение

Многие современные задачи биологии и медицины требуют знания геномов живых организмов, который состоит из нескольких нуклеотидных последовательностей молекул дезоксирибонуклеиновой кислоты (ДНК). В связи с этим возникает необходимость в дешевом и быстром методе секвенирования, то есть определения последовательности нуклеотидов в образце ДНК.

Существующие секвенаторы – устройства для чтения ДНК — не позволяют считать за один раз всю молекулу ДНК. Вместо этого они позво-

[1] Исследование поддержано в рамках Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России на 2009-2013 годы» (Государственный контракт № 16.740.11.0495, соглашение № 14.B37.21.0562)

ляют читать фрагменты генома небольшой длины. Длина фрагмента может быть разной, она является важным параметром секвенирования — от нее напрямую зависит стоимость секвенирования и время, затрачиваемое на чтение одного фрагмента: чем больше длина считываемого фрагмента, тем выше стоимость чтения и тем дольше это чтение происходит. Современные высокопроизводительные секвенаторы основаны на разных подходах со своими достоинствами и недостатками. Одним из возможных недостатков, с которым необходимо справляться, является наличие ошибок вставки и удаления — в прочитанных фрагментах в некоторых местах возможно удаление или вставка символа, например, одной из распространенных ошибок является добавление или удаление одного или нескольких из серии одинаковых подряд идущих нуклеотидов. Часто для сборки используют парные чтения, для которых известно примерное расстояние между ними. Это позволяет получить больше информации о геноме при сравнительно небольших затратах.

Геном обычно представляют в виде нескольких строк из букв A, G, C и T, соответствующих возможным нуклеотидам 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. Этот процесс делится, как правило, на следующие этапы:

1. Исправление ошибок в данных секвенирования.
2. Сборка *контигов* — максимальных непрерывных последовательностей нуклеотидов, которые удалось восстановить.
3. Построение *скэффолдов* — последовательностей контигов, разделенных промежутками, для длины которых найдены верхние и нижние оценки.

Одной из наиболее часто используемых при сборке генома математических моделей является так называемый граф де Брёйна [1]. На его использовании основаны следующие программные средства: *Velvet* [2], *Allpaths* [3], *AbySS* [4], *SOAPdenovo* [5], *EULER* [6].

Одним из недостатков, которым обладают перечисленные программные средства, является то, что эти программы были оптимизированы для работы с секвенаторами компании Illumina [7] и не поддерживают работу с большим числом ошибок вставки и удаления. Также эти методы требуют для своей работы большого объема оперативной памяти.

В настоящей работе предлагается метод, лишенный указанных недостатков. Построение скэффолдов в настоящей работе не рассматривается.

Предлагаемый метод

В работе предлагается модифицировать алгоритм, предложенный в работе [8].

Алгоритм исправления ошибок основан на анализе частот k-меров (строк длины k) и состоит из четырех шагов:

1. Разбиение k-меров на корзины согласно их префиксам. Фиксируется некоторая длина l. Для каждой из 4^l строк длины l выделяется множество k-меров, начинающихся с данного префикса.
2. Подсчет частот k-меров в одной корзине. Проходом по всем данным для каждого k-мера подсчитывается, сколько раз он встречается в наборе чтений, после чего все k-меры делятся на *надежные* (те, что встречаются много раз) и *ненадежные* (те, что встречаются мало раз).
3. Исправление ошибок в k-мерах одной корзины. Для каждого k-мера перебираются все возможные исправления, которые можно осуществить — на каждой позиции можно заменить, удалить или добавить символ. Максимальное число ошибок, исправляемых внутри одного k-мера, является параметром алгоритма. Внутри используемого при разбиении на корзины префикса ошибки не исправляются, благодаря чему исправленные k-меры лежат в той же корзине, что и исправляемые. Если ненадежный k-мер единственным образом превращается в надежный, найденное исправление записывается.
4. Применение исправлений. Для каждого чтения выписываются все k-меру, содержащиеся в нем. Те, к которым найдены исправления, заменяются на соответствующие надежные k-меры. Далее для каждой позиции по правилу консенсуса выбирается символ.

Разбиение на корзины позволяет проводить исправление ошибок и в небольшом объеме памяти.

После исправления ошибок происходит сборка контигов, выполняющаяся в два этапа: сборка квазиконтигов из парных чтений – фрагментов генома, из которых были получены парные чтения, и сборки контигов из квазиконтигов.

