ОТЗЫВ

официального оппонента 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.

Работа Солдатова Р.А. посвящена разработке компьютерных методов для предсказания структурированных участков молекул рибонуклеиновых кислот путем анализа их последовательностей. Вычислительная задача выявления функциональных элементов генома на основании его последовательности сегодня остается как никогда актуальной в связи с революционным удешевлением массовых технологий секвенирования. Наряду с прочтением геномов экспериментальные методы адаптируются и для массового анализа различных этапов реализации генетической информации, в первую очередь транскрипции и трансляции. Сегодня можно говорить, что ключевые аспекты работы клеток определены белками, как продуктами белок-кодирующих последовательностей, через призму регуляции экспрессии генов. Регуляция осуществляется на различных уровнях, и геномика как область знания все больше внимания уделяет некодирующим РНК и регуляторным районам мРНК. В этом смысле работа Руслана Андреевича безусловно находится в активном тренде современной молекулярной биологии и биоинформатики.

В силу вычислительной сложности задачи и неточности исходных данных прямой анализ последовательности генома имеет ограниченную применимость для полномасштабного выявления структурированных участков РНК. Тем не менее, разметка последовательности по степени структурированности делает возможной функциональную аннотацию геномов и, что даже более важно, транскриптомов и различных видов «производных» транскриптомных данных о стабильности и эффективности трансляции мРНК, которые с огромной скоростью появляются в открытом доступе и далеко не всегда полноценно анализируются в момент изначальной публикации. Резюмируя вышесказанное, практическая применимость разработанных в работе методов видится очень широкой.

Автор следует устоявшейся парадигме оценки структурированности путем вычисления статистической значимости свободной энергии РНК как Z-значений на основании фонового распределения, оцениваемого по последовательностям со схожим динуклеотидным составом. Важно отметить, что наряду с удачными техническими особенностями алгоритма и реализации, автором предложено принципиальное усовершенствование имеющихся методов предсказания структурированных участков РНК, которое снимает искусственное ограничение на фиксированный размер «окна» поиска, и показано, что снятие этого ограничения значительно улучшает качество распознавания известных аннотированных структурированных РНК. Более того, в работе сделан качественный переход от базового анализа последовательностей конкретного вида к сравнительно-геномному подходу, позволяющему учесть информацию о консервативности исследуемой области генома у близких видов. Что важно, используется не эмпирическое значение консервативности как «флага»-указателя функциональных районов. Напротив, использована изящная формализация эволюционного изменения Z-значений как диффузионного процесса, и предложен метод оценки статистической значимости наблюдаемого набора Z-значений. Таким образом, работа удачно сочетает классический анализ последовательностей, методы сравнительной геномики и теории случайных процессов. Наконец, автор показывает уверенное владение современными методами теории алгоритмов и информатики, не только представляя и обосновывая свой подход, но и предлагая исследовательскому сообществу полноценные компьютерные программы, доступные для практического использования.

Суммируя вышесказанное, актуальность выбранной темы и сильная научная сторона работы не вызывают сомнений. Тем не менее, работа не лишена множества недостатков, которые осложняют восприятие и правильную интерпретацию текста.

В первую очередь, речь идет о небрежном отношении к цитированию литературных источников, которые должны сопровождать ключевые положения. Страница 4: «К настоящему моменту не существует подходящих экспериментальных методов для предсказания и классификации новых структурных РНК». Это не так, высокопроизводительные методы существуют, и автор должен быть знаком с сопутствующей литературой, в т.ч. методикой SHAPE-Seq (Mortimer и др. 2012) и FragSeq (Underwood и др. 2010).

Далее, в обзоре литературы на странице 10 автор подряд пишет: «Так, вторичная структура рибозимов важна для выполнения их каталитической функции. МикроРНК осуществляют пост-транскрипционное подавление экспрессии мРНК…». Создается ложное впечатление, что МикроРНК и рибозимы это одно и то же. Страница 17, утверждение «тем не менее он работает крайне плохо» об алгоритме Нуссинов не подкреплено ссылками на литературные источники. На странице 28 автор пишет: «Для предсказания структурных РНК необходимо выяснить свойства, которые выделяют их на фоне остального транскриптома» и в следующем же абзаце «Геномы высших эукариот очень большие, например геном человека». **Вообще, полезность идеи о необходимости сканирования генома при наличии массовых транскриптомных данных требует обсуждения, и этот фрагмент авторского текста подталкивает к этому обсуждению.** Кроме того, в качестве «большого» генома опрометчиво приводится геном человека, который достаточно мал по сравнению с геномами растений. Путаница между геномом и транскриптомом преследует автора и далее, так, на странице 60 читаем о профиле «различных биологических свойств вдоль генома» (что само по себе стилистически неграмотно), и здесь же обсуждаются «плотность рибосом, уровень экспрессии, участки взаимодействия РНК-связывающих белков» (что явно имеет отношение к транскриптому).

