На правах рукописи

Карасев Дмитрий Алексеевич

Разработка метода протеохемометрики для предсказания взаимодействий белков и лигандов на основе их локального сходства

1.5.8. – Математическая биология, биоинформатика

Автореферат
диссертации на соискание учёной степени
кандидата биологических наук

Москва – 2023

Работа выполнена в Федеральном государственном бюджетном научном учреждении «Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича» (ИБМХ)

Научный руководитель: доктор биологических наук, профессор РАН

Лагунин Алексей Александрович

Официальные оппоненты: **Карягина-Жулина Анна Станиславовна**,
доктор биологических наук, профессор,
Федеральное государственное бюджетное учреждение «Федеральный научно-исследовательский центр эпидемиологии и микробиологии имени почетного академика Н.Ф. Гамалеи» Министерства здравоохранения Российской Федерации,
главный научный сотрудник

Орлов Юрий Львович,
доктор биологических наук, профессор РАН,
Федеральное государственное автономное образовательное учреждение высшего образования Первый Московский государственный медицинский университет имени И.М. Сеченова Министерства здравоохранения Российской Федерации (Сеченовский Университет), профессор

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова Российской академии наук

Защита состоится «26» октября 2023 г. в 11.00 часов на заседании диссертационного совета 24.1.172.01 (Д 001.010.01) при Федеральном государственном бюджетном научном учреждении «Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича» по адресу: 119121, Москва, ул. Погодинская, д. 10, стр. 8.

С диссертацией можно ознакомиться в библиотеке ИБМХ и на сайте www.ibmc.msk.ru.

Автореферат разослан _____ 2023 г.

Ученый секретарь Диссертационного совета
кандидат химических наук  Карпова Е.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы

В настоящее время компьютерные методы, основанные на анализе структуры лигандов (Ligand-Based Drug Design), широко используются при поиске и разработке новых лекарственных средств, что позволяет существенно повысить эффективность проводимых исследований (Раевский и др., 2006; Smith et al., 2021). Одним из наиболее эффективных подходов является (Q)SAR ((Quantitative) Structure–Activity Relationships), который подразумевает анализ взаимосвязи между структурой низкомолекулярного соединения и спектром его биологической активности. Для расширения области применимости в 2001 году было предложено наряду с информацией о структуре и биологической активности лигандов использовать данные о первичной структуре белков-мишеней (Lapinsh M. et al., 2001). Такой подход получил название «протеохемометрика». Было показано, что его применение позволяет повысить предсказательную способность по сравнению с классическими (Q)SAR моделями (Bongers B.J. et al, 2019). Из анализа литературы следует, что предложенные модели не являются универсальными; подходы к описанию фармакологических мишеней существенно отличаются (Bongers B.J. et al, 2019, Rifaioglu A.S. et al., 2019, Qiu T. Et al., 2017). Применение данных подходов к новым мишеням требует предварительной оптимизации имеющихся моделей или даже их создания *de novo* с учетом структурных особенностей мишеней и лигандов, что не всегда является тривиальной задачей и требует существенных временных затрат. Более того, «подстройка» модели под конкретную задачу может привести к снижению предсказательной способности вследствие «переобученности». Таким образом, актуальной проблемой является создание метода протеохемометрики, который был бы применим для широкого круга фармакологических мишеней.

Целью диссертационной работы является создание метода для широкомасштабного предсказания белок-лигандных взаимодействий на основе анализа локального сходства аминокислотных последовательностей белков и структур низкомолекулярных лигандов.

Для достижения цели исследования были поставлены и решены следующие **задачи**:

1. Сформировать обучающие выборки, содержащие информацию о структурах низкомолекулярных лигандов, аминокислотных последовательностях белков-мишеней и показателях аффинности для каждой пары «белок-лиганд».
2. Разработать метод для прогноза белок-лигандных взаимодействий на основе анализа локального сходства аминокислотных последовательностей белков-мишеней и структур низкомолекулярных лигандов.
3. Оценить эффективность разработанного метода при широкомасштабном прогнозе белок-лигандных взаимодействий на наборах данных, характеризующих взаимодействия лигандов с белками различных таксономических групп.
4. Реализовать веб-сервис для прогноза белок-лигандных взаимодействий на основе разработанного метода протеохемометрики.

Личный вклад автора

Автор самостоятельно провел поиск и анализ литературы по проблемной области, провел обобщение современных достижений в области протеохемометрики и сформулировал пути решения существующих проблем. Автор сформировал программный комплекс для прогноза белок-лигандных взаимодействий на основе нового протеохемометрического подхода, разработал методику сбора данных. Все расчёты, построение моделей и анализ полученных результатов выполнены лично автором.

Положения, выносимые на защиту:

1. Разработан оригинальный протокол сбора наиболее надежных данных из доступных информационных источников и их унификации для создания классификационных протеохемометрических моделей.
2. Разработанная и реализованная в виде программного комплекса методика позволяет предсказывать белок-лигандные взаимодействия в соответствии со сценариями, использующими в качестве входных данных структуры лигандов и аминокислотные последовательности. Сценарий протеохемометрики, при котором используются входные данные обоих типов, рассчитан на наиболее частую ситуацию, связанную с неполнотой обучающей информации.
3. Предложенный подход обеспечивает высокую точность предсказаний в широкой области применимости, которая включает наборы белков-мишеней с разной степенью

структурно-функционального сходства. При моделировании ситуации с неполными обучающими данными показана высокая эффективность разработанного нами протеохемометрического подхода.

4. Свободно доступный в сети Интернет веб-сервис предоставляет широкому кругу исследователей возможность проведения протеохемометрического анализа (http://way2drug.com/proteochemometrics/).

Научная новизна

Разработан оригинальный метод протеохемометрики, который позволяет прогнозировать белок-лигандные взаимодействия для различных групп белков-мишеней. Методика прогноза основана на поиске локальных соответствий между атомами низкомолекулярных лигандов и аминокислотными остатками белков-мишеней. При этом не требуется внесения изменений или оптимизации параметров при смене группы белков-мишеней, что является преимуществом в сравнении с существующими подходами. Метод показал эффективность при прогнозировании спектра лигандов на основе аминокислотной последовательности белка-мишени. Привлечение данных по структурному сходству лигандов позволяет предсказывать новые пары белок-лиганд в отсутствии сведений о спектрах взаимодействия для обоих компонентов.

Впервые разработан свободно доступный в сети Интернет веб-сервис, который предоставляет пользователям широкий спектр возможностей для компьютерной оценки белок-лигандных взаимодействий на основе протеохемометрики.

