МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

Кафедра системного анализа и компьютерного моделирования

МОЗОЛЬ Назар Русланович

РАЗРАБОТКА ПРОГРАММНОГО КОМПЛЕКСА ДЛЯ ПОИСКА СМЕЖНЫХ КОНТИГОВ ПРИ СБОРКЕ ГЕНОМОВ *DE NOVO*

Аннотация (реферат) к дипломной работе

Научный руководитель:
кандидат физико-математических наук,
доцент Н.Н. Яцков

Допущен к защите
«___» _____ 2025 г.
Зав. кафедрой системного анализа и компьютерного моделирования

кандидат физико-математических наук, доцент _____ В.В. Скакун

Минск, 2025

РЕФЕРАТ

В дипломной работе представлено 62 страницы, 14 рисунков, 1 таблица, 33 источника и 1 приложение.

Ключевые слова: сборка геномов *de novo*, смежные контиги, C++, OpenMP, R, Rcpp, Bioconductor, LQ-коэффициент, покрытие, GC-содержание.

Сборка геномов *de novo* является актуальной задачей биоинформатики, особенно при работе с живыми организмами без референсного генома. Один из ключевых этапов — поиск смежных контигов, фрагментов молекул ДНК потенциально находящихся рядом в исходной последовательности. Для решения задачи требуется использование параллельных вычислений и интеграция с современными инструментами анализа.

Цель работы — разработка алгоритма поиска смежных контигов при сборке геномов *de novo* с поддержкой мультипроцессорности.

Объект исследования — процесс поиска смежных контигов;

Предмет исследования — алгоритмы и программные средства, оптимизирующие поиск смежных контигов с применением распараллеленных вычислений.

В работе проведён обзор методов секвенирования и алгоритмов сборки, рассмотрены существующие инструменты, включая Combinator-FQ. Разработан и програмно реализован оптимизированный алгоритм поиска смежных контигов на языке программирования C++ с применением библиотеки OpenMP, включающей обработку FASTA-файлов экспериментальных данных геномного секвенирования, расчёт метрик (LQ-коэффициент, покрытие, GC-содержание) и экспорт результатов. Визуализация и анализ выполнены на языке R с использованием библиотеки ggplot2, реализована возможность для интеграции с биоинформатическими инструментами через Rcpp.

Программный комплекс протестирован на данных *E. coli* и *Bradyrhizobium liaoningense*. Получено ускорение до 30% по сравнению с аналогом, а применение библиотеки OpenMP позволило добиться прироста в 27% при 8 потоках, при индентичной точности определения смежных контигов.

Новизна работы — разработка алгоритма поиска смежных контигов с поддержкой параллелизма и интеграцией C++, R и Rcpp, что обеспечивает расширенные возможности анализа и аннотации собранных псоледовательностей молекул ДНК.

Практическая значимость заключается в применимости комплекса в геномных и метагеномных исследованиях, а также в задачах персонализированной медицины. Комплекс отличается высокой производительностью, масштабируемостью и удобством, облегчая процесс поиска смежных контигов и способствуя дальнейшему развитию геномных исследований.

РЭФЕРАТ

У дыпломнай рабоце прадстаўлена 62 старонки, 14 малюнкаў, 1 табліца, 33 крыніцы і 1 дадатак.

Ключавыя словы: зборка геномаў *de novo*, суседнія кантыгі, C++, OpenMP, R, Rcpp, Bioconductor, LQ-каэфіцыент, пакрыццё, GC-змесціва.

Зборка геномаў *de novo* з'яўляецца актуальнай задачай біёінфарматыкі, асабліва пры рабоце з жывымі арганізмамі, для якіх адсутнічае эталонны (рэферэнсны) геном. Адзін з ключавых этапаў — пошук суседніх кантыгаў, фрагментаў малекул ДНК, якія патэнцыйна знаходзяцца побач у зыходнай паслядоўнасці. Для вырашэння гэтай задачы патрабуецца выкарыстанне паралельных вылічэнняў і інтэграцыя з сучаснымі аналітычнымі інструментамі.

Мэта рабоце — распрацоўка алгарытму пошуку суседніх кантыгаў пры зборцы геномаў *de novo* з падтрымкай шматпрацэсарнасці.

