ОТЗЫВ

официального оппонента на диссертационную работу

Кулаковского Ивана Владимировича

«Регуляторные мотивы в геномах высших эукариот и их роль в экспрессии генов»,

предоставленную на соискание степени доктора биологических наук по специальности

03.01.09 – «Математическая биология, биоинформатика»

Вопреки расхожему мнению, что массовое прочтение геномных последовательностей не оправдало возлагаемых на него ожиданий, появление этой информации создало платформу для новой дисциплины - сравнительной биологии и спровоцировало развитие новых теоретических и экспериментальных методов анализа, благодаря которым были обнаружены совершенно неожиданные закономерности в структурной организации геномов и механизмов экспрессии генетической информации. К ним, в частности, относится огромное число разнообразных РНК, не кодирующих белки, что, скорее всего, приведёт к пересмотру представления о пропорции генов, кодирующих белки и РНК во всех геномах. Был обнаружен высокий уровень транскрипции практически всего содержимого геномов без явного участия многих продуктов в биосинтезе белков, что противоречит парадигме догеномной эры о том, что большая часть генов кодирует белки и их транскрипция строго контролируется, чтобы исключить расточительный синтез лишних РНК. В соответствии с «вездесущей транскрипцией» было обнаружено формирование транскрипционных комплексов, в том числе с участием регуляторных белков вдали от промоторных последовательностей генов. Многие из вновь выявленных закономерностей ещё долго будут находится в состоянии осмысления, и работа Ивана Владимировича вносит значительный вклад в этот процесс. Её актуальность тем более очевидна, что российская школа биоинформатики является одной из самых сильных в мире, поэтому проводимые здесь работы глубоко интегрированы в международные проекты и, следовательно, оказывают влияние и на стратегию современных геномных исследований и на их концептуальное осмысление.

Значительная часть рецензируемой работы посвящена разработке вычислительных методов анализа полногеномных данных, направленных на поиск сайтов связывания регуляторных белков в эукариотических геномах. В рамках этого исследования создана представительная коллекция мотивов, распознаваемых факторами транскрипции в геномах человека и мыши, а закономерности их структурно-функциональной организации изучены на ряде конкретных регуляторных систем. Весь материал изложен на 245 страницах хорошо структурированного машинописного текста. Работа иллюстрирована 48 рисунками и шестью таблицами. Список цитируемой литературы включает 541 ссылку.

Особого упоминания заслуживает нестандартный стиль изложения. Так, презентация начинается с короткого **Резюме**, отражающего суть проделанной работы. За ним следует **Предисловие**, в двух разделах которого даётся представление о биоинформатике, как полноценной и самостоятельной научной дисциплине, «порождающей новое биологическое знание» и делается акцент на ключевой роли различных методов полногеномного анализа для решения современных задач молекулярной биологии. В отдельном разделе приведён список англоязычных терминов и сокращений, снабжённый очень полезными комментариями.

Ключевым разделом объёмного **Введения** является характеристика факторов транскрипции, как основных регуляторов генной экспрессии, способных связываться с особыми последовательностями в промоторах или энхансерах и влиять на экспрессию контролируемых ими генов. Отмечается, что полногеномное позиционирование сайтов связывания факторов транскрипции в геноме, осуществляемое на данном этапе геномных

исследований, необходимо для моделирования и масштабной реконструкции генных сетей, которая в ближайшем будущем обязательно станет одной из рутинных задач системной биологии. Во Введении также сформулирована цель и конкретные задачи исследования, отражена новизна, теоретическое значение и научно-практическая ценность работы, а также приводится информация о публикациях автора и его личном вкладе.

