ОТЗЫВ официального оппонента
на диссертационную работу Цуканова Антона Витальевича
«Мультимодельный подход к эффективному картированию сайтов связывания транскрипционных факторов по данным ChIP-seq экспериментов»,
представленной на соискание учёной степени кандидата биологических наук по специальности 1.5.8. — математическая биология, биоинформатика

Актуальность исследования

Поиск сайтов связывания транскрипционных факторов (ССТФ) является важной задачей, поскольку ТФ являются ключевыми элементами в регуляции экспрессии генов, а идентификация ССТФ позволяет лучше понять, какие гены активируются или ингибируются в ответ на различные сигналы и условия. Автор работы поднимает важный вопрос насколько альтернативные модели применимы для поиска ССТФ в массовом анализе ChIP-seq данных. К настоящему моменту уже проведено огромное количество ChIP-seq экспериментов для разных организмов и обработанные данные в виде ChIP-seq пиков хранятся в базах данных (БД), таких, как например GTRD. Дальнейший анализ таких данных, а именно *de novo* поиск мотивов, преимущественно проводится с применением стандартной модели мотива – PWM. Существует множество баз данных, где представлены PWM для огромного числа разных транскрипционных факторов (ТФ) и во многом благодаря этому модель PWM является широко используемой в исследования ССТФ. Тем не менее модель PWM является только приближением при описании ССТФ, поскольку уже было неоднократно показано, что внутри мотива ТФ могут присутствовать зависимости между нуклеотидами, которые PWM не берёт в расчёт. Начиная с 2000-х годов было предложено множество альтернативных моделей, которые учитывают разного рода зависимости внутри мотива. Но несмотря на это альтернативные модели не получили широкого распространения, а авторы работ ставили себе цель сравнить точность моделей, а не понять какие сайты находят разные модели. Цуканов А.В. ставит в своём исследовании целью не только сравнение альтернативных моделей с традиционной PWM по точности предсказания сайтов, а также пытается дать ответ на вопрос о вкладах методологически разных моделей в общее разнообразие наблюдаемых в природе ССТФ. При этом большой интерес представляют результаты применения методологически разных моделей мотива с учётом классификации ТФ по структуре ДНК-связывающего домена (ДСД), а также результаты по анализу терминов генной онтологии, которые позволяют оценить функциональный вклад сайтов, предсказанных разными моделями.

Структура диссертации

Диссертационная работа Цуканова А.В. изложена на 204 страницах, хорошо иллюстрирована, содержит 46 рисунков и 16 таблиц. Работа имеет стандартную структуру: оглавление, список используемых сокращений, введение, литературный обзор, методы, результаты и обсуждение, заключение, выводы, список литературы, два приложения. Материал изложен последовательно, написан ясно.

Во введении обосновывается актуальность исследования, формулируется цель и задачи, научная новизна и практическая ценность работы. Перечисляются положения,

выносимые на защиту, апробация работы, личный вклад автора и публикации по теме работы

В литературном обзоре даётся представление о том какую роль играют ТФ в регулировании транскрипции генов. Детально разбирается структура ТФ и описываются основные домены, которые могут входить в состав белка. Большое внимание уделяется ключевому домену ТФ – ДНК-связывающему домену (ДСД), приводится самая известная на сегодняшний день классификация ТФ основанная на структуре ДСД и описываются все суперклассы, входящие в её состав. Даётся представление об основных этапах механизма взаимодействия ТФ с ДНК, как ТФ находит свой сайт и какие есть особенности у ССТФ. Автор даёт описание как традиционных моделей (консенсус и PWM), так и альтернативных моделей (BaMM, InMoDe, SiteGA и др.). В последнем разделе данной главы даётся общее представление о том, что такое ChIP-seq эксперимент и описываются ключевые этапы биоинформатической обработки ChIP-seq эксперимента. В завершении главы приводится обоснование необходимости сочетания методологически разных *de novo* методов поиска мотивов.

В методах автор работы приводит источники ChIP-seq данных, а именно из каких баз данных брались ChIP-seq пики в данной работе. Далее идёт общая схема с последующим подробным описанием всех этапов работы программного комплекса MultiDeNa, который разработал автор, для анализа ChIP-seq данных с помощью методологически разных моделей мотива. Автор подробно описывает метод по оценке точности распознавания разных моделей. Приводятся описание программ, которые применялись для оценки сходства мотивов, а также для аннотации пиков и проведения анализа генной онтологии (ГО). В конце упоминаются основные инструменты, используемые для статистического анализа и визуализации.

Первый раздел главы “Результаты и обсуждение” посвящен применению программного комплекса MultiDeNa на небольшой выборке данных только для одного ТФ FOXA2. На этой выборке показывается перспективность применения разработанного автором конвейера, и так же даётся характеристика моделям относительно эффективности работы с ССТФ данного ТФ.