На странице 81 автор пишет: «Так, если последовательность имеет высокий GC состав, то последующие изменения вероятней его уменьшат». Нельзя исключить случай, когда исследуемый геном имеет высокий фоновый GC-состав. Видимо, имеется в виду относительный GC-состав? В этой связи интересен и содержательный вопрос, насколько предсказания устойчивы при межвидовых сравнениях бактерий. **Правда ли для GC-богатого генома все предсказания структурированных участков будут «оштрафованы» нормировкой на динуклеотидный состав? Не будет ли это вызывать заметное «недопредсказание» функциональных элементов?**

На странице 25 автор делает замечание о сравнительно слабой структурированности тРНК. Возникает естественный вопрос, действительно ли тРНК «слабо структурированы» (что слабо соответствует нашему устоявшемуся знанию о тРНК), либо же это прямое ограничение используемого метода оценки свободной энергии, который не может учесть какие-то характерные элементы трехмерной структуры. На странице 29 тот же вопрос вызывает фраза

«что говорит о принципиальном ограничении сканирования окном». **Правда ли дело в «сканировании окном» или же мы наблюдаем принципиальные «ошибки» или ограничения используемого метода оценки свободной энергии?**

На странице 30 автор пишет: «как показали исследования и видно из Рисунок 1.3.4, при увеличении шага окна качество предсказания структурных РНК падает катастрофически». Во-первых, на рисунке отсутствует информация о доле ложных положительных предсказаний, следовательно, невозможно сделать какой бы то ни было вывод об общем качестве предсказаний. Во-вторых, рисунок является таблицей.

Очень слабым кажется раздел «2.1.4. Общая схема алгоритма и практическая реализация». Можно было бы ожидать блок-схемы алгоритма и описания особенностей практической реализации (что интересно, поскольку автор объединил существующие программы и собственный программный код в единый инструмент), можно предполагать, что авторская программа доступна в сети интернет под свободной лицензией. **Однако, ссылки на интернет-сайт программы в диссертации не приводится, а вместо практических ключевых особенностей в реализации алгоритма дана таблица с ключами командной строки для запуска программы.** Этот комментарий справедлив и для второй части диссертационной работы по использованию модели диффузионного процесса, практическая реализация, к сожалению, в подробностях не описана.

Вообще, по тексту диссертации технические подробности и обоснования в ряде мест опущены. Страница 50: «Тем не менее, эта поверхность хорошо приближается квадратичной регрессией», не обсуждаются альтернативные варианты и их точность. На страницах 52-54 вводится множество контрольных длин и разбиения множества Γ, но обоснования выбора конкретных вариантов не приводится. На странице 59 и далее никак не обсуждается преимущество используемого метода сглаживания по сравнению, например, с простым усреднением по ближайшим соседним ячейкам матрицы. **Непонятно, насколько предсказания устойчивы по отношению к выбору параметра сглаживания.** Наконец, на странице 94 и далее (для данных Таблицы 4) не описана методология тестирования при наличии нескольких геномов (в первую очередь, что именно считается истинным предсказанием). На странице 72 упоминается программа Infernal, но нет ни ссылки на сайт или литературу, ни детального описания применения.

