ПАРАЛЛЕЛЬНЫЙ АЛГОРИТМ DE NOVO СБОРКИ ГЕНОМА С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИИ MAPREDUCE

А.В. Александров, С.В. Казаков, С.В. Мельников, А.А. Сергушичев, П.В. Федотов, Ф.Н. Царев, А.А. Шалыто

Аннотация

В работе рассматриваются алгоритмы сборки протяженных фрагментов геномной последовательности (контигов).

Предложен алгоритм сборки контигов, основанный на технологии MapReduce, который является сверхмасштабируемым и параллельным.

Он применим, в том числе, на кластерах петафлопсного и экзафлопсного уровней производительности.

Введение

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

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.

Задача разработки методов сборки геномных последовательностей является, в определенном смысле, центральной среди всех 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.

В середине первого десятилетия XXI века широкое распространение получили так называемые технологии next generation sequencing (технологии секвенирования нового поколения). По оценкам экспертов [1], эти технологии в настоящее время развиваются существенно быстрее, чем компьютерные технологии и алгоритмы сборки геномных последовательностей.

Сборка генома из набора фрагментов, полученных на секвенаторе, – алгоритмически и вычислительно сложная задача, решение которой невозможно без использования кластеров. Например, каждая сборка генома печеночного сосальщика, основанная на данных нескольких запусков секвенатора, требует до недели работы кластера из двух десятков узлов по восемь ядер и 8 Гб оперативной памяти в каждом, объединенных по интерфейсу MPI [2]. Однако одного запуска почти всегда недостаточно — таких сборок может быть несколько из-за необходимости подбора оптимальных параметров алгоритма и добавления новых экспериментальных данных.

Данные секвенирования - чтения геномной последовательности (длина порядка 100 нуклеотидов), геном покрыт чтениями несколько десятков раз.

Длина генома может быть от нескольких миллионов (бактерии) до сотен миллиардов нуклеотидов.

Технология секвенирования следующая: сначала вычленяется случайно расположенный в геноме фрагмент, а затем происходит считывание двух последовательностей с его концов. Эти последовательности называются парными чтениями. Процесс повторяется такое число раз, чтобы обеспечить достаточно большое покрытие генома чтениями.

Заметим, что размеры фрагментов могут варьироваться от средних (около 200 нуклеотидов) до достаточно больших (10000 нуклеотидов), а размеры чтений в целом уменьшаются при увеличении размера фрагмента. Например, в секвенаторе Genome Analyser IIx компании Illumina [3], использующем такую технологию, размер фрагмента составляет примерно 500 нуклеотидов, а размер чтений – около 100.

На сегодня процедура работы секвенатора и сборки генома на кластере отличаются по времени в зависимости от конкретного оборудования и используемых алгоритмов, но в целом – это величины одного порядка. Выше было отмечено, что в ближайшие годы темпы роста производительности секвенаторов ожидаются более высокими по сравнению с ростом производительности кластеров, и поэтому "узким" местом в получении геномной последовательности будет именно процедура восстановления генома, выполняемая после получения результатов работы секвенатора.

Использование существующих в настоящее время алгоритмов на персональных компьютерах приведет к тому, что сборка одного генома займет месяцы, а может растянуться и на год. Для успешного решения этой задачи необходимо переходить на новые алгоритмы для кластеров, в том числе петафлопсного и экзафлопсного уровней производительности.

Сборщик ABySS

Сборщик ABySS основан на подходе Message Passing Interface. В качестве основной структуры данных используется граф де Брюина.

Подход к сборке контигов в программном средстве ABySS изложен в статье [4]. Этот подход состоит из двух этапов:

- сборка контигов без учета парной информации;
- разрешение неоднозначностей с помощью парной информации и наращивание контигов.

В основе всего подхода лежит хранение графа де Брюина в распределенной хеш-таблице.

Для того, чтобы собрать первоначальные версии контигов происходит объединение последовательностей смежных однозначных ребер — ребро называется однозначным, если исходящая степень его начальной вершины и входящая степень конечной вершины равны единице.