Обзор литературных данных даёт адекватное представление о современном понимании разнообразных механизмов регуляции генной экспрессии у эукариот, о спектре решаемых задач и о проблемах, способ разрешение которых пока не вполне понятен. Так как терминологию в этой области пока нельзя считать полностью устоявшейся, презентацию опубликованных работ автор предваряет чётким определением основных функциональных единиц генома, задействованных в регуляции генной экспрессии. В Обзоре достаточно подробно изложены не только современные методы компьютерного анализа полногеномных данных, но и экспериментальные подходы, используемые для их получения, а также потенциальные артефакты, обусловленные выбором конкретного подхода. В полном соответствии с основным направлением работы подробно описана стратегия позиционирования сайтов связывания регуляторных белков в локальных максимумах гистограмм распределения зарегистрированных прочтений вдоль геномной ДНК. Особенностью Обзора является многократное цитирование собственных публикаций, в том числе тех, которые отражены в защищаемой работе. Это является закономерным следствием интегрированности рецензируемой работы в современные геномные исследования, полноценный анализ которых невозможен без учёта этого вклада. В этой связи особенно ценными являются авторские оценки различных тенденций, которые, с одной стороны могут стать предметом конструктивной дискуссии, а с другой, безусловно, представляют большой интерес, так являются аргументированным мнением эксперта в данной области. Важно также, что в Обзоре приведены и критически обсуждаются доступные интернет ресурсы, достаточные для полноценного анализа результатов полногеномных экспериментов от прочтений до аннотированных сайтов связывания, а также охарактеризованы основные базы данных, содержащие информацию о сайтах связывания факторов транскрипции высших эукариот. Этот список, в частности включает и коллекцию HOCOMOCO, составленную в рамках рецензируемого исследования. Обзор полностью соответствует теме защищаемой работы и создаёт основу для адекватной оценки её значимости.

Раздел **Материалы и Методы** содержит подробное описание вычислительных методов, созданных в ходе выполнения рецензируемой работы, т.е., по сути, является такой же результативной частью диссертации, как и следующий раздел. К этим инструментам, прежде всего относится модифицированный алгоритм ChIPMunk, изначально предложенный авторами в 2009 г. и дополненный учётом профиля покрытия. В полном варианте эта программа ищет безделеционный мотив с наивысшим дискретным информационным содержанием, выявленный, во-первых, с использованием серии весовых матриц, построенных по наиболее представленным мотивам, а во-вторых, с использованием градиента их расположения в пиках экспериментальных гистограмм. Для выбора оптимальной длины мотива используются два пороговых значения информационного содержания: T и t, первый из которых соответствует граничным позициям, в которых равномерно распределены 3 из 4 возможных нуклеотидов, а второй позициям, в которых 2 одинаково представленных нуклеотида встречаются в 2 раза чаще, чем 2 других. По наличию внутри доминирующего мотива позиций с информационным содержанием меньше t, такой формализм позволяет легко идентифицировать двух-доменные сайты связывания. Такое сочетание опций позволило ChIPMunk успешно выступить в сравнительных тестированиях и интегрироваться в несколько программных комплексов, предназначенных для поиска мотивов в данных полногеномного

сканирования. Некоторое удивление вызывает отсутствие дифференциального режима у ChIPMunk. Несмотря на хорошо известную проблему с составлением негативных выборок, в каждом конкретном случае её с определённой степенью успешности можно решить и поисковые программы от этого обычно выигрывают. Тем не менее, наше тестовое использование этой программы для поиска мотива в сайтах связывания белка бактериального нуклеоида успешно выявило мотив, обнаруженный ранее MEME в режиме дифференциального поиска.

Для учёта корреляций между соседними позициями в сайтах связывания факторов транскрипции предложены динуклеотидные матрицы. Для сравнения качества алгоритмов ChIPMunk и diChIPMunk, кроме обычных тестовых критериев, авторы использовали такой параметр, как позиционирование выявленного мотива относительно максимумов покрытия в сайтах связывания. Лучшие ПВМ и диПВМ часто соответствовали мотивам, удаленным друг от друга более чем на 5 нп, но в случае динуклеотидных матриц эта разница оказалась меньшей, что, безусловно, свидетельствует о целесообразности их использования в дальнейшем.

