

Отзыв официального оппонента

о диссертации Ивана Владимировича Кулаковского

«Регуляторные мотивы в геномах высших эукариот

и их роль в экспрессии генов»,

представленной на соискание ученой степени доктора биологических

наук по специальности 03.01.09 - Математическая биология,

биоинформатика

Диссертационная работа представляет исследование по фундаментальной проблеме биоинформатики и молекулярной биологии - изучение механизмов контроля экспрессии генов эукариот на уровне транскрипции на основе анализа регуляторных мотивов и сайтов связывания транскрипционных факторов. Понимание регуляторных механизмов транскрипции, исследование их современными компьютерными методами, имеет не только большое теоретическое, но и практическое значение для задач биотехнологии и биомедицины. Продолжающееся быстрое развитие высокопроизводительных технологий секвенирования ДНК, методов экспериментального определения сайтов связывания транскрипционных факторов с помощью технологий иммунопреципитации хроматина опережает развитие компьютерных подходов анализа получаемых полногеномных данных. В связи с этим исследование И.В.Кулаковского по развитию методов биоинформатики для анализа нуклеотидных мотивов находится на передовом крае современной науки, имеет большое значение для развития специальности «Математическая биология, биоинформатика» в целом.

Актуальность темы. Работа И.В.Кулаковского представляет новые компьютерные методы исследования нуклеотидных мотивов связывания транскрипционных факторов и регуляторных районов генов эукариот, на

основе анализа данных экспериментальных технологий, использующих иммунопреципитацию хроматина и высокопроизводительное секвенирование ДНК (ChIP-chip, ChIP-seq и другие технологии, основанные на иммунопреципитации хроматина – Chromatin ImmoPrecipitation на английском языке). Представлены задачи поиска в геноме сайтов связывания на основе нуклеотидных мотивов и данных ChIP-seq, уточнения мотивов и построения моделей предсказания сайтов. Избранная диссертантом тема, несомненно, имеет большую актуальность.

Помимо компьютерной составляющей исследования в целом, можно выделить задачи, в которых применялись разработанные методы компьютерного анализа. Это изучение мотивов сайтов связывания транскрипционных факторов, ключевых факторов плюрипотентности OCT4, SOX2 и NANOG, определение давления отбора на соматические мутации в сайтах связывания транскрипционных факторов при раке, идентификация мотивов в тканеспецифичных промоторах, определенных в геномах мыши и человека. Компьютерное исследование локализации сайтов связывания транскрипционных факторов имеет большое значения для поиска их генов-мишеней, маркеров для диагностики заболеваний. Важная проблема - исследование механизмов транскрипционной регуляции генов, обеспечивающих поддержание плюрипотентного состояния стволовых клеток, относится к фундаментальным медицинским задачам. Исследование механизмов репрограммирования стволовых клеток, реконструкция регуляторных генных основано на анализе сайтов связывания транскрипционных факторов (прежде всего основных факторов Яманака - Nanog, Oct4, Sox2). В работе детально представлено исследование композитного мотива OCT4/SOX2/NANOG.

Автором создана коллекция мотивов связывания факторов транскрипции мыши и человека на основе опубликованных

экспериментальных данных, включая результаты экспериментов по иммунопреципитации хроматина ChIP-seq.

Степень обоснованности научных положений, выводов и рекомендаций.

В диссертационной работе И.В.Кулаковского рассмотрены известные данные по анализу сайтов связывания транскрипционных факторов и определению их мотивов в геномах эукариот. В целом в диссертации детально проработана литература как по компьютерным методам, моделям поиска нуклеотидных мотивов, так и по экспериментальным технологиям секвенирования ДНК, постгеномным технологиям (ChIP-seq).

Работы по определению сайтов связывания опирались на опубликованные данные, полученные с помощью технологий ChIP-on-chip, ChIP-seq, базы данных TRANSFAC, JASPAR, HOCOMOCO. В тоже время многие данные представленные в работе по определению мотивов сайтов связывания транскрипционных факторов, впервые были получены автором в его публикациях, и базе данных HOCOMOCO.

Объем диссертации составляет 245 машинописных страницы, включая 48 рисунков и 6 таблиц. Список литературы содержит 541 ссылку.

В работе диссертант грамотно использует математический и статистический аппарат, используются понятия нуклеотидного мотива, пиков профиля ChIP-seq, определения уровня значимости при поиске сайтов связывания транскрипционного фактора.

Практические рекомендации, сформулированные в диссертации, полностью обоснованы, работа апробирована в докладах на нескольких международных научных конференциях. Результаты использовалась при выполнении грантов и международных проектов, таких как FANTOM5. Публикации в высокорейтинговых международных научных журналах и в

рекомендованных ВАК изданиях (23 наименования) с достаточной полнотой отражают содержание диссертационной работы и позволяют оценить новизну работы, подтвердить ее практическую и научную значимость.