Сборка квазиконтигов осуществляется с использованием графа де Брёйна и состоит трех шагов:

1. Подсчет частот k-меров в исправленных чтениях. Выполняется так же, как и при исправлении ошибок.
2. Построение графа де Брёйна. В граф де Брёйна добавляются как вершины все k-меры, встречающиеся не меньше двух раз.
3. Поиск путей в графе де Брёйна, соответствующих фрагментам, из которых были получены парные чтения. Для каждой пары чтений в графе де Брёйна находится множество вершин и ребер, через которые проходит хотя бы один путь, соединяющий эти два чтения, подходящий по длине под априорные границы на расстояние между чтениями. В получившемся подграфе выделяется путь с наибольшим весом и проверяется, что все возможные пути, соединяющие парные чтения совпадают с этим путем с точностью до небольшого

числа ошибок замены, вставки или удаления. Если это так, то последовательность нуклеотидов на этом пути выводится как квазиконтиг.

Разбиение *k*-меров на корзине при подсчете частот и хранение в графе де Брёйна только ребер с их весами позволяет этому этапу также выполняться в небольшом объеме памяти.

Сборка контигов из квазиконтигов осуществляется с использованием графа перекрытий и метода Overlap-Layout-Consensus. Этот этап состоит из трех шагов:

1. Поиск перекрытий (в том числе неточных) между квазиконтигами. Выполняется с использованием суффиксного массива.
2. Построение и упрощение графа перекрытий – графа, в котором вершинами являются квазиконтиги, а ребрами – перекрытия между ними.
3. Вывод контигов. Для этого в упрощенном графе перекрытий находятся пути без ветвлений. Контиг получается консенсусом из множества квазиконтигов на этом пути с соответствующими сдвигами.

Экспериментальные исследования

Экспериментальные исследования проводились на библиотеке чтений бактерии p_stutzeri, геном которой имеет длину около 4 миллионов нуклеотидов. Данные получены секвенатором Ion Torrent. Всего использовалось три библиотеки, одна из которых содержала парные чтения со средним расстоянием между чтениями 2500 нуклеотидов. Общий объем исходных данных составлял 2,9 ГБ.

После сборки было получено 434 контига суммарным размером в 4,6 Мбаз, максимальным контигом длиной 57 Кбаз и с N50 равным 17000. На тех же данных с помощью программы AиуSS было получено 50 тысяч контигов суммарной длиной в 6,3 Мбаз, максимальной длиной 7 Кбаз и N50 равным 660.

Заключение

В работе предложен метод сборки генома, использующий небольшой объем памяти, поддерживающий работу с чтениями, содержащими ошибки вставки и удаления.

Литература

1. *Pevzner P.A.* 1-Tuple DNA sequencing: computer analysis // J. Biomol. Struct. Dyn. 1989. Vol. 7, pp. 63 – 73.
2. *Zerbino D.R., Birney E*. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research. 2008. Vol. 18, pp. 821 – 829.

3. *Butler J., MacCallum I., Kleber M., Shlyakhter I.A., Belmonte M.K., Lander E.S., Nusbaum C., Jaffe D.B.* ALLPATHS: De novo assembly of wholegenome shotgun microreads // Genome Research. 2008. Vol. 18, pp. 810–820.
4. *Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J., Birol I.* ABySS: A parallel assembler for short read sequence data // Genome Research. 2009. Vol. 19, pp. 1117–1123.
5. *Li R., Zhu H., Ruan J., Qian W., Fang X., Shi Z., Li Y., Li S., Shan G., Kristiansen K.,* et al. De novo assembly of human genomes with massively parallel short read sequencing // Genome Research. 2010. Vol. 20, pp. 265–272.
6. *Pevzner P.A., Tang H., Waterman M.S.* EULER: An Eulerian path approach to DNA fragment assembly // Proc. Natl. Acad. Sci. 2001. №98, pp. 9748–9753.
7. Illumina, Inc. [Электронный ресурс]. — Режим доступа: http://www.illumina.com/, свободный. Яз. англ. (дата обращения 23.04.2013).
8. *А.В. Александров, С.В. Казаков, С.В. Мельников, А.А. Сергушичев, Ф.Н. Царев, А.А. Шалыто*. Метод сборки контигов геномных последовательностей на основе совместного применения графов де Брюина и графов перекрытий. // Список-2012: Материалы всероссийской научной конференции по проблемам информатики. — 2012. — С. 415–418.