Министерство образования и науки Российской Федерации

УДК: 004.021
ГРНТИ: 20.01.01, 34.05.25
Инв. №: 310276

УТВЕРЖДЕНО:
Исполнитель: **федеральное государственное бюджетное образовательное учреждение высшего профессионального образования "Санкт- Петербургский национальный исследовательский университет информационных технологий, механики и оптики"**
Руководитель организации **_____/В. Н. Васильев/** **М.П.**

НАУЧНО-ТЕХНИЧЕСКИЙ ОТЧЕТ

о выполнении четвертого этапа Государственного контракта
№ 16.740.11.0495 от 16 мая 2011 г. и Дополнению от 25 октября 2011 г. № 1

Исполнитель: федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Санкт- Петербургский национальный исследовательский университет информационных технологий, механики и оптики»
Программа (мероприятие): Федеральная целевая программа «Научные и научно-педагогические кадры инновационной России» на 2009-2013 гг., в рамках реализации мероприятия № 1.2.1. Проведение научных исследований научными группами под руководством докторов наук.
Проект: Разработка метода сборки геномных последовательностей на основе восстановления фрагментов по парным чтениям
Руководитель проекта: **_____/Шалыто Анатолий Абрамович** (подпись)

Санкт-Петербург
2012 г.

СПИСОК ОСНОВНЫХ ИСПОЛНИТЕЛЕЙ

по Государственному контракту 16.740.11.0495 от 16 мая 2011 на выполнение поисковых научно-исследовательских работ для государственных нужд

Организация-Исполнитель: федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»

Руководитель темы:

доктор технических наук, профессор _____ Шалыто А. А.
подпись, дата

Исполнители темы:

доктор технических наук, профессор _____ Парфенов В. Г.
подпись, дата

кандидат технических наук, без ученого звания _____ Корнеев Г. А.
подпись, дата

кандидат технических наук, без ученого звания _____ Станкевич А. С.
подпись, дата

без ученой степени, без ученого звания _____ Царев Ф. Н.
подпись, дата

без ученой степени, без ученого звания _____ Федотов П. В.
подпись, дата

без ученой степени, без ученого звания _____ Буздалов М. В.
подпись, дата

без ученой степени, без ученого звания _____ Александров А. В.
подпись, дата

без ученой степени, без ученого звания _____ Казаков С. В.
подпись, дата

без ученой степени, без ученого звания _____ Мельников С. В.
подпись, дата

без ученой степени, без ученого звания _____ Сергушичев А. А.
подпись, дата

РЕФЕРАТ

Отчет 72 с., 1 ч., 6 рис., 3 табл., 14 источн., 3 прил.

Геном, сборка генома, контиг, квазиконтиг, оценка качества сборки.

В отчете представлены результаты исследований, выполненных по третьему этапу Государственного контракта № 16.740.11.0495 «Разработка метода сборки геномных последовательностей на основе восстановления фрагментов по парным чтениям» (шифр «2011-1.2.1-201-007») от 16 мая 2011 по направлению «Проведение научных исследований научными группами под руководством докторов наук в следующих областях: – биокаталитические, биосинтетические и биосенсорные технологии; – биомедицинские и ветеринарные технологии жизнеобеспечения и защиты человека и животных; – геномные и постгеномные технологии создания лекарственных средств; – клеточные технологии;- биоинженерия; – биоинформационные технологии» в рамках мероприятия 1.2.1. «Проведение научных исследований научными группами под руководством докторов наук», мероприятия 1.2. «Проведение научных исследований научными группами под руководством докторов наук и кандидатов наук» , направления 1 «Стимулирование закрепления молодежи в сфере науки, образования и высоких технологий» федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009 – 2013 годы.

Цель работы. Основной целью выполнения НИР в рамках мероприятия является обеспечение достижения научных результатов мирового уровня, подготовки и закрепления в сфере науки и образования научных и научно-педагогических кадров, формирование эффективных и жизнеспособных научных коллективов.

Целью выполнения НИР является разработка метода сборки геномных последовательностей на основе восстановления фрагментов по парным чтениям.

Целями четвертого этапа являются:

1. Составление плана проведения экспериментальных исследований.

2. Подготовка заявки на регистрацию программы для ЭВМ.

3. Проведение вычислительных экспериментов.

4. Анализ вычислительных экспериментов.

5. Подготовка статьи для публикации в журнале из перечня ВАК.

6. Подготовка отчетных документов и передача Заказчику.

При выполнении четвертого этапа НИР был использован следующий инструментарий:

1. Компьютер *Aquarius Elt E50 S46*. Инв. номер. 110104.7.0036527.
2. Компьютер *Aquarius Elt E50 S46*. Инв. номер. 110104.7.0036528.
3. Компьютер *Aquarius Elt E50 S46*. Инв. номер. 110104.7.0036529.
4. Компьютер *Aquarius Elt E50 S46*. Инв. номер. 110104.7.0036530.
5. Компьютер *Aquarius Elt E50 S46*. Инв. номер. 110104.7.0036531.
6. Компьютер *Aquarius Elt E50 S46*. Инв. номер. 110104.7.0036532.
7. Компьютер *Horse*. Имеет шестиядерный процессор *AMD Phenom™ II X6 1090T* с тактовой частотой 3,7 ГГц; оперативная память – 16 ГБ (четыре модуля объемом по 4 ГБ с частотой 1.3 ГГц).
8. Компьютер *Sphinx*. Имеет два 16-ти ядерных процессора *AMD Opteron* 6272 с тактовой частотой 2.1 ГГц; оперативная память – 128 Гб (16 модулей по 8 Гб, DDR3-1333 ECC Reg).

При подготовке научно-технического отчета были использованы следующие нормативные документы:

• Постановление Правительства Российской Федерации от 4 мая 2005 г. № 284 «О государственном учете результатов научно-исследовательских, опытно-конструкторских и технологических работ гражданского назначения»;

• Гражданский кодекс РФ;

• ГОСТ 7.32-2001 «Система стандартов по информации, библиотечному и издательскому делу. Отчет о научно-исследовательской работе. Структура и правила оформления»;

• ГОСТ 15.101.98 «Система разработки и постановки продукции на производство. Порядок выполнения научно-исследовательских работ»;

В ходе выполнения четвертого этапа научно-исследовательских работ были получены следующие результаты:

– составлен план проведения экспериментальных исследований;

– подготовлена заявка на регистрацию программы для ЭВМ;

– проведены вычислительные эксперименты;

– проведен анализ вычислительных экспериментов;

– подготовлена статья для публикации в журнале из перечня ВАК.

Многие современные задачи биологии и медицины требуют знания генома живых организмов, который состоит из нескольких нуклеотидных последовательностей ДНК. В связи с этим возникает необходимость в дешевом и быстром методе секвенирования – определения последовательности нуклеотидов в образце ДНК.

В работе представлен алгоритм восстановления фрагментов геномной последовательности по парным чтениям. Алгоритм состоит из нескольких этапов и включает в себя процедуру исправления ошибок в чтениях и построение достаточно длинных фрагментов геномной последовательности с помощью графа де Брюина. На четвертом этапе проводятся экспериментальные исследования разработанных алгоритмов.

СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ

В настоящем отчете о НИР применяются следующие термины с соответствующими определениями.

De novo сборка генома – определение геномной последовательности живого существа, геном которого неизвестен.

K-мер – строка длиной *K* символов над алфавитом {A, G, C, T}.

Бинарный поиск – алгоритм, позволяющий осуществлять поиск элемента в отсортированном массиве за время, пропорциональное двоичному логарифму длины массива.

Взвешенный граф – граф, в котором каждое ребро имеет вес – вещественное число, сопоставленное с ним.

Геном – совокупность генов организма. Представляет собой одну или несколько последовательностей нуклеотидов.

Граф – объект, состоящий из двух множеств. Первое множество – множество вершин графа. Второе множество – множество ребер – является подмножеством множества всех пар вершин. В ориентированных графах ребро соответствует неупорядоченной паре вершин, тогда как в неориентированных – упорядоченной.

Граф де Брюина – граф, вершинами которого являются *k*-меры, при этом из одной вершины есть ребро в другую, если существует такой *(k+1)*-мер, что *k*-мер, соответствующий первой вершине, является его префиксом, а *k*-мер, соответствующий второй, – суффиксом.

Граф перекрытий – граф, вершинами которого являются последовательности, причем ребро из одной вершины в другую сущеbствует в том случае, если суффикс последовательности, соответствующей первой вершине, совпадает с префиксом последовательности, соответствующей второй.

Квазиконтиг – протяженный непрерывный фрагмент геномной последовательности, префикс которого совпадает с одним парным чтением, а суффикс – со вторым.

Контиг – протяженный непрерывный фрагмент геномной последовательности, который не может быть расширен однозначным образом ни в одну из сторон.

Метрика N 50 – максимальная длина такого контига, что суммарная длина контигов с длиной не меньше этой составляет не меньше 50% длины генома.

Нуклеотид – химическое соединение, являющееся частью ДНК.

Обход в ширину – алгоритм, осуществляющий обход всех вершин, достижимых из заданной, в порядке увеличения расстояния.

Префикс строки – подстрока строки, начинающаяся с первого символа этой строки.

Секвенирование – определение последовательности нуклеотидов в образце ДНК.

Скэффолд – набор контигов, для которого с большой степенью уверенности определено его взаимное расположение в геномной последовательности.

Суффикс строки – подстрока строки, кончающаяся последним символом этой строки.

Хеш-таблица – структура данных, позволяющая хранить пары вида (ключ, значение). Для каждого ключа может храниться не более одного значения.

ВВЕДЕНИЕ

В настоящем отчете излагаются результаты выполнения *поисковых научно-исследовательских работ по теме «Разработка метода сборки геномных последовательностей на основе восстановления фрагментов по парным чтениям»*, выполняемых в рамках государственного контракта, заключенного между Министерством образования и науки Российской Федерации и федеральным государственным бюджетным образовательным учреждением высшего профессионального образования «Санкт-Петербургским национальным исследовательским университетом информационных технологий, механики и оптики» в соответствии с решением Конкурсной комиссии Министерства образования и науки Российской Федерации № 1 (протокол от 26.04.2011 г. № 3/01731000037110000032) по лоту шифр «2011-1.2.1-201-007» «Проведение научных исследований научными группами под руководством докторов наук в следующих областях: – биокаталитические, биосинтетические и биосенсорные технологии; – биомедицинские и ветеринарные технологии жизнеобеспечения и защиты человека и животных; – геномные и постгеномные технологии создания лекарственных средств; – клеточные технологии; – биоинженерия; – биоинформационные технологии» в рамках федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009 – 2013 годы, утвержденной постановлением Правительства Российской Федерации от 28 июля 2008 года № 568 «О федеральной целевой программе «Научные и научно-педагогические кадры инновационной России» на 2009 – 2013 годы».

Задачами этапа являются:

1. Составление плана проведения экспериментальных исследований.
2. Подготовка заявки на регистрацию программы для ЭВМ.
3. Проведение вычислительных экспериментов.
4. Анализ вычислительных экспериментов.
5. Подготовка статьи для публикации в журнале из перечня ВАК.
6. Подготовка отчетных документов и передача Заказчику.

В настоящее время исследования в области геномики ведутся в таких университетах и лабораториях мира, как, например, *Cold Spring Harbor Laboratory* (штат Нью-Йорк, США), Университет Мериленда (США), Национальный центр геномного анализа (Барселона, Испания). При изучении генома живого существа обычно выделяют три основных этапа:

а) секвенирование молекул ДНК, содержащих информацию о геноме (выполняется с использованием специальных устройств-секвенаторов);

б) сборка геномной последовательности (или коротко – сборка генома, выполняется с использованием компьютеров);

в) анализ и сравнение геномов (выполняется с использованием компьютеров).

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 (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) эти технологии в настоящее время развиваются существенно быстрее, чем компьютерные технологии и алгоритмы сборки геномных последовательностей – производительность компьютеров удваивается каждые два года, а

производительность геномных секвенаторов за тот же самый период увеличивается в 10 раз.

Использование существующих в настоящее время алгоритмов на персональных компьютерах приведет к тому, что сборка одного генома займет месяцы. Таким образом, актуальной является задача разработки новых алгоритмов сборки геномных последовательностей, соответствующих по своим параметрам существующим методам секвенирования. В рамках настоящей НИР будет рассматриваться так называемая задача *de novo* сборки генома – сборки генома живого существа, для которого геном еще не известен.

Сложность задачи сборки геномной последовательности обусловлена следующими факторами:

а) большой объем входных данных;
б) сложность структуры генома – наличие в нем повторов и полиморфизмов;
в) наличие ошибок в исходных данных, полученных с устройств-секвенаторов.

Для решения указанных проблем сборку геномной последовательности обычно разбивают на несколько этапов:

а) исправление ошибок в исходных данных – чтениях геномной последовательности;
б) сборка квазиконтигов – достаточно длинных непрерывных фрагментов искомой геномной последовательности;
в) сборка контигов – протяженных непрерывных фрагментов геномной последовательности, которые не могут быть расширены однозначным образом ни в одну из сторон.

