МЕТОД СБОРКИ ГЕНОМА С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИИ MAPREDUCE[1]

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 представлены алгоритмы сборки протяженных фрагментов геномной последовательности (контигов), основанные на технологии *MapReduce*. Данные алгоритмы являются сверхмасштабируемыми параллельными и предназначены для сборки геномных последовательностей. Они применимы, 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

[1] Исследования выполняются в рамках государственного контракта № 16.740.11.0495 (заключен в рамках Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России на 2009–2013 годы»).

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

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.

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.

В настоящей работе предлагается сверхмасшитабируемый параллельный метод сборки геномных последовательностей, который основан на использовании технологии *MapReduce* [4]. Сборку генома предлагается осуществлять в три этапа — исправление ошибок в чтениях, сборка квазиконтигов, сборка контигов. Алгоритмы для каждого из этапов отличаются от своих «последо-

вательных» версий, предложенных авторами в работах [5, 6]. 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 также состоит в распараллеливании по данным — 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 алгоритмом, аналогичным алгоритму сборки квазиконтигов [5]. 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* [3]. В каждую компоненту входит некоторая вершина этого графа и вершины расположенные на расстоянии, не превосходящем заданную величину.

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*. 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.

Предложенные алгоритмы были реализованы на программном фреймворке *Apache Hadoop* [4]. Были проведена сборка генома бактерии *E. Coli* [7] на кластере НИИ НКТ (НИУ ИТМО). В результате эксперимента были получены контиги с метрикой N50 более 2000, что показывает применимость данного алгоритма для сборки бактериальных геномов.

Также была проведена сборка чтений искусственного генома, использовавшегося в проекте *de novo Genome Assembly Assessment Project* (размер ге-

нома 1,8 миллиарда нуклеотидов) [8] на суперкомпьютере «Ломоносов» МГУ имени М. В. Ломоносова. При запуске на 30000 процессорных ядер 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.

Литература

1. *Зубов В. В.* Приборы для чтения ДНК // Химия и жизнь. 2010. № 7. С. 4–7; www.dubna-oez.ru/images/data/gallery/10_2948_.pps [дата просмотра: 20.04.2012].
2. *Прохорчук Е. Б.* Код жизни: прочесть не значит понять. http://biomolecula.ru/content/778/ [дата просмотра: 20.04.2012].
3. *Cohen J.* Graph Twiddling in a MapReduce World // Computing in Science & Engineering. 2009. Vol. 11. No. 4. Зр. 29–41.
4. Apache Hadoop. [Электронный ресурс]. — Режим доступа: http://hadoop.apache.org/, свободный. Яз. англ. [дата просмотра 11.04.2012].
5. Разработка метода сборки геномных последовательностей на основе восстановления фрагментов по парным чтениям. Отчет за второй этап. НИУ ИТМО. 2011.
6. *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.
7. NCBI: Experiment: SRX000429 — Illumina sequencing of Escherichia coli str. K-12 substr. MG1655 genomic paired-end library
8. De novo Genome Assembly Assesment Project. [Электронный ресурс]. — Режим доступа: http://cnag.bsc.es, свободный. Яз. англ. (дата обращения 11.04.2012).