Научно-практическая значимость

При создании новых лекарственных средств предложенный метод позволяет отбирать соединения наиболее перспективные для экспериментального тестирования в отношении не только уже известных фармакологических мишеней, но и в отношении новых белков-мишеней, т.е. таких, для которых неизвестны низкомолекулярные лиганды. Прогнозирование возможно с использованием различных входных данных в зависимости от цели исследования. Входной информацией являются либо аминокислотные последовательности, либо структуры химических соединений, либо данные обоих типов. Метод не требует оптимизации для новых групп белков, что позволит исследователям оперативно осуществлять прогноз для новых мишеней и отвечать на новые вызовы. Свободно доступный в сети Интернет веб-сервис позволяет использовать разработанный инструмент широкому кругу исследователей.

Апробация работы

Основные положения диссертации были представлены на российских и международных конференциях и симпозиумах, включая: The 13th International Conference

on Bionifomatics of Genome Regulation and Structure/Systems Biology (BGRS\SB-2022), Новосибирск (Россия), 2022; VII Съезд биохимиков России и X Российский симпозиум "Белки и пептиды", Дагомыс (Россия), 2021; Международный форум: Биотехнология: состояние и перспективы развития, Москва (Россия), 2020; 9-ая Московская конференция по вычислительной молекулярной биологии (MCCMB'19), Москва (Россия), 2019; 8-ая Московская конференция по вычислительной молекулярной биологии MCCMB'17, Москва (Россия), 2017; 43rd FEBS Congress, Biochemistry Forever, Прага (Чехия), 2018; VIII российский симпозиум «белки и пептиды», Москва (Россия), 2017; The 10th International Conference on Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS\SB-2016), Новосибирск (Россия), 2016; XXIII Российский национальный конгресс «Человек и лекарство», Москва (Россия), 2016.

Публикации

По материалам диссертации опубликовано 14 работ в российских и международных научных изданиях, в том числе 5 статей в рецензируемых научных журналах, входящих в Core Collection Web of Science и перечень, рекомендованный ВАК, и 9 публикаций в трудах конференций.

Объём и структура диссертации

Диссертационная работа состоит из введения, обзора литературы, материалов и методов исследования, результатов и обсуждения, заключения, выводов и списка литературы, включающего 155 источников. Работа изложена на 101 странице, содержит 23 рисунка и 7 таблиц.

МАТЕРИАЛЫ И МЕТОДЫ

1. Алгоритмы и методы

Методы протеохемометрики предназначены для прогнозирования взаимодействий белков-мишеней с лигандами. В общем случае используемые для обучения данные можно представить в виде комбинированного пространства признаков, характеризующих как белки, так и лекарственно-подобные соединения. Это позволяет сопоставлять оба компонента тестируемой пары с данными обучающей выборки и на этой основе получать прогностическую оценку возможного связывания.

Разработанный подход подразумевает реализацию трех сценариев, при этом третий сценарий соответствует центральной задаче протеохемометрики, когда необходимо оценить связывание пары «белок-лиганд», когда не аннотированы оба компонента (рис. 1).

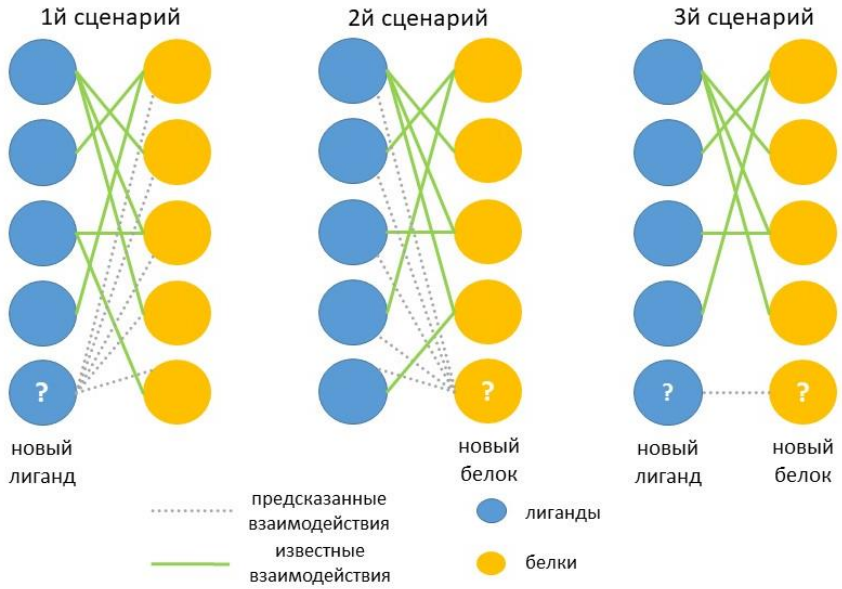

Рисунок 1. Сценарии, реализуемые при компьютерном прогнозе белок-лигандных взаимодействий.

Первый сценарий подразумевает предсказание спектра мишеней для нового лиганда путем сопоставления его структуры со структурами лигандов обучающей выборки. С этой целью в программном комплексе используется метод PASS (Prediction of Activity Spectra for Substances), успешно применяемый для прогноза биологической активности лекарственно-подобных соединений на протяжении тридцати лет (Поройков В.В. и др., 2019). В PASS структура лигандов представлена дескрипторами многоуровневых атомных окрестностей (MNA - Multilevel Neighborhoods of Atoms), которые служат входными данными для прогноза на основе наивного байесовского классификатора. Выходными данными являются величины *Pa* (вероятность того, что новый лиганд будет связываться с

белком-мишенью) и *Pi* (вероятность того, что связывания не будет). Если разность *Pa-Pi* больше нуля, то соединение признается активным в отношении рассматриваемой мишени.

Второй сценарий обеспечивает предсказание лигандов для нового белка-мишени путем сопоставления его аминокислотной последовательности с последовательностями белков обучающей выборки без учета структуры лигандов. Для этого используется оригинальный метод SPrOS (Specificity Projections on Sequence) (Karasev et al. 2020), основанный на локальном сопоставлении аминокислотных последовательностей.

Третий сценарий представляет собой центральную задачу протеохемометрики по оценке возможного связывания белка и лиганда, когда спектры взаимодействия обоих не установлены. В последнем случае входные данные должны представлять, как структуры лигандов, так и последовательности белков.

Схема прогноза при трех сценариях представлена на рисунке 2.