Наряду с презентацией алгоритмов ChIPMunk и diChIPMunk, в методическом разделе содержится описание использованных в работе мер сходства мотивов, способов выравнивания позиционно-весовых матриц и определения расстояния между ними, реализованных в пакете MACRO-APE. С использованием этого программного обеспечения был оценён уровень сходства по Жаккару для мотивов 85 факторов транскрипции, присутствующих в коллекциях HOCOMOTO и JSPAR, выявленных разными поисковыми программами в разных типах экспериментальных данных. Он оказался обескураживающе низким (30-50%), что указывает на необходимость дальнейшего совершенствования инструментария формализации мотивов связывания. Любопытно, однако, что средний уровень сходства увеличивался с увеличением порогового Р-значения, т.е. мотивы, предсказанные с меньшей достоверностью, оказались более похожими.

Кроме этого в методическом разделе представлен алгоритм SPRy-SARUS, решающий задачу поиска мотивов в заданной последовательности для расширенных моделей и алгоритм PERFECTOS-APE, оценивающий *P*-значения сайтов связывания в зависимости от аллельных вариантов, а также описан алгоритм сравнения качества распознаваемых сайтов с помощью ROC-кривых. Примечательно, что для идентификации мотивов и их тестирования были использованы ранжированные по высоте пики ChIP-Seq, с чётными и нечётными порядковыми номерами, соответственно. Для построения ROC-кривых потребовался подсчет ложноположительных сигналов, но для этого был использован не экспериментальный, а модельный подход, который базируется не на поиске соответствующих мотивов в контрольном наборе, а на вычислении вероятности их появления в тестовом наборе. Строго говоря это не очень корректно, но вполне приемлемо для массового тестирования. Все созданные методы реализованы в виде компьютерных программ с открытым доступом в сети Интернет (Вывод 1).

Раздел **Результаты и их Обсуждение** состоит всего из двух глав, в первой из которых описана авторская коллекции мотивов связывания факторов транскрипции человека и мыши HOCOMOCO, а во второй - результаты использования этих мотивов для анализа структурно-функциональной организации ряда генов или регуляторных систем. Коллекция HOCOMOCO в настоящее время является, одной из немногих депозитариев, который, с одной стороны, оперативно обновляется, а с другой, целенаправленно курируется на предмет снижения «несистематической избыточности». Большинство факторов транскрипции в HOCOMOCO представлено одним доминирующим мотивом, и

только для двух десятков белков с доказанной зависимостью селективности от лигандов указаны альтернативные сайты. Очевидно, что для решения некоторых задач это может быть лимитирующим фактором, но в таком случае всегда есть возможность обратиться к другим базам данных. Безусловной ценностью коллекции является единообразие способа её создания: для сотен наборов экспериментальных данных был использован один пакет программ (ChIPMunk и diChIPMunk), что устраняет вариации, обусловленные разнообразием методических приёмов поиска мотивов.

Особенностью коллекции HOCOMOCO является наличие позиционных матриц, построенных по моно- и динуклеотидам, а также ранжирование мотивов по качеству. С использованием ChIP-Seq данных, полученных в рамках международного проекта ENCODE, для нескольких десятков факторов транскрипции было проведено сравнительное тестирование качества моделей, представленных в TRANSFAC, JASPAR и HOCOMOCO. В качестве критерия была использована площадь под ROC-кривыми. Эти данные представлены на Рис. 26 и довольно любопытны: в большинстве случаев модели HOCOMOCO оказались более точными, но ~10% сайтов лучше моделируются инструментарием TRANSFAC. Жаль, что особенности этих мотивов никак не обсуждается. В любом случае, в рецензируемой работе создана наиболее полная по сравнению с ранее опубликованными коллекция мотивов сайтов связывания факторов транскрипции мыши и человека. Её можно использовать в качестве базовой для решения многих частных задач, а также в качестве источника информации для системного анализа эволюции регуляторных элементов. В работе, в частности, отмечаются различия в контексте и структурной организации сайтов связывания некоторых гомологичных факторов транскрипции человека и мыши, но вопрос о том, насколько они отражают эволюционный тренд пока остаётся открытым. Коллекция доступна в сети Интернет, а факт её создания отражён вторым Выводом диссертации.