На втором этапе между контигами устанавливаются связи, используя парную информацию. Пара чтений называется связывающей два контига, если первое чтение картируется на первый контиг, а второе — на второй. Между двумя контигами устанавливается связь, если число связывающих их чтений больше некоторой константы p (по умолчанию используется p = 5). Для каждого контига C строится множество связанных с ним контигов P. Затем в графе связей контигов ищется уникальный путь, проходящий через все контиги из P. В качестве ограничений при поиске выступают оценка на расстояния между контигами на основе принципа максимального правдоподобия и эвристическая оценка на число посещенных вершин. После того, как поиск таких путей для каждого контига завершился (успешно или нет), согласующиеся пути сливаются, образуя конечные контиги.

Сборщик Contrail

Сборщик Contrail описан в работе [5]. Этот сборщик использующет фреймворк MapReduce, что позволяет ему легко масштабироваться. Этот метод использует граф де Брюина в качестве основной структуры данных.

Для построения графа просматривается каждое чтение и создаются пары (u, v) — ребра графа де Брюина для двух последовательных k-меров u и v. Затем ребра для одного и того же k-мера группируются.

После этого производятся четыре типа упрощений графа: сжатие путей, удаление ошибок, раздвоение вершин, из которых выходит несколько ребер и, если доступны парные чтения, разрешение небольших повторов. Для сжатия неветвящихся путей используется параллельное рандомизированное ранжирование списков [6], для исправления ошибок – параллельный поиск шаблонов в сети [7].

Предлагаемый параллельный алгоритм сборки геномных последовательностей

В настоящей работе предлагается сверхмасшитабируемый параллельный метод сборки геномных последовательностей, который основан на использовании технологии MapReduce [8]. Сборку генома предлагается осуществлять в три этапа – исправление ошибок в чтениях, сборка квазиконтигов, сборка контигов. Архитектура сборщика изображена на рис. 1.

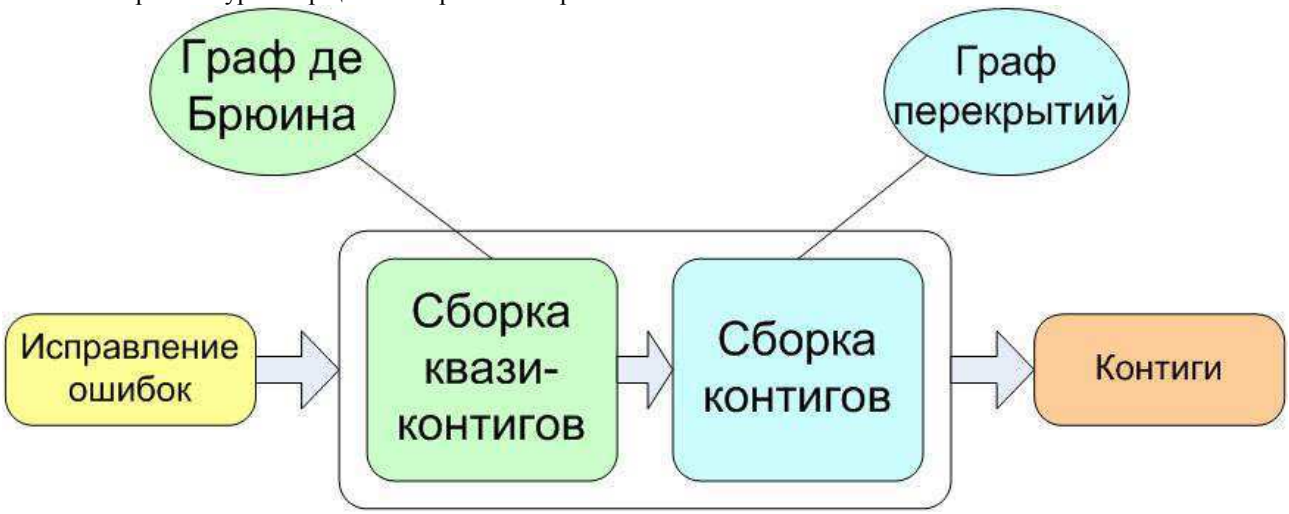

Рис. 1. Архитектура сборщика

Алгоритмы для каждого из этапов отличаются от своих «последовательных» версий, предложенных авторами в работах [9, 10]. Предлагаемые алгоритмы построены на основе распараллеливания по данным.

