Федеральное агентство научных организаций (ФАНО России)
Федеральное государственное бюджетное учреждение науки
ИНСТИТУТ БИООРГАНИЧЕСКОЙ ХИМИИ
им. академиков М.М. Шемякина и Ю.А. Овчинникова
Российской академии наук
(ИБХ РАН)

ул. Миклухо-Маклая, 16/10, ГСП-7, Москва, 117997. Для телеграмм: Москва В-437, Биоорганика
телефон: (495) 335-01-00 (канц.), факс: (495) 335-08-12, E-mail: *office@ibch.ru*, *www.ibch.ru*
ОКПО 02699487 ОГРН 1037739009110 ИНН/КПП 7728045419/772801001

ОТЗЫВ

на диссертационную работу на соискание учёной степени

кандидата физико-математических наук

Сафоновой Яны Юрьевны

«ИСПОЛЬЗОВАНИЕ ГРАФОВЫХ МОДЕЛЕЙ ДЛЯ БИОИНФОРМАТИЧЕСКОГО АНАЛИЗА ГИПЕРВАРИАБЕЛЬНЫХ БИОЛОГИЧЕСКИХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ»

По специальности 05.13.18 — «Математическое моделирование, численные методы и комплексы программ»

Автором разработано несколько программных пакетов для анализа данных массированного секвенирования: dipSPAdes (сборка полиморфных диплоидных геномов), IgRepertoireConstructor (сборка репертуаров иммуноглобулинов) и IgSimulator (симулятор данных секвенирования репертуаров антител).

Рецензент занимается изучением репертуаров рецепторов Т и В лимфоцитов, и соответственно уделил основное внимание последним двум разработкам.

Высоко-производительное секвенирование репертуаров рецепторов иммунных клеток быстро меняет ландшафт фундаментальных и прикладных исследований в современной иммунологии, и все плотнее подходит к широкой клинической практике (в частности, в области иммунотерапии онкологических заболеваний).

В этой быстро развивающейся и высоко-конкурентной области, автору удалось внести очень достойный вклад. Свежий взгляд на задачу построения репертуаров полноразмерных вариабельных фрагментов антител из данных массированного секвенирования позволил автору предложить новые эффективные алгоритмы, способные помочь в достижении оптимального результата: коррекции накопленных ошибок ПЦР и секвенирования при сохранении нативного разнообразия гипермутирующих подвариантов иммуноглобулинов в

образце. Верификация эффективности алгоритмов анализа таких данных требует конструирования искусственных репертуаров, максимально близко имитирующих реальную сложность образцов, включая накопленные ошибки ПЦР и секвенирования, типы и паттерны распределение которых также могут иметь сложную природу. В этом направлении автором тоже была проделана весьма продуктивная работа, и дальнейшее совершенствование алгоритмов симуляции данных по иммунным репертуарам и накапливаемых ошибок остается актуальной темой. Работы автора хорошо опубликованы и востребованы научным сообществом.

Предложенная работа соответствует выбранной специальности: представление данных секвенирования в виде строк, которые потом трансформируются в граф де Брюйна (в случае задачи сборки генома) или граф Хэмминга (в случае анализа иммунных рецепторов), соответствует математическому моделированию. Несмотря на то, что указанные модели уже применялись в биоинформатике новым, автор предложил оригинальные модификации такого применения. В качестве численных методов были предложены новые алгоритмы упрощения графа де Брюйна и поиска плотных подграфов в графе Хэмминга. Все предложенные алгоритмы доступны в виде программных комплексов с открытым кодом dipSPAdes, IgRepertoireConstructor и IgSimulator.

Таким образом, у рецензента не имеется никаких сомнений в том, что автор заслуживает присуждения искомой степени.

Тем не менее, рецензент хотел бы отметить ряд неточностей и незначительных но заметных недостатков, в надежде на то, что в дальнейшем эти замечания помогут автору повысить уровень своей работы.

Во введении, цитируя автора: «Предложенные алгоритмы стали первыми доступными инструментами для решения поставленных задач». Вероятно, это не совсем точно. Доступные инструменты для построения репертуаров рецепторов иммунных клеток из данных иммуносеквенирования, а также для симуляции иммунных репертуаров и данных иммуносеквенирования существовали и до настоящей работы, и часть ссылок на эти работы приведена автором.

