УТВЕРЖДАЮ

Директор
ФГБУН Институт математических
проблем биологии РАН

д.ф-м.н., проф. В. Д. Лахно

_____ 2015 г.

отзыв ведущей организации на диссертационную работу

Солдатова Руслана Андреевича

на тему " Методы предсказания структурных элементов РНК",

представленную на соискание ученой степени кандидата физико-математических наук по специальности 03.01.09 - математическая биология, биоинформатика.

Актуальность. Работа Руслана Андреевича Солдатова посвящена проблеме аннотирования некодирующих РНК в геномах различных организмов. Если задача аннотации белок-кодирующих генов практически решена для прокариотов, а для эукариотов близка к разрешению, в частности, благодаря появлению новых экспериментальных методик, то распознавание некодирующих РНК-генов и их локализация в геноме в настоящее время является открытой задачей. При этом наше представление о роли различных видов РНК, транскрибируемых в клетке, постоянно расширяется и сейчас становится понятным, что многие внутриклеточные процессы невозможны без участия РНК. Это определяет актуальность темы, выбранной диссертантом.

Основные научные результаты полученные в диссертации можно отнести к двум группам. Первая группа результатов касается предложенного автором метода распознавания участков генома потенциально имеющих термодинамически стабильную вторичную структуру (в терминологии диссертации - "локально-оптимальные структурированные сегменты"). Важным преимуществом метода является то, что он не требует предварительного задания размеров окна сканирования. Предложенная в диссертации формализация понятия сегмента,

потенциально обладающего вторичной структурой, основана на приписывании различным сегментам генома так называемого Z-значения, которое описывает степень потенциальной структурированности сегмента. В диссертации предложен метод эффективного вычисления Z-значений, который позволяет проводить массовый поиск структурированных сегментов в геномах и транскриптомах. Предложенный алгоритм реализован в виде программы RNASurface, которая снабжена интерфейсом для разметки потенциально структурных участков на геноме. Эффективность использования программы была показана на геноме *Bacillus subtilis*. Вторая группа результатов связана с одновременным исследованием структурированности ортологичных сегментов различных организмов. Предложен метод оценки статистической значимости набора Z-значений исследуемых ортологичных сегментов; метод основан на описании эволюции организмов с помощью диффузионного уравнения. Был предложен алгоритмы для вычисления необходимых характеристик и апробирован при исследовании группы геномов рода *Drosophila*. Предложенные в диссертации методы аннотации РНК превосходят по точности существующие аналоги и не уступают им по быстродействию. Тем не менее, при высоких значениях точности эти методы дают много избыточных предсказаний. Поэтому с практической точки зрения их следует рассматривать, как методы предварительной фильтрации генома, что также представляет большой интерес.

Научная новизна диссертации состоит как в полученных результатах, так и в предложенных подходах и постановках задач. Так, новой является постановка задачи на поиск фрагментов склонных к образованию вторичной структуры, не включающая в себя заранее определенную ширину окна, а также предложенный алгоритм решения этой задачи. Также новой является постановка задачи о склонности к формировании вторичной структуры одновременно для нескольких ортологичных участков ряда близких геномов. Оригинальным и перспективным является подход к решению этой задачи на основе диффузионного уравнения и алгоритм, реализующий этот подход. В контексте аннотации РНК подход, связанный с использованием диффузионных уравнений был применен впервые.

Практическая ценность диссертации состоит в разработанных алгоритмах и реализующих их программах. Программы могут применяться к различным геномным и транскрипционным данным для аннотации структурированных участков РНК. Программа RNASurface вместе с интерфейсом доступна по адресу http://bioinf.fbb.msu.ru/RNASurface/ и может свободно использоваться. В диссертации решена практически важная задача предварительной фильтрации генома при аннотации РНК-генов.

Структура работы. Диссертация имеет стандартную структуру и состоит из введения, обзора литературы (глава 1), двух глав с результатами (главы 2 и 3), выводов и библиографии. Каждая из двух основных глав соответствует одной из двух задач, рассматриваемой в диссертации. Общий объем диссертации 109 страниц, из них 94 страниц текста, включая 36 рисунков и 4 таблицы. Библиография включает 113 наименование на 10 страницах. Во введении описывается актуальность и цели задачи обнаружения структурированных РНК сегментов. Обзор литературы охватывает основные алгоритмические и статистические подходы к анализу вторичной структуры РНК и к предсказанию структурных элементов РНК в геномных последовательностях, а также содержит разбор основных моделей эволюции и изменения количественных характеристик генома в рмках этих моделей. В Главе 2 представлен подход по поиску локально-оптимальных структурированных сегментов, и описана его реализация в виде программы RNASurface. В основе подхода лежит оригинальная идея возможности быстрого вычисления не только свободных энергий всех промежуточных сегментов исследуемого участка генома, но и соответствующих Z-значений. Это, в свою очередь дает возможность отказаться от присущей другим методам необходимости априорно задавать ширину окна сканирования. Интерес представляют также проведенное автором с помощью разработанной программы исследование относительных плотностей потенциально структурированных фрагментов в различных областях генома.