Во второй главе раздела Результаты и их Обсуждение показаны примеры конструктивного применения авторских методов для анализа структурно-функциональной организации регуляторных элементов ряда генов, для изучения эволюционной стабильности сайтов связывания, для понимания механизмов ткань-специфического переключения генной экспрессии, а также для поиска корреляций между расположением сайтов связывания регуляторных белков и профилем метилирования ДНК.

Примером ген-специфического исследования стал анализ структурно-функциональной организации промоторных областей генов *OCT4/SOX2/NANOG*. Он был направлен на проверку гипотезы о возможном антагонизме между *OCT4* и *NANOG*. С использованием ROC-кривых были сопоставлены все предложенные модели для каждого фактора и лучшими оказались протяжённые мотивы, предложенные diChIPMunk, которые включали модули связывания и для *OCT4* и для *SOX2*, а мотив для *NANOG*, определённый по результатам независимых экспериментов, практически повторял мотив композитного элемента *OCT4-SOX2*. На этом основании была сформулирована идея тройственного композитного элемента, в котором *NANOG* связывает участок, перекрывающийся с боксом *SOX2*, и поэтому может препятствовать устойчивому связыванию гетеродимера *OCT4-SOX2*. Это не исключает возможность независимого связывания *NANOG* с сайтами, не входящими в композитный элемент, но соответствует гипотезе об антагонизме между OCT4 и *NANOG*. Результаты этого анализа адекватно сформулированы в Выводе 3а.

Масштабное исследование было предпринято для фактора транскрипции FoxA2. Его задачей было построение моделей с использованием доступных догеномных и ChIP-seq данных и их сравнительная верификация с использованием индивидуальных порогов, определённых методом задержки ДНК-белковых комплексов в геле. Весовые матрицы,

построенные с помощью ChIPMunk, diChIPMunk по ChIP-seq данным, матрица, построенная путём выравнивания мест связывания FoxA2 в экспериментах с ДНКазным футпринтингом, а также матрица, построенная SiteGA, вначале были использованы для тестирования их качества на тестовых пиках эксперимента ChIP-seq. Модель diChIPMunk обладала самым высоким предсказательным потенциалом. С использованием экспериментально определённых весовых порогов, для каждой модели был установлен процент распознаваемых сайтов в тестовом эксперименте и оказалось, что матрица diChIPMunk, находит сайты в 70-80% пиков, но её совместное использование с матрицей SiteGA, учитывающей удаленные зависимости и построенной по «догеномным» экспериментальным данным, позволяет увеличить долю распознаваемых сайтов. Это означает, что вопрос о возможности построения точных моделей только по данным ChIP-seq остаётся открытым. На мой взгляд это важно, так как допускает возможность закономерного искажения результатов картирования сайтов связывания используемыми в настоящее время методами полногеномного сканирования.

В следующем разделе на примере мишеней фактора транскрипции Spi1 исследована зависимость экспрессии регулируемого гена от взаимного расположения сайтов связывания. С использованием данных ChIP-Seq была построена модель распознаваемого мотива, а результаты экспрессионного профилирования позволили установить, что эффект, обусловленный нокаутом гена Spi1, зависит от взаимной ориентации сайтов связывания регуляторного белка: наличие мотивов с одинаковой ориентацией в промоторах, не пересекающихся с CpG-островками, характерно для активируемых генов, а наличие таких же сайтов в энхансерах, наоборот, характерно для ингибирования экспрессии. Это интересное наблюдение отражено в Выводе 3б.

