О Т З Ы В

официального оппонента на диссертационную работу Кулаковского Ивана Владимировича «Регуляторные мотивы в геномах высших эукариот и их роль в экспрессии генов», представленную на соискание ученой степени доктора биологических наук по специальности 03.01.09 – «Математическая биология, биоинформатика».

Разнообразие типов клеток у высших эукариот обеспечивается координированной работой сложнейшего клеточного механизма, управляющего экспрессией генов. Важнейшее звено этого механизма – этап транскрипции генов, в большой степени управляемый регуляторными белками – факторами транскрипции, связывающими специальные участки ДНК в некодирующих районах генома. Удешевление и значительное улучшение точности и скорости прочтения последовательностей ДНК и РНК позволило в последние 10-15 лет стремительно нарастить возможности и разнообразие высокопроизводительных методов молекулярной биологии. В частности, в большом объеме появляются данные по участкам, которые узнают в ДНК факторы транскрипции, как в искусственных системах *in vitro*, так и в живых клетках. Анализ беспрецедентного объема данных невозможен без использования современных компьютерных методов, и постоянный приток данных стимулирует активное развитие биоинформатики анализа последовательностей, в поле которой находится рассматриваемая работа. Тщательный анализ экспериментальных данных и построение вычислительных моделей ДНК-белкового узнавания на основе последовательностей ДНК позволяет затем проводить компьютерную функциональную аннотацию регуляторных районов генома, что далее может использоваться и в частных задачах, таких как изучение индивидуальных вариантов последовательностей, и в общих задачах, например, при моделировании генных сетей. Таким образом, методы и исследования, описанные в диссертационной работе, представленной Иваном Владимировичем Кулаковским, безусловно являются **актуальными** и соответствуют наиболее активно развивающимся направлениям в современной биоинформатике, причем **актуальность и востребованность их будут все более возрастать** в ближайшие годы. Стоит сразу отметить широту охвата проблемы (от базовых методов биоинформатического анализа ДНК-паттернов до полномасштабной базы данных, описывающей разнообразие регуляторных мотивов для факторов транскрипции человека и мыши), что безусловно является положительной чертой диссертации.

Обоснованность и достоверность научных результатов

Выполнение диссертационной работы потребовало разработки новых вычислительных методов и последующего масштабного вычислительного анализа экспериментальных данных. В результате получен многоплановый материал, который диссертанту удалось успешно систематизировать и изложить в сжатой форме. Хорошо прослеживается внутренняя логика исследования: от базовых биоинформатических методов анализа мотивов в последовательностях к сбору и систематизации информации о мотивах в виде базы данных и к применению разработанных компьютерных методов в различных практических задачах молекулярной биологии и биоинформатики. Диссертант продемонстрировал в работе глубокое понимание особенностей экспериментальных методов молекулярной биологии, что позволило улучшить компьютерные методы для поиска мотивов, а корректное применение методов математической статистики позволило сделать ряд важных утверждений и подкрепить их достоверными оценками *in silico* в отсутствие прямых экспериментальных доказательств. Материалы диссертации широко представлены научному сообществу на множестве профильных конференций и отражены и в автореферате, и в более чем 20 печатных публикациях, включая престижные международные журналы, такие как BMC Genomics, Nucleic Acids Research, и в приглашенных обзорах, подготовленных автором диссертации непосредственно по тематике диссертационной работы.

Научная новизна, практическая и теоретическая значимость исследования

Научная новизна исследования обусловлена и теоретической частью (разработка новых биоинформатических методов по анализу мотивов в последовательностях, создание новой систематической базы данных мотивов) и практической частью, где изучен ряд важных, ранее не исследованных вопросов в регуляторной геномике высших эукариот, затрагивающих как мотивы связывания факторов транскрипции в конкретных регуляторных системах, так и общие вопросы разнообразия регуляторных сигналов. **Практическая и теоретическая значимость** работы подтверждается активным использованием авторских методов и научным сообществом, и самим автором, что продемонстрировано в ходе работы и подтверждается активным цитированием публикаций по теме диссертации, судя по данным Scopus и Web-of-Science.