Второй и третьи раздела главы “Результаты и обсуждение” посвящены анализу двух выборок данных для *A. thliana* и *M. musculus*, соответственно, где автор сосредоточился на применении трёх методологически разных моделей мотива: PWM, BaMM и SiteGA. Автор сравнивает точности моделей, как качественно по визуализации ROC кривой, так и используя показатель pAUC, который является модификацией стандартной метрики AUC. В результате было показано, что модель BaMM превосходит модель PWM по показателю pAUC, при этом данный результат не зависит от класса ТФ по ДСД. Модель SiteGA может составлять конкуренцию PWM по точности для некоторых классов ТФ. Анализ сравнения функций распознавания моделей показал, что для пары моделей BaMM/PWM функции распознавания положительно коррелируют, а для пар PWM/SiteGA и BaMM/SiteGA корреляции нет. На основании этого утверждается, что модель BaMM имеет много общего с моделью PWM. Это хорошо согласуется с реализацией модели BaMM, которая расширяет модель PWM. С другой стороны, модель SiteGA, по-видимому, должна чаще выявлять сайты, которые другие модели не видят, поскольку она иначе оценивает аффинность ССТФ.

Проведен анализ совместной встречаемости мотивов разных моделей в пиках ChIP-seq. Показано, что альтернативные модели в значительной степени могут увеличивать долю пиков, содержащих предсказанные сайты в дополнение к результатам PWM, но, что более важно это увеличение существенно зависит от класса ТФ по ДСД. Полученные автором результаты для ТФ *A. thalina* и ТФ *M. musculus* согласуются между собой. Наибольшие

вклады альтернативных моделей у *A. thalina* и *M. musculus* наблюдается для классов ТФ Basic bHLH и bZIP. Это может быть связано с тем, что ТФ данных классов всегда димеризуются, а от состава димеров может зависеть структура мотива. Наименьший вклад альтернативные модели показали для класса C2H2 zinc finger factors, ТФ которого имеют более длинные сайты относительно других классов, а также практически не димеризуются. Поэтому автором работы было сделано предположение, что гипотеза об аддитивности вкладов позиций нуклеотидов лучше всего работает именно для C2H2 zinc finger factors.

Последний анализ, который был проведен – это анализ терминов ГО. Благодаря ему автор показал, что часть генов может иметь в промоторах только сайты одной из моделей, которые при этом связаны со специфическими терминами ГО и не выявляются другими моделями. Модели так же способны выявлять сайты, которые находятся в промоторах генов объединенных общими терминами ГО. Именно на общих терминах, было показано, что модель SiteGA имеет большее обогащение для терминов ГО по сравнению с моделями PWM и BaMM.

Замечания и рекомендации

В работе показана вариабельность вклада альтернативных моделей при распознавании разных ТФ как у *A. thaliana,* так и у *M. musculus*. Высказано предположение, что такая вариабельность обусловлена разной структурой ДНК–связываюших доменов, которое, однако, ничем не подкреплено и для наблюдаемой вариабельности можно предложить другие объяснения, например, непрямое действие ТФ, их низкую аффинность и т. п. Исходя из гипотезы об обусловленности вариабельности структурой ДНК–связывающего домена, диссертант приводит объяснения для такой зависимости только для двух классов ТФ – C2H2 zinc finger и Basic helix-loop-helix. Как можно объяснить увеличение доли вклада в распознавание от альтернативных моделей для других классов исследованных ТФ?

При картировании ССТФ на промоторы генов использвовались предсказания моделей на среднем пороге точности для ожидаемой частоты распознавания. Однко, при таком пороге точность модели SiteGA значительно ниже точности других моделей. Учитывая это, насколько разумно использовать эту модель для сравнительного анализа промоторов генов и не являются ли гены с уникальными промоторами, обнаруженные с использованием этой модели, ложно положительными предсказаниями?

Результаты работы суммированы в трех публикациях. Какая часть работы нашла отражение в публикации Жимулёв И.Ф., Ватолина Т.Ю., Левицкий В.Г., Колесникова Т.Д., Цуканов А.В. Развитие идеи Н.К. Кольцова о генетической организации междисков политенных хромосом *Drosophila melanogaster*. Онтогенез. 2023; 54(2), 172–175 ?

Также, как в тексте диссертации, так и в автореферате имеются многочисленные опечатки, неправильные согласования падежных окончаний прилагательного и сущиствительного и т. п.

Заключение

Несмотря на высказанные замечания, диссертация Цуканова Антона Витальевича является полноценной, законченной научно-квалификационной работой, посвященной важной проблеме – поиску сайтов связывания транскрипционных факторов с помощью методологически разных моделей. Таким образом, диссертация Цуканова Антона Витальевича полностью соответствует требованиям пп. 9 – 14 «Положения о присуждении

ученых степеней», утвержденного Постановлением Правительства Российской Федерации №842 от 24.09.2013, а её автор заслуживает присуждения учёной степени кандидата биологических наук по специальности «1.5.8. — математическая биология, биоинформатика».

Дата: 29 января 2024 г.

Официальный оппонент

Самсонова Мария Георгиевна

доктор биологических наук, профессор Высшей школы прикладной математики и вычислительной физики, заведующая научно-исследовательской лабораторией «Математическая биология и биоинформатика», ФГАОУ ВО «Санкт-Петербургский политехнический университет Петра Великого»

г. Санкт-Петербург

Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого»

Адрес: 29, Политехническая ул., Санкт–Петербург, 195251 Россия

Телефон: +7 812 290-9645

E-mail: m.samsonova@spbstu.ru