Основной идеей масштабирования алгоритма исправления ошибок в данных секвенирования (наборе чтений геномной последовательности) является независимая обработка групп k-меров с различными префиксами. Для каждой такой группы k-меров независимо определяются, какие k-меры являются достоверными, и в них нет ошибок чтения, а в каких вероятность ошибки высока. Для тех k-меров, которые не

являются достоверными, определяется, какие нуклеотиды в них были прочитаны с ошибками, и на какие нуклеотиды их необходимо исправить.

После этого для каждого исходного чтения геномной последовательности выполняется исправление ошибок на основании информации о том, как надо исправлять каждый k-мер.

Идея распараллеливания алгоритма последующих шагов (сборки квазиконтигов и сборки контигов) состоит в следующем - исходные данные (чтения геномной последовательности с внесенными в них исправлениями) разбиваются на группы, в каждой из которых они были прочитаны из близких позиций исходной геномной последовательности. Далее эти группы обрабатывается независимо друг от друга. Для такого разбиения разработан алгоритм кластеризации чтений геномной последовательности.

Он основан на построении графа общих k-меров чтений, в котором вершины соответствуют чтениям, а ребра числу 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 [8].

В каждую компоненту входит некоторая вершина этого графа и вершины расположенные на расстоянии, не превосходящем заданную величину.

На рис. 2 изображен граф для кластеризации для пяти чтений и k равным трем.

На рис. 3 показана компонента, выделенная в графе: исходное чтение обозначено красным цветом, чтения на расстоянии один – желтым, чтения на расстоянии два – зеленым.

