Отзыв официального оппонента

д.б.н., профессора Карягиной-Жулиной Анны Станиславовны 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

Актуальность исследования

Экспериментальное исследование белок-лигандных взаимодействий при разработке новых лекарственных соединений остается трудоемким и дорогостоящим процессом. Методы виртуального скрининга позволяют существенно сократить пространство поиска, уменьшив временные и финансовые затраты.

Прогноз белок-лигандных взаимодействий на основе структуры лигандов (моделирование структура активность – SAR) широко применяется в исследованиях направленных на поиск низкомолекулярных соединений активных в отношении известных белков-мишеней. Ограничения этого подхода связаны с недостаточным количеством данных о взаимодействиях белок-лиганд. Так на сегодняшний день низкомолекулярные лиганды определены не более чем для полутора тысяч белков человека. В то же время у человека идентифицировано около 25000 белок-кодирующих генов, а, с учётом сплайсинга, число возможных протеоформ превышает 100000. Этот массив данных содержит сведения о множестве белков, вовлеченных в патологические процессы, но с еще не установленным спектром взаимодействия. Модуляция этих регуляторных звеньев может стать основой для эффективной терапии. Однако SAR-методы не позволяют прогнозировать взаимодействия для еще не установленных мишеней.

Для расширения прогностических возможностей в модель, наряду с дескрипторами лиганда, можно включить описание аминокислотной последовательности белка-мишени. Такой поход получил название «протеохемометрика» (PCM). Существующие PCM-методы используют разнообразные способы описания белков, часто применимых только для решения конкретных задач. Изменение области исследования требует серьезной модификации метода. Таким образом, разработка метода с широкой областью применимости составляет актуальной задачу.

Диссертационная работа Карасева Д.А. посвящена разработке метода протеохемометрики, позволяющего проводить широкомасштабные предсказания белок-лигандных взаимодействий.

Структура и содержание

Диссертационная работа Карасева Дмитрия Алексеевича имеет классическую структуру и состоит из введения, обзора литературы, материалов и методов, результатов и обсуждения, заключения, выводов и списка литературы. Диссертация изложена на 101 странице и включает 155 литературных источников и семь таблиц.

Во вводной части диссертации автор подчеркивает актуальность исследования, формулирует цель и задачи исследования. Также автор обращает внимание на теоретическую и практическую значимость выполняемой работы, отмечает оригинальность предлагаемого подхода.

Литературный обзор включает в себя описание как классического моделирования структура-активность, так и особенностей, связанных с моделированием в условиях, когда используются дескрипторы белка и лиганда, т.е. моделирование с помощью протеохемометрики. Подробно рассмотрены различные подходы к описанию структур низкомолекулярных соединений. Существенная часть обзора посвящена подходам к описанию аминокислотных последовательностей белков мишеней. Кратко рассмотрены основные методы машинного обучения, применяемые для построения моделей протеохемометрики. Автор акцентирует внимание на специфике валидации таких моделей по сравнению с классическим моделированием структура-активность. Также вводится т.н. сценарный подход, который отражает различные ситуации при поиске активных соединений в отношении каких-либо белков-мишеней.

Обзор литературы написан достаточно лаконично, при этом фрагменты, описывающие разрабатываемые в мире подходы, имеющие непосредственное отношение к теме диссертационной работы, изложены подробно, давая полное представление об основных направлениях протеохемометрики.

В заключительной части литобзора соискатель критически рассматривает различные методы протеохемометрики, их особенности и ограничения, формулирует на этом фоне актуальность и значимость выполняемой работы.

В разделе Материалы и методы охарактеризованы основные информационные источники, использованные в работе. Представлены прогностические алгоритмы PASS и SPrOS. Дано исчерпывающее описание процедуры валидации моделей в соответствии со сценарием прогноза.

В разделе Результаты и обсуждение автор в первую очередь описывает разработанный им протокол по сбору обучающих данных из БД «ChEMBL» и процедуру извлечения информации из БД «Stanford HIV database». Следует отметить особую тщательность подхода к отбору данных, в частности, соискатель пришёл к выводу, что очень важна

обработка неформализованной информации, поскольку формализованные поля часто «не содержат важной информации, например, сведений о мутациях, которые являются безусловным поводом для исключения записи из дальнейшей обработки». Работа с неформализованной информацией редко используется при работе с базами данных, поскольку это очень трудоёмко, и требует дополнительных усилий и трудозатрат. Однако, анализ этой информации позволяет сделать обучающую выборку более адекватной и, соответственно, должно приводить к лучшим результатам предсказания.