В рамках четвертого этапа НИР выполняются экспериментальные исследования разработанного метода сборки геномных последовательностей на основе восстановления фрагментов из парных чтений. Разработанные на предыдущих этапах НИР алгоритмы позволяют осуществлять первые два из перечисленных выше этапов сборки геномной последовательности. В настоящем отчете проводится сравнение разработанного метода со сторонними методами, выполняющими те же задачи, а также анализируется применимость генерируемых

квазиконтигов для выполнения следующего этапа – сборки контигов сторонними средствами.

1 . СОСТАВЛЕНИЕ ПЛАНА ПРОВЕДЕНИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

В настоящем разделе описывается план проведения экспериментальных исследований.

Одним из недостатков, которым обладают существующие программные средства для сборки генома, является большой объем оперативной памяти, необходимый им для сборки генома, сходного по размерам с геномом человека (203 миллиарда нуклеотидов суммарно во всех чтениях). Так, например, SOAPdenovo [1] необходимо порядка 140 гигабайт оперативной памяти, а *ABySS* [2] – 21 компьютер с 16 гигабайтами каждый (всего – 336 гигабайт). 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 сборки является отсутствие внутреннего контроля качества сборки. В рамках настоящего исследования разрабатывается метод сборки генома [3], который лишен указанных недостатков.

При проведении экспериментальных исследований разработанный метод сравнивается с аналогами на различных исходных данных. При этом рассматриваются геномы различных размеров – от нескольких миллионов до нескольких миллиардов нуклеотидов. На основании результатов экспериментов разработанный метод сборки геномных последовательностей может быть доработан, после чего будет реализован прототип инструментального средства.

1.1 . Архитекутра метода сборки геномных последовательностей

В настоящем разделе описывается архитекутра разработанного метода сборки геномных последовательностей.

Сборка генома в предлагаемом методе осуществляется в три этапа (рис. 1):

1 Исправление ошибок в наборе чтений геномной последовательности на основе частотного анализа.

2 Восстановление фрагментов геномной последовательности по чтениям с помощью графа де Брюина (сборка квазиконтигов – последовательностей, по длине б*о*льших чтений, но не являющихся контигами в смысле невозможности наращивания).

3 Сборка контигов из восстановленных фрагментов.

Каждый следующий этап получает на вход результаты работы предыдущего.

При этом для первых двух этапов в рамках настоящей НИР разработаны алгоритмы и выполнена их программная реализация, а для третьего этапа – используются сторонние программные средства.

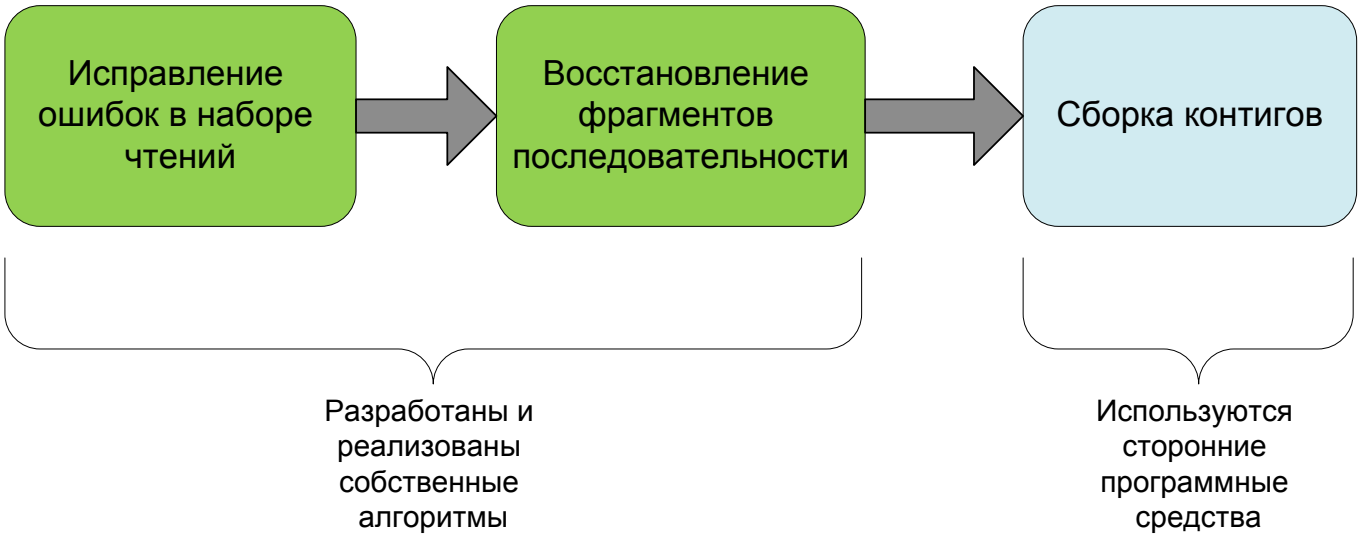

Рис. 1. Архитектура метода сборки геномных последовательностей

При выполнении предыдущих этапов настоящей НИР были представлены теоретические описания алгоритмов и их программная реализация на языке программирования *Java*.

1.2 . Общая схема проведения эксперимента

Разработанные в данной работе алгоритмы позволяют осуществлять сборку квазиконтигов – непрерывных протяженных фрагментов геномной последовательности. В рамках данного исследования проводится два ряда экспериментов.

В первом ряде экспериментов будет проводиться непосредственное сравнение результатов сборки квазиконтигов, собранными с помощью

разработанных в данной НИР алгоритмов, с результатами сборки квазиконтигов на тех же входных данных, выполненной другими сборщиками (рис. 2). Сборку квазиконтигов без сборки контигов осуществляет, в частности, такой сборщик как *GapFiller* [9]. С ним и будет проводиться сравнение в первом ряде экспериментов на различных входных данных и различных вычислительных системах.

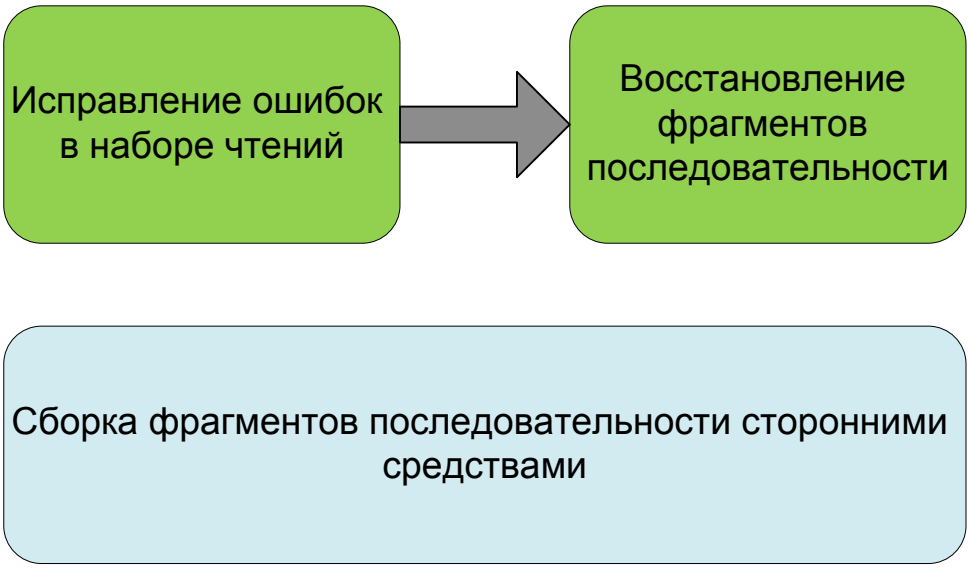

Рис. 2. Общая схема проведения эксперимента по сборке квазиконтигов

Многие современные сборщики геномных последовательностей могут принимать на вход как парные чтения, так и одиночные фрагменты геномных последовательностей [1, 2]. Более того, существует возможность сборки одного генома из данных нескольких библиотек, когда на вход сборщику подаются чтения с различными параметрами (например, «короткие» парные чтения и «длинные» одиночные).

Для второго ряда экспериментов предлагается следующая схема: для каждого набора чтений будут запущены реализованные алгоритмы исправления ошибок в наборе чтений и восстановление фрагментов геномной последовательности. Полученные таким образом фрагменты будут использоваться в качестве библиотеки «длинных» одиночных чтений для запуска сторонних сборщиков контигов. Результаты работы такой последовательности алгоритмов будут сравниваться с результатами сборки контигов сторонних сборщиков непосредственно из парных чтений (рис. 3).

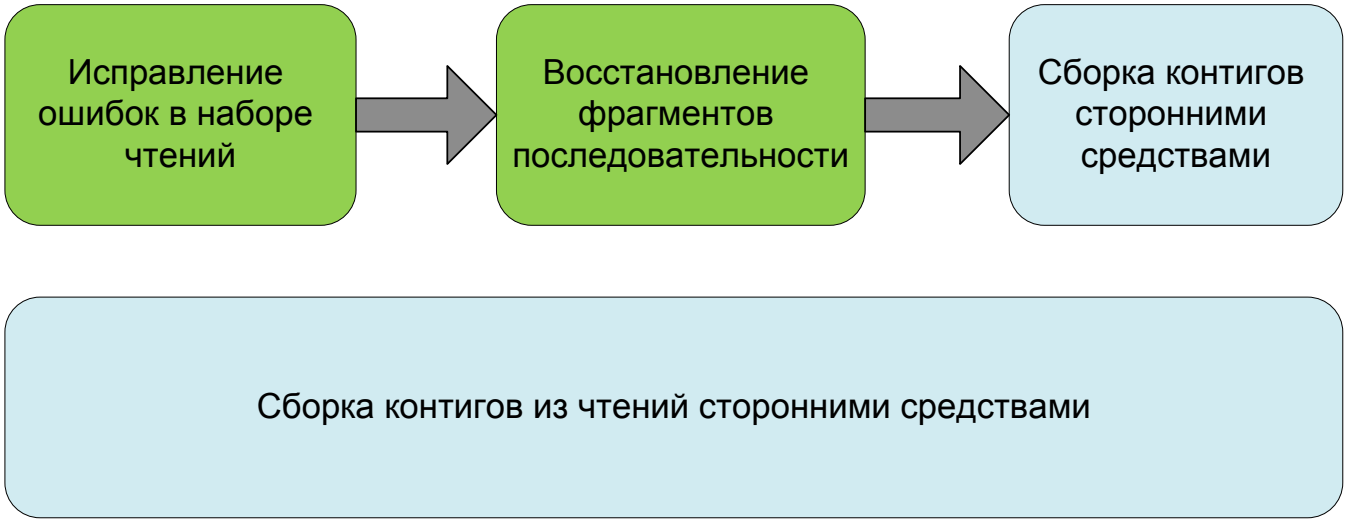

Рис. 3. Общая схема проведения эксперимента по сборке контигов

1.3 . Наборы исходных данных

В данном разделе приводится описание исходных данных, используемых в экспериментах. Для каждого набора исходных данных приводится описание следующих параметров:

- природа данных (какой организм или сгенерированные данные);
- длина чтений (число нуклеотидов в одном фрагменте);
- средний размер вставки (число нуклеотидов между двумя фрагментами одного пары чтений);
- размер генома (число нуклеотидов в полном геноме).
- покрытие генома;
- суммарный объем данных (объем информации в ГБ);

Используются парные чтения геномов:

- бактерии *Escherichia coli* [7];
- насекомого *Drosophila melanogaster* [8];
- искусственного генома, использовавшегося в проекте *de novo Genome Assembly Assessment Project* [10].

1.3.1 . Escherichia coli

Escherichia coli (кишечная палочка) – бактерия, геном которой является «небольшим» и одним из наиболее хорошо изученых на сегодняшний день. В данном исследовании используется набор данных *SRR001665* [7] из *NCBI Sequence Read Archive*, полученных по штамму *K-12*.

Чтения имеют следующие параметры:

- длина чтений: 36;
- средний размер вставки: 200;
- размер генома: 4.5 млн.
- покрытие генома: 166;
- суммарный объем данных: 4 ГБ.

1.3.2 . Drosophila melanogaster

Drosophila melanogaster (чернобрюхая дрозофила) – насекомое, являющееся модельным организмом в биомедицинских исследованиях [11, 12]. Имеет средний по размеру геном. Чтения имеют следующие параметры:

- длина чтений: 146;
- средний размер вставки: 350;
- размер генома: 139.5 млн.
- покрытие генома: 39;
- суммарный объем данных: 15 ГБ.

1.3.3 De novo Genome Assembly Assessment Project

Проект *de novo Genome Assembly Project* (*dnGASP*, http://cnag.bsc.es) [10] организован Национальным центром геномного анализа, Барселона, Испания) имеет целью оценку и сравнение алгоритмов и методов сборки генома. В рамках проекта dnGASP участникам требовалось за время с 15 декабря 2010 года по 15 февраля 2011 года (в дальнейшем этот срок был продлен до 1 марта 2011 года) выполнить сборку генома из данных, предоставленных организаторами проекта. Организаторами проекта был подготовлен искусственный геном размером в 1,8

млрд нуклеотидов, из которого были симулированы чтения секвенатора нового поколения. Выбор искусственного генома обосновывается тем, что в таком случае проще проводить сравнение результатов (так как геном известен организаторам) и проще обеспечить честность соревнования (так как геном неизвестен участникам и отсутствует в геномных базах данных). Основным отличием проекта *dnGASP* от другого аналогичного проекта *Assemblathon* является то, что в проекте *dnGASP* рассматривается задача сборки «большого» генома (1,8 млрд нуклеотидов), а в проекте *Assemblathon* – «среднего» генома (100 млн нуклеотидов).

