

МЕТОД СБОРКИ КОНТИГОВ ГЕНОМНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ НА ОСНОВЕ СОВМЕСТНОГО ПРИМЕНЕНИЯ ГРАФОВ ДЕ БРЮИНА И ГРАФОВ ПЕРЕКРЫТИЙ¹

А. В. Александров

магистрант кафедры компьютерных технологий; alexandrov@rain.ifmo.ru

С. В. Казаков

магистрант кафедры компьютерных технологий; svkazakov@rain.ifmo.ru

С. В. Мельников

студент кафедры компьютерных технологий; melnikov@rain.ifmo.ru

А. А. Сергушичев

магистрант кафедры компьютерных технологий; alserg@rain.ifmo.ru

Ф. Н. Царев

аспирант кафедры компьютерных технологий; fedor.tsarev@gmail.com

А. А. Шалыто

*д.т.н., проф., зав. кафедрой технологий программирования;
shalyto@mail.ifmo.ru*

**Санкт-Петербургский государственный университет
информационных технологий, механики и оптики**

Аннотация: В работе предлагается метод сборки контигов геномных последовательностей. Особенностью этого метода является разбиение процесса сборки контигов на два этапа — сборка квазиконтигов из чтений и сборка контигов из квазиконтигов. На первом этапе используется граф де Брюина, на втором — граф перекрытий. Описываются результаты экспериментального исследования разработанного метода на чтениях генома рыбы *Maylandia zebra*, размер генома которой оценивается в один миллиард нуклеотидов. С помощью разработанного метода контиги генома этой рыбы были собраны за пять суток на компьютере с двумя 4-ядерными процессорами и 32 ГБ оперативной памяти.

¹ Исследования выполняются в рамках государственных контрактов № 07.514.11.4010 (заключен в рамках Федеральной целевой программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 годы») и № 16.740.11.0495 (заключен в рамках Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России на 2009–2013 годы»).

Введение

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

Существующие секвенаторы — устройства для чтения ДНК — не позволяют считать за один раз всю молекулу ДНК. Вместо этого они позволяют 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. В связи с этим сейчас получил распространение следующий дешевый и эффективный подход: сначала вычленяется случайно расположенный в геноме фрагмент длиной около 500 нуклеотидов, а затем считываются его префикс и суффикс (длиной порядка 80–120 нуклеотидов каждый). Эти префикс и суффикс называются *парными чтениями*. Описанный процесс повторяется такое число раз, чтобы обеспечить достаточно большое покрытие генома чтениями. Указанным образом работают, например, секвенаторы компании *Illumina* [1].

Отметим, что описанные выше префикс и суффикс читаются с разных нитей ДНК: один — с прямой, другой — с обратно-комплементарной, причем неизвестно, который откуда. Поэтому удобно рассматривать геном и чтения, дополненные своими обратно-комплементарными копиями.

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.

Предлагаемый метод

Процесс сборки 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 Брюина [2]. На его использовании основаны следующие программные средства: *Velvet* [3], *Allpaths* [4], *AbySS* [5], *SOAPdenovo* [6], *EULER* [7].

Одним из недостатков, которым обладают перечисленные программные средства, является большой объем оперативной памяти, необходимый им для сборки генома, сходного по размерам с геномом человека (2–3 миллиарда нуклеотидов). Так, например, *SOAPdenovo* необходимо порядка 140 ГБ оперативной памяти, а *ABuSS* — 21 компьютер с 16 ГБ каждый (всего — 336 ГБ). Такие затраты памяти обусловлены наличием ошибок секвенирования в исходных данных (такие ошибки ведут к увеличению размера графа де Брюина), а также неоптимальным методом хранения этого графа. Другим недостатком существующих методов является отсутствие внутреннего контроля качества сборки.

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

Сборка контигов в предлагаемом методе выполняется в два этапа:

4. Сборка квазиконтигов из чтений геномной последовательности. Квазиконтигами называются последовательности, которые, с одной стороны, длиннее чтений, но, с другой стороны, не являются контигами в смысле невозможности наращивания вправо и влево. Этот этап выполняется с использованием графа де Брюина.
5. Сборка контигов из квазиконтигов. Выполняется с использованием графа перекрытий и метода *Overlap-Layout-Consensus* [8].

Экспериментальные исследования

Экспериментальные исследования разработанного метода проводились в рамках проекта *Assemblathon 2* [9], организованного Калифорнийским университетом в Дэвисе (University of California, Davis), на одном из наборов данных, который был подготовлен организаторами, — наборе чтений рыбы *Maylandia zebra*. Размер генома этой рыбы оценивается примерно в один миллиард нуклеотидов.

Для сборки контигов использовался набор чтений со средним размером фрагмента 180 и 60-кратным покрытием. Общий объем исходных данных составлял 140 ГБ (в сжатом виде), из них авторами были использованы только 44 ГБ.

Алгоритмы сборки генома были реализованы на языке программирования *Java*. Для запуска программ использовался компьютер с 32 ГБ оперативной памяти и двумя 4-ядерными процессорами. Суммарное время работы всех трех этапов — исправления ошибок, сборки квазиконтигов и сборки контигов — составило пять суток. Опишем подробнее результаты каждого из этапов.

Перед исправлением ошибок чтения были обрезаны, чтобы вероятность отдельной ошибки в каждом нуклеотиде не превышала 10%. После этого длина всех чтений в среднем уменьшилась на 20%. Исправление ошибок работало в течение 42 часов. В результате было найдено 150 миллионов ис-

правлений. Всего чтений было 600 миллионов, поэтому было исправлено в среднем каждое четвертое чтение. Сборка квазиконтигов заняла 38 часов. Квазиконтиги были получены из 60% чтений. Сборка контигов выполнялась за 26 часов. В результате было получено 734 165 контигов, суммарный размер которых составляет $680 \cdot 10^6$ нуклеотидов. Длина максимального составляет 23 514 нуклеотидов, средняя длина — 927, значение метрики N50 — 1799.

Заключение

Предложен метод сборки контигов геномных последовательностей, основанный на совместном использовании графа де Брюина и графа перекрытий. Экспериментальное исследование этого метода проведено в рамках проекта *Assemblathon 2*. Это экспериментальное исследование показало, что с помощью разработанного метода можно собирать геномы размером в миллиард нуклеотидов быстрее чем за неделю.

Л и т е р а т у р а

1. Illumina, Inc. [Электронный ресурс]. — Режим доступа: <http://www.illumina.com/>, свободный. Яз. англ. (дата обращения 17.04.2012).
2. Pevzner P. A. 1-Tuple DNA sequencing: computer analysis // J. Biomol. Struct. Dyn. 1989. Vol. 7. Pp. 63–73.
3. 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.
4. 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.
5. 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.
6. 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.
7. Pevzner P. A., Tang H., Waterman M. S. EULER: An Eulerian path approach to DNA fragment assembly // Proc. Natl. Acad. Sci. 2001. No. 98. Pp. 9748–9753.
8. International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome // Nature. Vol. 409. No. 6822. Pp. 860–921.
9. Проект Assemblathon 2. [Электронный ресурс]. — Режим доступа: <http://www.assemblathon.org>, свободный. Яз. англ. (дата обращения 17.04.2012).