Роль пиримидиновых треков в мРНК генов, регулируемых сигнальным каскадом mTOR, исследована на примере РНК-продуктов TCT-промоторов. Особенностью 5'-терминальных Y-треков является участие в регуляции не только транскрипции, но и трансляции. Поэтому для анализа были использованы данные рибосомного футпринтинга в совокупности с результатами кэп-анализа генной экспрессии. Сначала, пиримидиновый мотив был идентифицирован в мРНК генов мишеней mTOR человека с помощью ChIPMunk, но его локализация относительно аннотированного 5'-конца, даже для mTOR-зависимых генов, менее, чем в 50% случаев оказалась терминальной. В результате более детального анализа, области стартов транскрипции были классифицированы на «узкие» и «широкие». Мотивы пиримидиновых треков для них немного отличались, а часть РНК-продуктов, синтезируемых с «широких» стартов не содержала терминального пиримидинового трека. Это важно, так как означает, что от профиля транскрипции таких генов зависит их подверженность регуляции на уровне трансляции (Вывод 3в).

Закономерной частью работы является оценка эволюционной стабильности сайтов связывания факторов транскрипции. Для этого были использованы мутации в разных типах раковых клеток, а возможные изменения аффинности сайтов оценивались с помощью PERFECTOS-APE. В результате были обнаружены сайты не только с повышенной, но и со сниженной частотой мутаций, что свидетельствует о стабилизирующем давлении отбора, направленном на сохранение критических участков регуляторных сетей (Вывод 3г). Любопытно, что для геномных локусов, доступных действию ДНКазы 1, частота мутационных изменений аффинности оказалась меньшей, что, с одной стороны, свидетельствует о большем давлении на них отрицательного отбора, а с другой, - против простого объяснения эволюционной стабильности экранированием сайтов от мутагенных факторов.

Совсем другая по масштабу и стратегии задача описана с следующем разделе: полногеномная идентификация сайтов связывания регуляторных белков в каталоге промоторов, составленном в рамках проекта FANTOM5. Мотивы связывания регуляторных белков искались по данным кэп-скрининга 4-мя разными алгоритмами, включая ChIPMunk в модификации, использующей градиент ChIP-seq пика. Не последней целью такого глобального анализа было получение ответа на вопрос: насколько полно уже известные коллекции мотивов связывания регуляторных белков отражают всю природную популяцию регуляторных сигналов в промоторах человеческого генома. Конкретной задачей стало адекватное сравнение тысяч *de novo* найденных мотивов с уже известными, для которого был использован алгоритм TomTom с пороговым уровнем для E=0.5. В результате использования очень правильной стратегии было установлено, что доля потенциально новых мотивов, распознаваемых пока не идентифицированными факторами транскрипции меньше 10% и, следовательно, имеющиеся каталоги идентифицированных сайтов практически полностью отражают доступный репертуар регуляторных сигналов (Вывод 3д).

Масштабность предыдущего вывода настолько высока, что его вполне можно было бы сделать итоговым для всей диссертации, тем не менее, в последней главе раздела приводятся результаты систематического анализа позиционных данных по метилированию, ткань-специфичной активности промоторов и расположению предсказанных сайтов связывания факторов транскрипции. Так как метилирование является относительно плохо изученным фактором регуляции генной экспрессии и в равной степени может влиять на и зависеть от связывания факторов транскрипции, такая логика постановки задачи на потенциально полностью аннотированном (в плане распределения регуляторных сигналов) геноме представляется разумной. Как и ожидалось, с использованием очень большой выборки полногеномных данных, позитивную корреляцию между уровнями метилирования CpG-сайтов и экспрессии генов зарегистрировали менее чем в 1% случаев, в то время как негативная корреляция была обнаружена в 10-15% случаев. CpG-сайты, ассоциированные с репрессией транскрипции получили название CpG-светофоров. Дальнейший анализ показал явное избегание перекрытия сайтов связывания факторов транскрипции с CpG-светофорами (Вывод 3е). Высказано предположение, что «это вызвано давлением отбора, исключающим систематическое выключение функциональных сайтов связывания» регуляторных белков, т.е. отключение опосредованной факторами транскрипции системы регуляторных связей. Такое предположение, по сути, имеет не менее масштабное биологическое значение, чем предыдущий вывод.