Искусственный геном в *dnGASP* был составлен из 14 хромосом, из которых 11 были взяты из реальных геномов живых существ (человек, курица, муха-дрозофила, морской слизень, нематода, арабидопсис, дрожжи), а три были созданы искусственно. Их структура была выбрана таким образом, чтобы иметь возможность оценить способность программ-сборщиков генома собирать различные типы повторов, весьма важные в прикладной геномике.

При симуляции чтений в них вносились ошибки в соответствие с величинами качества, полученными при реальном запуске секвенатора Illumina GAIIx.

Полученные чтения имеют следующие параметры:

- длина чтений: 114;
- средний размер вставки: 500;
- размер генома: 1.8 млрд.
- покрытие генома: 44;
- суммарный объем данных: 172 ГБ.

1.4 . Описание программных средств, используемых в экспериментальных исследованиях

В данном разделе приводится описание сторонних программных средств, обрабатывающих геномные данные, которые используются в рамках данного экспериментального исследования.

1.4.1 . Сборка квазиконтигов

Для сборки квазиконтигов в данном исследовании используется программное средство *GapFiller* [9].

GapFiller является локальным сборщиком, основанным на методе постепенного наращивания контигов. В качестве первого приближения квазиконтига используется чтение, затем на каждой итерации производятся попытки увеличения квазиконтига. Для этого из всех чтений выбираются те, которые лучше всего перекрываются с концом текущей версии контига. Внутри этой группы производится выбор консенсусной последовательности. В зависимости от того, насколько часто нуклеотид из консенсусной последовательности встречается в соответствующих позициях чтений, некоторые из чтений обрезаются и удаляются. После этого консенсус считается еще раз и квазиконтиг наращивается.

Остановка цикла наращивания квазиконтига происходит в одном из четырех случаев:

- Число чтений перекрывающихся с концом квазиконтига меньше заранее заданного порога m. В этом случае квазиконтиг не может быть продолжен.
- Число чтений, оставшихся после удаления, становится меньше m. Это обычно означает, что квазиконтиг нельзя продолжить из-за повтора.
- Длина квазиконтига превышает заранее заданный порог $L_{\max}$, что означает, что чтение, парное к исходному, не было найдено, и квазиконтиг был продолжен неправильно.
- Было найдено чтение, парное к тому, которое было использовано в качестве первогого приближения квазиконтига. Это означает, что скорее всего квазиконтиг был получен правильно.

Для того, чтобы быстро искать чтения, перекрывающиеся с суффиксом текущего квазиконтига применяется хеш-таблица. Перед запуском алгоритма выбирается параметр b – длина префикса, по которому будут индексироваться чтения. Все чтения добавляются в хеш-таблицу, отображающую префикс чтения в список чтений с таким префиксом.

1.4.2 . Сборка контигов

Для сборки контигов из парных чтений и из квазиконтигов, полученных с помощью сборщика, разработанного в рамках данной НИР, используется программное средство *AbySS*. В статье [2] изложен подход к сборке контигов в программном средстве *ABySS*. 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 распределенный граф де Брюина.

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_i строится множество связанных с ним контигов P_i. Затем в графе связей контигов ищется уникальный путь, проходящий через все контиги из P_i. 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.

1.5 . Описание вычислительных систем, используемых в экспериментальных исследованиях

Для проведения экспериментов были выделены серверы со следующими конфигурациями:

а) *horse* (конфигурация 1):
   1) шестиядерный процессор *AMD Phenom™ II X6 1090T* с тактовой частотой 3,7 ГГц;
   2) оперативная память – 16 ГБ (четыре модуля объемом по 4 ГБ с частотой 1.3 ГГц);
   3) шесть жестких дисков объемом 2 ТБ, скорость вращения шпинделя 7200 об./мин, объединенных в массив *RAID 5*;

б) *godzilla* (конфигурация 2):
   1) два четырехядерных процессора *Intel® Xeon® E5410* с тактовой частотой 2.33 ГГц;
   2) оперативная память — 32 ГБ (8 модулей по 4 ГБ с частотой 667 МГц);
   3) 14 жестких дисков объемом 750ГБ (скорость вращения шпинделя — 15000 об./мин.), объединенных в *RAID 1+0*;

в) *sphinx* (конфигурация 3):
   1) два 16-ти ядерных процессора *AMD Opteron 6272* с тактовой частотой 2.1 ГГц;
   2) оперативная память – 128 Гб (16 модулей по 8 Гб, *DDR3-1333 ECC Reg*);
   3) восемь жестких дисков объемом 2 Тб, скорость вращения шпинделя 7200 об./мин объединенные в массив *RAID 5*.

Выводы по главе 1

1. Составлен план проведения экспериментальных исследований.
2. Определены наборы тестовых данных, на которых будет проводиться экспериментальное исследование. В эксперименте будут использованы чтения геномов небольшого, среднего и большого размеров.
3. Определены методы, с которыми будет сравниваться разработанный метод.

2 . ПРОВЕДЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

В данном разделе приводится описание проведенных вычислительных экспериментов.

2.1 . Сборка квазиконтигов

2.1.1 . Сборка квазиконтигов сборщиком, разработанным в данной НИР

В качестве исходных данных использовались чтения генома бактерии *E. Coli*, чтения мушки *D. Melanogaster* и чтения искусственного генома *dnGASP*.

Сборка квазиконтигов генома *E. Coli* была запущена на вычислительной машине с шестью ядрами (конфигурация 1) в шесть потоков (первый эксперимент). Сборка была произведена путем запуска команды:

```
do_all.sh \
-F 2 \
-c config.properties \
-k 18 \
-q sanger \
-w /data/genome/eColi_work_k18/ \
--min-size 1 \
--max-size 550 \
/data/genome/eColi/SRR/SRR001665_1.fastq \
/data/genome/eColi/SRR/SRR001665_2.fastq
```

Время работы составило 1999 секунд, из которых 631 секунд выполнялся первый этап – исправление ошибок и 1368 секунд – сборка квазиконтигов из исправленных чтений. Полный протокол эксперимента приведен в приложении А (протокол эксперимента № 1). В приведенном эксперименте задано значение *k*-мера равное 18. Также производились запуски с другими значениями параметра *k*. Полученные описанным выше образом квазиконтиги использовались в качестве одиночных чтений в эксперименте со сборкой квазиконтигов.

Для сравнения результатов сборки квазиконтигов с результатами работы GapFiller использовалась следующая модификация в разработанном алгоритме: в случае, если на определенной позиции не удается однозначно определить нуклеотид, то на его место записывался символ *N*. Значение длины *k*-мера в этом запуске было выбрано равным 21.

Сборка квазиконтигов генома *D. Melanogaster* была запущена на вычислительной машине с шестью ядрами (конфигурация 1) в шесть потоков (второй эксперимент). Сборка была произведена путем запуска команды:

Запуск был произведен путем запуска команды:

```
do_all.sh \
-F 2 \
-c config.properties \
-k 30 \
-q sanger \
-w /data/genome/dMelanogaster_work/ \
--min-size 1 \
--max-size 550 \
/data/genome/dMelanogaster_unpacked/SRR094875_1.fastq \
/data/genome/dMelanogaster_unpacked/SRR094875_2.fastq
```

Время работы составило 109550 секунд, из которых 62826 секунд выполнялся первый этап – исправление ошибок и 46724 секунд – сборка квазиконтигов из исправленных чтений. Полный протокол эксперимента приведен в приложении А (протокол эксперимента №2).

Сборка квазиконтигов генома *dnGASP* была запущена на вычислительной машине с 32 ядрами (конфигурация 3) в 32 потока (третий эксперимент). Сборка была произведена путем запуска команды:

```
do_all.sh \
-F 2 \
-c config.properties \
-k 30 \
```

```
-q illumina \
-w /home/data/dnGasp_work \
--min-length 1 \
--max-length 550 \
-M 2 \
/home/data/dnGasp/*.fastq
```

Время работы составило 179562 секунд, из которых 80338 секунд выполнялся первый этап – исправление ошибок и 99224 секунд – сборка квазиконтигов из исправленных чтений. Полный протокол эксперимента приведен в приложении А (протокол эксперимента № 3).

2.1.2 . Сборка квазиконтигов с использованием *GapFiller*

Сборка квазиконтигов с использованием *GapFiller* была запущена на машине с процессором *Intel Xeon E5410* 2.33 ГГц и 32 ГБ оперативной памяти (конфигурация 2, четвертый эксперимент). В настоящее время сборщик *GapFiller* не поддерживает распараллеливания, поэтому сборка проводилась в один поток. Запуск был произведен следующей командой:

```
IGAassembler \
    --k 15 \
    --short-1
/mnt/raid/genome/fixed_ecoli/SRR001665_1.fastq \
    --short-2
/mnt/raid/genome/fixed_ecoli/SRR001665_2.fastq \
    --output fixed_coli \
    --statistics fixed_coli_stat \
    --short-ins 200 \
    --short-var 50 \
    --overlap 20 \
    --slack 15 \
    --read-length 36 \
    --global-mismatch 3 \
```

```
--no-read-cycle \
--max-length 250
```

Время работы составило 27 часов, при этом было использовано 7 ГБ оперативной памяти. Полный протокол эксперимента приведен в приложении А (протокол эксперимента № 4).

С учетом того, что даже для небольшого генома *E. Coli* время работы *GapFiller* составило более суток, закономерным оказалось то, что на геномах среднего и большого размера положительного результата сборки квазиконтигов добиться не удалось при использовании *GapFiller*.

2.2 . Сборка контигов

В настоящей НИР сборка контигов осуществляется сторонними сборщиками. При этом квазиконтиги, полученные с помощью сборщика, разработанного в рамках НИР, используются в качестве одиночных чтений для сборщиков контигов.

2.2.1 . Сборка контигов из квазиконтигов

В данном ряде экспериментов в качестве исходных данных использовались собранные квазиконтиги (разд 2.1.1) совместно с исходными чтениями генома бактерии *E. Coli* и чтениями мушки *D. Melanogaster*. Для сборки контигов использовалось программное средство *ABySS* (версия 1.3.2).

Сборка контигов генома *E. Coli* была запущена на вычислительной машине с шестью ядрами в шесть потоков (шестой эксперимент). Сборка была произведена путем запуска команды:

```
abyss-pe \
j=6 np=6 \
k=30 \
se='SRR001665.fasta' \
in='SRR001665_1.fastq SRR001665_2.fastq' \
name=ecoli_from_out_quasicontigs_with_pe
```

Время работы составило пять минут, полный протокол эксперимента приведен в приложении А (протокол эксперимента № 6).

Сборка контигов генома *D. Melanogaster* была запущена на вычислительной машине с шестью ядрами в шесть потоков (девятый эксперимент). Сборка была произведена путем запуска команды:

```
abyss-pe \
j=6 np=6 \
k=30 \
se='SRR094875.fasta' \
in='../SRR094875_1.fastq ../SRR094875_2.fastq' \
name=dMelanogaster_from_pe
```

Время работы составило 40 минут, полный протокол эксперимента приведен в приложении А (протокол эксперимента № 9).

Были проведены еще два эксперимента (приложение А, протоколы 5, 8), в которых сборка контигов осуществлялась непосредственно из квазиконтигов без использования парных чтений. Данные о результатах сборки приведены в главе 3.

2.2.2 . Сборка контигов из парных чтений с использованием *AbySS*

В данном ряде экспериментов в качестве исходных данных использовались парные чтения такие же, как и в предыдущем разделе. Отличие состояло в том, что квазиконтиги не подавались на вход сборщику. Для сборки контигов использовалось программное средство *ABySS* (версия 1.3.2).

Сборка контигов генома *E. Coli* из парных чтений была запущена на вычислительной машине с шестью ядрами в шесть потоков (протокол эксперимента № 7). Сборка была произведена путем запуска команды:

```
abyss-pe \
j=6 np=6 \
k=30 \
in='SRR001665_1.fastq SRR001665_2.fastq' \
name=ecoli_from_out_quasicontigs_with_pe
```

Сборка контигов генома *D. Melanogaster* из парных чтений была запущена на вычислительной машине с шестью ядрами в шесть потоков (протокол эксперимента № 10). Сборка была произведена путем запуска команды:

```
abyss-pe \
j=6 np=6 \
k=30 \
in='../SRR094875_1.fastq ../SRR094875_2.fastq' \
name=dMelanogaster_from_pe
```

Выводы по главе 2

1. Проведены вычислительные эксперименты по сборке квазиконтигов с использованием сборщика, разработанного в рамках данной НИР.
2. Проведены вычислительные эксперименты по сборке квазиконтигов с использованием программного средства *GapFiller*.
3. Проведены вычислительные эксперименты по сборке контигов из квазиконтигов и парных чтений с использованием программного средства *ABySS*.

3 . АНАЛИЗ РЕЗУЛЬТАТОВ ВЫЧИСЛИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

В данном разделе приводится анализ результатов вычислительных экспериментов. Сравнение разработанного метода проводится по таким параметрам, как время работы, объем используемой памяти и качество сборки.