Достоверность и новизна исследования, полученных результатов, выводов и рекомендаций, сформулированных в диссертации

Диссертационная работа в целом содержит оригинальные компьютерные методы анализа мотивов связывания сайтов связывания транскрипционных факторов в геноме на основе данных ChIP-seq - иммунопреципитации хроматина, сопряженной с последующим высокопроизводительным секвенированием фрагментов ДНК. Исследования представляют срез современного научного фронта биоинформатики, были новыми на момент публикации, их достоверность подтверждена рядом последующих публикациями.

Впервые на основе данных по иммунопреципитации хроматина систематически идентифицированы тройственные композитные элементы сайтов связывания факторов транскрипции OCT4/SOX2/NANOG; установлено избегание ключевых позиций мотивов сайтов связывания относительно CpG-светофоров (CpG районов, прежде всего промоторов, метилирование которых коррелирует с экспрессией генов); выявлено действие отрицательного отбора на соматические мутации, возникающие в сайтах связывания ряда семейств факторов транскрипции в геномах раковых клеток; показана контрастная роль кластеров сайтов связывания белка Sp1 в регуляции экспрессии генов при эритролейкемии.

Диссертантом в качестве новых научных результатов выдвинуты ряд положений, включающих указанные выше позиции, впервые построена новая, наиболее полная коллекция мотивов ДНК-белкового узнавания для факторов транскрипции мыши и человека.

В целом, результаты, полученные автором, являются новыми научными знаниями по специальности «Математическая биология, биоинформатика». Основные результаты диссертации опубликованы в 27 печатных работах, они неоднократно обсуждались на различных конференциях и симпозиумах и получили одобрение ведущих специалистов.

По теме диссертации в целом представлено 27 публикации, из них 23 статей (по списку ВАК), есть также устные доклады (тезисы международных конференций).

Значимость для науки и практики полученных автором результатов

Значимость работы для науки связана с разработкой компьютерной модели сайтов связывания транскрипционных факторов, позволяющей определять локализацию сайтов связывания транскрипционного фактора по координатам секвенированных прочтений в геноме. Представлена компьютерная модель динуклеотидных матриц для описания мотивов сайтов связывания.

Практическая ценность разработанных компьютерных методов состоит в возможности поиска регуляторных районов генов по данным секвенирования в масштабе генома эукариот. Разработан комплекс новых биоинформатических методов для анализа мотивов в нуклеотидных последовательностях и база данных НОСОМОСО, активно используемая и цитируемая в настоящее время.

В целом по тематике диссертационной работы выполнен ряд грантов РФФИ и РНФ, было участие в масштабных международных проектах, в частности FANTOM5, что свидетельствует о практической направленности исследования.

Конкретные рекомендации по использованию результатов и выводов диссертации

Результаты работы могут использоваться в курсе биоинформатики для студентов университетов. Разработанные методы анализа мотивов сайтов связывания по экспериментам ChIP-seq могут быть использованы при планировании и анализе экспериментов ChIP-seq для новых транскрипционных факторов и модельных организмов эукариот.

Полученные результаты и компьютерные методы могут быть использованы в дальнейших исследованиях в научных учреждениях РАН, ФАНО РФ, занимающихся полногеномным анализом, секвенированием, и обработкой данных экспериментов ChIP-seq. Практические рекомендации, сформулированные в диссертации, обоснованы проведенными исследованиями и могут служить руководством в работе.

Оценка содержания диссертации, ее завершенность

Диссертация построена по стандартной схеме с тремя ключевыми разделами: «Обзор литературы», «Материалы и методы», «Результаты и обсуждение».

Структура диссертации всесторонне и в логической последовательности позволяет соискателю изложить свои научные взгляды, сформулировать цели работы и описать методы и результаты исследований.

Введение содержит обоснование актуальности темы исследования, автором сформулированы цели и научные задачи исследования. Детально показан личный вклад соискателя в работу в совместных публикациях с другими соавторами, соответствует специальности диссертационной работы. Приведены основные научные результаты, выносимые на защиту. Отмечается актуальность анализа экспериментальных данных

секвенирования в масштабе генома, использования современных постгеномных (омиксных) технологий.

Дан обзор литературы, современных исследований регуляторных районов транскрипции в геноме человека, включая определение сайтов связывания с помощью технологий иммунопреципитации хроматина (ChIP). Представлены исследования по регуляции экспрессии генов. Описаны подходы к изучению эмбриональных стволовых клеток человека и мыши, показана роль исследованных транскрипционных факторов, таких как Oct4, в поддержания плюрипотентности.