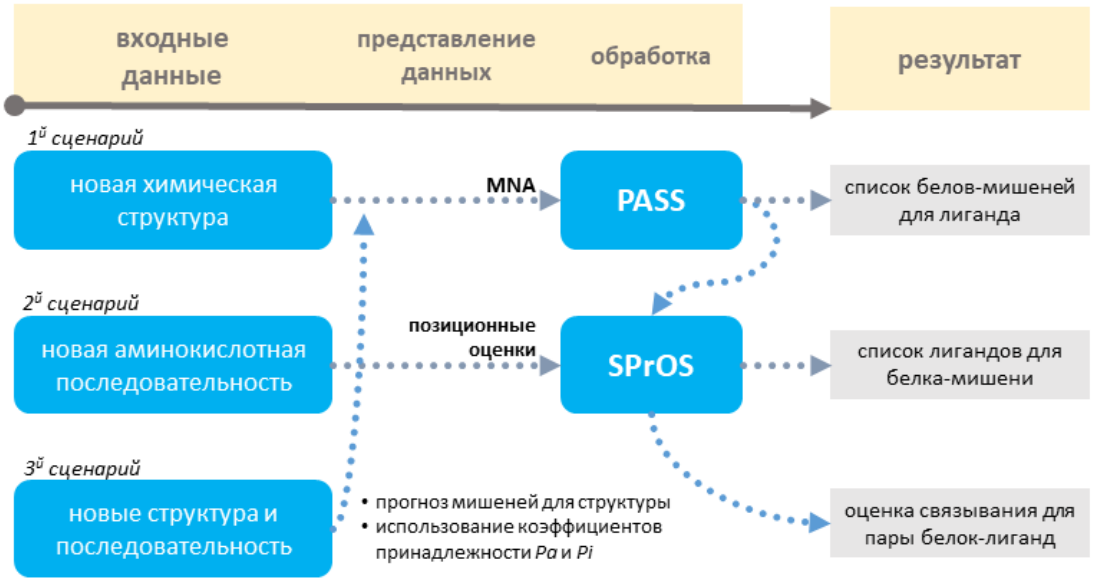

Рисунок 2. Схема прогноза с помощью разработанного нами метода протеохемометрики в зависимости от задачи и входных данных.

Для прогноза по второму и третьему сценариям наш комплекс включает оригинальный метод SPrOS.

Программа SPrOS может работать в двух режимах:

1) **Позиционный**, предусматривающий выявление аминокислотных остатков белка, значимых для связывания лиганда.
2) **Прогностический**, предусматривающий оценку взаимодействия белок-лиганд на основе позиционных оценок по второму или третьему сценарию.

Оба режима основаны на расчете оценок специфичности отдельных аминокислотных остатков белка по отношению к лиганду путем сопоставления тестовой

последовательности с белками обучающей выборки. Последние разбиваются на два класса - взаимодействующих (положительные примеры) и невзаимодействующих (отрицательные примеры) с данным лигандом. Разбиение производится на основе выбранного порога аффинности в соответствии с используемой методикой детекции взаимодействия. Принадлежность белка к числу положительных примеров задается коэффициентом *a*, а к числу отрицательных примеров коэффициентом *b*. При втором сценарии используются бинарные коэффициенты, которые определяют четкие классы лигандной специфичности путем присвоения каждому белку значений [*a*=1, *b*=0] либо [*a*=0, *b*=1]. При третьем сценарии нечеткие классы принадлежности определяются с помощью вещественных коэффициентов, которые рассчитываются на основе структурного сходства и принимают значения в пределах от 0 до 1.

В обоих режимах рассчитываются величины, оценивающие вклад аминокислотных остатков тестовой последовательности в формирование комплекса белок-лиганд.

При позиционном режиме в результате выполнения программы рассчитываются *p*-значения, которые показывают статистическую значимость позиционных оценок для отдельных аминокислотных остатков. Это осуществляется за счет случайного перемешивания последовательностей по классам лигандной специфичности, с последующим расчетом набора псевдослучайных оценок для каждой позиции. *P*-значения определяются на основе полученного распределения оценок. Валидация результатов проводилась путем их сопоставления с трехмерными структурами комплексов белок-лиганд.

При прогностическом режиме оценки отдельных аминокислотных позиций используются при вычислении общей оценки взаимодействия белок-лиганд. Последняя варьирует от -1 (вероятнее всего связывания нет) до +1 (вероятнее всего связывание есть).

При реализации второго сценария оценки взаимодействия рассчитываются с использованием бинарных коэффициентов.

При третьем сценарии алгоритм работает поэтапно. Программа PASS, следуя первому сценарию, предсказывает спектр возможных мишеней. Полученные вероятностные оценки *Pa* и *Pi* используются в программе SPrOS как нечеткие коэффициенты лигандной специфичности, отражающие принадлежность пары «белок-лиганд» к взаимодействующим и невзаимодействующим примерам.

Валидацию при тестировании в прогностическом режиме проводили на основе скользящего контроля с исключением по одному (Leave-one-out cross-validation). В этом случае каждая пара белок-лиганд последовательно исключается из обучающей выборки с последующим прогнозом ее взаимодействия. Полученные результаты использовались для

оценки точности прогноза, выраженных в значениях IAP (Invariant Accuracy Prediction). Значения IAP эквивалентны величинам AUC, получаемым при ROC-анализе. Все значения точности, приведенные ниже, соответствуют величинам IAP.

2. Формирование обучающих данных.

Для тестирования метода в позиционном режиме были сгенерированы наборы искусственных аминокислотных последовательностей путем симуляции эволюционного процесса. Каждый из наборов разбивался на группы, которые различались между собой специфичными «мутациями» (рис. 3).

В остальных случаях тестирование метода проводили на выборках, каждая из которых представляла результаты экспериментального исследования по взаимодействию определенной группы белков с группой лигандов.

Если выборка создавалась на основе прямых измерений аффинности связывания, то для разделения пар белок-лиганд на взаимодействующие (положительные примеры) или невзаимодействующие (отрицательные примеры) использовался показатель K_i (константа ингибирования). Значение ниже выбранного порога позволяло отнести каждую пару белок-лиганд к числу положительных примеров с бинарным коэффициентом 1, и к числу отрицательных примеров с бинарным коэффициентом 0, в противоположном случае. В данной работе использовались два пороговых значения Ki – 1 и 10 μM. Верхний порог лучше подходит для задач с менее жесткими критериями отбора, а нижний – при более строгих требованиях.

Значительная доля выборок была сформирована на основе сведений из базы данных ChEMBL (https://www.ebi.ac.uk/chembl/), которая в настоящее время является одним из крупнейших ресурсов, содержащих экспериментальную информацию о биологической активности химических соединений. Использовалась локальная версия под управлением СУБД MySQL. Записи ChEMBL содержат сведения, как о белке-мишени, так и о структуре лигандов. Поскольку более чем для половины пар «белок-лиганд» было найдено несколько записей с разными значениями одного и того же показателя связывания (K_i), был применен подход к оценке, основанный на сопоставлении порога с медианным значением показателя.