При оценивании качества сборки традиционно используют следующие параметры:

cnt – число фрагментов – квазиконтигов;

sum – суммарная длина квазиконтигов;

max – максимальная длина квазиконтига;

min – минимальная длина квазиконтига;

avg – средняя длина квазиконтига;

genome length – длина генома;

N 50 – такая длина фрагмента, что все фрагменты такой или больше*й* длины суммарно составляют не менее половины (50 %) длины сборки. Существуют модификации этой метрики: *Nx*, в которой суммарная длина должна быть не менее *x* %;

eq100 – число фрагментов, не содержащих ошибок (такие фрагменты, которые входят в референсный геном как подстрока);

lessX – число фрагментов, у которых процент нуклеотидов, совпадающих с соответствующим фрагментом референсного генома меньше *x* % .

3.1 . Сборка квазиконтигов

Приведем основные характеристики полученных фрагментов с использованием сборщика, разработанного в рамках данной НИР и сборщика *GapFiller*.

а) Результаты работы сборщика, разработанного в рамках данной НИР

1) геном небольшого размера *E. Coli*. Время работы составило семь минут. Было использовано 1.5 ГБ оперативной памяти. Приведем основные характеристики результатов сборки фрагментов геномной последовательности (в указанной ниже

статистике покрытие генома вычислялось по 90 % совпадению квазиконтига с референсным геномом):

- *cnt*: 9812301
- *sum*: 2111556268
- *max*: 519
- *min*: 19
- *avg*: 215.1948119
- *N50*: 216
- *N 90*: 202
- *eq100*: 9660914
- *less100*: 151193
- *less99*: 81466
- *less90*: 2984
- *less50*: 4
- собранными квазиконтигами не покрыто 0.51% референсного генома.

В модифицированной версии сборки квазиконтигов, использующей символ *N* в позициях с неоднозначностью выбора нуклеотида, были получены следующие результаты (в данном случае при вычислении статистики учитывалось только идеально совпадающее покрытие):

- *eq100*: 9756990
- *less100*: 2567
- *less99*: 81466
- *less90*: 554
- 0.38% референсного генома не покрыто.

2) геном среднего размера *D. Melanogaster*. Время работы составило 30 часов. Было использовано 4 ГБ оперативной памяти. Приведем основные характеристики результатов сборки фрагментов геномной последовательности.

- *cnt*: 15903917
- *sum*: 6592407673

– *max*: 550
– *min*: 30
– *avg*: 414.514718167
– *N 50*: 472
– *N 90*: 329
– *eq100*: 2259877
– *less100*: 13452248
– *less99*: 7183454
– *less90*: 751073
– *less50*: 83136

Собранными квазиконтигами не покрыто 6.13% референсного генома.

3) геном большого размера *dnGASP*. Время работы составило 50 часов. Было использовано 30 ГБ оперативной памяти.

б) Результаты работы сборщика *GapFiller*:

1) геном небольшого размера *E. Coli*. Время работы составило 27 часов. Было использовано 7 ГБ оперативной памяти.

В результате было получено 9,3 миллиона квазиконтигов для которых было найдено парное чтение (остальные при вычислении статистики были отброшены). Из них 0,7% содержали ошибки. Квазиконтигами, которые содержали не больше 10% ошибок, было не покрыто 0,1% референсного генома. Квазиконтигами, не содержащими ошибок, было не покрыто 0,3%. Более точно, значение метрик было следующим:

– *eq100*: 9322364
– *less100*: 68947
– *less99*: 14146
– *less90*: 916
– 0.29% референсного генома не покрыто.

2) геном среднего размера *D. Melanogaster* – работа сборщика не была корректно завершена;

3) геном большого размера *dnGASP* – работа сборщика не была корректно завершена.

При сравнении результатов можно сделать следующие выводы: разработанный в данной научно-исследовательской метод сборки геномных последовательностей использует меньшее количество ресурсов (как времени работы вычислительной системы, так и используемой оперативной памяти), чем программное средство *GapFiller*, и может быть применен на геномных данных среднего и большого размеров в случаях, когда *GapFiller* оказывается неприменим. При оценке качества сборки можно отметить, что *GapFiller* имеет лучшие показатели в части покрытия генома квазиконтигами, тем не менее, разработанный сборщик оказывается лучше в части идеально совпадающих с референсом квазиконтигов к их общему числу.

3.2 . Сборка контигов

В данном разделе приводится сравнение результатов работы сборщика *ABySS*, использующего в качестве входных данных квазиконтиги и парные чтения.

В начале приведем результаты работы *ABySS* на исходных данных, состоящих исключительно из квазиконтигов. Основные характеристики качества сборки демонстрируют относительно низкие результаты (в сравнении с использованием парных чтений). Приведем основные параметры статистики на каждом из наборов данных:

E. Coli (время работы семь минут при работе в 6 потоков, протокол № 5):

- *cnt*: 3029
- *sum*: 4616489
- *max*: 60385
- *min*: 30
- *avg*: 1524.09673159
- *N 50*: 13143
- *N 90*: 3339

D. Melanogaster (время работы 1 час 50 минут при работе в itcnm потоков, протокол № 8):

- *cnt*: 245249
- *sum*: 118237756
- *max*: 44727
- *min*: 25
- *avg*: 482.113101379
- *N 50*: 3382
- *N 90*: 394

dnGASP (время работы шесть часов при работе в 24 потока):

- *cnt*: 3016527
- *sum*: 1546375214
- *max*: 657870
- *min*: 30
- *avg*: 512.634302295
- *N* 50: 2928
- *N* 90: 360

Далее приводится более детальное сравнение результатов работы *ABySS* при использовании парных чтений. Сначала приводятся результаты работы на чтениях *E. Coli*. Указанные ниже статистические данные получены с помощью программного средства *QUAST* [13, 14]. В табл. 1 приведена базовая статистика. В местах, где в скобрах указано количество баз, соответствующая метрика вычисляется только на контигах, длина которых не менее указанного в скобках числа. *#N* – число непокрытых контигами нуклеотидов, N's (%) – их процент от общего числа.

Таблица 1 . Базовая статистика

Исходные данные	Только парные чтения	Квазиконтиги и парные чтения
Cnt	111	140
cnt (>= 500 баз)	108	133
cnt (>= 1000 баз)	107	119
Max	268 389	268 390
Sum	4 623 045	4 596 270
sum (>= 500 баз)	4 622 131	4 593 980
sum (>= 1000 баз)	4 621 340	4 583 966
Длина генома	4 639 675	4 639 675
N 50	87 805	80 047
N 75	45 735	42 438
L 50	17	19
L 75	35	41
# N	106	233
N's (%)	0.00229	0.00507

В табл. 2 приводится статистика несовпадений нуклеотидов в контигах с референсным геномом. В табл. 3 приводится дополнительная статистика результатов сборки контигов.

Таблица 2 . Статистика несовпадений

Инверсий	0	0
Локальных несовпадений	3	6
Несовпадений на 100 КБ	5.21	2.52
Вставок	44	66
Вставок на 100 КБ	0.95	1.44

Таблица 3 . Дополнительная статистика

Число неоднозначно выравненных контигов	6	14
Длина неоднозначно выравненных контигов	4634	7357
Покрытие генома (%)	99.165	99.056
Дублирующее покрытие генома (%)	1.005	1.000
Число генов (частичных)	4234 (58 ч.)	4212 (81 ч.)
Число оперонов (частичных)	849 (30 ч.)	844 (37 ч.)

Далее приводятся несколько графиков. На каждом графике синим точкам соответствуют результаты сборки контигов с использованием квазиконтигов и парных чтений, а красным точкам – результаты сборки исключительно из парных чтений.

На рис. 4 приведен график зависимости суммарной длины контигов от их количества. На рис. 5 – график значений метрики *Nx* при возможных значениях *x*. На рис. 6 – график числа генов во множестве контигов (накопительным итогом).

