ОТЗЫВ

официального оппонента на диссертацию

Цуканова Антона Витальевича

«Мультимодельный подход к эффективному картированию сайтов связывания транскрипционных факторов по данным ChIP-seq экспериментов»

Актуальность исследования

В работе Цуканова А.В. исследуются вопросы о применимости альтернативных моделей мотивов сайтов связывания транскрипционных факторов для полногеномного анализа. В настоящее время подавляющее число исследований, направленных на полногеномное картирование сайтов связывания транскрипционных факторов (ССТФ), используют инструменты *de novo* поиска мотивов, основанных на применении модели традиционной модели мотива PWM. Такое положение сложилось ещё до начала эпохи массового секвенирования (в частности, появления технологии ChIP-seq). Вместе с этим, методология PWM изначально возникла ещё в 80-х годах прошлого столетия именно как упрощённая модель сложного процесса, и было неоднократно показано, что PWM, в целом, является хотя и достаточно точным, но всё же явным приближением для расчёта экспериментально измеряемой аффинности транскрипционных факторов (ТФ) и ДНК. В данной работе целью является не только сравнение альтернативных моделей с традиционной PWM. Автор ставит вопрос о вкладах моделей разного вида в общее разнообразие наблюдаемых в природе взаимодействий транскрипционных факторов с ДНК *in vivo*. Отдельный интерес для дальнейшего развития представляют результаты применения классификации ТФ по структуре ДНК-связывающего домена (ДСД), а также применение методологии генной онтологии для выяснения достоинств и недостатков разных моделей мотивов сайтов связывания транскрипционных факторов.

Структура диссертации

Диссертационная работа Цуканова А.В. изложена на 204 страницах, из которых 43 страницы занимает приложение. Работа имеет стандартную структуру: оглавление, список используемых сокращений, введение, обзор литературы, методы, результаты и обсуждение, заключение, выводы, список литературы, приложения А и Б. Материал изложен последовательно, написан ясно. Список литературы включает 239 источников.

В главе «Введение» автор кратко формулирует актуальность, цель и задачи проводимого исследования, а также его научную новизну и практическую значимость.

В главе «Обзор литературы» приводится описание роли ТФ и основные механизмы регуляции транскрипции генов за счёт ТФ. В работе дано представление о том, каким образом ТФ связываются с ДНК, а также особенности ССТФ. Развернуто описаны разные модели, описывающие ТФ, начиная от консенсуса и заканчивая моделями, учитывающими зависимости

позиций нуклеотидов внутри мотива. Коротко описаны основные этапы, необходимые для проведения ChIP-seq эксперимента и его биоинформатического анализа. Дано обоснование необходимости сочетания методологически разных *de novo* методов поиска мотивов.

В главе «Методы» даётся представление об использованных в исследовании данных. Автор работы подробно описывает разработанный конвейер программ для анализа данных ChIP-seq и его основные этапы работы, включающие подготовку данных, оценку точности моделей и выбор их параметров, установку порога распознавания, классификацию пиков ChIP-seq. Подробно описан подход оценки точности распознавания разных моделей. Перечислены инструменты, используемые для статистического анализа и визуализации данных.