Эталонная выборка белков и лигандов с бинарными характеристикам их взаимодействия (Gold Standard, Yamanishi et al., 2008) была сформирована специально для тестирования протеохемометрических методов на основе разнородных источников.

Для тестирования разработанного подхода были собраны группы выборок, которые позволили:

- Провести валидацию позиционного режима на последовательностях, полученных путем симуляции эволюционного процесса.
- Идентифицировать критичные для связывания лиганда позиции аминокислотных остатков в реальных белках.
- Протестировать подход на эталонных данных, которые использовали разработчики других методов при их валидации (Таблица 1).
- Исследовать пригодность метода при работе с данными, отражающими разную степень филогенетического родства белков-мишеней: эволюционно дивергировавшие белковые семейства и пары «белок-лиганд», собранные без учета филогенетического родства (Таблицы 2-4).

При формировании обучающих выборок были использованы скрипты, написанные на языке программирования Python, а также СуБД MySQL для обработки информации из базы данных ChEMBL. Для нормализации структур химических соединений использовали пакет RDKit имплементированный в платформе KNIME (https://www.knime.com/). Для картирования результатов работы программы SPrOS в позиционном режиме использовался пакет молекулярной графики PyMOL.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

1. Тестирование в позиционном режиме

Для того чтобы оценить на формальном уровне эффективность метода SPrOS в позиционном режиме, был использован набор из 21 искусственной последовательности с «мутациями», специфичными для отдельных групп (рис. 3).

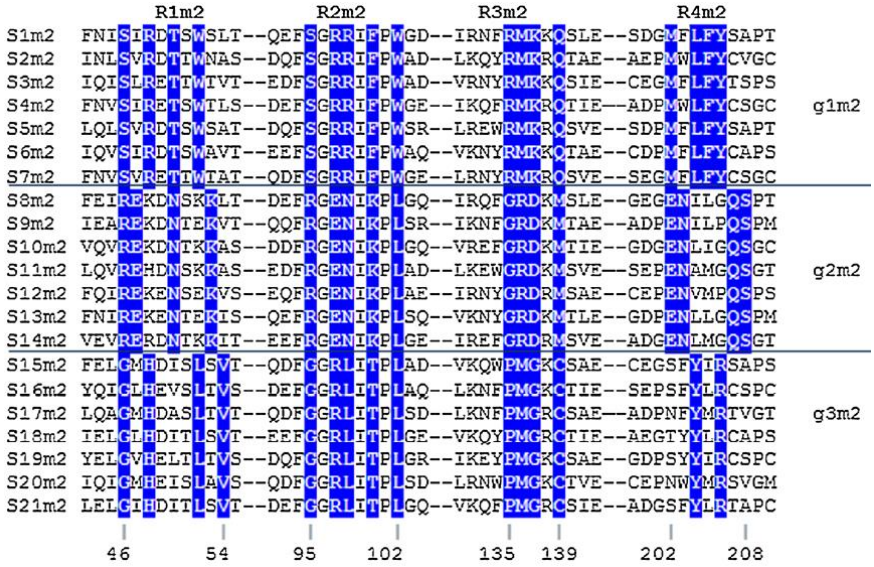

Рисунок 3. Участки искусственных последовательностей R1m2, R2m2, R3m2 и R4m2 с «мутациями» специфичными для подгрупп g1m2, g2m2 и g3m2 выделены цветом.

Во всех искусственных последовательностях специфичные для ее группы позиции идентифицировались со значимыми *p*-значениями ниже 10^{-3} (рис. 4). Точность определения групп-специфичных остатков (IAP) на фоне всех остальных составила 0,999.

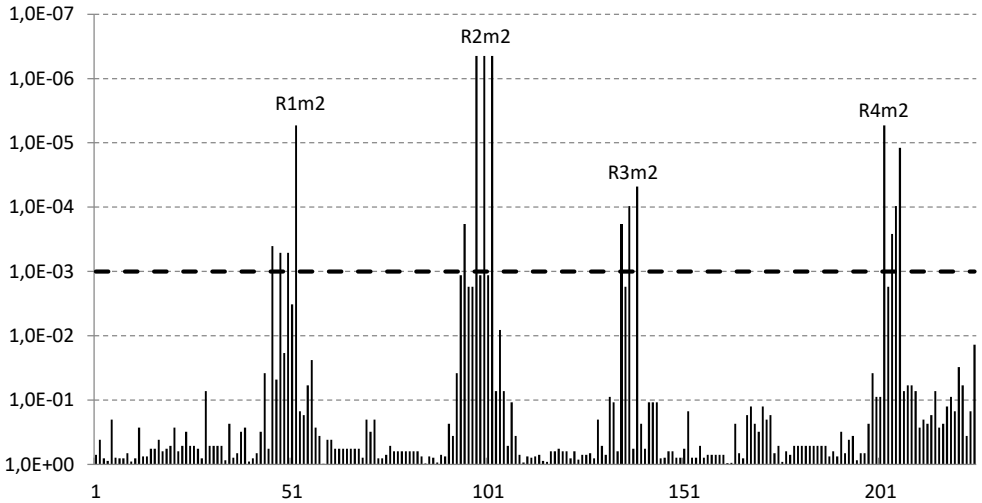

Рисунок 4. Специфичность позиций последовательности S1m2 (X-ось) для группы g1m2. *P*-значения (Y-ось) даны в обратной логарифмической шкале.

Полученные результаты показали, что позиции, определяющие принадлежность к группе S1m2, получили статистические значимые оценки (низкие *p*-значения).

При валидации метода на основе информации по реальным белок-лигандным взаимодействиям мы использовали сведения, представленные в публикации Karaman и др. (Karaman et al., 2008). Результаты были верифицированы по трехмерным структурам соответствующих комплексов. Так, например, в структуре, содержащей протеинкиназу AurkB и ингибитор VX-680 (идентификатор PDB – 4AF3) найдено шесть аминокислотных остатков, принимающих участие во взаимодействии, с *p*-значениями ниже 10^{-3} (рис. 5).