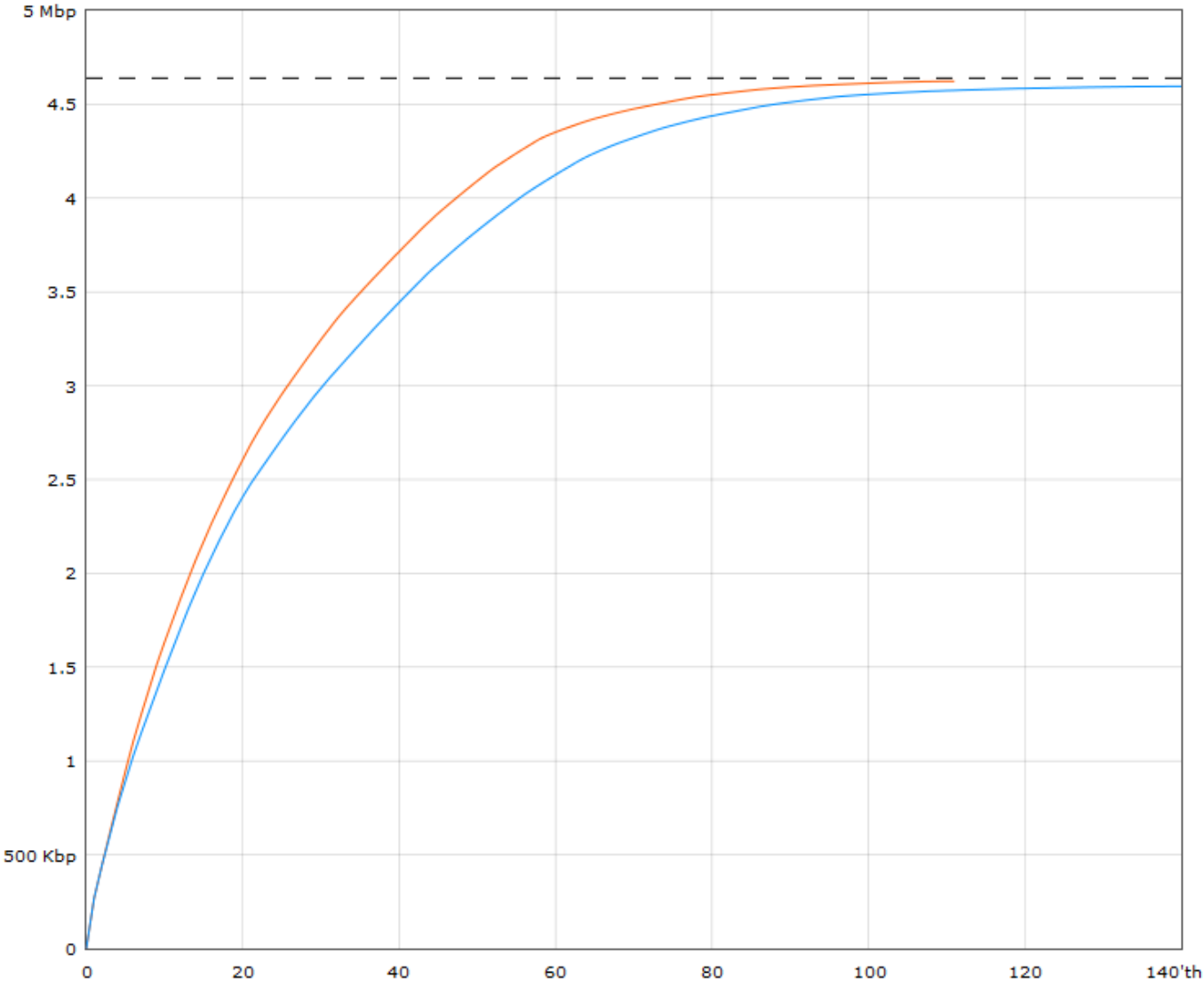

Рис. 4. Суммарная длина контигов

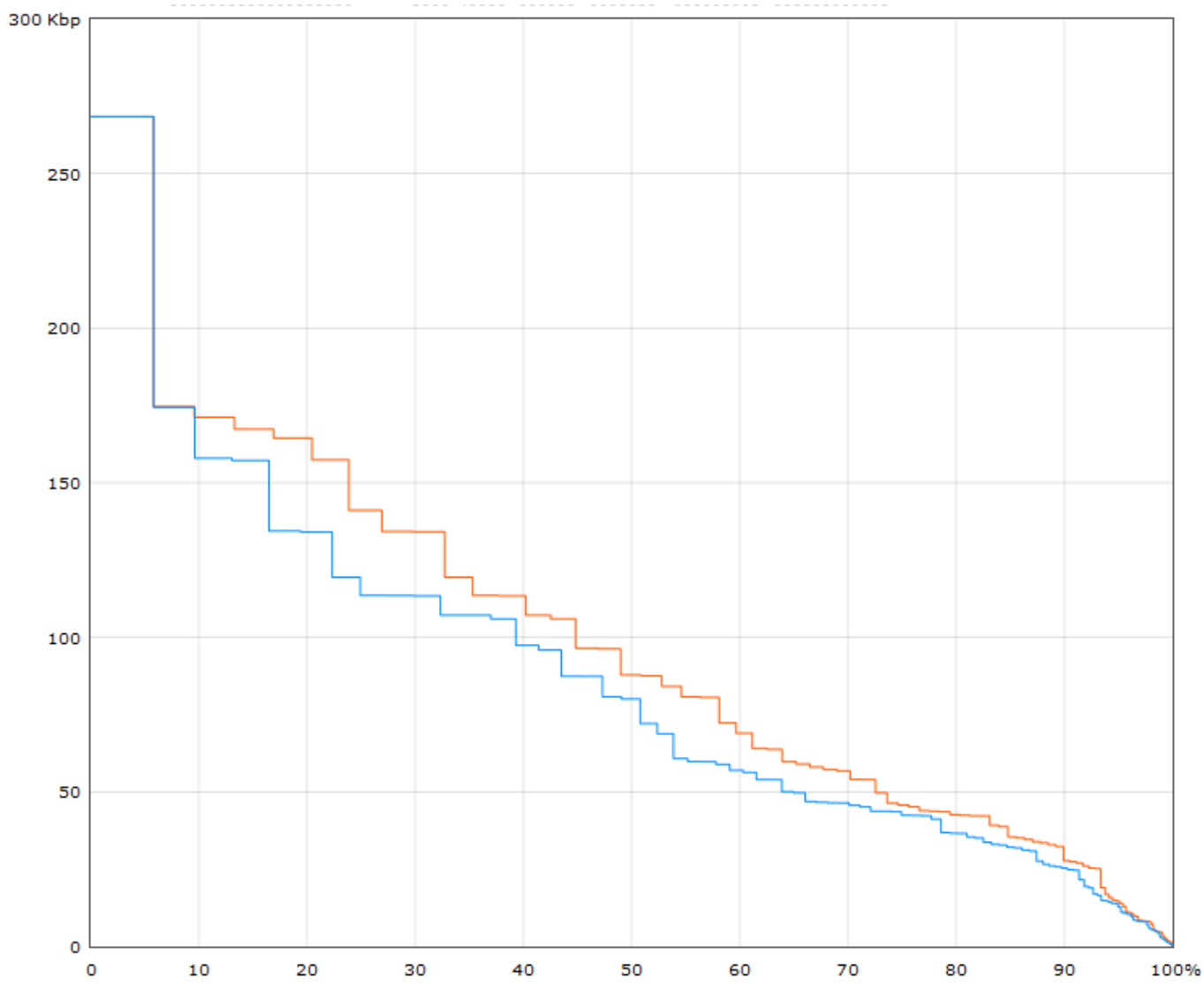

Рис. 5. Значение метрики Nx

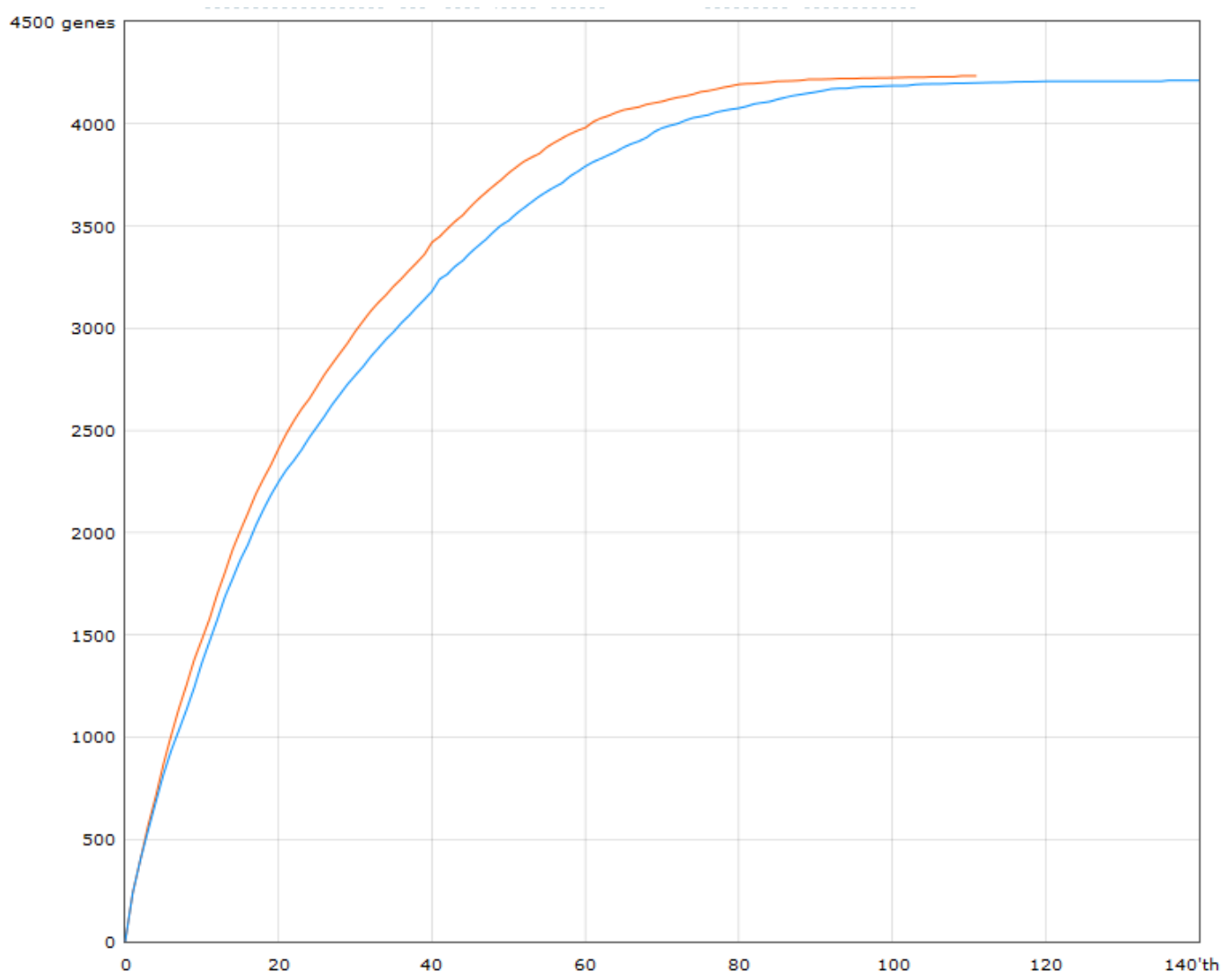

Рис. 6. Число генов

Из приведенных таблиц и графиков можно сделать вывод о том, что использование квазиконтигов в дополнение к парным чтениям, несущественно влияет на итоговое качество сборки контигов программным средством *ABySS*. Величина некоторых параметров при использовании квазиконтигов стала несколько хуже, что указывает на то, что несмотря на наличие этапа исправления ошибок, построенные квазиконтиги содержат ошибки.

Отметим также, что использование парных чтений существенно повысило качество сборки контигов по сравнению со сборкой исключительно из квазиконтигов.

Опишем далее основные характеристики результатов работы *ABySS* на чтениях *D. Melanogaster* с использованием и без использования квазиконтигов.

Сборка с использованием квазиконтигов и парных чтений (время работы 2 ч. 25 мин. в шесть потоков):

- *cnt*: 212434
- *sum*: 124866245
- *max*: 157892
- *min*: 30
- *avg*: 587.788418991
- *N 50*: 16943
- *N 90*: 842

Сборка непосредственно из парных чтений (время работы 1 ч. 40 мин. в шесть потоков):

- *cnt*: 224510
- *sum*: 126855238
- *max*: 157876
- *min*: 30
- *avg*: 565.031570977
- *N 50*: 16862
- *N 90*: 645.

Можно отметить, что как и в случае со сборкой *E. Coli*, использование квазиконтигов не оказывает существенного эффекта на качество сборки контигов программным средством ABySS.

Выводы по главе 3

1. Выполнен анализ результатов вычислительных экспериментов. Сравнение разработанного метода проводилось по таким параметрам, как время работы, объем используемой памяти и качество сборки.
2. Результаты экспериментов показали, что разработанный метод работает быстрее и использует меньше оперативной памяти, чем программное средство GapFiller.

3. Результаты экспериментов показали, что разработанный метод применим для сборки квазиконтигов геномов больших и средних размеров, в случаях когда программному средству *GapFiller* не хватает вычислительных ресурсов.
4. Результаты экспериментов показали, что в связке с программным средством *ABySS* разработанный метод не дает выигрыша во времени работы или качестве сборки.

ЗАКЛЮЧЕНИЕ

В результате исследований, выполненных на четвертом этапе работ по контракту, были получены следующие результаты:

а) Составлен план проведения экспериментальных исследований.

б) Определены наборы тестовых данных, на которых будет проводиться экспериментальное исследование. В эксперименте будут использованы чтения геномов небольшого, среднего и большого размеров.

в) Определены методы, с которыми будет сравниваться разработанный метод.

г) Проведены вычислительные эксперименты по сборке квазиконтигов с использованием сборщика, разработанного в рамках данной НИР.

д) Проведены вычислительные эксперименты по сборке квазиконтигов с использованием программного средства *GapFiller*.

е) Проведены вычислительные эксперименты по сборке контигов из квазиконтигов и парных чтений с использованием программного средства *ABySS*.

ж) Выполнен анализ результатов вычислительных экспериментов. Сравнение разработанного метода проводилось по таким параметрам, как время работы, объем используемой памяти и качество сборки.

з) Результаты экспериментов показали, что разработанный метод работает быстрее и использует меньше оперативной памяти, чем программное средство *GapFiller*.

и) Результаты экспериментов показали, что разработанный метод применим для сборки квазиконтигов геномов больших и средних размеров, в случаях когда программному средству GapFiller не хватает вычислительных ресурсов.

к) Получено свидетельство о регистрации программы для ЭВМ «Программное средство для сборки квазиконтигов из парных чтений» (приложение Б).

л) Подготовлена статья для публикации в журнале из перечня ВАК (Александров А.В., Казаков С.В., Мельников С.В., Сергушичев А.А., Царев Ф.Н. Метод сборки контигов геномных последовательностей на основе совместного применения графов де Брюина и графов перекрытий, приложение В).

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Li R., Zhu H., Ruan J., Qian W., Fang X., Shi Z., Li Y., Li S., Shan G., Kristiansen K., Li S., Yang H., Wang J., Wang J.* De novo assembly of human genomes with massively parallel short read sequencing // Genome Research. 2010. Vol. 20. No. 2, pp. 265 – 272.

2 *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. 2009. Vol. 19. No. 6, pp. 1117 – 1123.

3 *Разработка и программная реализация алгоритма исправления ошибок в данных секвенирования*. Промежуточный отчет по этапу III «Разработка метода сборки геномных последовательностей на основе восстановления фрагментов по парным чтениям». НИУ ИТМО. 2011.

4 *Cock P., Fields C., Goto N., Heuer M., Rice P.* The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants // Nucleic Acids Research, 2010. Vol. 38. No. 6, pp. 1767–1777.

5 *FASTA format*. [Электронный ресурс]. – Режим доступа: http://zhanglab.ccmb.med.umich.edu/FASTA/, свободный. Яз. англ. (дата обращения 05.06.2012).

6 *International Human Genome Sequencing Consortium*. 2001. Initial sequencing and analysis of the human genome // Nature. 409: 860 – 921.

7 *NCBI*: Experiment: SRX000429 – Illumina sequencing of Escherichia coli str. K-12 substr. MG1655 genomic paired-end library.

8 *SRA Run*: SRR094875: Illumina Genome Analyzer II paired end sequencing; ZI342_1-HE. [Электронный ресурс]. – Режим доступа: http://www.ebi.ac.uk/ena/data/view/SRR094875, свободный. Яз. англ. (дата обращения 11.10.2012).

9 *Nadalin F., Vezzi F., Policriti A.* GapFiller: a de novo assembly approach to fill the gap within paired reads // BMC Bioinformatics. 2012. Vol. 13 (Supl. 14).

10 *De novo Genome Assembly Assesment Project*. [Электронный ресурс]. – Режим доступа: http://cnag.bsc.es, свободный. Яз. англ. (дата обращения 11.04.2012).

11 *Model Organisms for Biomedical Research*. [Электронный ресурс]. – Режим доступа: http://www.nih.gov/science/models/, свободный. Яз. англ. (дата обращения 11.10.2012).

12 *FlyBase*. A Database of Drosophila Genes & Genomes. [Электронный ресурс]. – Режим доступа: http://flybase.org/, свободный. Яз. англ. (дата обращения 11.10.2012).

13 *QUAST*. QUality ASsesment Tool for Genome Assembly. [Электронный ресурс]. – Режим доступа: http://bioinf.spbau.ru/quast, свободный. Яз. англ. (дата обращения 11.10.2012).

14 *Гуревич А. А.* Разработка методов оценивания качества сборки геномных последовательностей. Магистерская диссертация. СПбАУ НОЦ НТ РАН. 2012.

ПРИЛОЖЕНИЕ А

ПРОТОКОЛЫ ЭКСПЕРИМЕНТОВ

В настоящем приложении приводятся протоколы проведения экспериментов.

Протокол эксперимента № 1

Общая информация об эксперименте	
Цель эксперимента	Исследование характеристик программного модуля, реализующего алгоритм сборки квазиконтигов
Объект эксперимента	Программный модуль, реализующий алгоритм сборки квазиконтигов
Дата и время начала эксперимента	26 сентября 2012 г. 15:17
Дата и время окончания эксперимента	26 сентября 2012 г. 15:50

Информация об используемой вычислительной системе	
Название вычислительной системы	Сервер кафедры КТ НИУ ИТМО
Тип вычислительной системы	Шестиядерный сервер
Характеристики вычислительной системы	Один сервер: а) один шестиядерный процессор *AMD® Phenom® II X6 1090T* с тактовой частотой 3200 МГц; б) оперативная память – 16 ГБ (четыре модуля по 4 ГБ с частотой 1333 МГц); в) 6 жестких дисков объемом 2 ТБ, скорость вращения шпинделя 7200 об./мин. (диски объединены в *RAID5*). г) скорость последовательного чтения/записи

	на жесткий диск: скорость чтения – до 600 МБ/с, скорость записи – от до 600 МБ/с.

