Сборка генома и технология MapReduce

Текст А. А. Сергушичев, Ф. Н. Царёв

Иллюстрация В. Камаев

Сборка генома – одна из центральных задач биоинформатики. Это объясняется тем, что без ее решения нельзя приступить к детальному изучению генома живого существа и его анализу с применением других алгоритмов биоинформатики.

Изучение генома человека и других живых существ имеет важное прикладное значение. На основании результатов сборки генома конкретного человека возможна реализация персонифицированной медицины – определения предрасположенности человека к различным болезням, создание индивидуальных лекарств и т. д. Кроме этого, на основе результатов исследования геномов растений и животных с использованием методов биоинженерии могут быть выведены новые их виды, обладающие определенными свойствами.

Исследования в области секвенирования и сборки геномов ведутся в мире с 70–80-х годов XX века. Основная сложность состоит в том, что молекулы дезоксирибонуклеиновой кислоты (ДНК), в которых записана генетическая информация, обычно имеют очень большую длину (миллионы нуклеотидов даже для простейших организмов). Современные технологии не позволяют считывать всю молекулу целиком. Поэтому единственным способом получить информацию о геноме является чтение большого числа небольших фрагментов из разных частей молекулы ДНК (рис. 1). Так как при этом информация о взаимном расположении фрагментов теряется, возникает задача сборки генома – восстановление всей последовательности по ее фрагментам. Задача усложняется тем, что в геномах обычно есть большое число повторяющихся частей, что приводит к тому, что полностью восстановить достаточно большой геном не удается.

Одним из наиболее значительных проектов в этой области «Геном человека» (Human Genome Project) был запущен в 1990 году и завершен в 2003 году получением 99% геномной последовательности человека. В рамках этого проекта для секвенирования использовались так называемые технологии первого поколения, в частности, секвенирование методом обрыва цепи (*http://en.wikipedia.org/wiki/Sanger_sequencing*). Интересным фактом является то, что практически одновременно с этим проектом частной компанией Celera Genomics под руководством Крейга Вентера (Craig Venter) был завершен аналогичный проект.

В середине первого десятилетия XXI века широкое распространение получили технологии секвенирования второго поколения (они также часто называются технологиями следующего поколения, или next generation sequencing). Эти технологии позволяют существенно быстрее и дешевле получать на порядок большие объемы данных о геномной последовательности. Их существенным отличием от технологий первого поколения является то, что длины чтений при использовании составляют 36–100 нуклеотидов, а не 500–1000, как при использовании технологий первого поколения. Технологии секвенирования второго поколения активно развиваются, поэтому требуется постоянная разработка новых алгоритмов сборки генома по данным секвенирования и совершенствование существующих.

По оценкам экспертов эти технологии в настоящее время развиваются существенно быстрее, чем растет производительность компьютеров. В соответствии с законом Мура производительность компьютеров удваивается каждые полтора года, а производительность геномных секвенаторов за тот же самый период увеличивается в десять раз. График, приведенный на рис. 2 и иллюстрирующий падение стоимости секвенирования генома человека, теперь можно увидеть практически в каждой презентации, посвященной задаче сборки генома. Отметим, что стоимость секвенирования не включает в себя стоимость сборки генома и его последующего анализа.

В настоящее время ведущими корпорациями, разрабатывающими и производящими секвенаторы нового поколения, являются Illumina, Life Technologies и Roche. Среди секвенаторов можно выделить Illumina HiSeq 2000, который способен за один рабочий цикл длительностью порядка недели считать 200 миллиардов нуклеотидов, разбитых на чтения по 100 символов, и Roche FLX Junior, который за 10 часов считывает 35 миллионов нуклеотидов, разбитых на чтения по 400 символов.

