МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
(МИНОБРНАУКИ 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

«УТВЕРЖДАЮ»
……
Директор Федерального государственного
бюджетного учреждения науки
Институт биоорганической химии им. академиков
М.М. Шемякина и Ю.А. Овчинникова
Российской академии наук (ИБХ РАН)

академик А.Г. ГАБИБОВ

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

на диссертационную работу Карасева Дмитрия Алексеевича
на тему: «Разработка метода протеохемометрики для предсказания взаимодействий белков и лигандов на основе их локального сходства», представленную на соискание ученой степени кандидата биологических наук по специальности 1.5.8. - «Математическая биология, биоинформатика»

В настоящее время компьютерные технологии конструирования биологически активных соединений – CADD, Computer-Aided Drug Design, - включая прототипы новых лекарств, стали неотъемлемой частью биомедицины и биотехнологии. Учитывая большие сложности получения моделей пространственной структуры молекулярных клеточных мишеней (как правило, белков), перспективных с фармакологической точки зрения, большинство работ в этой области по-прежнему связано с использованием методов, основанных на анализе данных по структуре и активности лигандов (Ligand-Based Drug Design. LBDD). Среди последних наиболее часто применяют вычислительные подходы, направленные на выявление количественных взаимосвязей «структура-активность» - Quantitative Structure-Activity

Relationships (QSAR) - и их различные модификации, в том числе и самые современные, задействующие элементы искусственного интеллекта – машинное обучение и смежные технологии. Хорошо известно, что эффективность указанных приложений потенциально может быть повышена за счет добавления к информации о свойствах лигандов сведений о способных взаимодействовать с ними белках-мишенях. Это сравнительно новое (относительно классических QSAR-моделей) направление получило название «протеохемометрика» (Proteochemometrics, PCM). Исследования в данной области активно ведутся около 20 лет, но предложенные модели и специализированные базы данных все еще не являются универсальными и общепринятыми, поэтому использование PCM-подходов в каждом конкретном случае требует отдельного рассмотрения. Это сдерживает необходимое на современном уровне CADD-приложений «потоковое» применение PCM на базе больших баз данных и множественных сценариев анализа взаимодействий белок-лиганд. Поэтому назревшей важной задачей является создание унифицированных методов PCM и их автоматизированная реализация, например, в виде удобных веб-сервисов и пр. Решению этих задач и посвящена диссертационная работа Карасева Д.А., что определяет ее актуальность и значимость для развития данной предметной области.

Диссертация изложена на 101 странице и включает: Введение, Обзор литературы (Глава 1), Раздел «Материалы и методы» (Глава 2), описание результатов и их обсуждение (Глава 3), Заключение, Выводы, Список работ, опубликованных по теме диссертации, Список цитируемой литературы из 155 источников. Диссертация содержит 23 рисунка и 7 таблиц.

Во Введении обоснована актуальность темы диссертации, сформулированы цель, задачи, методология и методы исследования, охарактеризованы научная новизна, теоретическая и практическая значимость работы. Кроме того, здесь же показан личный вклад автора, описаны методология и методы исследования, представлены основные положения, выносимые на защиту, приведены сведения о достоверности полученных результатов, об апробации работы и о публикациях автора по теме диссертации, а также кратко описана структура работы. Глава 1 представляет собой литературный обзор, включающий 8 разделов, которые затрагивают следующие основные тематики: 1) Компьютерные методы, используемые при разработке лекарственных соединений (раздел 1.1); 2) Анализ взаимосвязи структура-активность (Q)SAR (раздел 1.2); 3) Построение прогностических моделей на основе структур низкомолекулярных соединений и аминокислотных последовательностей белков-мишеней. Протеохемометрическое моделирование (раздел 1.3); 4) Применение PCM к решению биомедицинских задач. Мишени, исследуемые при PCM-моделировании. Данные для построения моделей (раздел 1.4); 5) Подходы к описанию структур лигандов и белков-мишеней

(раздел 1.5); 6) Методы машинного обучения, применяемые в PCM-моделировании (раздел 1.6); 7) Процедура валидации PCM-моделей (раздел 1.7); 8) Заключение (раздел 1.8).

Из обзора литературы следует, что выбранная автором область исследования является актуальной и перспективной, хотя и развивается не такими быстрыми темпами, как того требуют современные практические задачи в области компьютерного конструирования лекарств с использованием LBDD-подходов.