Описывая методы секвенирования нового поколения, автор умалчивает о последних разработках, таких как Nanopore (https://nanoporetech.com/) и PacBio (http://www.pacb.com/), уже активно применяющихся на практике.

Утверждение что массированное секвенирование позволяет провести «полноразмерное сканирование … гипервариабельных последовательностей *циркулирующих* антител» неточно, так как секвенирование ДНК либо РНК образца В лимфоцитов, который может включать в

себя наивные В лимфоциты и клетки памяти не эквивалентно набору циркулирующих антител, продуцируемых плазматическими В клетками, преимущественно локализованными в костном мозге. Более точная формулировка: «репертуар В-клеточных рецепторов», либо «репертуар иммуноглобулинов», без указания на то что эти иммуноглобулины циркулирующие, так как в большинстве случаев это неверно.

Утверждение, что «для того, чтобы иметь возможность определять как антитела взаимодействуют с определенными потенциально вредоносными агентами или антигенами необходимо провести анализ, использующий комбинацию СНП и МС технологий» неточно. Возможно, комбинация масс-спектрометрии и секвенирования нового поколения поможет точнее определять клональный состав продуцируемых циркулирующих антител. Однако от задачи определения того как именно эти антитела взаимодействуют с антигенами эти методы очень далеки. Кроме того не вполне ясно, из чего следует что все антигены интереса непременно вредоносны. В качестве контр-примера, задача идентификации вредоносных антител, атакующих нормальные антигены организма при аутоиммунных заболеваниях также актуальна.

Утверждение что «В некоторых случаях адаптивная система не в состоянии идентифицировать инородные клетки, что может привести к таким недугам как сенная лихорадка, астма и другие виды аллергий» сделано видимо на очень скорую руку.

Утверждение что «Антитела с большой кратностью чаще мутируют, создавая новые антитела отличающиеся от исходных только небольших набором отличий» спорно. Мы чаще можем наблюдать достоверные гипермутирующие подварианты когда речь идет о крупных клональных экспансиях, но это скорее результат ограниченной чувствительности доступных методов. «Крупность» клонов может также коррелировать с присутствием в образце плазматических В клеток, например через 8 дней после иммунизации, в периферической крови. Ввиду высокого уровня экспрессии мРНК иммуноглобулинов такими клетками их клоны могут выглядеть весьма крупными при получении кДНК библиотек (оставаясь относительно «мелкими» в терминах числа клеток). Появление таких клеток может быть сопряжено с процессом интенсивной гипермутации, это также верно. Однако и «мелкие» клоны могут при этом накапливать гипермутации, и возможно с не меньшей частотой.

Цитируя автора: «Кратность отдельных антител в репертуаре после вакцинации, состоящем из ≈ 2.3 миллионов антител, варьируется от 1 до ≈ 33,000. При этом, самые представленные из них могут составлять всего лишь ≈ 1% всех прочтений.» Репертуаре каких именно В клеток? Взятых откуда? Анализируемых каким методом – по крайней мере – на уровне ДНК или РНК? Здесь и в целом по ходу работы отсутствие необходимых деталей иногда не позволяет читателю разобраться в сути описываемых процессов.

Касательно замеченных опечаток – их относительно немного. Однако рецензента, в его собственной практике, всегда умиляло количество грамматических ошибок, допускаемых непосредственно в ходе рассуждении об ошибках секвенирования. В настоящей работе автор также остается верна тенденции:

«среднЯЯ уровень ошибок», «предоставлЯЯющий … референс для анализа качества», «сгруппировать в кластерА», «для всех пары строк», и проч.

В целом же, цитируя автора («биоинформатика, огромная междисциплинарная наука, объединяющая компьютерные науки, математику, статистику и разработку программного обеспечения для решения разнообразных биологических задач») хочется добавить, что степень рациональности биоинформатических подходов неотрывно связана с глубоким пониманием природы анализируемых данных, как с точки зрения биологии объекта, так и с точки зрения молекулярных технологий, задействованных для получения этих данных. Хочется пожелать автору в дальнейшей работе проводить больше времени в беседах с «мокрыми» биологами, непосредственно производящими данные, и постараться максимально проникаться сутью и важными деталями по их сторону работы.

Чудаков Дмитрий Михайлович, д.б.н.,

зав. лаб. Геномики Адаптивного Иммунитета ИБХ РАН

30 марта 2017г.