Информация об исходных данных для сборки геномной последовательности	
Источник данных	NCBI: Experiment: SRR001665
Объем данных	4 ГБ
Характеристики данных	Парные чтения бактерии *Escherichia coli str. K-12 substr. MG1655* с длиной чтения 36 нуклеотидов и размером вставки 200 нуклеотидов Произведены секвенатором *Illumina Genome Analyzer*. Уникальный идетификатор: 071112_SLXA-EAS1_s_4
L – размер рассматриваемого генома	4.5 миллиона нуклеотидов
C – покрытие генома чтениями	166

Информация о вычислениях	
Дата и время начала вычислений	27 марта 2012 г. 17:25
Дата и время окончания вычислений	27 марта 2012 г. 17:32
Общее время выполнения приложения	7 минут

Эксперимент провел:

Программист кафедры ПИиВП, НИУ ИТМО, А. В. Александров

_____ 26 сентября 2012 года.

Протокол эксперимента № 2

Общая информация об эксперименте	
Цель эксперимента	Исследование характеристик программного модуля, реализующего алгоритм сборки квазиконтигов
Объект эксперимента	Программный модуль, реализующий алгоритм сборки квазиконтигов
Дата и время начала эксперимента	27 сентября 2012 г. 21:12
Дата и время окончания эксперимента	29 сентября 2012 г. 03:37

Информация об используемой вычислительной системе	
Название вычислительной системы	Сервер кафедры КТ НИУ ИТМО
Тип вычислительной системы	Шестиядерный сервер
Характеристики вычислительной системы	Один сервер: а) один шестиядерный процессорй *AMD® Phenom® II X6 1090T* с тактовой частотой 3200 МГц; б) оперативная память – 16 ГБ (четыре модуля по 4 ГБ с частотой 1333 МГц); в) 6 жестких дисков объемом 2 ТБ, скорость вращения шпинделя 7200 об./мин. (диски объединены в RAID5). г) скорость последовательного чтения/записи на жесткий диск: скорость чтения – до 600 МБ/с, скорость записи – до 600 МБ/с.

Информация об исходных данных для сборки геномной последовательности	
Источник данных	NCBI: Experiment: SRR094875
Объем данных	15 ГБ
Характеристики данных	Парные чтения мушки Drosophila melanogaster с длиной чтения 146 нуклеотидов и размером вставки 200 нуклеотидов. Произведены секвенатором *Illumina Genome Analyzer*. Уникальный идетификатор: HWI-EAS66_0016_FC62L6H
L – размер рассматриваемого генома	137 миллионов нуклеотидов
C – покрытие генома чтениями	39

Информация о вычислениях	
Дата и время начала вычислений	27 сентября 2012 г. 21:12
Дата и время окончания вычислений	29 сентября 2012 г. 03:37
Общее время выполнения приложения	30 часов 25 минут 50 секунд

Эксперимент провел:

Программист кафедры ПИиВП, НИУ ИТМО, А. В. Александров

_____ 29 сентября 2012 года.

Протокол эксперимента № 3

Общая информация об эксперименте	
Цель эксперимента	Исследование характеристик программного модуля, реализующего алгоритм сборки квазиконтигов
Объект эксперимента	Программный модуль, реализующий алгоритм сборки квазиконтигов
Дата и время начала эксперимента	29 сентября 2012 г. 14:10
Дата и время окончания эксперимента	01 октября 2012 г. 16:02

Информация об используемой вычислительной системе	
Название вычислительной системы	Сервер кафедры КТ НИУ ИТМО
Тип вычислительной системы	Тридцатидвухъядерный сервер
Характеристики вычислительной системы	Один сервер: а) два шестнадцатииядерных процессора *AMD® Opteron® 6272* с тактовой частотой 2100 МГц; б) оперативная память – 128 ГБ (шестнадцать модулей по 8 ГБ с частотой 1333 МГц); в) 8 жестких дисков объемом 2 ТБ, скорость вращения шпинделя 7200 об./мин (диски объединены в RAID5). г) скорость последовательного чтения/записи на жесткий диск: скорость чтения –до 600 МБ/с, скорость записи – до 600 МБ/с

Информация об исходных данных для сборки геномной последовательности	
Источник данных	Конкурс *dnGASP*
Объем данных	172 ГБ
Характеристики данных	Парные чтения сгенерированного генома *dnGASP* с длиной чтения 114 нуклеотидов и размером вставки 300 нуклеотидов Произведены секвенатором *Illumina Genome Analyzer*. Уникальный идетификатор: pe500_179316_SLIU
L – размер рассматриваемого генома	1,8 миллиардов нуклеотидов
C – покрытие генома чтениями	44

Информация о вычислениях	
Дата и время начала вычислений	29 сентября 2012 г. 14:10
Дата и время окончания вычислений	01 октября 2012 г. 16:02
Общее время выполнения приложения	49 часов 52 минуты 42 секунды

Эксперимент провел:

Программист кафедры ПИиВП, НИУ ИТМО, А. В. Александров

_____ 01 октября 2012 года.

Протокол эксперимента № 4

Общая информация об эксперименте	
Цель эксперимента	Исследование характеристик программного средства *GapFiller*
Объект эксперимента	Программное средство *GapFiller*
Дата и время начала эксперимента	26 сентября 2012 г. 15:17
Дата и время окончания эксперимента	27 сентября 2012 г. 18:17

Информация об используемой вычислительной системе	
Название вычислительной системы	Сервер кафедры КТ НИУ ИТМО
Тип вычислительной системы	Шестиядерный сервер
Характеристики вычислительной системы	Один сервер: а) два четырехядерных процессора *Intel® Xeon® E5410* с тактовой частотой 2.33 ГГц; б) оперативная память – 32 ГБ (8 модулей по 4 ГБ с частотой 667 МГц); в) 14 жестких дисков объемом 750ГБ (скорость вращения шпинделя – 15000 об/мин), объединенных в *RAID 1+0*; г) скорость последовательного чтения: 400 МБ/с, записи: 400 МБ/с

Информация об исходных данных для сборки геномной последовательности	
Источник данных	NCBI: Experiment: SRR001665
Объем данных	4 ГБ

Характеристики данных	Парные чтения бактерии *Escherichia coli str. K-12 substr. MG1655* с длиной чтения 36 нуклеотидов и размером вставки 200 нуклеотидов Произведены секвенатором *Illumina Genome Analyzer*. Уникальный идетификатор: 071112_SLXA-EAS1_s_4
L – размер рассматриваемого генома	4.5 миллиона нуклеотидов
C – покрытие генома чтениями	166

Информация о вычислениях	
Дата и время начала вычислений	26 сентября 2012 г. 15:17
Дата и время окончания вычислений	27 сентября 2012 г. 18:17
Общее время выполнения приложения	27 часов

Эксперимент провел:

Программист кафедры ПИиВП, НИУ ИТМО, А. А. Сергушичев

_____ 26 сентября 2012 года.

Протокол эксперимента № 5

Общая информация об эксперименте	
Цель эксперимента	Исследование характеристик совместного применения программного модуля, реализующего алгоритм сборки квазиконтигов, и программного средства *ABySS*
Объект эксперимента	Программный модуль, реализующего алгоритм сборки квазиконтигов. Программное средство *ABySS*
Дата и время начала эксперимента	12 октября 2012 г. 16:32
Дата и время окончания эксперимента	12 октября 2012 г. 16:39

Информация об используемой вычислительной системе	
Название вычислительной системы	Сервер кафедры КТ НИУ ИТМО
Тип вычислительной системы	Шестиядерный сервер
Характеристики вычислительной системы	Один сервер: а) один шестиядерный процессор *AMD® Phenom® II X6 1090T* с тактовой частотой 3200 МГц; б) оперативная память – 16 ГБ (четыре модуля по 4 ГБ с частотой 1333 МГц); в) 6 жестких дисков объемом 2 ТБ, скорость вращения шпинделя 7200 об./мин. (диски объединены в *RAID5*). г) скорость последовательного чтения/записи на жесткий диск: скорость чтения – до 600

	МБ/с, скорость записи – от до 600 МБ/с.

Информация об исходных данных для сборки геномной последовательности	
Источник данных	NCBI: Experiment: SRR001665
Объем данных	4 ГБ
Характеристики данных	Парные чтения бактерии *Escherichia coli str. K-12 substr. MG1655* с длиной чтения 36 нуклеотидов и размером вставки 200 нуклеотидов Произведены секвенатором *Illumina Genome Analyzer*. Уникальный идетификатор: 071112_SLXA-EAS1_s_4
L – размер рассматриваемого генома	4.5 миллиона нуклеотидов
C – покрытие генома чтениями	166

Информация о вычислениях	
Дата и время начала вычислений	12 октября 2012 г. 16:32
Дата и время окончания вычислений	12 октября 2012 г. 16:39
Общее время выполнения приложения	7 минут

Эксперимент провел:

Программист кафедры КТ, НИУ ИТМО, С. В. Мельников

_____ 12 октября 2012 года.

Протокол эксперимента № 6

Общая информация об эксперименте	
Цель эксперимента	Исследование характеристик совместного применения программного модуля, реализующего алгоритм сборки квазиконтигов, и программного средства *ABySS* с использованием парных чтений
Объект эксперимента	Программный модуль, реализующий алгоритм сборки квазиконтигов. Программное средство *ABySS*
Дата и время начала эксперимента	12 октября 2012 г. 18:43
Дата и время окончания эксперимента	12 октября 2012 г. 18:47

Информация об используемой вычислительной системе	
Название вычислительной системы	Сервер кафедры КТ НИУ ИТМО
Тип вычислительной системы	Шестиядерный сервер
Характеристики вычислительной системы	Один сервер: а) один шестиядерный процессор *AMD® Phenom® II X6 1090T* с тактовой частотой 3200 МГц; б) оперативная память – 16 ГБ (четыре модуля по 4 ГБ с частотой 1333 МГц); в) 6 жестких дисков объемом 2 ТБ, скорость вращения шпинделя 7200 об./мин. (диски объединены в *RAID5*). г) скорость последовательного чтения/записи

	на жесткий диск: скорость чтения – до 600 МБ/с, скорость записи – от до 600 МБ/с.

Информация об исходных данных для сборки геномной последовательности	
Источник данных	NCBI: Experiment: SRR001665
Объем данных	4 ГБ
Характеристики данных	Парные чтения бактерии *Escherichia coli str. K-12 substr. MG1655* с длиной чтения 36 нуклеотидов и размером вставки 200 нуклеотидов Произведены секвенатором *Illumina Genome Analyzer*. Уникальный идетификатор: 071112_SLXA-EAS1_s_4
L – размер рассматриваемого генома	4.5 миллиона нуклеотидов
C – покрытие генома чтениями	166

Информация о вычислениях	
Дата и время начала вычислений	12 октября 2012 г. 18:43
Дата и время окончания вычислений	12 октября 2012 г. 18:47
Общее время выполнения приложения	4 минуты

Эксперимент провел:

Программист кафедры КТ, НИУ ИТМО, С. В. Мельников

_____ 12 октября 2012 года.

Протокол эксперимента № 7

Общая информация об эксперименте	
Цель эксперимента	Исследование характеристик программного средства *ABySS*
Объект эксперимента	Программное средство *ABySS*
Дата и время начала эксперимента	13 октября 2012 г. 16:55
Дата и время окончания эксперимента	13 октября 2012 г. 16:59

Информация об используемой вычислительной системе	
Название вычислительной системы	Сервер кафедры КТ НИУ ИТМО
Тип вычислительной системы	Шестиядерный сервер
Характеристики вычислительной системы	Один сервер: а) один шестиядерный процессор *AMD® Phenom® II X6 1090T* с тактовой частотой 3200 МГц; б) оперативная память – 16 ГБ (четыре модуля по 4 ГБ с частотой 1333 МГц); в) 6 жестких дисков объемом 2 ТБ, скорость вращения шпинделя 7200 об./мин. (диски объединены в *RAID5*). г) скорость последовательного чтения/записи на жесткий диск: скорость чтения – до 600 МБ/с, скорость записи – от до 600 МБ/с.

