ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертацию Карпулевича Евгения Андреевича

«Построение программного конвейера для выравнивания последовательностей в приложениях биоинформатики», представленную на соискание ученой степени кандидата физико-математических наук по специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Работа Карпулевича Е.А. посвящена актуальной проблеме развития методов выравнивания последовательностей. Существует большое количество методов выравнивания последовательностей, которые находят применение в таких прикладных областях как сжатие данных, информационный поиск, обработка естественных языков и анализ генетических последовательностей. В частности, для анализа генетических данных задача выравнивания генетических последовательностей решается при помощи алгоритмов двух классов: выравнивание на линейную последовательность и выравнивание на последовательности представленные в виде графа. Первый класс алгоритмов обладает высокой скоростью выравнивания, а второй класс алгоритмов обладает большей точностью. Целью работы Е.А.Карпулевича является разработка методов и алгоритмов выравнивания, которые сочетают в себе преимущества обоих классов алгоритмов. Таким образом, разработка подобного метода является актуальной задачей.

Содержание диссертации и автореферата. Диссертация содержит 123 страницы и состоит из введения, трех глав, заключения и списка литературы, включающего 96 источников.

Во введении обосновывается актуальность темы диссертации, определяются цель и задачи исследования, приводятся научная новизна и практическая значимость работы, перечисляются положения, выносимые на защиту. Также во введении приведены данные об апробации, личном вкладе и публикациях автора.

Первая глава посвящена обзору существующих алгоритмов выравнивания в применении к задаче анализа генетических данных. Рассмотрены алгоритмы точного выравнивания последовательностей, подход seed-chain-align и метод выравнивания на пангеномный граф. Для оценки качества подходов к анализу генетических данных с использованием алгоритмов выравнивания последовательностей приведены способы оценки качества вычислительного анализа генетических последовательностей. Подчеркивается актуальность создания биоинформатических программных конвейеров для решения задач анализа генетических данных.

Во второй главе описан предложенный автором метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах, приведено описание алгоритмов в составе метода и доказаны три теоремы о пространственной и вычислительной сложности алгоритма. Оценки, полученные в результате доказательства теорем, показывают, что вычислительная сложность алгоритмов построения модифицированного индекса референсной генетической последовательности остается линейной, а вычислительная сложность алгоритмов выравнивания генетических последовательностей на модифицированный индекс референсной генетической последовательности не изменяется по сравнению с выравниванием на стандартный индекс референсного генома. Теорема об оценке пространственной сложности позволяет оценить количество оперативной памяти, необходимой для работы алгоритмов.

В третьей главе описана реализация системы анализа данных NGS на базе программного конвейера, реализующего предложенный метод выравнивания генетических последовательностей. Обоснован выбор набора прикладных инструментов для реализации программного конвейера оценки генетических данных, описана реализация метода в виде библиотеки minimap2_index_modifier, приведены результаты экспериментов, выполненных на реальных данных.

В заключении приводятся основные результаты работы.

Новизна и значимость результатов. Научная новизна диссертационного исследования заключается в реализации нового метода выравнивания генетических последовательностей, который сочетает в себе преимущества методов выравнивания на линейную последовательность и выравнивания на граф. Результаты экспериментальных исследований подтверждают работоспособность предложенного метода.

Практическая значимость заключается в разработке архитектуры системы анализа данных NGS на базе программного конвейера, реализующего предложенный метод выравнивания генетических последовательностей. Реализация программного конвейера анализа данных NGS секвенирования ДНК человека с использованием модифицированного инструмента minimap2 позволило снизить количество ложноотрицательных срабатываний на 25% (274 SNP) по сравнению с программным конвейером bgallagher-sentieon, победившем в конкурсе PrecisionFDA Truth Challenge. Результаты работы могут быть использованы в масштабных научных исследованиях и промышленных проектах в области генетики и в других областях.

К сильным сторонам работы можно отнести:

1. Универсальность предложенного метода и возможность его применения для задач из других доменов.
2. Возможность использования предложенной архитектуры системы анализа генетических данных для широкого класса задач анализа генетических данных
3. Возможность применения результатов работы в существующих коммерческих и научных решениях с поддержкой выравнивания с помощью инструмента minimap2

Замечания

К содержанию и оформлению работы имеются следующие замечания:

1. В главе 3 в разделе 3.5.1 не указаны условия подбора параметров k и w инструмента minimap2
2. В главе 3 следовало бы подробнее описать реализацию метода в виде библиотеки minimap2_index_modifier, какие компоненты были реализованы и как они связаны друг с другом.
3. В диссертации имеются опечатки и неточности оформления

Замечания не являются критическими, хотя и требуют внимания со стороны автора.

Таким образом, диссертационное исследование Карпулевича Евгения Андреевича «Построение программного конвейера для выравнивания последовательностей в приложениях биоинформатики» является законченной научно-квалификационной работой. Выполненная диссертация отвечает требованиям ВАК, предъявляемым к кандидатским диссертациям, 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».

Официальный оппонент
кандидат технических наук, старший преподаватель
кафедры программной инженерии
Института информационных технологий и интеллектуальных систем
Федерального государственного автономного
об] ия высшего образования
«К федеральный университет» И.С. Алимова
17 октября 2023