Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Космических и информационных технологий
институт
Высокопроизводительные вычислительные системы
кафедра

УТВЕРЖДАЮ
Заведующий кафедрой
_____ Кузьмин Д.А.
подпись фамилия, инициалы
«____» _____ 2017г.

# МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Разработка алгоритмов
обработки геномных данных
тема

09.04.01 Информатика и вычислительная техника
код и наименование направления

09.04.01.01 Высокопроизводительные вычислительные системы
код и наименование магистерской программы

| | | | |
|---|---|---|---|
| Научный руководитель | _____ | _____ | Д.А. Кузьмин |
| | подпись, дата | должность, ученая степень | инициалы, фамилия |
| Выпускник | _____ | | А.Н. Цыбин |
| | подпись, дата | | инициалы, фамилия |
| Рецензент | _____ | _____ | _____ |
| | подпись, дата | должность, ученая степень | инициалы, фамилия |

Красноярск 2017

# РЕФЕРАТ

Выпускная квалификационная работа по теме «Разработка алгоритмов обработки геномных данных» содержит 76 страниц текстового документа, 2 приложения и 33 использованных источников.

БИОИНФОРМАТИКА, ПРОГРАММНЫЙ КОМПЛЕКС, ПОВТОР, ФИЛЬТРАЦИЯ, СБОРКА, ОБРАБОТКА ГЕНОМНЫХ ДАННЫХ, СИРК.

Цели работы:

- оптимизация алгоритма СИРК с целью уменьшения времени работы и расхода оперативной памяти;
- модификация СИРК для обеспечения возможности сборки повторов до сборки генома;
- разработка и реализация алгоритма генерации генома с повторами;
- разработка и реализация алгоритма сборки контигов.

В результате работы был оптимизирован алгоритм СИРК, что позволило добиться более чем 2х-кратного ускорения при обработке больших объемов данных. СИРК также подвергся модификации, что позволило собирать повторы после этапа фильтрации. При тестировании СИРК использовался разработанный генератор генома. Программа по сборке контигов хоть и обеспечила быструю и почти не подверженную ошибкам сборку контигов, но не смогла соперничать по эффективности со стандартными алгоритмами сборки.

В итоге была модифицирована 1 программа, составлен 1 пайплайн и разработано 2 программы, используемых для различных биологических задач в рамках мега-проекта «Геномные исследования основных бореальных лесообразующих хвойных видов и их наиболее опасных патогенов в Российской Федерации».

**СОДЕРЖАНИЕ**

ВВЕДЕНИЕ.....5
1 Анализ предметной области.....7
1.1 Фильтрация повторов.....7
1.1.1 Требования к решению.....8
1.2 Сборка повторов по результатам СИРК.....9
1.2.1 Требования к решению.....10
1.3 Генерация генома с повторами.....11
1.3.1 Требования к решению.....11
1.4 Сборка контигов.....13
1.4.1 Требования к решению.....14
1.5 Выбор технологий и средств для решения поставленных задач.....15
1.5.1 Выбор языка.....15
1.5.2 Выбор средств параллелизма.....16
1.5.2.1 CUDA.....16
1.5.2.2 MPI.....17
1.5.2.3 Потоки в C++11.....17
2 Алгоритмы решения задач.....18
2.1 СИРК.....18
2.1.1 Идея.....18
2.1.2 Основные этапы работы алгоритма.....21
2.1.2.1 Этап предобработки.....21
2.1.2.2 Этап кластеризации.....22
2.1.3 Дополнения.....24
2.1.3.1 Оптимальная длина ридов.....24
2.1.3.2 Похожесть ридов.....24
2.1.3.3 Случайный выбор кмеров.....24
2.1.4 Оптимизация алгоритма.....28
2.1.4.1 Параллельное построение СИРК.....28
2.1.4.2 Параллельное нахождение похожих ридов.....29
2.2 СИРК-2.....30
2.2.1 Работа с памятью.....30
2.2.2 Динамическое распределение данных потокам.....31

2.2.3 Lock-free решение при параллельном построении СИРК.....33
2.3 Сборка повторов.....36
2.3.1 Вывод ридов повторов.....36
2.3.2 Сборка базовых ридов.....37
2.4 Генератор генома.....38
2.4.2 Генерация мест для вставки повторов.....38
2.4.3 Дробление генома на риды.....40
2.5 Сборка контигов.....42
2.5.1 Техника сборки по частям.....44
2.5.1.1 Виды сборщиков.....44
2.5.2 Идея.....47
2.5.3 Алгоритм.....47
2.5.3.1 Кластеризация.....47
2.5.3.2 Нахождение контигов для объединения.....48
2.5.3.3 Тип выравнивания.....49
2.5.3.4 Алгоритм выравнивания.....50
2.5.3.5 Объединение контигов на основе профиля и выравнивания.....52
2.5.4 Оптимизации.....54
2.5.4.1 Выравнивание.....54
2.5.4.2 Построение консенсуса.....55
3 Результаты.....57
3.1 СИРК-2.....57
3.2 Сборка повторов.....59
3.3 Генератор генома.....61
3.4 Сборка контигов.....63
ЗАКЛЮЧЕНИЕ.....68
СПИСОК СОКРАЩЕНИЙ.....69
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....70
ПРИЛОЖЕНИЕ А Блок-схема алгоритма СИРК.....74
ПРИЛОЖЕНИЕ Б Блок-схема алгоритма сборки контигов.....75

**Произведено изъятие текста ВКР с 5-76 стр.**

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Космических и информационных технологий
институт
Высокопроизводительных вычислений
кафедра

УТВЕРЖДАЮ
Заведующий кафедрой
_____ Кузьмин Д.А.
подпись фамилия, инициалы
«13» 06 2017г.

# МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Разработка алгоритмов
обработки геномных данных
тема

09.04.01 Информатика и вычислительная техника
код и наименование направления

09.04.01.01 Высокопроизводительные вычислительные системы
код и наименование магистерской программы

Научный руководитель _____ зав.каф. к.т.н., доцент Д.А. Кузьмин
подпись, дата должность, ученая степень инициалы, фамилия

Выпускник _____ 13.06.17 А.Н. Цыбин
подпись, дата инициалы, фамилия

Нормоконтролер _____ Д.А. Кузьмин
подпись, дата инициалы, фамилия

Рецензент _____ в.н.с, к.б.н. Н.В. Орешкова
подпись, дата должность, ученая степень инициалы, фамилия

Красноярск 2017