Ряд претензий, содержательных и технических, возникает в связи с иллюстрациями. На рисунке 1.3.2. вертикальная пунктирная линия, по всей видимости, не соответствует подписи, где сказано «только 2% случайных последовательностей имеют энергию ниже, чем у тРНК». На рисунке 1.3.3. не помечены «пики», обсуждаемые в подписи к рисунку. **Визуальный осмотр показывает, что всплески графика на обсуждаемых координатах соответствуют высоким положительным Z-значениям, что в корне противоречит утверждению об отрицательных Z-значениях структурированных участков.** Рисунок 2.1.5. иллюстрирует профили предсказания структурированных участков РНК, но в подписи к рисунку не указано, действительно ли и если да, то какие структурированные РНК представлены на рисунке. Где-то в этом же разделе нарушена нумерация, рисунок 2.1.1. находится позже рисунка 2.1.5; присутствуют два рисунка 2.2.1. (на странице 68 и на странице 80). На рисунке 2.2.1. упоминаются “p-value”, но нигде не объясняется, как именно эти значения вычислены. Ряд рисунков имеет низкое разрешение (в т.ч. Рисунок 1.2.2. и 1.2.5.), что снижает их читаемость и просто выглядит неопрятно. Рисунок 2.1.1. (страница 68) озаглавлен как «схема алгоритма», но на самом деле не содержит блок-схемы алгоритма, а схематично иллюстрирует последовательные этапы выполнения программы. На рисунке 2.2.7. время выполнения программы измеряется в неведомых величинах (размерности оси Y не подписана) и достаточно анекдотично выглядит упоминание модели процессора «Intel Xeon Processor E5506» без указания детальных технических подробностей сравнительного тестирования (операционная система, компиляторы, число используемых вычислительных потоков).

Наконец, встречаются и опечатки и стилистические недочеты, в том числе вызванные использованием профессионального жаргона, особенно в разделе «Обзор литературы». Страница 5: «Разработка на основе модели сравнительно-геномного подхода к предсказанию … структурных элементов РНК». Страница 6: «Традиционные подходы к выявлению структурных РНК сканируют геном». Там же упоминается «профиль структурированности РНК вдоль последовательности». Страница 7: требует пояснений, что значит «вдоль филогенетического дерева». Страница 8: «Теоретическое и практическое время работы и занимаемая память RNASurface не уступает самым эффективным программам данного класса». Страница 10: «Долгое время центральную роль в изучении клетки занимали белки». Страница 11: «днРНК не имеют кодирующий потенциал». Страница 12: «В дальнейшем анализе другие неканонические взаимодействия не будут учитываются». Страница 28:

«оказать решающие значение на качество предсказание». Страница 32: «Структуру со спаренными крайними нуклеотидами будем называть закрытыми». Страница 56: «из геномов человека и дрозофилы» вместо Homo sapiens и Drosophila melanogaster. Что характерно, на странице 64 в слове Sapiens сделано две опечатки. Страница 59: «Примитивные подход вычисления». Не расшифровываются полноценно и английские акронимы (PPV, FPR, и т.д.). С одной стороны, они знакомы большинству читателей, с другой стороны автор не поленился дать их русскую расшифровку. Страница 83: «ортологичных участков генома» (имеются в виду геномы). Там же: «Ошибка! Источник ссылки не айден.-Ошибка! Источник ссылки не найден». На странице 93 склеены две ссылки на рисунки «(Рисунок 3.2.2Рисунок 2.2.1)».

Перечисленные выше недостатки, безусловно, усложняют чтение и понимание работы. Однако, они никоим образом не снижают ее высокого научного уровня. Выводы и положения, сделанные в работе, хорошо обоснованы, предложены новые идеи и практические соображения; личный вклад соискателя в работу несомненен, а содержание работы отражено в достойных научных публикациях по теме. Автореферат полноценно отражает содержание диссертации. Таким образом, в заключение отмечаю, что работа Солдатова Руслана Андреевича полностью соответствует требованиям «Положения о присуждении ученых степеней», утвержденным Постановлением № 842 Правительства РФ от 24 сентября 2013 г. В свою очередь, автор, Солдатов Р.А., заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 03.01.09 – математическая биология, биоинформатика.

кандидат физико-математических наук,
старший научный сотрудник
Лаборатории вычислительных методов системной биологии
Федерального государственного бюджетного учреждения науки
Института молекулярной биологии имени В.А. Энгельгардта
Российской академии наук (ИМБ РАН)

/ И.В. Кулаковский /

24 ноября 2015 года

Личную подпись Кулаковского
УДОСТОВЕРЯЮ
ЗАВ. КАНЦЕЛЯРИЕЙ ИМБ РАН
«24» ноября 2015 г.