Глава «Результаты и обсуждение» состоит из трёх разделов. Первый раздел посвящен апробации разработанного автором конвейера на небольшой выборке данных для ССТФ FOXA2, где показана целесообразность применения методологически разных моделей для поиска ССТФ. Второй и третий раздел главы «Результаты и обсуждение» посвящены массовому анализу двух коллекций ChIP-seq данных для ССТФ растений (*A. thaliana*) и млекопитающих (*M. musculus*) с использованием разработанного автором конвейера. Автор проводит сравнения точности моделей с помощью стандартного подхода ROC кривой, используя версию pAUC стандартной метрики AUC. В результате в работе показано, что модель BaMM превосходит по метрике pAUC стандартную модель PWM, при этом данный результат не зависит от класса ТФ по ДСД. Модель SiteGA имеет сравнимую точность с моделью PWM только для некоторых классов ТФ. Анализ результатов распознавания моделей показал, что у пары моделей BaMM/PWM предсказания значимо положительно коррелируют, а для пар PWM/SiteGA и BaMM/SiteGA корреляции нет. Автор предположил, что модель SiteGA будет чаще выявлять отличные от других моделей ССТФ, так как она иначе оценивает аффинность ССТФ. Проведен анализ совместной встречаемости мотивов разных моделей в пиках ChIP-seq. Показано, что альтернативные модели существенно расширяют результаты PWM, увеличивая долю пиков, содержащих сайты, при этом вклад альтернативных моделей существенно зависит от класса ТФ по ДСД. Важно, что данный вывод согласован для ТФ удалённых таксонов растений *A. thalina* и млекопитающих *M. musculus*. Например, наибольшие и наименьшие вклады для обоих видов организмов наблюдаются для классов ТФ Basic helix-loop-helix factors (bHLH) и C2H2 zinc finger factors, соответственно. Следовательно, структурное разнообразие мотивов и вклад зависимостей позиций в информационное содержание нуклеотидного контекста мотива (оцениваемое как аффинность ССТФ) показывает явную зависимость от структуры ДСД. В заключительных частях второго и третьего разделов главы «Результаты и обсуждение» проведён анализ с помощью подхода генной онтологии (ГО) коллекций данных *A. thaliana* и *M. musculus*. Сайты, предсказанные моделями PWM, BaMM и SiteGA, были картированы на промоторы генов. Показано, что часть генов в своих промоторах могут иметь сайты, предсказанные только одной из моделей. Было предположено, что

группы таких генов могут быть объединены специфической функцией. Анализ терминов ГО показал, что модели могут выявлять специфические термины ГО для отдельных моделей. Для общих терминов ГО, с которыми связаны модели PWM, BaMM и SiteGA, было показано, что именно модель SiteGA имеет большее обогащение предсказанных ССТФ для генов, помеченных данными терминами ГО, по сравнению с моделями PWM и SiteGA.

В разделе «Заключение» суммируются полученные данные и дается общее заключение, содержащее концентрированное изложение сути работы.

Замечания

Возникшие замечания носят преимущественно редакционный характер. Например, на титульных листах автореферата и диссертации в качестве даты защиты указан 2023 год. Кроме того, в тексте замечен ряд пунктуационных ошибок и опечаток. Однако эти замечания не отражаются на высокой положительной оценке диссертационной работы Антона Витальевича.

Заключение

Диссертация Цуканова Антона Витальевича является полноценной, законченной научно-квалификационной работой, посвященной важной проблеме – поиску сайтов связывания транскрипционных факторов с помощью методологически разных моделей. По поставленным задачам, уровню их решения, актуальности и научной новизне полученных результатов диссертационная работа Антона Витальевича полностью соответствует требованиям п.п. 9-14 «Положения о присуждении учёных степеней» (утверждено Постановлением Правительства РФ от 24 сентября 2013 г. №842 в редакции от 26.01.2023 № 101), а её автор Цуканов Антон Витальевич заслуживает присуждения степени кандидата биологических наук по специальности 1.5.8. - математическая биология, биоинформатика.

Официальный оппонент:
зав. лабораторией клеточного деления Федерального государственного бюджетного учреждения науки Института молекулярной и клеточной биологии Сибирского отделения Российской академии наук (ИМКБ СО РАН)
кандидат биологических наук Омелина Евгения Сергеевна

Контактная информация:
Адрес: 630090, г. Новосибирск, пр. Ак. Лаврентьева, д. 8/2
Тел.: +7-383-363-90-42; e-mail: omelina@mcb.nsc.ru

Федеральное государственное бюджетное учреждение науки Институт молекулярной и клеточной биологии Сибирского отделения Российской академии наук Россия г.Новосибирск
Для документов (ИМКБ СО РАН)

Федеральное государственное бюджетное учреждение науки Институт молекулярной и клеточной биологии Сибирского отделения Российской академии наук
Подпись Омелиной Е. С.
ЗАВЕРЯЮ
Зав. канцелярией
«05» 02 2024 г.