ОТЗЫВ

председателя диссертационного совета на диссертацию Банкевича Антона Викторовича, на тему: «Модели, численные методы и комплекс программ для сборки геномов из стандартных данных секвенирования», представленную на соискание ученой степени кандидата физико-математических наук по специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ

В середине XX века была открыта функция молекул ДНК как основного хранилища генетической информации. Это открытие не только совершило революцию в биологии, но и впервые позволило представить такой сложный объект как геном в виде строки над алфавитом из четырех букв. Однако эффективные методы анализа геномной строки появились только в начале XXI века с появлением технологий секвенирования: биохимичеких методов, позволяющим «читать» короткие подстроки геномной строки (так называемые «прочтения»). Поскольку каждое прочтение содержит только частичную информацию о геноме, появилась необходимость разработка методов массовой совместной обработки и анализа данных секвенирования.

Анализ данных секвенирования стал основой многих передовых направлений как в современной биологии, так и в медицине и экологии, таких как синтетическая биология, персональная медицина, консервационная геномика и многих других. Интенсивное развитие биологических методов и появление новых экспериментальных средств повлекло за собой развитие вычислительных методов.

Таким образом, тема диссертации Банкевича А.В., посвящённой анализу данных секвенирования, полученных с использованием современных технологий, несомненно, является актуальной и востребованной.

Основными задачами, рассмотренными в работе, являются задачи геномной сборки (восстановления геномной строки из данных секвенирования) и предсказания общей длины геномных строк в метагеномном сообществе. Особо необходимо отметить, что в обеих задачах объектом моделирования стали именно данные секвенирования. Однако использованные подходы были принципиально различны: графовая модель, основанная на графе де Брюйна, и вероятностная модель. В результате для решения этих задач были использованы методы из различных областей математики: теории графов и оценка параметра случайной величины, соответственно. Следует отметить, что аналогичные

модели ранее уже использовались для представления данных геномного секвенирования. Однако, именно значительные нововведения, предложенные автором, отражающие особенности экспериментальных данных (например, изменение способа моделирования ошибочных рёбер в графе де Брюйна), позволили разработать максимально эффективные вычислительные методы для решения поставленных задач.

В качестве численных методов автором предложены следующие два эвристических алгоритма: алгоритм поиска ошибочных ребер в графе де Брюйна и алгоритм оценки суммарной длины геномных строк в метагеноме. В диссертации проведен теоретический анализ эффективности предложенных алгоритмов, а также приведены результаты вычисленных экспериментов, подтверждающих эффективность работы программного комплекса на реальных данных.

Важным достоинством диссертации является высокая эффективность разработанных алгоритмов, уже признанная на международном уровне. Программный комплекс SPAdes, в рамках которого были реализованы предложенные алгоритмы, широко используется в мире для решения задач сборки бактериальных геномов.

Среди недостатков работы можно отметить эвристический характер большинства предложенных алгоритмов. Необходимость дополнять (корректировать) точные решения рассматриваемых вычислительных задач, связанных с неучтенными особенностями входных данных, вызвано недостаточно точным представлением объекта в математической модели. Например, основной алгоритм поиска ошибочных ребер напрямую можно применить только в предположении отсутствия разрывов покрытия, что и привело к необходимости введения ряда эвристических корректировок, позволяющих применять алгоритм к реальным данным. Однако, подобная ситуация может быть объяснена высокой сложностью объекта моделирования. Хотя предложенный автором работы метод не имеет строго формализованного доказательства его корректности, однако именно использование эвристических оценок является в определенной степени изюминкой данной работы.

Следует также особо отметить особенность представленной к защите работы, которая заключается в симбиозе трех компонент: биологической составляющей (предмет исследования), математической модели, базирующейся на теории графов, и программного продукта. Следует также отметить, что разработанное и реализованное программное обеспечение