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»

Актуальность темы.

Диссертационная работа Карпулевича Евгения Андреевича посвящена исследованию методов выравнивания последовательностей в приложениях биоинформатики. Построение выравниваний - это классическая задача геномных и протеомных исследований. Объемы выравниваемых последовательностей экспоненциально растут с развитием технологий секвенирования, в последнее время речь идет о сотнях терабайт или петабайтах последовательностей. Построение достоверных выравниваний является основным алгоритмом так называемых постгеномных технологий - стремительно становящихся одним из основных методов биологических исследований. В диссертационной работе автором предложен метод выравнивания генетических последовательностей. На базе предложенного метода реализованы модификация инструмента выравнивания генетических последовательностей minimap2 и программный конвейер анализа данных полногеномного секвенирования.

Достоверность и степень обоснованности научных положений, выводов и рекомендаций, сформулированных в диссертации

Достоверность данного исследования основана на нескольких факторах. В первую очередь, она обеспечивается путем тщательного теоретического обоснования свойств предлагаемого метода и его алгоритмов, а также через проведение вычислительного эксперимента с использованием разработанного программного конвейера для анализа данных полногеномного секвенирования.

Все утверждения, выводы и рекомендации, представленные в данной диссертации, имеют надежное научное обоснование. Обоснование опирается на детальный анализ существующей литературы, связанной с темой исследования. В диссертации приведен список основных работ отечественных и зарубежных авторов в данной области, которые были подвергнуты внимательному анализу и изучению.

Новизна исследования, полученных результатов, выводов и рекомендаций, сформулированных в диссертации

Следующие основные результаты диссертации являются новыми:

1. Разработан метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах.
2. Разработаны алгоритмы в составе метода выравнивания генетических последовательностей и аналитические оценки их вычислительной и пространственной сложности через доказательство соответствующих теорем. 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 не изменяется по сравнению с выравниванием на индекс референсного генома. Теорема об оценке пространственной сложности позволяет оценить количество оперативной памяти, необходимой для работы реализации алгоритмов.
3. Разработана архитектура системы анализа данных NGS на базе программного конвейера, реализующего предложенный метод выравнивания генетических последовательностей на основе популярного инструмента выравнивания ридов minimap2. Программный конвейер апробирован на данных проекта "The Genome in a Bottle".
4. Разработанный 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 с использованием стандартного индекса по качеству (по полноте, Recall) и количеству качественно выровненных ридов, при этом значения остальных метрик (Precision и F–score) не уменьшаются.

Значимость для науки и практики полученных автором результатов

Теоретическая значимость исследования заключается в разработке нового метода выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах, который сочетает в себе преимущества методов выравнивания на линейный референсный геном и выравнивания на граф, составленный по ДНК нескольких организмов.

Практическая значимость исследования заключается в разработке и реализации архитектуры системы анализа данных NGS на базе программного конвейера, реализующего предложенный метод выравнивания генетических последовательностей. Экспериментально показано, что предложенная реализация программного конвейера анализа данных NGS секвенирования ДНК человека с использованием модифицированного инструмента minimap2 позволяет снизить количество ложноотрицательных срабатываний на 25% (274 SNP) по сравнению с программным конвейером bgallagher-sentieon, победившем в конкурсе PrecisionFDA Truth Challenge.

Рекомендации по использованию результатов диссертации

Результаты работы могут быть использованы в генетических научных исследованиях и промышленных проектах, которые предполагают массовое секвенирование ДНК с помощью технологии NGS. Также предложенный метод может быть использован в применении к задачам выравнивания последовательностей в других предметных областях, например, в задаче обработки естественных языков.

Оценка содержания диссертации, ее завершенность

Диссертация написана последовательным и понятным языком и состоит из введения, трех глав, заключения и приложения.

В первой главе приводится обзор существующих подходов к выравниванию последовательностей, рассмотрены методы точного выравнивания последовательностей и методы двухэтапного выравнивания. Приведен обзор способов построения биоинформатических программных конвейеров, сделан акцент на важности воспроизводимости результатов работы программных конвейеров.