Отличительной особенностью секвенаторов Illumina и Life

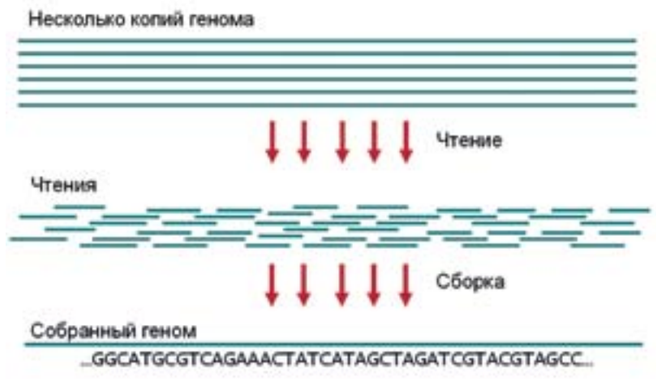

Рис. 1. Чтения и сборка генома

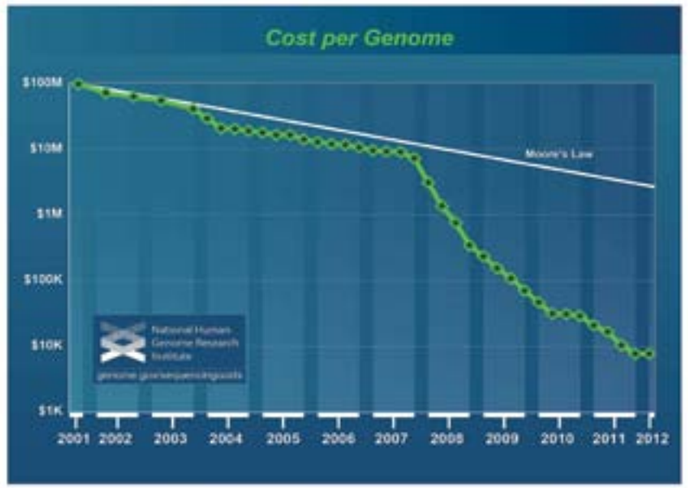

Рис. 2. Падение стоимости секвенирования

Technologies SOLiD является их возможность генерировать так называемые парные чтения. Это означает, что чтения длиной 50–100 нуклеотидов разбиты на пары и в каждой паре известно примерное расстояние между концами чтений, которое может достигать 500 нуклеотидов (рис. 3). Это позволяет иметь компромисс между низкой стоимостью секвенирования и длиной чтений.

Существует два основных подхода к сборке генома:

- *Сборка с использованием референсного генома – применяется, когда известен геном того же или близкого биологического вида.*
- *De novo сборка – сборка генома без использования референсного генома. Она, в частности, находит применение при изучении раковых клеток.*

Данная статья посвящена задаче de novo сборки. В целом задача сборки генома сложна по нескольким причинам:

- *достаточно большой обьем входных данных – типичный входной обьем данных для de novo сборки генома человека составляет 100 ГБ;*
- *сложность структуры генома – наличие в нем повторов и полиморфизмов;*
- *наличие ошибок в исходных данных.*

Стоит отметить, что геном человека размером в 3 миллиарда нуклеотидов вовсе не является самым большим в природе. Есть целый ряд растений, используемых в народном хозяйстве и имеющих большие геномы: ячмень (5.3 миллиарда нуклеотидов), сосна (от 18 до 40 миллиардов нуклеотидов) и др. Кроме того, некоторые организмы имеют существенно больший размер генома: цветок *Pieris japonica* имеет размер генома в 150 миллиардов нуклеотидов, геном рыбы *Protopterus aethiopicus* – 130 миллиардов нуклеотидов, что

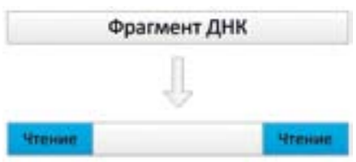

Рис. 3. Парные чтения

может быть интересно с научной точки зрения.

Существенным отличием задачи сборки генома от традиционно решаемых на суперкомпьютерах (гидродинамические расчеты) является то, что она ориентирована на данные (data-intensive), то есть большую часть времени занимают не вычисления, а операции ввода/вывода, сортировка, поиск и др.

Одним из подходов к решению задач, ориентированных на данные, является использование технологии MapReduce – подхода к написанию распределенных программ, предложенного компанией Google для упрощения работы над большими объемами данных на больших кластерах.

Технология MapReduce предполагает представление алгоритма в виде двух функций: map и reduce, а данных – в виде пар <ключ, значение>. Функция map принимает одну пару <ключ, значение> и выдает несколько, функция reduce принимает на вход ключи и несколько значений и тоже выдает несколько пар. MapReduce-фреймворк сначала для каждой пары <ключ, значение> в исходных данных вызывает функцию map, затем группирует полученные пары по ключу и для каждой группы вызывает функцию reduce. То, что выдали функции reduce, является результатом. В общем случае программа, использующая MapReduce, может состоять из нескольких шагов, каждый из которых содержит map-фазу и reduce-фазу.

Важным достоинством MapReduce является то, что он устойчив к отказу одного или нескольких вычислительных узлов, так как невыполненные вычисления можно провести повторно на других узлах.