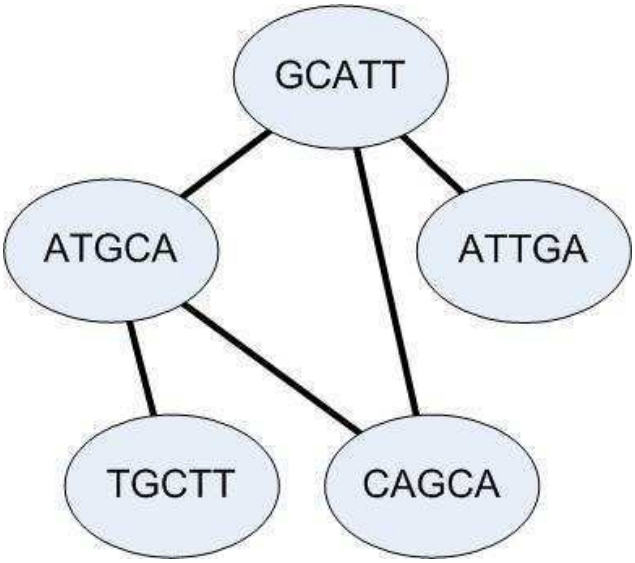

Рис. 2. Граф для кластеризации

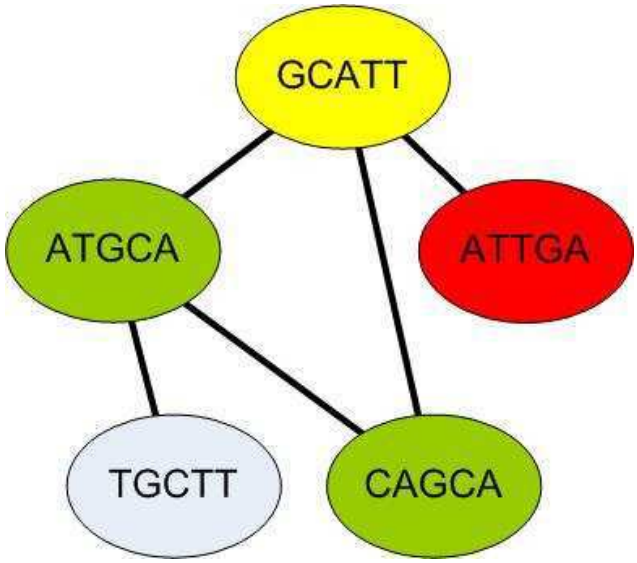

Рис. 3. Граф с выделенной компонентой

Для сборки контигов из квазиконтигов последние разбиваются на группы, близких по положению в геноме. Для каждой из таких групп применятся алгоритм, основанный на подходе Overlap-Layout-Consensus [11]. Для осуществления разбиения квазиконтигов на группы применяется алгоритм, аналогичный описанному выше алгоритму кластеризации чтений геномной последовательности.

Для увеличения размера получаемых контигов выполняется несколько итераций сборки контигов, при этом контиги, полученные на предыдущей итерации, используются в качестве входных данных для следующей.

Экспериментальные исследования

Описанный сверхмасштабируемый параллельный алгоритм сборки геномных последовательностей был реализован на языке программирования Java с использованием программного фреймворка Apache Hadoop [12].

В рамках его экспериментальных исследований были проведена сборка генома бактерии E. Coli [13] на кластере НИИ НКТ (НИУ ИТМО).

В качестве исходных данных были использованы парные чтения генома бактерии Escherichia Coli с покрытием 40.

Данные парные чтения имеют длину 36 нуклеотидов и расстояние между концами 200 нуклеотидов. Всего было использовано 2 миллиона чтений.

Исследования проводились на вычислительном кластере научно-исследовательского института наукоемких компьютерных технологий (НИИ НКТ) НИУ ИТМО.

Кластер НИИ НКТ НИУ ИТМО состоял из следующих вычислительных узлов:

головного сервера:

четыре процессора Intel® Xeon® Processor X5570 с тактовой частотой 2.93 ГГц;

оперативная память – 24 ГБ с тактовой частотой 1.3 ГГц;

десяти вычислительных узлов со следующей конфигурацией:

два процессора Intel® Xeon® Processor X5570 с тактовой частотой 2.93 ГГц;
оперативная память – 24 ГБ с тактовой частотой 1.3 ГГц;

Все узлы были соединены в Gigabit Ethernet сеть.

Для проведения экспериментов использовался фреймворк Hadoop версии 0.20.205.

Каждый из десяти вычислительных узлов был:

- узлом распределенной файловой системы Hadoop Distributed File System (сервис DataNode), в качестве хранилища данных использовался локальный жесткий;
- вычислительным узлом (сервис TaskTracker), на каждом узле одновременно выполнялось не более 8 операций Map, и не более 4 операций Reduce.

Время работы описанного алгоритма составило 180 минут. Значение метрики N50 [14] 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%.

Также была проведена сборка чтений искусственного генома, использовавшегося в проекте de novo Genome Assembly Assessment Project (размер генома 1,8 миллиарда нуклеотидов) [15] на суперкомпьютере "Ломоносов" МГУ имени М.В. Ломоносова [16]. При запуске на 30000 процессорных ядрах были исправлены ошибки в чтениях указанного генома, однако проблемы с записью в файловую систему не позволили провести остальные этапы сборки. Данный эксперимент показал, что предложенный метод исправления ошибок масштабируется на большое число узлов.

Заключение

В работе предложен параллельный алгоритм сборки геномных последовательностей, основанный на технологии MapReduce.

Экспериментальные исследования показали, что для бактериальных геномов значение метрики N50 у контигов, полученных при помощи разработанного алгоритма, существенно превосходит значение этой метрики у контигов, полученных с помощью сборщика генома Contrail. Однако доля генома, не покрытого контигами, оказалась немного больше, чем доля у результатов сборщика Contrail.