На основании анализа литературных данных сделан вывод о том, что PCM-моделирование имеет большой потенциал применимости в предметной области, уже существуют надежные компьютерные методы, успешно примененные для разнообразных биологических мишеней и их лигандов, а также существуют обширные базы обучающих данных для разработки предсказательных моделей. Наиболее распространенными подходами остаются (Q)SAR-исследования, которые продемонстрировали свою эффективность и при решении задач PCM-моделирования. В рамках этих методов накоплен богатый и находящийся в свободном доступе инструментарий для генерации основных типов дескрипторов для химических соединений. Это позволяет быстро проводить генерацию наиболее популярных молекулярных дескрипторов. Одним из важных моментов является вывод автора о начале широкого применения аминокислотных последовательностей белков-мишеней для задач прогнозирования с помощью машинного обучения. Вместе с тем, существуют методологические проблемы, связанные с отсутствием унифицированных методик работы с представлениями белков в ходе PCM-анализа, с разнородным характером соответствующих выборок данных и т.д. Это позволило автору заключить, что назрела необходимость в создании нового, более универсального метода с широкой областью применимости, который бы позволил анализировать взаимодействия лигандов с белками различных таксономических групп, т.е. сформулирована основная цель диссертационной работы. Кроме того, предложены требования к создаваемому PCM-подходу и к его валидации, которая должна проводиться с учетом возможных сценариев, реализующихся при компьютерной оценке белок-лигандных взаимодействий. Представленный материал непосредственно связан с темой работы, содержит свежую (вплоть до источников 2023 года) информацию. Эти сведения будут полезны как начинающим исследователям, так и специалистам в данной области.

Изложенный в обзоре литературы материал подводит читателя к логически обоснованным выводам относительно целей диссертационной работы и к формулировке конкретных задач, которые необходимо решить. Непосредственно за литературным обзором следует Глава 2, посвященная описанию использованных автором в работе вычислительных методов и подходов. Изложение технических деталей постановки и проведения компьютерных экспериментов – неотъемлемая и крайне важная часть любой работы по хемоинформатике, тем

более, если речь идет о таких сложных многокомпонентных системах, какими являются комплексы белок-лиганд. Материал изложен четко и дает подробное представление об использованных подходах, что позволяет заинтересованным исследователям самостоятельно воспроизвести вычисления, проделанные автором. Это крайне важно для специалистов, занимающихся разработкой новых методических подходов и их программной реализацией.

Представленные в диссертационной работе научные результаты изложены в Главе 3. Они относятся к подготовке тестовых и обучающих данных, анализу их областей применимости, обучающих данных. Кроме того, большое значение уделено оценке точности прогноза в различных сценариях разработанного автором вычислительного подхода – при использовании прогностического и позиционного режимов. В заключительном разделе главы описан созданный автором веб-сервис для предсказания белок-лигандных взаимодействий.

В ходе выполнения работы автором были получены следующие основные научные результаты:

- Разработан оригинальный протокол сбора наиболее надежных данных из доступных информационных источников, проведена их унификации для создания классификационных протеохемометрических моделей.
- Показано, что разработанная методика позволяет предсказывать взаимодействия белок-лиганд в соответствии с тремя сценариями, использующими в качестве входных данных структуры лигандов (сценарий 1) и аминокислотные последовательности (сценарий 2), а также их комбинацию (сценарий 3).
- Приведены доказательства того, что созданный подход обеспечивает высокую точность предсказаний в широкой области применимости, которая включает наборы белков-мишеней с разной степенью структурно-функционального сходства. Среди них – рецепторы из семейства GPCR, протеинкиназы, ионные каналы и др. При моделировании ситуации с неполными обучающими данными показана высокая эффективность созданного PCM-подхода.
- Разработанная методика реализована в виде программного комплекса, свободно доступного в сети Интернет. Указанный веб-сервис (http://way2drug.com/proteochemometrics/) предоставляет широкому кругу исследователей возможность самостоятельного проведения протеохемометрического анализа.

На мой взгляд, одной из наиболее интересных и перспективных как с фундаментальной, так и прикладной точек зрения является предложенный автором алгоритм прогноза взаимодействия белок-лиганд в позиционном режиме, без доступной *a priori* из баз данных и литературы информации о каждом из партнеров. Именно подобный сценарий, предполагаю,

будет наиболее востребован в будущем при проведении как поисковых исследований, так и ориентированных на биомедицинское применение научных проектов. Сильной стороной работы Карасева Д.А. считаю четкое понимание проблем и «узких мест» в области протеохемометрики, на решении которых необходимо было сосредоточить внимание. Кроме того, автору удалось умелое сочетание методологической, но очень важной работы по созданию выверенных, отвечающих единым критериям источников информации о белках-мишенях с разработкой новых алгоритмов прогноза межмолекулярных взаимодействий. Учитывая вышесказанное, считаю, что соискатель продемонстрировал научный, а не только технологический подход к компьютерному моделированию. Это говорит о хороших перспективах результатов диссертационной работы и о соискателе как о вполне самостоятельном исследователе.

