ОТЗЫВ
официального оппонента
на диссертацию
на соискание ученой степени кандидата биологических наук
Пензара Дмитрия Дмитриевича
на тему: «Вычислительное предсказание эффектов мутаций в регуляторных районах генов»
по специальности 1.5.8 «Математическая биология, биоинформатика»

Диссертационная работа Пензара Дмитрия Дмитриевича посвящена **актуальной теме** - вычислительному предсказанию эффектов мутаций в регуляторных последовательностях. Даже наиболее современные модели не справляются с оценкой вклада малых изменений, таких как однонуклеотидные варианты, в регуляцию экспрессии генов. Для решения этой задачи, автором разработан **новый** вычислительный метод на основе глубокого обучения, существенно превосходящий существующие аналоги. **Достоверность научных выводов и рекомендаций, сформулированных в диссертации,** подтверждается тем, что по результатам исследования опубликовано 6 печатных работ, в том числе 6 статей в рецензируемых научных журналах, индексируемых в WoS и Scopus. **Положения, выносимые на защиту, являются полностью обоснованными**, поскольку методология исследования отвечает современным принципам, особое внимание уделяется проблеме переобучения и утечки данных, точность моделей проверялась на результатах независимых экспериментов.

Диссертационная работа состоит из титульного листа, оглавления, списка сокращений и условных обозначений, введения, обзора литературы, материалов и методов, результатов, заключения, выводов, списка литературы, списка публикация по теме диссертации и приложений. Работа изложена на 166 страницах, иллюстрирована 55 рисунками, 6 таблицами и 1 приложением. К достоинствам работы можно отнести обширный обзор литературы. Список литературы состоит из 334 источников.

Тем не менее, есть ряд вопросов и замечаний к работе:

- На рис. 31 (B,C) отсутствуют подписи по осям. В легенде указано, что "На графике B изображены ROC-кривые, на графике C – PR-кривые", поэтому специалисты могут догадаться, что отображено по осям. Тем не менее, подписи стоило бы привести обязательно.

- Смущают термины на рис. 32 (E,F): "мотивные признаки", "ДНКазное покрытие". Они используются в русскоязычной литературе?

- На рис. 32 (C,D) стоило бы обозначить отдельные (тонкие) линии на панелях цветами и привести легенду на картинке. Одна линия соответствует одной хромосоме, верно? Интересно, что несколько линий расположены ниже диагонали на панели D. Интересно, что это за хромосомы? Почему так получилось? Это самые длинные хромосомы, удаление которых так сильно ухудшает предсказание?

- На рис. 32 (C,D) приведены значения AUC для усреднения. Но хорошо бы видеть полную картину - в каком диапазоне изменяется AUC в каждом случае? При удалении какой хромосомы получаются самые плохие значения, и насколько (численно) они плохи?

- В подписи к рис. 32 указано: "... для 3 транскрипционных факторов и типов клеток с наибольшим количеством ASB." Что значит ASB? Это сокращение не вводилось ранее.

- В диссертации присутствует достаточно большое количество опечаток, англицизмов и жаргонизмов. Например: "применение этих моделей применение" (стр.87), "для каждого группы" (стр. 112), "задизайненные" (стр. 116), "отдельные выборки фиксированных размер" (рис. 38), "группы активности с номерам" (стр. 125), и многие другие.

- Вызывает вопрос валидность использование теста зависимых корреляций Сильвера (рис. 39 C,D). Я не была знакома с этим тестом ранее, и поиск по литературе показывает, что этот тест редко используется. Хорошо, что автор приводит ссылку на работу, где вводится этот тест. В этой работе указано, что сила теста зависит от количества наблюдений. На рис. 39 C,D, автор генерирует 10000 наблюдений искусственно, методом бутстрепа. В таких случаях некорректно использовать статистические тесты, результат которых зависит от количества наблюдений. Ведь если бы автор использовал 1000000 итераций бутстрепа вместо 10000, полученное p-value было бы ещё ниже. Таким образом, увеличивая количество итераций, можно получить какое угодно p-value. В подобных случаях обычно используют пермутационный тест.

- Такое же замечание к рис. 42.

- В легенде к рис. 42 указана неверная ссылка - номер 32 вместо 332.

- Похожий вопрос возникает к рис. 44, на котором представлен другой анализ, но проблема та же: автор получает распределения путем случайной выборки 10000 промоторов, а затем использует тест Колмогорова-Смирнова для оценки статистической значимости отличий этих распределений. Этот тест очень чувствителен к количеству наблюдений. Если взять выборку размером 1000000, получатся ещё более значимые p-values.

- Аналогичная проблема на рис. 50A и рис. 53: t-test чувствителен к количеству наблюдений. Можно было бы использовать 20 фолдов кросс-валидации вместо 10, и p-value получились бы меньше.

Вместе с тем, отличия, показанные всех упомянутых рисунках, настолько драматические, что, несмотря на некорректный выбор статистических тестов, я не сомневаюсь в корректности сделанных на их основе выводов. Уверена, что применение более подходящих статистических процедур привело бы к получению принципиально таких же результатов.

Поэтому указанные замечания не умаляют значимости диссертационного исследования. Диссертация отвечает требованиям, установленным Московским государственным университетом имени М.В.Ломоносова к работам подобного рода. Содержание диссертации полностью соответствует специальности 1.5.8 «Математическая биология, биоинформатика» (по биологическим наукам), а именно следующим ее направлениям «Компьютерная системная биология», «Разработка и применение новых вычислительных алгоритмов для анализа экспериментальных данных в биологии и медицине», «Разработка и применение методов машинного обучения и искусственного интеллекта для анализа и прогнозирования свойств биологических объектов на основе анализа больших биомедицинских данных», а также критериям, определенным пп. 2.1-2.5 Положения о присуждении ученых степеней в Московском государственном университете имени М.В.Ломоносова, а также оформлена согласно требованиям Положения о совете по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук Московского государственного университета имени М.В.Ломоносова.

Таким образом, соискатель Пензар Дмитрий Дмитриевич заслуживает присуждения ученой степени кандидата биологических наук по специальности 1.5.8 «Математическая биология, биоинформатика».

Официальный оппонент:

доктор биологических наук
доцент Центра молекулярной и клеточной биологии Автономной некоммерческой образовательной организации высшего образования «Сколковский институт науки и технологий»

Храмеева Екатерина Евгеньевна

Контактные данные:

тел.: +7(495) 280-14-81 доб. 3413, e-mail: e.khrameeva@skoltech.ru
Специальность, по которой официальным оппонентом защищена диссертация:
1.5.8 – Математическая биология, биоинформатика

Подпись ... удостоверяю

Руководитель отде...
кадрового админис...

Адрес места работы:

121205, г. Москва, Большой бульвар д.30, стр.1

Автономная некоммерческая образовательная организация высшего образования «Сколковский институт технологий», Центр молекулярной и клеточной биологии

Тел.: +7(495) 280-14-81 ; e-mail: inbox@skoltech.ru