В главе «Материалы и методы» дано описание алгоритмов поиска нуклеотидных мотивов, компьютерных моделей сайтов связывания транскрипционных факторов в эукариотическом геноме, представлены алгоритмы анализа данных ChIP-seq и базы данных по сайтам связывания, в том числе, разработанные соискателем. Таким образом, эта глава относится скорее к описанию алгоритмов.

Глава «Результаты и обсуждение» содержит описание применения разработанных компьютерных методов к исследованию в геноме человека, результаты применения разработанных автором программ к реконструкции мотивов и локализации сайтов связывания транскрипционных факторов, полученных по экспериментальным данным ChIP-seq. Здесь же приводятся основные результаты диссертационной работы, обсуждение полученных результатов.

Достоинства и недостатки в содержании и оформлении диссертации, оценка научной работы соискателя в целом

Работа хорошо структурирована, каждая глава содержит подразделы, иногда даже избыточные (четырёхуровневая нумерация, подразделы вида 3.1.2.1), логически представляющие шаги исследования.

Достоинством, и одновременно недостатком, данной работы является ее богатый иллюстративный материал, кроме рисунков представленный «врезками» - схемами и рисунками без нумерации.

Замечания

Текст диссертации написан на хорошем языке, на удивление для многих современных диссертационных работ без грамматических ошибок. Однако стиль изложения, использование англоязычных и не общепринятых терминов, сносок и ссылок дают противоречивое впечатление о работе. С одной стороны яркое и образное представления материала, использование «врезок» - дополнительных рисунков, иногда просто поясняющих формулы, придает наглядность, с другой стороны мешает логическому изложению, строгости формулировок, не соответствуя принятым стандартам изложения материала.

Глава «Материалы и методы» фактически содержит ряд результатов, и часть обзора литературы. Стоило озаглавить ее как «Алгоритмы», или перенести в обсуждение, а собственно сами материалы и данные (используемые базы данных) коротко описать в небольшом разделе.

Избыточное введение, относящееся к современному развитию технологий, но не биоинформатике - первую страницу можно бы и вовсе опустить. С другой стороны, яркое, с пафосом, введение, эпиграфы к каждой главе, показывают большую личную заинтересованность автора, и раскрывает сам предмет, таким, каким видят его современные студенты. Умение популярно объяснить суть предмета - признак зрелости ученого.

Использован ряд терминов не в общепринятом написании, например правильнее омиксные технологии, или омиксы, а не просто «омики».

Определения терминов и сокращений должны быть даны при первом упоминании в тексте - так CpG-светофоры описаны с 210-й страницы, но впервые упоминаются в задачах исследования в начале работы.

Теоретическое значение и научно-практическая ценность диссертации подтверждаются активным цитированием ключевых статей.

Есть оформительские замечания - рисунки должны быть даны при первом упоминании в тексте. Так, после упоминания рисунков 3 и 4 через несколько страниц приведен рисунок 2.

Перевод фамилии - Маргарет Дэйхофф (автора матриц замен, M.Dayhoff) - обычно или Дайхофф или Дайхофф.

В технических замечаниях автор пишет: «форматирование ссылок на литературу в тексте и записей в списке литературы приближено к стандарту ГОСТ, за исключением прямых ссылок на Интернет-ресурсы: в силу ограниченного срока жизни они не включены в основной список литературы, а приведены в виде гиперссылок в сносках». Это прямой ответ, в то же время несоответствие ГОСТу является отклонением от стандарта в квалификационной диссертационной работе - должно быть строгое следование стандарту оформления, а не приближение. Вынужден сделать замечание.

Сноски на Интернет-ресурсы можно привести в списке литературы, а не просто подстрочно, в зависимости от вида ресурса, наличия публикации есть и другие формы цитирования (например - Wetterstrand KA. DNA Sequencing Costs...).

Автор пишет: «Поддерживаемые сообществом сайты-каталоги программ (21,22) часто оказываются существенно актуальнее, чем научные статьи, описывающие устаревающие или уже забытые подходы». Да. Это действительно так, такие ресурсы можно упоминать как ссылки в тексте, и совместно с литературой - например, сноска 48 - HGNC 48 (human gene names

consortium, [Gray и др., 2013]), может быть дана просто в тексте, иначе идет и расшифровка литературы, и сноска, и цитирование одновременно, что затрудняет восприятие.

Термины на английском языке должны поясняться - цитирование это, или параметр программы, или многократно используемый термин - например «(ни одного либо одно вхождение мотива в последовательность, zero or one occurrence per sequence)» - это перевод или параметр цитируемой в тексте программы?

Рисунок 6 - нагляден для восприятия, но термины Р-значения можно выразить формулой, это упрощение (разделить одну площадь под кривой на другую), которое больше подходит для обучающих работ студентов.