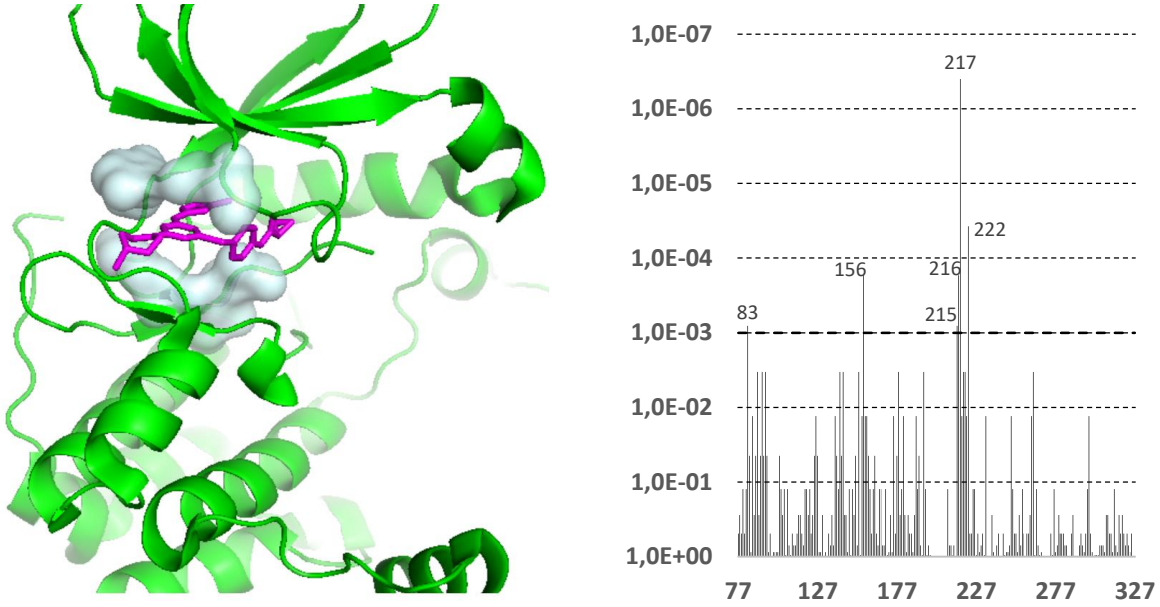

Рисунок 5. Слева: структура протеинкиназы AurkB (зеленая) в комплексе с ингибитором VX-680 (фиолетовый). Серой поверхностью обозначены аминокислотные остатки, получившие *p*-значение менее 10^{-3}. Справа: позиционные оценки специфичности белка к лиганду: *p*-значения в обратной логарифмической шкале (Y-ось).

Пять из них (L83, Y156, K215, I216, A217) принимают участие в формировании полости связывания ингибитора (АТФ-связывающий карман) и расположены на расстоянии не более восьми ангстрем от ингибитора. Остаток S222, локализованный в активационной петле, попадает в нерасшифрованную область.

Таким образом, позиционные оценки отражают специфичность белка к лиганду. Из этого также следует, что расчет интегральной оценки сродства белка к лиганду на основе позиционных оценок выглядит вполне обоснованным.

2. Тестирование в прогностическом режиме

Тестирование подхода с использованием эталонной обучающей выборки

Yamanishi и соавторы собрали выборку «Gold Standard» из различных источников (Yamanishi et al., 2008), которая используется в качестве эталонной рядом авторов для оценки протеохемометрических методов. Мы сравнили результаты тестирования нашего

подхода по второму сценарию, с результатами тестирования других методов при том же сценарии. Значения точности, полученные нами, были близки к оценкам других авторов или превосходили их. (табл. 1).

Таблица 1. Значения ROC AUC, полученные при использовании разных методов.

Белки-мишени	**Метод**							
	NetLapRLS	WNN-GIP	RLScore	KBMF2K	CMF	NRLMF	TMF	SPrOS
Ферменты	0,91	0,95	0,93	0,88	0,92	0,97	0,98	0,94
Ионные каналы	0,91	0,95	0,94	0,94	0,91	0,96	0,97	0,92
GPCR	0,77	0,93	0,85	0,88	0,84	0,93	0,96	0,94
Ядерные рецепторы	0,66	0,94	0,74	0,67	0,68	0,85	0,93	0,99

Наиболее высока точность, существенно превосходящая таковую у других авторов, была получена для группы мишеней «ядерные рецепторы». При этом точность прогноза для других групп мишеней была также высокой с минимальным значением IAP = 0,92 для «Ионных каналов».

Тестирование подхода на наборах данных, отражающих основные области применимости подхода

Важной характеристикой метода протеохемометрики является область применимости – то есть пространство данных, которые используются для построения моделей, пригодных для предсказания новых взаимодействий белок-лиганд. В существующих методах область применимости связана с группой белков-мишеней, связанных разной степенью филогенетического родства. При разработке метода, тестирование проводят на определенной группе мишеней, а смена мишени требует модификации алгоритма. Нами были подготовлены наборы данных, соответствующие основным областям применимости (белковые семейства, белки из различных семейств), при этом тестирование осуществлялось без какой-либо дополнительной оптимизации под конкретный объект исследования.

С помощью оригинальной процедуры сбора данных удалось сформировать представительные выборки по активностям лекарственно-подобных соединений в отношении четырех наборов белков (Таблица 2). Обучающие выборки получились весьма разнородными по количеству взаимодействующих пар, что позволило оценить эффективность подхода в различных ситуациях.

Тестирование на выборках белковых семейств и их лигандов

При выполнении первого сценария были получены высокие значения точности (IAP) от 0,98 до 0,99 при обоих порогах K_i (Таблица 2). Это послужило основанием для того, чтобы при выполнении третьего сценария применить оценки *Pa* и *Pi*, как коэффициенты принадлежности к взаимодействующим и не взаимодействующим парам, соответственно.

В случае второго сценария, когда сопоставление аминокислотных последовательностей по группам лигандной специфичности осуществлялось без учета структуры лигандов, также были продемонстрированы высокие значения IAP от 0,94 до 0,99 (Таблица. 2).

Третий сценарий – моделирование ситуации, в которой неизвестны спектры взаимодействия для обоих компонентов тестируемой пары. Каждому тестируемому лиганду присваивали нечеткие коэффициенты принадлежности ко всем мишеням выборки на основе сопоставления структуры тестируемого лиганда с установленными лигандами соответствующих мишеней. Поскольку в этом случае мы вносили неопределенность в характеристику тестируемой пары, точность прогноза закономерно снижалась по сравнению со вторым сценарием, оставаясь достаточно высокой от 0,86 до 0,99. Наиболее сильное снижение точности получено для семейства протеинкиназ при обоих порогах. Большинство ингибиторов протеинкиназ связываются с одним и тем же структурно консервативным сайтом. Незначительные структурные различия в этой области приводят к парадоксальной селективности ингибиторов [Thaimattam et al., 2007], что обуславливает существенное расхождение между филогенетическим деревом и спектрами связывания ряда лигандов. Это возможная причина меньшей точности, полученной для этого семейства белков. Тем не менее, предложенный подход оправдал себя, будучи применимым и в таком сложном случае. При тестировании на других группах мишеней также отмечалось снижение эффективности, оставаясь на высоком уровне с минимальным значением 0,9.