Аб'ект даследавання — працэс пошуку суседніх кантыгаў.

Прадмет даследавання — алгарытмы і праграмныя сродкі, што аптымізуюць пошук суседніх кантыгаў з выкарыстаннем паралельных вылічэнняў.

У рабоце праведзены агляд метадаў секвенавання і алгарытмаў зборкі, разгледжаны існуючыя інструменты, уключаючы Combinator-FQ. Распрацаваны і рэалізаваны аптымізаваны алгарытм пошуку суседніх кантыгаў на мове праграмавання C++ з выкарыстаннем бібліятэкі OpenMP, які ўключае апрацоўку FASTA-файлаў з эксперыментальных дадзеных геномнага секвенавання, разлік метрык (LQ-каэфіцыент, пакрыццё, GC-змесціва) і экспарт вынікаў. Візуалізацыя і аналіз выкананы на мове R з выкарыстаннем бібліятэкі ggplot2, рэалізавана магчымасць інтэграцыі з біёінфарматычнымі інструментамі праз Rcpp.

Праграмны комплекс пратэставаны на дадзеных *E. coli* і *Bradyrhizobium liaoningense*. Атрыманая хуткасць выканання павялічылася да 30% у параўнанні з аналагам, а выкарыстанне бібліятэкі OpenMP дазволіла дасягнуць прыросту прадукцыйнасці на 27% пры выкарыстанні 8 патокаў без страты дакладнасці вызначэння суседніх кантыгаў.

Навізна рабоце — распрацоўка алгарытму пошуку суседніх кантыгаў з падтрымкай паралелізму і інтэграцыяй C++, R і Rcpp, што забяспечвае пашыраныя магчымасці для аналізу і анатацыі сабраных паслядоўнасцей малекул ДНК.

Практычная значнасць заключаецца ў магчымасці прымянення комплексу ў геномных і метагеномных даследаваннях, а таксама ў задачах персаналізаванай медыцыны. Комплекс вызначаецца высокай прадукцыйнасцю, маштабаванасцю і зручнасцю, што спрашчае працэс пошуку суседніх кантыгаў і спрыяе далейшаму развіццю геномных даследаванняў.

STRUCTURAL ABSTRACT

The thesis comprises 62 pages, 14 figures, 1 table, 33 references, and 1 appendix.

Keywords: *de novo* genome assembly, adjacent contigs, C++, OpenMP, R, Rcpp, Bioconductor, LQ coefficient, coverage, GC content.

De novo genome assembly is a pressing task in bioinformatics, particularly when working with living organisms lacking a reference genome. One of the key stages is identifying adjacent contigs—fragments of DNA molecules that are potentially located next to each other in the original sequence. Solving this task requires the use of parallel computing and integration with modern analytical tools.

The goal of the work is to develop an algorithm for identifying adjacent contigs in *de novo* genome assembly with multiprocessor support.

Object of the study: the process of finding adjacent contigs.

Subject of the study: algorithms and software tools that optimize adjacent contig search using parallel computation.

The thesis provides an overview of sequencing methods and assembly algorithms, and examines existing tools, including Combinator-FQ. An optimized algorithm for identifying adjacent contigs was developed and implemented in C++ using the OpenMP library. It processes FASTA files from experimental genome sequencing data, computes key metrics (LQ coefficient, coverage, GC content), and exports results. Visualization and analysis were conducted in R using the ggplot2 library, and integration with bioinformatics tools was enabled through Rcpp.

The software was tested on data from *E. coli* and *Bradyrhizobium liaoningense*. The solution achieved up to a 30% speedup compared to an existing tool, with OpenMP yielding a 27% performance increase using 8 threads, while maintaining the same accuracy in identifying adjacent contigs.

Novelty of the work: development of an algorithm for adjacent contig identification with support for parallelism and integration of C++, R, and Rcpp, providing extended capabilities for analysis and annotation of assembled DNA sequences.

Practical significance: the system can be applied in genomic and metagenomic research, as well as in personalized medicine. It is distinguished by high performance, scalability, and user-friendliness, facilitating the search for adjacent contigs and contributing to the advancement of genomic research.