В Главе 3 описан метод обнаружения эволюционно консервативных структурированных РНК по набору ортологичных последовательностей. Выше

было указано, что уровень избыточных предсказаний программы RNASurface, хотя и лучше, чем у существующих аналогов, достаточно велик. Эту характеристику можно значительно улучшить, если одновременно исследовать несколько родственных геномов. Сравнительная геномика - одно из наиболее востребованных направлений биоинформатики. До появления работ автора диссертации применение методов сравнительной геномики основывалось на поиске общей структуры анализируемых гомологичных фрагментов. Предложенный в диссертации подход позволяет охватить более широкое множество случаев - достаточно, чтобы все (или значительная часть) гомологичных сегментов обладали способностью к образованию вторичной структуры, однотипность структур для различных организмов не требуется. Как и в случае главы 2, предложенные алгоритмы реализованы в виде программы; программа была апробирована на 8 геномах рода *Drosophila*.

Замечания к работе. В данном случае недостатки работы являются продолжением её достоинств, а именно глубины и нетривиальности разработанных алгоритмов. К сожалению, описание этих алгоритмов излишне формально.

1. В Главе 2 недостаточно полно проанализировано, почему программа не распознает некоторые некодирующие РНК-гены.
2. В Главе 3 недостаточно полно обоснована применимость подхода связанного с описанием эволюции диффузионным уравнением. Сам факт применимости такого подхода сомнений не вызывает, однако приведенное в диссертации обоснование желательно бы сделать более полным и свободным от использования узко-специальных терминов.
3. При описании математических моделей автор употребляет неточные выражения. Так, например, на странице 81 автор употребляет следующее выражение: "признак имеет некоторое распределение значений p(x)"; правильно было бы "распределение вероятностей". Далее употребляется выражение "случайная нарезка из генома", что тоже не является математически точным.

4. При описании оценки параметров диффузионной модели на странице 83 автор оперирует выражением "малое количество замен", не давая количественной оценки.
5. На странице 89 автор пишет об использовании алгоритма Метрополиса-Гастингса, нигде до этого не описывая этот алгоритм и не сопровождая его ссылкой.
6. На странице 33 автор обсуждает метод ускорения алгоритма Зукера на порядок по длине последовательности, в то время как это эмпирическое наблюдение и не имеет математического доказательства.
7. На рисунке 2.1.1а, который описывает заполнение матрицы свободных энергий, закрашена не та часть матрицы.
8. В нескольких местах диссертации происходит путаница с нумерацией рисунков: есть два рисунка с номером Рисунок 2.1.1, на странице 83 вместо номера рисунка появляется надпись «Ошибка! Источник ссылки не айден.-Ошибка! Источник ссылки не найден».
9. В "Обзоре литературы" несколько раз употребляется термин "закрытая структура", не являющийся общепринятым. П
10. одпись к рисунку 2.2.1 На странице 68 содержит аббревиатуру FRP, по-видимому автор имел ввиду FPR.

Все эти замечания носят технический характер и не влияют на общую высокую оценку работы.

Рекомендации по практическому использованию. Разработанные программы могут быть использованы для обнаружения новых структурных элементов РНК в вирусах, бактериях и эукариотах, в том числе - совместно с различными экспериментальными данными о геноме в целом и его отдельных фрагментах. Описанные в диссертации алгоритмы, программы и результаты могут быть использованы в Федеральном государственном бюджетном учреждении науки Институт математических проблем биологии РАН, Федеральном государственном бюджетном учреждении науки Институт общей генетики им. Н.И. Вавилова РАН, Федеральном государственном бюджетном учреждении науки Институт молекулярной генетики РАН, Федеральном государственном бюджетном

учреждении науки Институт биологии гена РАН, Федеральном государственном бюджетном учреждении науки Институт проблем передачи информации им. А.А. Харкевича РАН, Московском государственном университете имени М.В. Ломоносова, Новосибирском государственном университете и других учебных и научно-исследовательских организациях. Результаты, представленные в диссертации, могут быть использованы при чтении курсов по биоинформатике и структуре РНК, которые читаются в Московском государственном университете имени М.В. Ломоносова, Московском физико-техническом институте и других ВУЗах страны.

Заключение. Результаты диссертации представлены в двух статьях, опубликованных в рецензируемых журналах, а также в тезисах пяти конференций. Диссертационная работа Солдатова Руслана Андреевича "Методы предсказания структурных элементов РНК" полностью соответствует критериям "Положения о порядке присуждения ученых степеней", утвержденного Постановлением №842 Правительства РФ от 24 сентября 2013 г., а её автор заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 03.01.09 - "математическая биология, биоинформатика".

Отзыв рассмотрен на расширенном семинаре лаборатории прикладной математики ИМПБ РАН 23.11.2015 г., протокол №8.

Научный сотрудник лаборатории прикладной математики

ФГБУН Института математических проблема биологии РАН

к.ф-м.н. В. В. Яковлев

25 ноября 2015 г. Тел. (4967) 318530

Подпись Яковлева В.В. заверяю.
Нач-к ОК /Галушко Т.А./