Таблица 2. Результаты тестирования разработанного подхода при трех сценариях прогноза.

Группа белков	Порог* (μM)	IAP**			Размер обучающей выборки (белки/лиганды)
		1й сценарий	2й сценарий	3й сценарий	
GPCR	1	0,99	0,98	0,90	110/4754
	10	0,98	0,98	0,92	112/6411
Протеинкиназы	1	0,98	0,96	0,89	72/277
	10	0,98	0,94	0,86	77/339
Ионные каналы	1	0,99	0,98	0,96	16/15
	10	0,99	0,97	0,96	16/20
Ядерные рецепторы	1	0,99	0,99	0,98	22/55
	10	0,99	0,99	0,99	23/89

* В соответствии с выбранным порогом Ki осуществляется разделение на взаимодействующие и невзаимодействующие пары белок-лиганд.

** Точность прогноза оценивалась с помощью величины IAP, рассчитанной с использованием процедуры скользящего контроля с исключением по одному, которая численно равна величине ROC AUC.

Тестирование предложенного метода продемонстрировало высокую прогностическую эффективность для всех наборов, характеризующихся различным количеством пар белок-лиганд, а также разным уровнем дивергенции последовательности в пределах семейства.

Тестирование на объединенной выборке

Созданный программный комплекс был применён к выборкам, содержащим белки из различных семейств (Таблица 3). Данные выборки являются наиболее представительными наборами из собранных для белков человека, как в отношении мишеней, так и в отношении лигандов.

Таблица 3. Результаты тестирования разработанного подхода в трех сценариях прогноза.

Порог аффиности Ki (μM)	*IAP*			Размер обучающей выборки
	1й сценарий	2й сценарий	3й сценарий	белки/лиганды
1	0.98	0.98	0.91	252/6443
10	0.98	0.98	0.86	313/9200

При первом и втором сценарии получены высокие оценки точности. При третьем сценарии для обоих порогов отмечается снижение точности, как и в случае с отдельными семействами, но она остается достаточно высокой. Таким образом, подход позволяет

решать задачи прогноза в самом общем виде без предварительного разбиения мишеней на филогенетические группы.

3. Веб-сервис для прогноза взаимодействий белок-лиганд

Разработанные ранее протеохемометрические методы часто оказываются недоступными для широкого круга исследователей, в особенности для пользователей, которые не владеют навыками программирования. В связи с этим, созданный подход был реализован на платформе “Way2Drug” в виде свободно доступного веб-сервиса http://way2drug.com/proteochemometrics/. Помимо центральной задачи протеохемометрики – прогноза взаимодействий на основе структуры лиганда и белка-мишени, пользователю также предоставляются возможность прогнозирования по первому сценарию – только на основе структуры лиганда и второму сценарию – только на основе аминокислотных последовательностей (рис. 6).

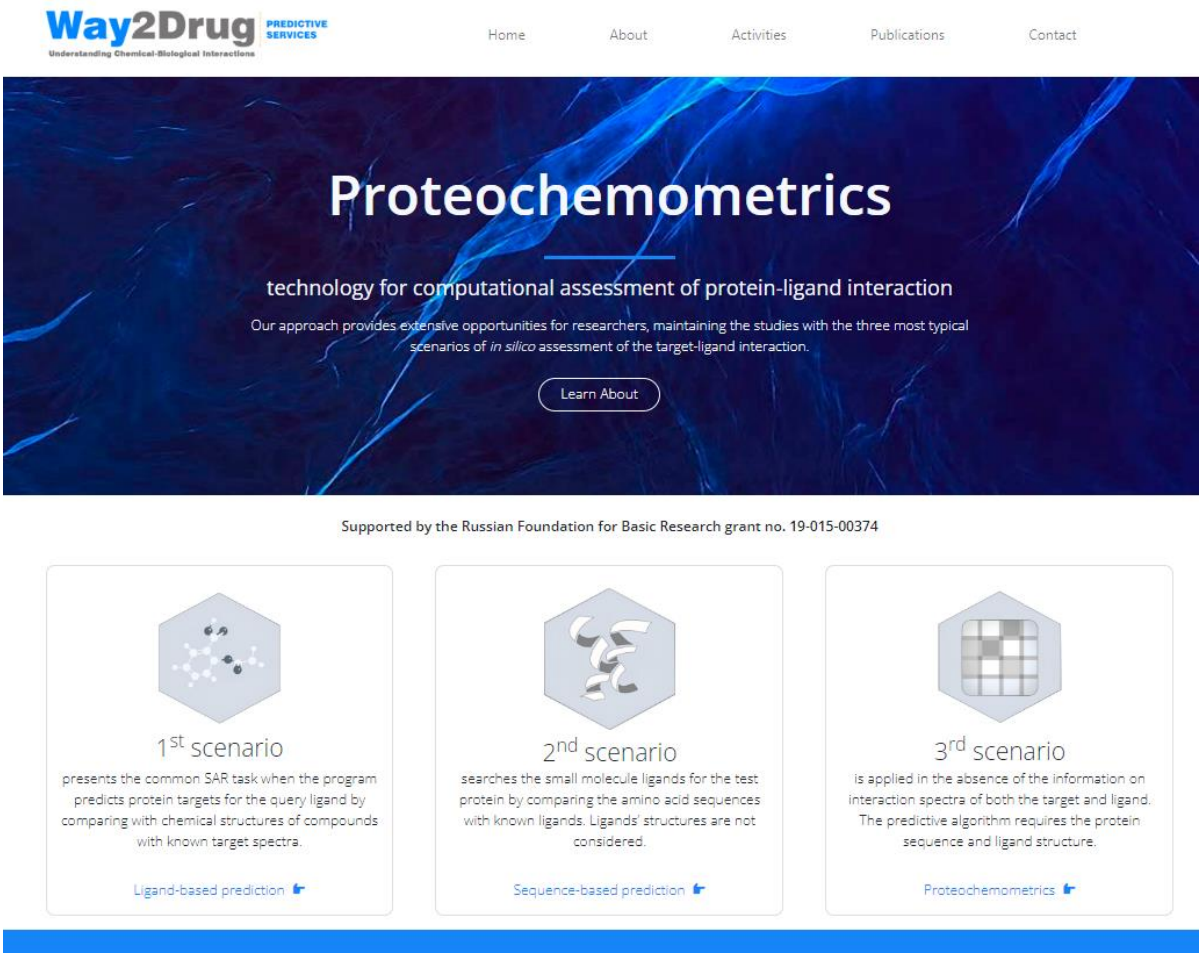

Рисунок 6. Главная страница веб-сервиса по прогнозу белок-лигандных взаимодействий.