Цитата из работы «...необходимые стандартные и хорошо зарекомендовавшие себя инструменты собраны в единый интерфейс, пусть и лишенный гламурного лоска». Слова «гламурный лоск» совершенно лишние, достаточно написать просто «визуализация».

Или такая цитата: «эволюционный биолог, скандально известный за едкие публикации как в научных журналах, так и в личном блоге, издевательски сравнивал (27) тон публикаций ENCODE...». Автор, видимо, любит изящную словесность, но зачем ссылаться на блоги - диссертационная работа представлена отнюдь не по гуманитарным наукам - это стилистическое замечание.

Формулы пронумерованы, но не все. Например, после формулы (13) идет блок пояснения по формулам КДИС без нумерации.

«Формализация алгоритма ChIPMunk в виде псевдокода» (стр. 103) может быть оформлена как рисунок или схема, с соответствующим пояснением. Подписи не приведены.

Выводы содержат три основных пункта, каждый из которых представляет несколько научных положений.

Формулировка вывода 3 не совсем удачна «С помощью новых ... методов получен ряд важных результатов...» - слово «важных» избыточно, автор не должен сам оценивать свои результаты, корректнее формулировка - просто «С помощью разработанных в рамках получены следующие результаты».

Стр. 104 - Пример визуализации для мотива связывания фактора транскрипции EBF1 (SOE1) приведен на врезке. Врезка очень маленькая - почему не сделать отдельным рисунком?

Стр. 122 - 4.3.1. Сходство мотивов по Жаккару.

Визуальное пояснение формулы (пересечение множеств) - можно оформить как отдельную схему или рисунок.

Врезка на странице 130 перед рисунком 24 (соотношение P-значений) - почему бы не сделать отдельным рисунком?

Рисунок 38 - лого пары сайтов слишком мелки.

В текст диссертации и в Выводах многократно упомянуты разработанные автором «программы в открытом доступе». Стоило бы привести в явном виде таблицу ссылок на эти программы, хотя бы в качестве приложения.

Список литературы очень обстоятелен, включает 545 наименований. В то же время он не упорядочен ни по алфавиту, ни по первому упоминанию в тексте. Упорядочение по алфавиту начинается с источника 38 - видимо это техническая ошибка. Оформление списка нескольких авторов в статьях на английском языке указано как «и др.», когда следует «et al».

В огромном списке литературы процитировано всего две статьи на русском языке. Но статьи есть, в том числе авторов, чьи статьи цитируются только на английском - например, Левицкий и соавт., Ощепков и соавт., статьи которых на русском языке, более ранние и по той же теме нуклеотидных мотивов не упомянуты.

Вывод 3(д) «В ходе проекта FANTOM5 успешно проведена идентификация...» - Вывод должен показывать работу автора - наверное, точнее будет «Успешно проведена идентификация... на основе данных проекта FANTOM5».

В выводах рекомендуется избегать сокращений и аббревиатур, так вывод 3(а) «...композитном OSN-элементе...» достаточно «композитном элементе...», без OSN.

Отмеченные недостатки и технические замечания не влияют на главные основные и практические результаты диссертации. Научная работа соискателя в целом должна быть оценена на самом высоком уровне.

Заключение

Диссертация является законченным научно-исследовательским трудом, выполненным автором самостоятельно на высоком научном уровне. В работе приведены научные результаты, позволяющие квалифицировать их как теоретические положения, совокупность которых можно квалифицировать как научное достижение. Полученные автором результаты достоверны, выводы и заключения обоснованы.

Работа базируется на достаточном числе исходных данных, примеров и расчетов, представленных в высокорейтинговых публикациях. По каждому результату диссертации и работе в целом сделаны четкие выводы.

Материалы диссертации полностью отражены в публикациях автора, автореферат соответствует основному содержанию диссертации.

Диссертационная работа Кулаковского Ивана Владимировича «Регуляторные мотивы в геномах высших эукариот и их роль в экспрессии генов» соответствует п.9 Положения «О порядке присуждения ученых степеней», утвержденного Постановлением Правительства Российской Федерации от 24 сентября 2013 г. №842, с изменениями Постановления Правительства Российской Федерации от 21 апреля 2016 года №335, а ее

автор заслуживает присуждения искомой степени доктора биологических наук по специальности «03.01.09 - Математическая биология, биоинформатика».



Орлов Юрий Львович

Доктор биологических наук,

Профессор РАН,

Старший научный сотрудник Лаборатории эволюционной биоинформатики и теоретической генетики,

Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук» (ИЦиГ СО РАН)

630090, Новосибирск, Россия, пр.ак.Лаврентьева,10

Тел./факс: +7(383)3331278

Эл.почта: orlov@bionet.nsc.ru

« 9 » 10 2017 г.

Подпись Ю.Л. Орлова заверяю



Ученый секретарь ИЦиГ СО РАН

ж.б.н. Г.В. Орлова