Из этих сравнений можно сделать следующие выводы:

- результаты предложенного алгоритма и алгоритма сборщика Contrail в целом схожи;
- предложенный метод превосходит метод Contrail'a по качеству сборки контигов, ввиду того, что основной характеристикой качества сборки является значение метрики N50.

Кроме этого, экспериментальная проверка показала масштабируемость алгоритма исправления ошибок в данных секвенирования для вычислительной системы с 30000 процессорными ядрами.

Исследования проводились в рамках государственного контракта № 07.514.11.4010 от 19.08.2011 г. (заключен в рамках Федеральной целевой программы "Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы").

При проведении работ был использован суперкомпьютер "Ломоносов" МГУ имени М.В. Ломоносова.

ЛИТЕРАТУРА:

1. Зубов В. В. Приборы для чтения ДНК // Химия и жизнь. 2010. № 7, с. 4 – 7. [Электронный ресурс]. – Режим доступа: www.dubna-oez.ru/images/data/gallery/10_2948_.pps, свободный. Яз. рус. (дата обращения: 01.06.2012).
2. Прохорчук Е. Б. Код жизни: прочесть не значит понять. [Электронный ресурс]. – Режим доступа: http://biomolecula.ru/content/778/, свободный. Яз. рус. (дата обращения: 01.06.2012).
3. Illumina, Inc. [Электронный ресурс]. – Режим доступа: http://www.illumina.com/, свободный. Яз. англ. (дата обращения: 01.06.2012).
4. Simpson J. T., Wong K., Jackman S. D., Schein J. E., Jones S. J. M., Birol I. Abyss: a parallel assembler for short read sequence data. // Genome Res. Jun 2009. Vol. 19, no. 6. Pp. 1117–1123.
5. Schatz M., Sommer D., Kelley D., Pop M. Contrail: Assembly of Large Genomes using Cloud Computing. [Электронный ресурс]. – Режим доступа: http://sourceforge.net/apps/mediawiki/contrail-bio/index.php?title=Contrail, свободный. Яз. англ. (дата обращения: 05.07.2012).
6. Wyllie J. C. The Complexity of Parallel Computation. Ph.D. thesis, Department of Computer Science, Cornell University. 1979.
7. Schatz M., Cooper-Balis E., Bazinet A. Parallel network motif finding. 2008.
8. Dean J., Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters. [Электронный ресурс]. – Режим доступа: http://research.google.com/archive/mapreduce.html, свободный. Яз. англ. (дата обращения: 05.07.2012).

9. Разработка метода сборки геномных последовательностей на основе восстановления фрагментов по парным чтениям. Отчет за второй этап. НИУ ИТМО. 2011.
10. Александров А. В., Казаков С. В., Мельников С. В., Сергушичев А. А., Царев Ф. Н., Шалыто А. А. Метод исправления ошибок в наборе чтений нуклеотидной последовательности //Научно-технический вестник Санкт-Петербургского государственного университета информационных технологий, механики и оптики. 2011. № 5, с. 81–84.
11. International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860–921.
12. Apache Hadoop. [Электронный ресурс]. – Режим доступа: http://hadoop.apache.org/, свободный. Яз. англ. (дата обращения: 01.06.2012).
13. NCBI: Experiment: SRX000429 – Illumina sequencing of Escherichia coli str. K-12 substr. MG1655 genomic paired-end library.
14. N50 statistic. [Электронный ресурс]. – Режим доступа: http://en.wikipedia.org/wiki/N50_statistic, свободный. Яз. англ. (дата обращения: 05.07.2012).
15. De novo Genome Assembly Assesment Project. [Электронный ресурс]. – Режим доступа: http://cnag.bsc.es, свободный. Яз. англ. (дата обращения: 01.06.2012).
16. Суперкомпьютер "Ломоносов". Общая характеристика. [Электронный ресурс]. – Режим доступа: http://parallel.ru/cluster/lomonosov.html, свободный. Яз. рус. (дата обращения: 01.06.2012)