Структура и содержание диссертации

Диссертационная работа И.В. Кулаковского оформлена в соответствие с требованиями, предъявляемыми к докторским диссертациям и построена по традиционному плану, изложена на 245 страницах машинописного текста, включает 48 рисунков и 6 таблиц. Список литературы включает 541 процитированных источников.

Во «Введении» автор вводит основные понятия, описывает роль факторов транскрипции в регуляции экспрессии генов, формулирует цели и задачи работы, обосновывает актуальность темы, научную новизну и значимость работы.

Обзор литературы последовательно излагает современное состояние темы и содержит подразделы, посвященные рассмотрению (1) общих принципов участия факторов транскрипции в регуляции транскрипции и мотивов в структуре регуляторных некодирующих последовательностей, (2) вычислительному представлению и биоинформатическим методам анализа мотивов, (3) экспериментальным методам ДНК-белкового узнавания, с фокусом на современных методах на основе иммунопреципитации хроматина с последующим секвенированием. Обзор литературы покрывает ключевые источники по теме. Раздел **«Материалы и методы»** описывает авторские методы анализа мотивов, использованные в работе: идентификацию мотивов в больших выборках нуклеотидных последовательностей, в том числе, с учетом корреляций между соседними позициями в мотивах, оценку сходства мотивов на основе меры Жаккара, сопутствующие методы анализа мотивов для поиска вхождений и оценки роли однонуклеотидных вариантов. Методы адекватны поставленным задачам, а их техническая реализация в виде кросс-платформенных программ с открытым исходным кодом соответствует хорошим современным практикам, обеспечивающим воспроизводимость биоинформатических результатов.

Раздел **«Результаты»** изложен достаточно подробно, он посвящен описанию авторской коллекции и базы данных мотивов ДНК-белкового узнавания для факторов транскрипции человека и мыши, а также использованию методов в задачах регуляторной геномики, в том числе, вопросам организации гомо- и гетеротипических композитных элементов сайтов связывания, и взаимосвязанной регуляции транскрипции и трансляции, реализуемых через вхождения соответствующих перекрывающихся регуляторных мотивов в ДНК и РНК.

Помимо важных методических достижений, хочется особенно отметить несколько **принципиально важных и интересных результатов** и соответствующих им выводов, сделанных в работе, а именно: (1) на основании анализа современных данных ChIP-Seq выявлены структурные особенности композитных элементов для мотива фактора Spi1 в мышиной модели эритролейкемии и для мотивов SOX2/OCT4/NANOG в регуляции плюрипотентности; (2) изучена ко-локализация сайтов связывания факторов транскрипции и некоторых «точечных» объектов в геноме – CpG-светофоров (избегание) и соматических мутаций в раковых клетках (избегание и предпочтение); (3) исследовано разнообразие промоторных регуляторных элементов на основе широчайшего спектра промоторов, активных в различных клеточных типах.

Тем не менее, не смотря на **высокий уровень представленной диссертации**, имеются замечания и вопросы к тексту работы:

Нет полного названия беловых факторов транскрипции, название фактор Spi1 – ничего не говорит об объекте, хотя в роли фактора может выступать как белок, так и белковый комплекс с зашифрованным названием. Только на стр. 177 появляется описание фактора транскрипции Spi1. Оказывается, что это очень важный и интересный белок, у которого много функций и есть уникальные характеристики. На рисунке 32 приведены основные структурные семейства факторов транскрипции, но они даны на английском языке, а должны быть переведены на русский.

Известны ли области связывания белка с ДНК? Проводилась ли работа по выявлению характеристик для белковых участков связывания?

Почему так мало общего пересечения по факторам транскрипции для разных баз данных (Рис.25) ?

Используя комплекс биоинформатических методов для анализа мотивов, можно ли найти сайты связывания транскрипционных факторов для геномов растений или бактериальных геномов, изучать структуру CRISPR локусов?

