«УТВЕРЖДАЮ»
науке и инновациям
иальный исследовательский
ударственный университет
бачевского, к.ф.-м.-н,
[redacted]знов Михаил Юрьевич

«27» октября 2023г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ
Национального исследовательского
Нижегородского государственного университета
им. Н. И. Лобачевского (ННГУ)

на диссертационную работу
Карпулевича Евгения Андреевича

«Построение программного конвейера для выравнивания последовательностей в приложениях биоинформатики»

представленную к защите на соискание ученой степени кандидата физико-математических наук по специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Диссертация Карпулевича Е.А. посвящена исследованию и разработке математических алгоритмов и программного обеспечения в биоинформатических приложениях. Конкретной целью стала разработка алгоритмов и метода выравнивания последовательностей для решения задачи секвенирования ДНК, а также разработка и реализация архитектуры воспроизводимых биоинформатических программных конвейеров обработки данных секвенирования ДНК человека. Диссертация имеет общий объем 123 страницы и состоит из введения, трех глав, заключения, списка литературы , рисунков, таблиц, листингов и приложения.

Актуальность темы

Данные в ряде прикладных и научных областей могут быть представлены в виде последовательностей. Задача "выравнивания" последовательностей находит применение в таких прикладных областях как сжатие данных, информационный поиск, обработка естественных

языков и анализ генетических последовательностей. Методы решения задачи выравнивания последовательностей стали активно развиваться во второй половине 20-го века и нашли свое применение в различных областях, в частности, в обработке генетических данных. В инструментах, которые реализуют процедуру выравнивания генетических последовательностей, могут применяться алгоритмы двух классов: выравнивание на линейный референсный геном и выравнивание на граф, составленный по ДНК нескольких организмов. Первый класс алгоритмов обладает высокой скоростью выравнивания, в то время как второй характеризуется большей точностью. Разработка метода и алгоритмов выравнивания, которые сочетают в себе преимущества обоих подходов, является актуальной задачей.

Основные результаты. В диссертации предложены: новый метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах, алгоритмы в составе метода выравнивания генетических последовательностей и аналитические оценки их вычислительной и пространственной сложности через доказательство соответствующих теорем, архитектура и реализация системы анализа генетических данных на базе программного конвейера для обработки данных секвенирования ДНК человека с использованием модифицированного индекса референсной геномной последовательности.

В диссертационной работе Карпулевича Е.А. получены следующие основные результаты:

1. Автором разработан метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах.
2. Автором разработаны алгоритмы в составе метода выравнивания генетических последовательностей и сделаны аналитические оценки их вычислительной и пространственной сложности через доказательство соответствующих теорем.
3. Дана оценка потребления памяти для алгоритма модификации индекса на модельном примере.
4. Разработана архитектура системы анализа данных NGS на базе программного конвейера, реализующего предложенный метод выравнивания генетических последовательностей.
5. В рамках предложенной архитектуры реализован программный конвейер на основе популярного инструмента выравнивания прочитанных последовательностей minimap2. Программный конвейер апробирован на данных проекта "The Genome in a Bottle".
6. Разработанный автором программный конвейер для обработки данных секвенирования ДНК человека с использованием модифицированного индекса превосходит существующие реализации по полноте (Recall) идентификации однонуклеотидных полиморфизмов. Кроме того, разработанный программный конвейер с использованием модифицированного индекса превосходит аналогичный конвейер, использующий

стандартный индекс, по качеству (полноте, Recall) и количеству качественно выровненных ридов, при этом значения остальных метрик (Precision и F–score) не уменьшаются.

Новизна полученных результатов заключается в том, что разработаны:

- новый метод выравнивания генетических последовательностей, который сочетает в себе преимущества методов выравнивания на линейный референсный геном и выравнивания на граф;
- разработаны новые алгоритмы в составе метода выравнивания генетических последовательностей,доказаны теоремы об их вычислительной и пространственной сложности, на их основе получены асимптотические оценки сложности и требуемой оперативной памяти.

Автором выполнен большой объем исследований в области выравнивания последовательностей в приложениях биоинформатики. Предложенный метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах реализован посредством модификации функций существующего инструмента выравнивания генетических последовательностей на референсный геном minimap2. Сам инструмент для выравнивания последовательностей minimap2 достаточно известен и используется, в том числе, в коммерческих решениях (например, MGI MegaBOLT). Результаты работы могут найти применение в научных исследованиях и промышленных проектах, которые предполагают массовое секвенирование ДНК с помощью технологии NGS.

Замечания.

1. Валидация разработанных алгоритмов и реализации на базе программного конвейера, равно как и демонстрация преимуществ по сравнению с существующими инструментами проведена на эталонных данных. Степень общности этого результата для иных данных в тексте диссертации разъяснена не вполне.
2. Притом что продемонстрированы количественные преимущества разработанных алгоритмов и методов, из текста диссертации не ясно, были ли с их помощью получены качественно новые биоинформатические результаты, получить которые в рамках существовавших ранее подходов не представлялось возможным, а если нет, то каковы перспективы получить подобные результаты?
3. На рисунке 3.2 приведена схема программного конвейера для анализа данных полногеномного секвенирования из набора конвейеров "Best Practices Workflows". В тексте не описано влияние шагов по удалению дубликатов и перекалибровке качества прочтения нуклеотидов на качество итогового результата.

4. В приложении А.1 в таблице 9 приведены оценки работы программного конвейера при разных параметрах вычисления минимизаторов k и w, не до конца понятно, являются ли значения, выбранные по результатом проведенных экспериментов, оптимальными и насколько важно в контексте исследования, чтобы k и w были оптимальными.

5.

Достоверность полученных результатов подтверждается экспериментальной и теоретической проверкой работоспособности предложенного подхода, а также апробацией на научных конференциях и публикациями в рецензируемых журналах.

Таким образом, диссертационная работа Карпулевича Евгения Андреевича «Построение программного конвейера для выравнивания последовательностей в приложениях биоинформатики» удовлетворяет всем требованиям ВАК, предъявляемым к кандидатским диссертациям, а автор работы заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 2.3.5 – математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей.

Диссертационная работа и отзыв рассмотрены и утверждены на заседании кафедры прикладной математики ННГУ, протокол №5 от 27.10.2023г. На заседании семинара присутствовало 10 человек, включая двух докторов и четырех кандидатов наук.

Сведения о ведущей организации: Национальный исследовательский Нижегородский государственный университет им. Н. И. Лобачевского (ННГУ).

Адрес: 603022, г. Нижний Новгород, пр.Гагарина, 23
Электронная почта: unn@unn.ru
Телефон: +7 (831) 462-30-03
ОКПО 02068143; ОГРН 1025203733510
ИНН/КПП 5262004442/526201001

Заведующий кафедрой прикладной
математики Института информационных
технологий, математики и механики ННГУ,
доктор физико-математических наук,
(01.04.03 Радиофизика), доцент Иванченко Михаил Васильевич