Во второй главе предложен метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах, а также описаны алгоритмы, разработанные автором для использования в составе метода выравнивания генетических последовательностей, и получены аналитические оценки их вычислительной и пространственной сложности. Метод выравнивания генетических последовательностей на референсный геном с использованием данных об известных генетических вариантах состоит в последовательном решении двух задач: задачи создания модифицированного индекса референсной генетической последовательности с добавлением данных об известных генетических вариантах и задачи выравнивания генетических последовательностей на модифицированный индекс.

Приведено теоретическое обоснование свойств предложенного метода, в частности, доказаны теоремы для оценки вычислительной сложности предложенного алгоритма построения модифицированного индекса, оценки потребления памяти предложенным алгоритмом построения модифицированного индекса и оценки вычислительной сложности предложенного алгоритма выравнивания ридов на модифицированный индекс.

В третьей главе описана разработка и программная реализация архитектуры системы анализа данных NGS на базе программного конвейера, реализующего предложенный в главе 2 метод выравнивания генетических последовательностей. Определены оптимальные требования для работы программного конвейера анализа данных полногеномного секвенирования человека.

Приведены численные эксперименты оценки качества программного конвейера для обработки данных секвенирования ДНК человека с использованием модифицированного и стандартного индексов референсного генома на данных соревнования "The precisionFDA Truth Challenge". Для дополнительного исследования работы программного конвейера для обработки данных секвенирования ДНК человека с использованием модифицированного индекса были проведены эксперименты по оценке качества разработанного программного конвейера с использованием модифицированного индекса референсного генома по сравнению с программным конвейером, который использует стандартный индекс инструмента minimap2 для разной глубины покрытия. Для оценки этапа предобработки программного конвейера приведено распределение значений метрики качества выравнивания.

Работу Карпулевича Е.А. можно квалифицировать как завершенную научно-исследовательскую работу, содержащую новые результаты, имеющие научную и практическую ценность, вносящие существенный вклад в разработку методов выравнивания последовательностей в приложениях биоинформатики и других прикладных областях.

Замечания к содержанию и оформлению диссертации

В качестве замечаний к тексту диссертации следует отметить:

1. Автор использует одно и то же условное обозначение k для обозначения нескольких разных сущностей.
2. Приведенные на рисунках 3.3 и 3.4 графики результатов профилировки программного конвейера требуют дополнительных пояснений.
3. В диссертации показана работа реализации метода только на одном наборе тестовых данных, было бы интересно исследовать результаты анализа на нескольких наборах данных.

4. Список сокращений и условных обозначений не отсортирован по алфавиту

Указанные недостатки не являются принципиальными и не ставят под сомнение ценность работы в целом.

Заключение о соответствии диссертации критериям, установленным Положением о порядке присуждения ученых степеней

Результаты диссертационной работы представлены в трех работах, опубликованных в изданиях, рекомендованных ВАК, кроме того, получено свидетельство о государственной регистрации программы для ЭВМ.

Материал диссертации изложен последовательно, структурные разделы диссертационной работы позволяют получить полное представление о проделанных исследованиях и полученных результатах. Автореферат отражает содержание диссертации и дает представление об актуальности темы, целях, задачах, новизне и полученных результатах работы.

Таким образом, диссертация Карпулевича Евгения Андреевича является научно-квалификационной работой, в которой содержатся подходы к решению задачи выравнивания последовательностей, которые сочетают в себе преимущества методов выравнивания на линейный референсный геном и выравнивания на граф, имеющие значение для развития математики, что соответствует требованиям п.9 «Положения о порядке присуждения ученых степеней», утвержденного постановлением Правительства Российской Федерации от 24.09.0213 г. № 842, предъявляемым к диссертациям ученой степени кандидата наук, а ее автор 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
Всеволод Юрьевич Макеев
доктор физико-математических наук, член-корреспондент РАН, г.н.с.
Федерального государственного бюджетного учреждения науки
Институт общей генетики им. Н.И.Вавилова Российской академии наук
Дата «09» ноября 2023 г.