Входными данными являются: в первом сценарии структура низкомолекулярного соединения, во втором - аминокислотная последовательность. Для первого и второго сценария пользователь получает список предполагаемых белков-мишеней или лигандов,

соответственно. В третьем же сценарии входными данными являются как структура низкомолекулярного соединения, так и аминокислотная последовательность (рис. 7). В связи с этим пользователь получает только одну оценку, отражающую вероятность взаимодействия лиганда с белком-мишенью.

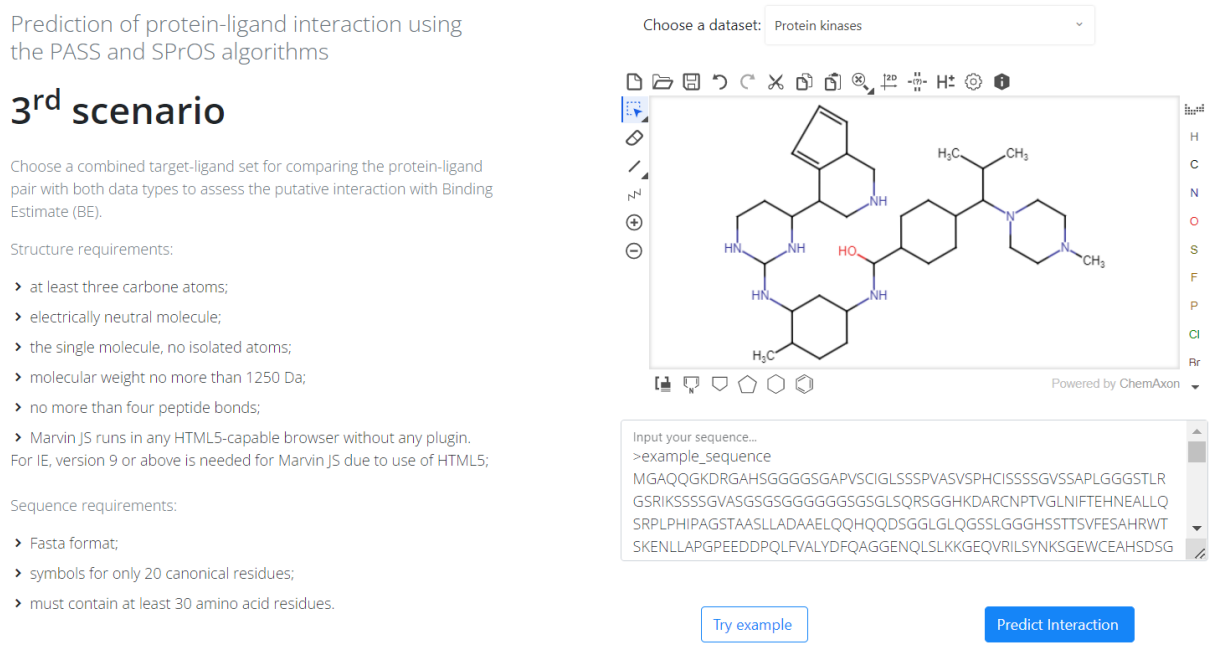

Рисунок 7. Интерфейс веб-сервиса, работающего в режиме третьего сценария с введенной структурной формулой соединения и последовательностью белка в FASTA формате для оценки их взаимодействия.

Сервис предоставляет пользователю возможность предсказания белок-лигандного взаимодействия для семейств белков, представляющих перспективные лекарственные мишени, включая протеинкиназы, ядерные рецепторы и рецепторы, связанные с G-белками. Таким образом, возможен выбор модели взаимодействия для более специфичной, в случае белковых семейств, и более общей в случае объединённого набора белков.

Заключение

В ходе выполнения диссертационной работы разработан подход к широкомасштабному предсказанию взаимодействий «белок-лиганд». Прогноз может быть осуществлен, следуя наиболее популярным сценариям, которые возникают при компьютерной оценке белок-лигандных взаимодействий. Так, с помощью разработанного подхода можно осуществлять прогноз только на основе структуры лиганда, при этом могут быть предсказаны мишени, для которых уже известны какие-либо лиганды. Осуществлять прогноз можно также только на основе последовательности белка-мишени, за счет сопоставления с белками обучающей выборки. При таких сценариях могут быть предсказаны лиганды, для которых уже известны какие-либо мишени. Основным преимуществом метода является возможность прогноза для пар белок-лиганд, у которых не известен спектр взаимодействий ни для одного компонента. Таким образом, разработанный подход позволяет эффективно исследовать обобщенное пространство белков и низкомолекулярных соединений.

Эффективность метода продемонстрирована на наборах известных данных, которые отражают наиболее важные области применимости при поиске лекарственно подобных соединений и идентификации их белков мишеней. Первой областью применимости является компьютерная оценка активности лигандов в отношении белковых семейств. В работе продемонстрирована высокая точность в отношении семейств белков, представляющих перспективные лекарственные мишени, к ним относятся протеинкиназы, ядерные рецепторы, ионные каналы, и рецепторы, связанные с G-белком. Собранные данные существенно различаются по количеству как лигандов, так и белков-мишеней, это и относительно малые наборы данных, где матрица взаимодействий включает около 400 значений, так и весьма крупные наборы – около 700 тысяч значений. Оценка метода на выборках белков из различных семейств также показала высокую точность. При этом данные такого рода наиболее представительны, матрица взаимодействий содержит около 2,7 миллионов записей. Представленные результаты подчеркивают, что разработанный метод обладает наибольшей областью применимости среди подходов протеохемометрики.

Важной особенностью метода является позиционный режим, благодаря которому у исследователя появляется возможность детально оценивать вклад единичных аминокислотных остатков в аффинность связывания белков с низкомолекулярными лигандами. Эффективность такого режима продемонстрирована как на модельных аминокислотных последовательностях, так и на естественных последовательностях протеинкиназ человека.

Подход реализован в виде первого в мире свободно доступного веб-сервиса, который предоставляет возможности применения протеохемометрических методов для широкого круга пользователей.