Автор проводит демонстрацию работы разработанного подхода в т.н. позиционном режиме. Протестированы искусственные аминокислотные последовательности, при этом автор отмечает, что все «функционально значимые» остатки были идентифицированы. При тестировании на естественных последовательностях с помощью разработанного метода автору также удается определить остатки, которые могут быть ассоциированы со специфичностью к класс-образующим лигандам.

При тестировании в прогностическом режиме на эталонной обучающей выборке «Gold standard» достигнуты результаты, сопоставимые или превосходящие по точности результаты других авторов.

Автор демонстрирует эффективность подхода при тестировании на различных таксономических группах белков, объединяющих такие перспективные лекарственные мишени как протеинкиназы, рецепторы, связанные с G-белком, ионные каналы и ядерные рецепторы. Для всех мишеней автор отмечает высокую точность прогноза (для большинства тестовых выборок IAP превышает 0,9). Тестирование на обобщенной выборке, включающей белки из различных семейств также показало высокую точность.

В качестве набора данных, представляющего близкие гомологи, автор использовал белки ВИЧ с единичными заменами, обуславливающими различную чувствительность к ингибиторам. Автору удалось построить модели, которые позволили с высокой точностью предсказывать резистентность к большинству ингибиторов.

В конце раздела подробно описан веб-сервис, который представляет свободный доступ к разработанному автором программным комплексом.

В разделе Выводы автор резюмирует полученные результаты, полностью релевантные поставленной цели и задачам диссертационной работы.

Автореферат полностью отражает результаты диссертации.

Научная новизна

Предлагаемый подход позволяет прогнозировать белок-лигандные взаимодействия для разнородных групп белков-мишеней. Оригинальная методика описания аминокислотных

остатков на основе их окружения показала эффективность при работе с группами белков-мишеней, которые характеризовались разной степенью филогенетического родства. Важной особенностью предлагаемого подхода является то, что метод не требует оптимизации при смене исследуемой группы белков-мишеней, что выгодно отличает разработанный подход от других известных программ.

Разработанный соискателем подход реализован в виде веб-сервиса. Сервис в настоящее время функционирует, не требует регистрации, достаточно удобен в использовании, имеет хороший дизайн и интуитивно понятный интерфейс. Можно скачать списки используемых белков-мишеней и лигандов в виде таблиц Excell. Имеется достаточно подробное описание алгоритма и ссылки на статьи с описанием используемых программ. Использование веб-сервиса не требует навыков программирования – он может использоваться широким кругом пользователей, например, экспериментаторов для предсказания лигандов для своих белков, мишеней для исследуемых лигандов или оценки связывания конкретной мишени и лиганда. На тестовых примерах работает достаточно быстро. Из пожеланий по доработке – хорошо было бы поставить счётчик посещений. Было бы также полезно предоставить возможность пользователям вводить не по одной структуре лиганда и не по одному белку, а списками, и выдачу, соответственно, присылать на электронную почту.

В сети интернет не удалось найти веб-сервис с подобной функциональностью. Поэтому данный сервис можно считать первым программным комплексом, который предоставляет широкому кругу пользователей возможность PCM-прогноза согласно трем типовым сценариям.

Теоретическая и научно-практическая значимость

Поиск низкомолекулярных модуляторов активности белков имеет как фундаментальное, так и прикладное значение. Селективные лиганды могут использоваться в качестве т.н. химических зондов, а также при разработке лекарственных соединений. Предлагаемый автором подход позволяет подбирать потенциальные лиганды как для известных белков-мишеней, так и для белков с неизвестным спектром лигандов.

Разработанная 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 используемой при обучении информации. Метод обеспечивает высокую точность предсказаний в широкой области применимости, которая покрывает наборы белков-мишеней с разной степенью филогенетического родства.

Позиционный режим метода позволяет исследовать механизм белок-лигандного взаимодействия, исследовать причины той или иной селективности низкомолекулярного соединения. На основе данной информации совместно с результатами молекулярного моделирования становится возможным вносить изменения в структуру лиганда с целью достижения необходимой специфичности.

Полнота освещения результатов диссертации в печати

По материалам диссертации опубликовано 14 печатных работ в российских и международных изданиях, пять из которых – статьи в рецензируемых международных журналах. Содержание опубликованных работ полностью отражает основные положения диссертационной работы Карасева Дмитрия Алексеевича.

Вопросы и замечания

В целом, текст диссертации написан грамотно, логично, хорошим литературным языком, однако, к сожалению, автору не удалось избежать некоторого количества ошибок и опечаток (ошибки выделены полужирным шрифтом и подчёркнуты), например:

Стр. 11. Раздел «Методология и методы диссертационного исследования»:

«Оценка эффективности **предложенного выполнена** с помощью процедуры скользящего контроля в соответствии с современными требованиями к валидации **результов** прогноза. Оценка области применимости предложенного **похода** ... ».

Стр. 23. Раздел 1.5.1 «Описание структур лигандов».

Трехмерные дескрипторы учитывают пространственную организацию молекулы и часто требуют дополнительной подготовки с **использование** подходов молекулярного моделирования [Cramer и соавт., 2007].

Стр. 66.

3.4 Оценка точности прогноза в позиционном режиме с использованием модельных аминокислотных **последовательностях**

Стр. 73, нижняя строка

... **интерпритируемые** результаты…

Стр. 79

Стендфорская **базы** данных (Standford HIV database)…

Стр. 80

Поэтому **вы** провели тестирование с двумя значениями длины сопоставляемых фрагментов F – 7 и 30.

Так, в случае протеазы оценка точности варьировала от 0,93 (Типранавир) до 0,98 (**Лопинавира**).

Таблица 4 – не содержит данных, однако данные обсуждены в тексте.

По диссертационной работе Карасева Д.А. имеется ряд вопросов и замечаний:

1. Обучающие выборки при тестировании подхода в прогностическом режиме получились весьма разнородными по количеству взаимодействующих пар. Вероятно, с этим связаны не очень хорошие значения точности для некоторых групп белков. Существует ли возможность ещё более расширить или качественно изменить обучающую выборку с целью выровнять значения точности для различных групп белков в пределах одной стратегии? Например, добавив данные, появившиеся в базах данных с 2021 года?

2. Схема работы разработанного подхода, представленная на рисунке 8, недостаточно подробна и не даёт полного представления, о том, каким именно образом реализуется разработанный подход в различных ситуациях. Нужно было нарисовать её более детально или, возможно, представить в виде трёх схем для разных вариантов использования подхода. Указать, что подаётся на вход, что является выходом. В каких случаях реализуются прогностический и позиционный режимы. На схеме эти термины не употребляются. Вместо этого используются термины «позиционные оценки» и «прогноз мишеней для структуры». Это затрудняет понимание схемы работы подхода на этом этапе чтения диссертации.

3. Рисунок 14. – В подписи к рисунку не указано, какие программы использованы для получения изображений структур и филогенетических деревьев, в случае деревьев – с какими параметрами?

4. Рисунок 16. – Что означают подписи «N-доля» и «C-доля»? N-концевой и C-концевой домены белка?

Заключение

Диссертация Карасева Дмитрия Алексеевича «Разработка метода протеохемометрики для предсказания взаимодействий белков и лигандов на основе их локального сходства» является законченной научно-квалификационной работой. По актуальности поставленных задач, научной новизне, теоретической и практической значимости диссертационная работа соответствует требованиям пп. 9-14 «Положения о порядке присуждения ученых степеней», утвержденным Постановлением правительства Российской Федерации №842 от 24.09.2013 г. (в редакции с актуальными изменениями), предъявляемым к диссертациям на соискание ученой степени кандидата наук, а ее автор Карасев Дмитрий Алексеевич заслуживает присуждения ученой степени кандидата биологических наук по специальности 1.5.8. – «Математическая биология, биоинформатика».

Оппонент
Главный научный сотрудник
лаборатории биологически активных наноструктур
федерального государственного бюджетного учреждения
«Национальный исследовательский центр эпидемиологии и
микробиологии имени почетного академика Н.Ф. Гамалеи»
Министерства здравоохранения Российской Федерации,
доктор биологических наук,
профессор

_____/Карягина-Жулина Анна Станиславовна/

Специальность: 03.01.03 – «Молекулярная биология»,
Почтовый адрес: 123098, г.Москва, ул. Гамалеи, д. 18
Телефон: +7 (499) 193-30-01
Адрес электронной почты: akaryagina@gmail.com

Подпись Карягиной-Жулиной А.С. заверяю
Ученый секретарь
федерального государственного бюджетного учреждения
«Национальный исследовательский центр эпидемиологии
и микробиологии имени почетного академика Н.Ф. Гамалеи»
Министерства здравоохранения Российской Федерации,
кандидат биологических наук

_____/Сысолятина Елена Владимировна/

09.10.2023
_____
дата