Как делался анализ на выявление положительного или отрицательного отбора для сайтов и ТФ?

В диссертации есть отпечатки, использование английских слов вместо русских:

«сэмплирования по Гиббсу» стр. 44;

«в то числе» стр. 45;

«конкретных факторов транскрипции, которые склонны связывать единичные сайты.» стр. 50;

«ROC-кривая хорошего классификатора стремится находится», стр. 52;

«осуществляемому in vivo фьюжном фактора транскрипции», стр.64;

«фрагментов ДНК после соникации», стр.69;

«например комплексов сайленсинга», стр. 81;

«Трудоемкий способ – добиваться стабильных экспериментальных **повторностей** и использовать оценку невоспроизводимости в биологических **повторностях**», стр. 82 (есть слово повторяемость);

«Гиббсовского сэмплера», стр.84;

«впоследствие даже портированная для вычисления на графических процессорах GPGPU», стр. 84;

«средняя точность весовых матриц, основанных на ChIP-Seq данных [Kulakovskiy и др., 2013a; Kulakovskiy и др., 2016; Wang и др., 2012b; Wang и др., 2013a], стабильно и заметно превосходит мотивы из ранних коллекций [Dabrowski и др., 2015; Kibet, Machanick, 2015].», стр.85. Ссылки относятся к 2012 годам, а ранние коллекции к 2015 годам, несоответствие времён;

«позиционно-весовых матриц», где-то встречается полностью, а где-то сокращённое название, например стр. 114;

«базового релиза коллекции мотивов HOCOMOCO», стр. 139;

«кор мотива», «при удалении от кора» стр. 141;

«Обзор первого релиза коллекции», стр. 142;

«мотивов A-C качества» стр. 142. В автореферате стоит вообще непонятная фраза «моделей высокого качества (ABC , курированных в HOCOMOCO v9 …)»;

«сравнительное тестирование-«бенчмарк»», стр. 145;

«останавливая траверс дерева в точке, когда минимальное попарное сходство между членами кластера становилось меньше 0.05» стр.143;

«релиз сентября 2013 года», стр. 146;

«мы использовали AUC ROC», стр. 150;

«динуклеотидных мотивов мотивов для оценки», стр. 150;

«Так, у мотивов STAT1 отличается ориентация боксов: тандемный повтор у мыши и, чаще всего, палиндром у человеческого фактора, см. Рисунок 31.», стр. 159;

«Идентификация мотива извлекает известный коровый участок, и достаточно тяжелые G-богатые фланки.», стр. 159, идентификация ничего не извлекает;

«аннотированный и верифицированный старты», стр. 183;

«давление отбора изучается», стр. 191; «Оценка давления отбора» стр. 192;

«делает невозможным различение», стр. 198;

«выбиваться из общего тренда», стр. 199.

Параграф «Оценка силы отбора» стр. 201 должен быть в методах, а не перед заключением. Названия белков на рисунках 44 и 32 должны быть представлены на русском языке.

Однако, в целом, эти замечания носят дискуссионный характер и не снижают позитивного впечатления от представленной работы.

Диссертационная работа Кулаковского Ивана Владимировича представляет собой **завершенную научно-исследовательскую работу**, представляющую **решение крупной научной проблемы**, имеющей как **теоретическое, так и практическое значение**. Автореферат **полно и достоверно** отражает содержание диссертации, выводы диссертации **достоверны, обоснованны и подкреплены фактическим материалом**. Таким образом, 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».

Галзитская Оксана Валериановна
Главный научный сотрудник, к.ф.-м.н., д.ф.-м.н.,
Руководитель группы биоинформатики Института белка РАН,
142290, Россия, Московская обл., г. Пущино
Ул. Институтская д.4
Тел.: +7 495 5140218, Факс: +8 495 318435
Эл.почта: ogalzit@vega.protres.ru

Галзитская О.В.
«16» октября 2017 г.

Подпись Галзитской О.В. заверяю зав. канцелярией 16 [illegible]