Как же MapReduce может помочь при сборке генома? Одной из подзадач является задача исправления ошибок в данных секвенирования. Ее можно решать на основе частот появлений подстрок фиксированной длины k (k-меров) в исходных данных. Это следует из того, что геном покрыт чтениями многократно и, соответственно, k-меры, присутствующие в нем, встречаются в чтениях достаточно много раз, а отсутствующие могут появиться только из-за ошибок и поэтому будут встречаться редко (рис. 4, на графике по оси абсцисс отложено число вхождений k мера в чтения, а по оси ординат – число k-меров, (шкала по оси ординат – логарифмическая). При исправлении ошибок ставится цель из каждого k-мера, встречающегося редко, получить k-мер, встречающийся часто, заменой одного или нескольких символов.

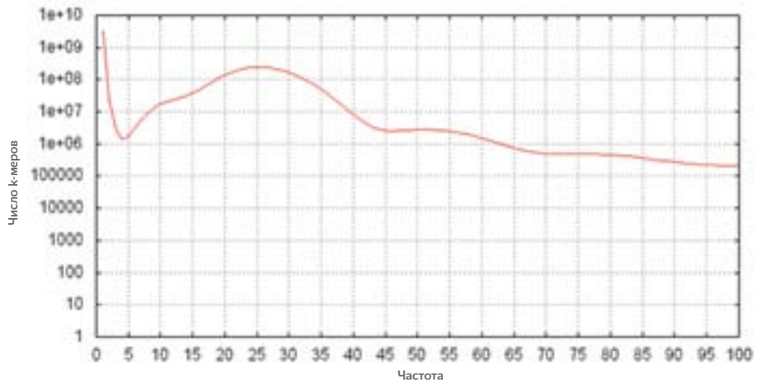

Рис. 4. График зависимости числа k-меров от частоты их вхождений

Первым шагом является подсчет частот k-меров, который может быть выполнен вариацией классического MapReduce-алгоритма подсчета числа слов *(http://hadoop.apache.org/docs/r0.20.2/mapred_tutorial.html#Example%3A+WordCount+v1.0)*. Поиск исправлений также может быть записан «на языке» MapReduce. При использовании такого алгоритма исправление ошибок в чтениях объемом 100 ГБ требует около 30 минут на 9000 восьмиядерных узлах суперкомпьютера «Ломоносов» МГУ имени М. В. Ломоносова (к сожалению, этот эксперимент не был доведен до конца, так как распределенная файловая система не выдержала нагрузки). Для сравнения, исправление ошибок на 24-ядерной машине с 24 ГБ оперативной памяти заняло сутки. Вероятно, эффективность распределения может быть еще больше увеличена.

Для распределенной сборки генома требуется распределить чтения по разным компьютерам так, чтобы их можно было обрабатывать более или менее независимо. Для этого можно применить следующий алгоритм. Это можно сделать, если ввести понятие близости чтений.

Идея распараллеливания алгоритма последующих шагов состоит в следующем: исходные данные (чтения с внесенными в них исправлениями) разбиваются на группы, в каждой из которых они были прочитаны из близких позиций.

Далее эти группы обрабатываются независимо друг от друга.

Для такого разбиения можно применить следующий алгоритм кластеризации чтений. Построим граф общих k-меров чтений, в котором вершины соответствуют чтениям, а ребра – наличию общих k-меров у соответствующих чтений, и последующем выделении компонент с большим числом ребер внутри них. Для выделения таких компонент применяется аналог алгоритма обхода в ширину, реализованный при помощи технологии MapReduce. В каждую компоненту входит некоторая вершина этого графа и вершины, расположенные на расстоянии, не превосходящем заданную величину. В рамках экспериментальных исследований была проведена сборка генома бактерии E. Coli на кластере НИИ НКТ (НИУ ИТМО).

Всего было использовано 2 миллиона чтений.

Кластер НИИ НКТ НИУ ИТМО состоял из десяти вычислительных узлов с четырьмя процессорами Intel Xeon Processor X5570 с тактовой частотой 2.93 ГГц и оперативной памятью 24 ГБ. Все узлы были соединены в Gigabit Ethernet сеть.

Время работы описанного алгоритма составило 180 минут. Значение метрики N50 составило 4718. Доля генома, не покрытая контигами, составила 6.69%.

Также был проведен запуск сборщика Contrail. Время его работы составило 100 минут. Значение метрики N50 составило 672. Доля генома, не покрытая контигами, составила 4.91%. При сборке на одном узле с помощью современных сборщиков достигается значение N50, большее 50 000, что говорит о том, что над MapReduce-алгоритмами сборки генома надо еще работать и работать.

Результаты работы MapReduce-алгоритма исправления ошибок позволяют надеяться, что эта работа завершится успехом. Работа выполняется в лаборатории «Алгоритмы сборки геномных последовательностей» кафедры «Компьютерные технологии» Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики.