МЕТОД *DE NOVO* СБОРКИ КОНТИГОВ ГЕНОМНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ НА ОСНОВЕ СОВМЕСТНОГО ПРИМЕНЕНИЯ ГРАФОВ ДЕ БРЮИНА И ГРАФОВ ПЕРЕКРЫТИЙ

А.В. Александров, С.В. Казаков, С.В. Мельников, А.А. Сергушичев
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики
E-mail: alserg@rain.ifmo.ru

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. 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].

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:

- Исправление ошибок в данных секвенирования.
- Сборка контигов — максимальных непрерывных последовательностей нуклеотидов, которые удалось восстановить.
- 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*, *Allpaths*, *AbySS*, *SOAPdenovo*, *EULER*.

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 *ABySS* – 21 компьютер с 16 Гб каждый (всего – 336 Гб). Такие затраты памяти обусловлены наличием ошибок секвенирования в исходных данных, что ведет к увеличению размера графа де Брюина, а также неэкономным методом его хранения.

В настоящей работе предлагается метод, лишенный указанного недостатка. Построение скэффолдов в настоящей работе не рассматривается.

Сборка контигов в предлагаемом методе выполняется в два этапа:

- 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.
- Сборка контигов из квазиконтигов. Выполняется с использованием графа перекрытий и метода Overlap-Layout-Consensus [3].

Экспериментальные исследования

Экспериментальные исследования разработанного метода проводились в рамках проекта *Assemblathon 2* [4], 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 исправлений. 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 выполнилась за 26 часов. В результате было получено 734165 контигов, суммарный размер которых составляет 680321319 нуклеотидов. Длина максимального составляет 23514 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.

Литература

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. International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome // Nature. Vol. 409. № 6822, pp. 860 – 921.
4. Проект Assemblathon 2. [Электронный ресурс]. – Режим доступа: http://www.assemblathon.org, свободный. Яз. англ. (дата обращения 17.04.2012).