В конце диссертационной работы имеется краткое **Заключение**, которое не повторяет результативную часть, а скорее, формулирует авторское представление о современных тенденциях в геномных исследованиях. Восемь **Выводов** полностью отражают результативную часть работы.

Диссертация хорошо написана и легко читается. Наличие эпиграфов очень оживляет изложение, а оперативное использование сносок с пояснениями или ссылками на сетевые ресурсы и некоторые публикации, облегчает знакомство с материалом. В тексте диссертации содержится минимальное количество технических погрешностей (стр. 45, 68, 83, 101, 141, 149, 150, 190), но некоторые требуют уточнения:

- нет ли ошибки в утверждении «как минимум 5% сходства по Жаккару» (стр. 148)?
- на стр. 109 ошибочно указана ссылка на Рис. 8;
- в тексте и подписи к Рис. 18 не указано, чем отличаются режимы поиска ChIPMunk 1-3;

- Таблица 6 называется «Известные мотивы связывания ключевых факторов плюрипотентности», но в реальности мотивы в ней не указаны;
- на стр. 139 есть несоответствие между приведённым в тексте числом исследованных факторов транскрипции (474) и числами, указанными на Рис. 25.
- во фразе: «… перекрытие сайтов связывания с CpG-светофорами было обнаружено для 271 фактора транскрипции, и для 100 из них была обнаружена значимая недопредставленность CpG-светофоров в сайтах связывания» вместо слова «обнаружено» (стр. 212), по-видимому, должно стоять слово «исследовано», иначе полностью меняется смысл.

Думаю, что необоснованно краткой является описание результатов эксперимента EMSA. Даже в опубликованной статье мне не удалось найти информацию о том, какие олигонуклеотиды были использованы для конкурентного формирования ДНК-белковых комплексов. Мало понятной является подпись к рисунку 37 в этом разделе, а также к рисунку 43, содержащему информацию о частоте мутаций в некодирующих районах геномов раковых клеток. Более серьёзным замечанием является уже упоминавшееся отсутствие дифференциального режима у алгоритмов ChIPMunk и diChIPMunk.

Диссертационная работа является цельным, хорошо спланированным и законченным исследованием. Достоверность результатов никаких сомнений не вызывает, а выводы отражают суть полученных данных. Автореферат соответствует содержанию диссертации, а в публикациях автора отражены все ключевые результаты. По актуальности темы, объему и важности проведенных исследований и сделанных выводов рецензируемая работа, несомненно, соответствует требованиям «Положения о порядке присуждения учёных степеней», утвержденного Постановлением Правительства Российской Федерации от 24 сентября 2013 г. №842, с изменениями Постановления Правительства Российской Федерации от 21 апреля 2016 года №335, предъявляемым к диссертациям на соискание ученой степени доктора наук, а сам диссертант, Кулаковский Иван Владимирович, несомненно, заслуживает присуждения учёной степени доктора биологических наук по специальности 03.01.09 – "Математическая биология, биоинформатика".

Зав. лабораторией функциональной
геномики и клеточного стресса
Института биофизики клетки РАН
д.б.н., профессор Озолинь О.Н.

Адрес: 142290, г. Пущино, ул. Институтская, д. 3,
Телефон: +7(496)773-91-40
E-mail: ozoline@rambler.ru

Подпись Озолинь О.Н.
Удостоверяю зав. канц.
15.09.2017