Информация об исходных данных для сборки геномной последовательности	
Источник данных	NCBI: Experiment: SRR001665

Объем данных	4 ГБ
Характеристики данных	Парные чтения бактерии *Escherichia coli str. K-12 substr. MG1655* с длиной чтения 36 нуклеотидов и размером вставки 200 нуклеотидов Произведены секвенатором *Illumina Genome Analyzer*. Уникальный идетификатор: 071112_SLXA-EAS1_s_4
L – размер рассматриваемого генома	4.5 миллиона нуклеотидов
C – покрытие генома чтениями	166

Информация о вычислениях	
Дата и время начала вычислений	13 октября 2012 г. 16:55
Дата и время окончания вычислений	13 октября 2012 г. 16:59
Общее время выполнения приложения	4 минуты

Эксперимент провел:

Программист кафедры КТ, НИУ ИТМО, С. В. Мельников

_____ 13 октября 2012 года.

Протокол эксперимента № 8

Общая информация об эксперименте	
Цель эксперимента	Исследование характеристик совместного применения программного модуля, реализующего алгоритм сборки квазиконтигов, и программного средства *ABySS*
Объект эксперимента	Программный модуль, реализующий алгоритм сборки квазиконтигов. Программное средство *ABySS*
Дата и время начала эксперимента	12 октября 2012 г. 22:04
Дата и время окончания эксперимента	12 октября 2012 г. 23:55

Информация об используемой вычислительной системе	
Название вычислительной системы	Сервер кафедры КТ НИУ ИТМО
Тип вычислительной системы	Шестиядерный сервер
Характеристики вычислительной системы	Один сервер: а) один шестиядерный процессор *AMD® Phenom® II X6 1090T* с тактовой частотой 3200 МГц; б) оперативная память – 16 ГБ (четыре модуля по 4 ГБ с частотой 1333 МГц); в) 6 жестких дисков объемом 2 ТБ, скорость вращения шпинделя 7200 об./мин. (диски объединены в *RAID5*). г) скорость последовательного чтения/записи на жесткий диск: скорость чтения – до 600

	МБ/с, скорость записи – от до 600 МБ/с.

Информация об исходных данных для сборки геномной последовательности	
Источник данных	NCBI: Experiment: SRR094875
Объем данных	15 ГБ
Характеристики данных	Парные чтения мушки Drosophila melanogaster с длиной чтения 146 нуклеотидов и размером вставки 200 нуклеотидов. Произведены секвенатором *Illumina Genome Analyzer*. Уникальный идетификатор: HWI-EAS66_0016_FC62L6H
L – размер рассматриваемого генома	137 миллионов нуклеотидов
C – покрытие генома чтениями	39

Информация о вычислениях	
Дата и время начала вычислений	12 октября 2012 г. 22:04
Дата и время окончания вычислений	12 октября 2012 г. 23:55
Общее время выполнения приложения	1 час 50 минут

Эксперимент провел:

Программист кафедры КТ, НИУ ИТМО, С. В. Мельников

_____ 12 октября 2012 года.

Протокол эксперимента № 9

Общая информация об эксперименте	
Цель эксперимента	Исследование характеристик совместного применения программного модуля, реализующего алгоритм сборки квазиконтигов, и программного средства *ABySS* с использованием парных чтений
Объект эксперимента	Программный модуль, реализующий алгоритм сборки квазиконтигов. Программное средство *ABySS*
Дата и время начала эксперимента	12 октября 2012 г. 18:43
Дата и время окончания эксперимента	12 октября 2012 г. 18:47

Информация об используемой вычислительной системе	
Название вычислительной системы	Сервер кафедры КТ НИУ ИТМО
Тип вычислительной системы	Шестиядерный сервер
Характеристики вычислительной системы	Один сервер: а) один шестиядерный процессор *AMD® Phenom® II X6 1090T* с тактовой частотой 3200 МГц; б) оперативная память – 16 ГБ (четыре модуля по 4 ГБ с частотой 1333 МГц); в) 6 жестких дисков объемом 2 ТБ, скорость вращения шпинделя 7200 об./мин. (диски объединены в *RAID5*). г) скорость последовательного чтения/записи

	на жесткий диск: скорость чтения – до 600 МБ/с, скорость записи – от до 600 МБ/с.

Информация об исходных данных для сборки геномной последовательности	
Источник данных	NCBI: Experiment: SRR094875
Объем данных	15 ГБ
Характеристики данных	Парные чтения мушки Drosophila melanogaster с длиной чтения 146 нуклеотидов и размером вставки 200 нуклеотидов. Произведены секвенатором *Illumina Genome Analyzer*. Уникальный идетификатор: HWI-EAS66_0016_FC62L6H
L – размер рассматриваемого генома	137 миллионов нуклеотидов
C – покрытие генома чтениями	39

Информация о вычислениях	
Дата и время начала вычислений	12 октября 2012 г. 18:43
Дата и время окончания вычислений	12 октября 2012 г. 18:47
Общее время выполнения приложения	4 минуты

Эксперимент провел:

Программист кафедры КТ, НИУ ИТМО, С. В. Мельников

_____ 12 октября 2012 года.

Протокол эксперимента № 10

Общая информация об эксперименте	
Цель эксперимента	Исследование характеристик программного средства *ABySS*
Объект эксперимента	Программное средство *ABySS*
Дата и время начала эксперимента	15 октября 2012 г. 14:24
Дата и время окончания эксперимента	15 октября 2012 г. 16:05

Информация об используемой вычислительной системе	
Название вычислительной системы	Сервер кафедры КТ НИУ ИТМО
Тип вычислительной системы	Шестиядерный сервер
Характеристики вычислительной системы	Один сервер: а) один шестиядерный процессор *AMD® Phenom® II X6 1090T* с тактовой частотой 3200 МГц; б) оперативная память – 16 ГБ (четыре модуля по 4 ГБ с частотой 1333 МГц); в) 6 жестких дисков объемом 2 ТБ, скорость вращения шпинделя 7200 об./мин. (диски объединены в *RAID5*). г) скорость последовательного чтения/записи на жесткий диск: скорость чтения – до 600 МБ/с, скорость записи – от до 600 МБ/с.

Информация об исходных данных для сборки геномной последовательности	
Источник данных	NCBI: Experiment: SRR094875

Объем данных	15 ГБ
Характеристики данных	Парные чтения мушки *Drosophila* melanogaster с длиной чтения 146 нуклеотидов и размером вставки 200 нуклеотидов. Произведены секвенатором *Illumina Genome Analyzer*. Уникальный идетификатор: HWI-EAS66_0016_FC62L6H
L – размер рассматриваемого генома	137 миллионов нуклеотидов
C – покрытие генома чтениями	39

Информация о вычислениях	
Дата и время начала вычислений	15 октября 2012 г. 14:24
Дата и время окончания вычислений	15 октября 2012 г. 16:05
Общее время выполнения приложения	1 час 40 минут

Эксперимент провел:

Программист кафедры КТ, НИУ ИТМО, С. В. Мельников

_____ 15 октября 2012 года.

ПРИЛОЖЕНИЕ Б

СВИДЕТЕЛЬСТВО О РЕГИСТРАЦИИ ПРОГРАММЫ ДЛЯ ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2012616774

Программное средство для сборки квазиконтигов из парных чтений

Правообладатель(ли): ***федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики» (RU)***

Автор(ы): ***Александров Антон Вячеславович, Казаков Сергей Владимирович, Мельников Сергей Вячеславович, Сергушичев Алексей Александрович, Федотов Павел Валерьевич, Царев Федор Николаевич (RU)***

Заявка № **2012614488**

Дата поступления **4 июня 2012 г.**

Зарегистрировано в Реестре программ для ЭВМ

27 июля 2012 г.

Руководитель Федеральной службы по интеллектуальной собственности

Б.П. Симонов

ПРИЛОЖЕНИЕ В

ПОДГОТОВКА СТАТЬИ К ПУБЛИКАЦИИ В ЖУРНАЛЕ ИЗ ПЕРЕЧНЯ ВАК

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

№ 130 от "12" 10 2012 г.

На _____ от "___" _____ 201__ г.

СПРАВКА

дана Александрову А.В., Казакову С.В., Мельникову С.В., Сергушичеву А.А., Цареву Ф.Н.в том, что представленная в редакцию статья

Александров А.В., Казаков С.В., Мельников С.В., Сергушичев А.А., Царев Ф.Н. «Метод сборки контигов геномных последовательностей на основе совместного применения графов Де Брюина и графов перекрытий»

Принята к опубликованию в номере 6 за 2012 год журнала «Научно-технический вестник информационных технологий, механики и оптики». Выход номера из печати – ноябрь 2012 года.

Заместитель главного редактора Н.С. Кармановский

Почтовый адрес 197101, Санкт-Петербург, Кронверкский пр., д.49, комн. 330
Тел./факс +7 (812) 2334551
Эл. почта karmanov@mail.ifmo.ru
Сайт books.ifmo.ru/ntv

УДК 004.021

МЕТОД 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, Ф.Н. Царев

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 из этапов используется граф де Брюина, на втором – граф перекрытий. Описываются результаты экспериментального исследования разработанного метода на чтениях генома рыбы Maylandia zebra, размер генома которой составляет примерно миллиард нуклеотидов. Преимущество разработанного метода состоит в том, что для его работы требуется существенного меньше оперативной памяти по сравнению с существующими программными средствами для сборки генома.

Ключевые слова: сборка генома, контиги, граф де Брюина, граф перекрытий.

Введение

Многие современные задачи биологии и медицины требуют знания геномов живых организмов, которые состоят из нескольких нуклеотидных последовательностей молекул дезоксирибонуклеиновой кислоты (ДНК). В связи с этим возникает необходимость в дешевом и быстром методе определения последовательности нуклеотидов в образце ДНК. Существующие 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 в настоящее время распространение получил следующий дешевый и эффективный подход: сначала выделяется 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, какой откуда. По этой причине 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. Этот процесс делится, как правило, на следующие этапы:

1. Исправление ошибок в данных секвенирования.
2. Сборка *квазиконтигов* – фрагментов, префиксы и суффиксы которых были получены на этапе секвенирования.
3. Сборка *контигов* — максимальных непрерывных последовательностей нуклеотидов, которые удалось восстановить.
4. Построение *скэффолдов* – последовательностей контигов, разделенных промежутками, для длин которых известны верхние и нижние оценки.

Одной из наиболее часто используемых при сборке генома математических моделей является граф де Брюина [2]. На его использовании основаны следующие программные средства сборки генома: 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 размером в миллиард нуклеотидов. Так, например, SOAPdenovo необходимо порядка 140 ГБ оперативной памяти, а ABySS – 21 компьютер с 16 ГБ оперативной памяти каждый (всего – 336 ГБ). Такие затраты памяти обусловлены наличием ошибок секвенирования в исходных данных, которые ведут к увеличению размера графа де Брюина, а также неоптимальным методом хранения этого графа.

Целью настоящей работы является разработка алгоритма сборки контигов геномной последовательности, использующего меньший объем оперативной памяти по сравнению с существующими. Входные данные для алгоритма представляют собой набор парных чтений, полученных на секвенаторе, а выходными данными – набор контигов.

Для исправления ошибок в данных секвенирования в настоящей работе используется алгоритм, описанный в работе [8]. Метод сборки контигов базируется на методе, предложенном в работе [9], для сборки бактериальных геномов размером в несколько миллионов нуклеотидов. Отличие предлагаемого метода от известного состоит в том, что предлагаемый метод рассчитан на сборку геномов в несколько миллиардов нуклеотидов. Построение скэффолдов в настоящей работе не рассматривается.

Архитектура метода сборки контигов

Сборка контигов в предлагаемом методе выполняется в два этапа:

1. Сборка квазиконтигов из чтений геномной последовательности.
2. Сборка контигов из квазиконтигов. Осуществляется с использованием графа перекрытий и метода Overlap-Layout-Consensus [10].

Сборка квазиконтигов из чтений геномной последовательности. Для сборки квазиконтигов осуществляется построение графа де Брюина, в котором множество ребер состоит только из «надежных» (k+1)-меров – тех, которые встречаются в чтениях достаточно большое число раз, не меньшее некоторого порогового значения, для того чтобы их можно было с очень большой вероятностью считать входящими в геном. Рассмотрим работу алгоритма на примере 10 пар чтений генома из 25 символов (рис. 1).

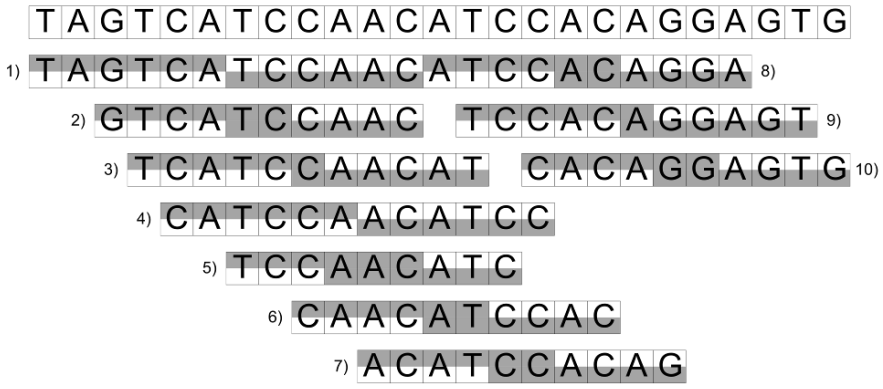

Рис. 1. Геном и его парные чтения

Если взять порог частоты для вхождения в граф равным единице, а k=3, то получится граф де Брюина, изображенный на рис. 2 (для простоты обратно-комплементарные ребра на рисунках не показаны).