Считаю, что цель и задачи, сформулированные в работе, достигнуты в ходе её выполнения. Автором самостоятельно получены и описаны оригинальные результаты, которые были доложены на российских и международных специализированных научных встречах. Выводы и заключения, представленные в работе Карасева Д.А., логически обоснованы и убедительно показывают новизну и применимость полученных результатов.

По материалам диссертации считаю необходимым сформулировать ряд вопросов и замечаний.

1. Наиболее сложным с точки зрения достижения корректного прогноза, на мой взгляд, является разработанный автором позиционный режим (сценарий 3), поэтому оценка точности предсказания в данном случае особенно важна, причем наиболее эффективно ее проводить на естественных последовательностях белков и экспериментально установленных моделях пространственной структуры их комплексов с лигандами. Вместе с тем, для такой оценки были выбраны лишь два комплекса из обширной базы данных, причем оба они относятся к семейству протеинкиназ, хотя и далеко отстоящих друг от друга на филогенетическом древе (Раздел 3.5, стр. 68-69). При этом результаты предсказания, полученные для первой пары белок-лиганд, оказались довольно скромными – из 29 предсказанных остатков лишь 5 в действительности находятся в активном центре белка. Удаление из обучающей выборки ряда белков – близких гомологов целевой киназы - позволило улучшить результат, но ведь на практике, для новых объектов, подобной «подсказки» не будет. В связи с этим возникают следующие вопросы: 1) Почему проверку проводили лишь на двух белках, причем одного семейства, хотя сотни комплексов доступны для анализа? 2) Почему в обучающей выборке содержались высокогомологичные белки вообще и близкие гомологи целевых белков - в частности? Обычно стараются проводить обучение на невырожденных наборах белков, например, часто – с уровнем

попарной гомологии, не более 20%. 3) При выявлении остатков активного центра можно ли оценить, являются они лишь сближенными с лигандом или же непосредственно взаимодействуют с ним? 4) Насколько оптимальным (т.е. подтвержденным на этапе параметризации модели) является порог p = 0,001? Считаю, что для использования в дальнейшей практике необходимо иметь результаты проверки этого подхода на большом числе (~10^2 шт.) тщательно подобранных комплексов белок-лиганд разных классов (и мишеней, и лигандов), на основании которой можно было бы сделать статистически обоснованные выводы о точности метода.

2. Учитывая сказанное выше, считаю чересчур расплывчатым заключение автора о том, что «позиционный режим позволяет наглядно оценить идею идентификации локального соответствия между белками и лигандами, при этом получая интерпретируемые результаты» (стр. 73). Что значит: «локальное соответствие»? Геометрическая близость лиганда и определенных остатков активного центра? Характер межмолекулярных взаимодействий между ними? Кроме того, непонятно, как можно получать «неинтерпретируемые результаты», можно ли их вообще называть «результатами»?

3. Раздел 3.8: При анализе точности прогноза по сценариям 2 и 3 в зависимости от выбора параметра F - длины скользящего участка последовательности – высокая точность (> 85%) достигалась уже при F = 3, возрастая при дальнейшем увеличении длины окна. Насколько трипептиды способны выявить сходство в структурной организации и физико-химических свойствах белков, включая информацию о сайтах связывания лигандов? Ведь такие сайты формируются гораздо более протяженными участками белка, в т.ч. и сильно удаленными остатками с F >> 30?

4. Какова погрешность при расчете точности предсказания – приведенные значения даны с указанием второго знака после запятой, но насколько надежно отличается, например, результат, полученный с точностью 0,94, от такового со значением 0,92? Какие статистические критерии были использованы в подобных оценках? Это, в свою очередь, ставит следующий вопрос: практически во всех случаях прогностические модели демонстрируют очень высокую точность (> 80%), причем как созданные автором, так и известные из литературы. Имеет ли смысл подобная «борьба за малые поправки»?