Выводы

1. Разработан оригинальный подход к извлечению данных из открытых источников о белок-лигандных взаимодействиях. Информация включает в себя структуры низкомолекулярных соединений, аминокислотные последовательности белков-мишеней, а также показатели взаимодействия для каждой пары. Полученные матрицы белок-лигандных взаимодействий включали от 400 и более значений для отдельных белковых семейств и до 2,7 миллионов значений в случае мишеней, неклассифицированных по филогении.
2. Создан метод для прогноза белок-лигандных взаимодействий на основе анализа локального сходства аминокислотных последовательностей и структур низкомолекулярных лигандов. Разработанный программный комплекс обеспечивает прогноз в соответствии с тремя типовыми сценариями: (1) новый лиганд – известный белок-мишень, (2) новый белок–– известный лиганд, (3) новый белок – новый лиганд.
3. Проведена валидация метода на наборах, представляющих различные группы белков-мишеней, различающихся по филогенетическим отношениям. Высокая точность прогноза (IAP>0,85), достигнутая при всех трех сценариях, свидетельствует о широкой области применимости предложенного подхода, включая случаи с неполными обучающими данными.
4. Разработан свободно доступный в сети Интернет веб-сервис (http://www.way2drug.com/proteochemometrics/), позволяющий пользователю осуществлять прогноз во всех упомянутых сценариях для компьютерной оценки белок-лигандных взаимодействий.

Список работ, опубликованных по теме диссертации

Статьи в рецензируемых журналах

1. Karasev D. A., Sobolev B. N., Lagunin A. A., Filimonov D. A., Poroikov V. V. The method predicting interaction between protein targets and small-molecular ligands with the wide applicability domain // Computational biology and chemistry. – 2022. – V. 98, P. 107674.
2. Karasev D.A. Sobolev B.N., Lagunin A.A., Filimonov D.A., Poroikov V.V. Prediction of Protein-ligand Interaction Based on Sequence Similarity and Ligand Structural Features // International journal of molecular sciences. – 2020. – V. 21, № 21, P. 8152.
3. Karasev D.A. Sobolev B.N., Lagunin A.A., Filimonov D.A., Poroikov V.V. Prediction of Protein-Ligand Interaction Based on the Positional Similarity Scores Derived from Amino Acid Sequences // International journal of molecular sciences. – 2020. – V. 21(1), P. 24.
4. Карасев Д.А., Веселовский А.В., Лагунин А.А., Филимонов Д.А., Соболев Б.Н. Распознавание аминокислотных остатков, обуславливающих специфичное взаимодействие протеинкиназ с низкомолекулярными ингибиторами // Молекулярная биология. – 2018. – Т. 52(3), С. 555–564
5. Karasev D. A., Veselovsky A. V., Oparina N. Y., Filimonov D. A., Sobolev, B. N. Prediction of amino acid positions specific for functional groups in a protein family based on local sequence similarity // Journal of molecular recognition. – 2016. – V. 29(4), P. 159–169.

Работы, опубликованные в сборниках материалов научных конференций

6. Karasev D.A., Sobolev B.N., Lagunin A.A., Filimonov D.A., Poroikov V.V. Predicting the protein-ligand interactions based on ligands' structures and proteins' sequences // The 13th International Conference on Bionifomatics of Genome Regulation and Structure/Systems Biology (BGRS\SB-2022). Novosibirsk, Russia, 2022. P. 345.
7. Карасев Д.А., Соболев Б.Н., Лагунин А.А., Филимонов Д.А., Поройков В.В. Прогноз связывания белков с низкомолекулярными лигандами на основе химических структур и аминокислотных последовательностей // VII Съезд биохимиков России и X Российский симпозиум "Белки и пептиды". Дагомыс, Россия, 2021. С. 173.
8. Карасев Д.А., Филимонов Д.А., Соболев Б.Н., Лагунин А.А. Прогноз белок-лигандных взаимодействий in silico // Международный форум: Биотехнология: состояние и перспективы развития. Москва, Россия, 2020. С. 254.
9. Karasev D.A., Filimonov D.A., B.N. Sobolev, Lagunin A.A. Detection of the protein targets for small molecular ligands with use of the local sequence similarity estimation // Moscow Conference on Computational Molecular Biology (MCCMB'19). Moscow, Russia, 2019.

10. Karasev D.A., Lagunin A.A., Filimonov D.A., Veselovsky A.V., Sobolev B.N. // 43rd FEBS Congress, Biochemistry Forever. Prague, Czech Republic, 2018. P. 287-288.
11. Karasev D.A., Savosina P.I., Veselovsky A.V., Filimonov D.A., Sobolev B.N. Identification of amino acid residues affecting on the specificity of interaction of protein kinases and small molecular inhibitors // Moscow Conference on Computational Molecular Biology (MCCMB'17). Moscow, Russia, 2017.
12. Карасев Д.А., Савосина П.И., Веселовский А.В., Филимонов Д.А., Соболев Б.Н. Определение лиганд-специфичных аминокислотных остатков в последовательностях протеинкиназ // VIII Российский симпозиум «белки и пептиды». Москва, Россия. 2017. С. 122.
13. Karasev D.A., Veselovsky A.V., Oparina N.Yu., Rudik A.V., Filimonov D.A., Sobolev B.N. Based on the local sequence similarity method for prediction of amino acid positions related to the protein-ligand specificity // The 10th International Conference on Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS\SB-2016). Novosibirsk, Russia, 2016. P. 117
14. Карасев Д.А., Веселовский А.В., Опарина Н.Ю., Рудик А.В., Филимонов Д.А., Соболев Б.Н. определение аминокислотных остатков, определяющих селективность ингибиторов к протеинкиназам человека // XXIII Российский национальный конгресс «Человек и лекарство». Москва, Россия 2016. С. 181.

Благодарности

Автор выражает благодарность научному руководителю д.б.н., профессору РАН Алексею Александровичу Лагунину за поставленную актуальную интересную задачу и руководство на всех этапах исследования, к.б.н. Борису Николаевичу Соболеву за помощь при подготовке специализированной версии программы SPrOS и участие в интерпретации результатов, к.ф.-м.н. Дмитрию Алексеевичу Филимонову за консультации при выработке общей концепции исследования и подготовку специализированной версии программы PASS, к.б.н. Опариной Нине Юрьевне за участие в создании набора искусственных последовательностей. Автор благодарит к.б.н. Дмитрия Сергеевича Дружиловского, к.б.н. Анастасию Владимировну Рудик и м.н.с. Никиту Сергеевича Ионова за помощь в разработке веб-сервиса.

Отдельную благодарность автор выражает своей семье, которые поддерживали его на протяжении выполнения диссертационной работы.

Автор выражает благодарность Российскому фонду фундаментальных исследований за поддержку данной работы (гранты № 16-04-00491 и № 19-015-00374).