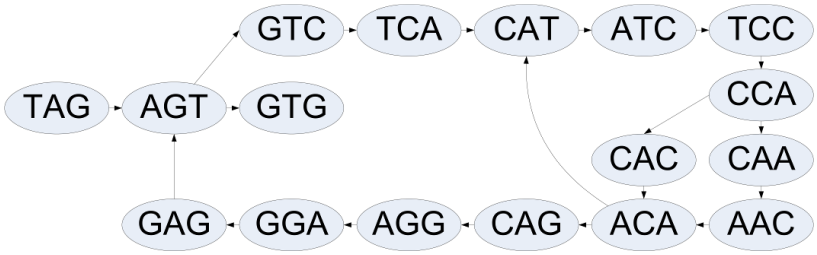

Рис. 2. Граф де Брюина при k=3

Одним из важных свойств графа де Брюина является наличие в нем пути, соответствующего геному, при условии достаточного покрытия чтениями. В частности, это означает наличие пути для каждого из фрагментов, из которых были получены парные чтения. Предлагаемый метод сборки квазиконтигов основан на поиске таких путей.

Из всех путей, начало и конец которых совпадают с парой чтений, вызывают интерес только те, которые укладываются в априорные границы длин фрагментов, поэтому слишком короткие и слишком длинные пути можно отбросить. Оставшиеся пути – «хорошие» кандидаты на роль пути, соответствующего фрагменту в действительности. Если такой путь (рис. 3) – единственный, то можно с очень большой уверенностью сказать, что он соответствует реальной подстроке геномной последовательности, поэтому этот фрагмент считается восстановленным, а найденный путь выводится как квазиконтиг.

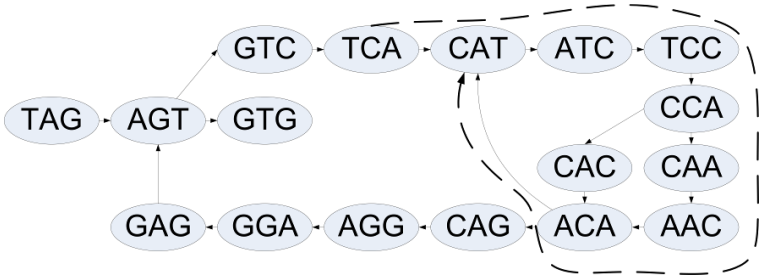

Рис. 3. Путь в графе де Брюина, соответствующий паре чтений номер 3

В случае если паре чтений в графе де Брюина соответствуют несколько путей (рис. 4), то из такой пары чтений квазиконтиг не генерируется.

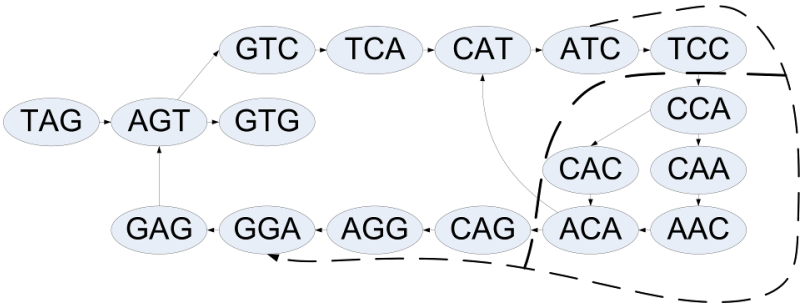

Рис. 4. Пути в графе де Брюина, соответствующие паре чтений номер 8

Приведем формальное описание алгоритма. Пусть путь p_1 длины l_1 ведет из вершины v_1 в вершину v_2, а путь p_2 длины l_2 ведет из вершины v_2 в вершину v_3. Будем обозначать конкатенацию этих путей, т.е. путь длины $l_1 + l_2$, соединяющий вершины v_1 и v_3, проходящий сначала по пути p_1, а затем — по p_2, как $p_1 \cdot p_2$. Рассмотрим множества путей P_1 и P_2.

С помощью $P_1 \cdot P_2$ будем обозначать все пути, которые можно получить конкатенацией путей p_1 и p_2 из P_1 и P_2 соответственно, т.е. $P_1 \cdot P_2 = \{ p_1 \cdot p_2 \,|\, p_1 \in P_1, p_2 \in P_2 \}$. Заметим, что конкатенировать можно только те пары путей p_1 и p_2, у которых последняя вершина p_1 совпадает с первой вершиной p_2. Например, если P_1 состоит из одного пути $v_1 \to v_2$, а множество P_2 состоит из путей $v_2 \to v_3$ и $v_4 \to v_5$, то множество $P_1 \cdot P_2$ будет состоять из одного пути $v_1 \to v_3$.

Задача, решаемая алгоритмом, состоит в поиске всех путей, соединяющих две заданные вершины v_1 и v_2, длины которых лежат в промежутке $[l_{\min}; l_{\max}]$. Будем обозначать множество всех путей из v_1 в v_2 длины l как P^l, тогда искомое множество всех путей из v_1 в v_2 будет получаться объединением множеств P^l: $P = \bigcup_{l=l_{\min}}^{l_{\max}} P^l$.

Для поиска таких путей будем применять двунаправленный поиск [11], в котором происходит одновременный поиск путей, ведущих из первой вершины, и путей, ведущих во вторую вершину. Это позволяет сократить время работы с $O(d^{l_{\max}})$ $O(d^{l_{max}})$ до $O(d^{l_{\max}/2})$ $O(d^{\frac{l_{max}}{2}})$, где d – средняя исходящая степень вершины графа (для графов де Брюина, встречающихся на практике, эта величина несущественно превышает единицу).

Применимость двунаправленного поиска объясняется тем, что любой путь p длины l из v_1 в v_2 можно разбить на два более коротких пути p_1 и p_2 длиной l_1 и l_2 так, чтобы выполнялись равенства $p = p_1 \cdot p_2$, $l = l_1 + l_2$.

Для реализации двунаправленного поиска параллельно запустим два обхода в ширину: из вершины v_1 по прямым ребрам и из v_2 – по обратным. Тогда на каждом шаге l можно поддерживать следующий инвариант: для первой вершины будем хранить множество $P_1^{l_1}$ всех исходящих из нее путей длины l_1, а для второй – множество $P_2^{l_2}$ $P_2^{l_2}$ всех входящих путей длины l_2, причем $l_1 + l_2 = l$. Таким образом, на l-ом шаге можно получить все пути длины l из v_1 в v_2 путем конкатенацией путей из множеств $P_1^{l_1}$ $P_1^{l_1}$ и $P_2^{l_2}$ $P_2^{l_2}$: $P^l = P_1^{l_1} \cdot P_2^{l_2}$ $P^l = P_1^{l_1} \cdot P_2^{l_2}$.

На начальном шаге этого алгоритма l, l_1 и l_2 равны нулю, а P_1^0 P_1^0 и P_2^0 P_2^0 содержат по одному пути нулевой длины (эту пути состоят соответственно из вершин v_1 и v_2).

Если E – это множество всех ребер графа, то шаг в первом обходе осуществляется по формуле $P_1^{l_1+1} = P_1^{l_1} \cdot E$, а шаг во втором обходе по формуле $P_2^{l_2+1} = E \cdot P_2^{l_2}$.

Для того чтобы сократить потребление памяти при применении предлагаемого метода, необходимо иметь компактное представление используемого подграфа графа де Брюина. Для этого достаточно хранить только множество его ребер, что можно эффективно делать, используя, например, хеш-таблицу с открытой адресацией [12]. Преимуществами такого подхода хранения перед другими являются его простота в реализации и достаточно высокое быстродействие.

Также для уменьшения требуемого объема оперативной памяти используется то, что каждый (k+1)-мер входит в граф вместе с обратно-комплементарным. Тогда вместо пары (k+1)-меров s и s^{rc} можно хранить только один, определяемый по некоторому правилу – например, таким правилом может быть выбор лексикографически минимального (k+1)-мера. В этом случае необходимый объем памяти уменьшается примерно в два раза для четных k и ровно в два раза – для нечетных (только для четных k существуют обратно-комплементарные себе (k+1)-меры).

Сборка контигов из квазиконтигов. Сборка контигов из квазиконтигов основана на подходе Overlap-Layout-Consensus и состоит из нескольких этапов:

1. построение графа перекрытий между квазиконтигами;
2. уточнение графа перекрытий;
3. поиск контигов в графе перекрытий.

Для поиска перекрытий построим строку вида $C_1\$C_2\$C_3\$...C_n\$$, где C_i – i-й квазиконтиг, а n – число квазиконтигов. После этого необходимо построить суффиксный массив [13] для этой строки – отсортированный массив всех суффиксов строки. С его помощью можно найти все квазиконтиги, в которых встречается заданная подстрока. Это можно сделать, например, с помощью бинарного поиска суффиксов в суффиксном массиве, которые начинаются с заданной подстроки. Они будут располагаться рядом за счет сортировки.

Зафиксируем квазиконтиг, все перекрытия с которым требуется найти. Будем рассматривать его префиксы в порядке увеличения длины, начиная с минимального порога. Для каждого префикса будем проверять, входит ли такая подстрока с добавленным в конце $ в суффиксный массив. Для того чтобы учесть еще и неточные перекрытия, проверяются не только сами префиксы, но и префиксы, в которые внесены небольшие изменения.

На следующем этапе происходят анализ найденных перекрытий, а также добавление и удаление перекрытий. Добавление происходит в случае, если квазиконтиг *A* перекрывается с квазиконтигами *B* и *C*, причем квазиконтиги *B* и *C* должны перекрываться, но такое перекрытие найдено не было. Удаление происходит, если квазиконтиг *A* перекрывается с квазиконтигом *B*, но *B* не похож на большинство квазиконтигов, с которыми перекрывается *A*.

Для поиска контигов выполняется поиск в ширину [14] в графе перекрытий. Он прерывается, если после текущего квазиконтига нет консенсуса – это означает, что квазиконтиги, перекрывающиеся с ним, различаются в большом числе позиций.

Экспериментальное исследование

Экспериментальные исследования разработанного метода проводились в рамках проекта Assemblathon 2, организованного университетом Калифорнии Санта-Круз [14]. Одним из наборов данных, который был подготовлен организаторами, являлся набор чтений рыбы Maylandia zebra. Размер генома этой рыбы оценивается примерно в 1 млрд нуклеотидов. Чтения геномной последовательности были разбиты на несколько библиотек, характеристика которых приведена в таблице.

Средний размер фрагмента в библиотеке	Покрытие
180	60
2500	62
5000	14
7000	16
9000	15
11000	12
40000	2,5

Таблица. Характеристика чтений генома рыбы Maylandia zebra

Общий объем исходных данных составлял 140 ГБ. Для сборки контигов использовалась только одна из библиотек чтений – со средним размером фрагмента в 180 нуклеотидов и 60-кратным покрытием. 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 5 суток. Опишем подробнее результаты каждого из этапов.

Перед исправлением ошибок чтения были обрезаны таким образом, 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 ч. В результате было найдено 150 млн исправлений. Всего чтений было 600 млн, поэтому было исправлено в среднем каждое четвертое чтение.

Сборка квазиконтигов заняла 38 ч. Квазиконтиги были получены из 60% парных чтений.

Сборка контигов выполнилась за 26 ч. В результате было получено 734165 контигов, суммарный размер которых составляет 680×10^6 нуклеотидов. Длина максимального контига составляет 23514 нуклеотидов, средняя длина – 927, значение метрики N50 – 1799.

Отметим, что при использовании программного средства SOAPdenovo необходимый для сборки этого генома объем оперативной памяти составил 140 ГБ, а при использовании программного средства ABySS – 336 ГБ. Это позволяет говорить о том, что у разработанного метода требования к объему оперативной памяти существенного меньше.

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 исследование показало, что разработанный метод обладает существенно меньшими требованиями к объему оперативной памяти, чем существующие.

Исследования выполнялись в рамках федеральных целевых программ «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 годы» (государственный контракт № 07.514.11.4010) и «Научные и научно-педагогические кадры инновационной России на 2009–2013 годы» (государственный контракт № 16.740.11.0495).

Литература

1. Illumina, Inc. [Электронный ресурс]. – Режим доступа: http://www.illumina.com/, свободный. Яз. англ. (дата обращения 21.03.2012).
2. Pevzner P.A. 1-Tuple DNA sequencing: computer analysis // J. Biomol. Struct. Dyn. – 1989. – V. 7. – P. 63–73.
3. Zerbino D.R., Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs // Genome Research. – 2008. – V. 18. – P. 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. – V. 18. – P. 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. – V. 19. – P. 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. – V. 20. – P. 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. – V. 98. – P. 9748–9753.
8. 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 СПбГУ ИТМО. – 2011. – № 5 (75). – С. 81–84.
9. Исенбаев В.В. Разработка системы секвенирования ДНК с использованием paired-end данных. Бакалаврская работа. СПбГУ ИТМО, 2010 [Электронный ресурс]. – Режим доступа http://is.ifmo.ru/genom/_isenbaev_thesis.pdf, свободный. Яз. англ. (дата обращения 21.03.2012).
10. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome // Nature. – 2001. – V. 409. – № 6822. – P. 860–921.
11. Рассел С., Норвиг П. Искусственный интеллект: современный подход: Пер. с англ. – 2-е изд. – М.: Вильямс. 2006. – 1408 с.
12. Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К. Алгоритмы: построение и анализ. – М.: Вильямс, 2011. – 1296 с.
13. Гасфилд Д. Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология. – СПб: Невский диалект, 2003. – 654 с.
14. Проект Assemblathon 2 [Электронный ресурс]. – Режим доступа: www.assemblathon.org, свободный. Яз. англ. (дата обращения 21.03.2012).