Ознакомление с диссертационной работой вызывает и ряд вопросов, связанных с форматом изложения материала:

5. В обширном обзоре современного состояния проблемы (Глава 1) акцент часто сделан на перечислении методов, предложенных в различных работах, но не хватает результатов критического анализа автора достоинств/недостатков этих подходов, встреченных проблем и предлагаемых путей их решения. В частности, это относится к изложению литературных данных о разработке и применении технологий протеохемометрики (PCM) – именно той темы, которой и посвящена настоящая работа (Раздел 1.4).

К недостаткам работы относятся и некоторые погрешности оформления. Видимо, из-за ошибки форматирования некорректно отображается Таблица 4 (стр. 75) – в ней нет данных. При рецензировании была использована аналогичная таблица из автореферата (Табл. 1). Кроме того, не все методы, указанные в этой таблице, расшифрованы в тексте. Автор использует ряд неудачных, жаргонных и некорректных выражений, например: «исследование расшифрованного комплекса» (стр. 6.), «интегральные характеристики белка, такие как … автокорреляция» (стр. 27), «значение Z-шкалы» (стр. 31), «поверхностные взаимодействия» (стр. 34). Однако отмечу, что подобных случаев совсем немного, в отличие от большинства известных рецензенту диссертационных работ.

Вместе с тем, отмеченные недостатки не снижают моей высокой оценки работы Карасева Д.А. Высказанные замечания носят рекомендательный характер и служат для того, чтобы подчеркнуть сложность поставленной задачи, которая была успешно решена автором. Работа выполнена на высоком методическом уровне, содержит новые интересные научные данные, хорошо оформлена и легко читается. Полученные автором результаты, наряду с богатым справочным материалом, несомненно, будут полезны не только исследователям, занимающимся вычислительной медицинской химией, хемо- и биоинформатикой, рациональным конструированием лекарств, но и специалистам в других областях – молекулярной биофизике, структурной биологии, биоинженерии. Подобные лаборатории и группы существуют в ВУЗах и научно-исследовательских организациях – на ряде естественнонаучных факультетов МГУ им. М.В. Ломоносова, СПбГУ и др., в Институте химической биологии и фундаментальной медицины СО РАН, в Институте биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова РАН, в Институте цитологии и генетики СО РАН и др.

Заключение. Диссертационная работа на тему «Разработка метода протеохемометрики для предсказания взаимодействий белков и лигандов на основе их локального сходства» полностью соответствует критериям пп. 9-14 «Положения о порядке присуждения ученых

степеней», утвержденного постановлением Правительства РФ от 24 сентября 2013 года № 842 (с изменениями в редакции постановлений Правительства РФ от 21.04.2016 г. № 335, от 02.08.2016 г. № 748, от 28.08.2017 г. № 1024 и от 01.10.2018 г. № 1168), и представляет собой завершенную научно-квалификационную работу. Выносимые на защиту результаты опубликованы в виде пяти статей в реферируемых международных журналах, индексируемых в библиографической базе данных Web of Science, докладывались на российских и международных научных конференциях. Отмечу, что во всех статьях Карасев Д.А. является первым автором. Рукопись автореферата соответствует содержанию рассматриваемой диссертации, результатам и положениям, выносимым на защиту. Диссертационная работа Карасева Д.А. была рассмотрена на расширенном семинаре лаборатории моделирования биомолекулярных систем ИБХ РАН (Протокол № 1 от 02.10.2023 г.). Содержание диссертации соответствует паспорту специальности 1.5.8. – «математическая биология, биоинформатика» (по биологическим наукам), а ее автор Карасев Дмитрий Алексеевич заслуживает присуждения ученой степени кандидата биологических наук по специальности 1.5.8. – «Математическая биология, биоинформатика».

Главный научный сотрудник,
Заведующий лабораторией моделирования биомолекулярных систем
Федерального государственного бюджетного учреждения науки
Институт биоорганической химии
им. академиков М.М. Шемякина и Ю.А. Овчинникова
Российской академии наук (ИБХ РАН)

доктор физико-математических наук,
профессор Ефремов Роман Гербертович Р.Г. Ефремов

Адрес: ИБХ РАН, ул. Миклухо-Маклая, д. 16/10, г. Москва, 117997, Россия
Телефон: +7 903 743-16-56
Адрес электронной почты: r-efremov@yandex.ru

06 октября 2023 г.

Подпись проф. Р.Г. Ефремова удостоверяю

Ученый секретарь ИБХ РАН
д.ф.-м.н. В.А. Олейников