ОТЗЫВ

официального оппонента, доктора биологических наук, заведующего лабораторией информационных технологий в фармакологии и компьютерного моделирования лекарств Научного центра инновационных лекарственных средств с опытно-промышленным производством Федерального государственного бюджетного образовательного учреждения высшего образования «Волгоградский государственный медицинский университет» Министерства здравоохранения Российской Федерации Васильева Павла Михайловича на диссертацию Столбова Леонида Алексеевича на тему «Разработка подходов к виртуальному скринингу антивирусных соединений с учетом гетерогенности информации», представленной к защите на соискание ученой степени кандидата биологических наук по специальности 1.5.8. – Математическая биология, биоинформатика

Актуальность проблемы

По данным Всемирной организации здравоохранения, в 2022 г. в мире насчитывалось не менее 39 млн. ВИЧ-инфицированных, а еще у 4 млн. человек, живущих с ВИЧ, вирус до сих пор не диагностирован. ВИЧ по-прежнему является серьезной проблемой общественного здравоохранения.

Кроме того, в настоящее время наблюдается резкий рост общемировой заболеваемости COVID-19. По данным ВОЗ на 30.08.2023 г., во всем мире зарегистрировано более 769 млн. случаев заражения и около 7 млн. обусловленных COVID-19 смертей.

Таким образом, создание новых лекарственных препаратов для лечения указанных вирусных инфекций является весьма актуальной задачей.

Как известно, полная стоимость разработки нового лекарственного препарата очень высока и может достигать 2 млрд. долларов. Больше половины этой суммы составляют доклинические исследования, что обуславливает необходимость оптимизации данного этапа. Важную роль в этом играют доэкспериментальные методы поиска новых лекарственных веществ. Методы *in silico* позволяют существенно сократить временные,

материальные и финансовые затраты на экспериментальные исследования за счет предварительного отсева малоэффективных структур, оптимизируя таким образом направленный поиск активных веществ.

Вместе с тем, общеизвестно, что многие из существующих в настоящее время вычислительных методов не обеспечивают достаточной точности выявления высокоактивных соединений.

Таким образом, разработка новых универсальных подходов к компьютерному конструированию лекарств, обладающих высокой точностью и предсказательной способностью и применимых, в том числе, для поиска антивирусных соединений, активных в отношении ВИЧ-1 и SARS-CoV-2, является актуальной задачей.

Диссертация Столбова Л.А. посвящена разработке на основе теории самосогласования новой методологии виртуального скрининга антивирусных соединений, с учетом гетерогенности данных, на примере ингибиторов релевантных белков ВИЧ-1 и SARS-CoV-2, что, в контексте вышеизложенного, обуславливает высокую актуальность проведенной работы.

Обоснованность и достоверность результатов исследования

Работу отличает корректно поставленная цель, грамотно сформулированные задачи, адекватно подобранные методы исследования.

Проведенное исследование характеризуется масштабностью и тщательностью подготовки исходного материала, выполненной с использованием широко известных и надежных баз данных: 1) из БД Tox21 получено и проанализировано 10 000 записей по структуре и активности химических соединений, по которым сформировано 38 обучающих и тестовых выборок общим объемом 149 444 структур; 2) из БД NIAID ChemDB HIV, ChEMBL и Integrity получено и проанализировано 35 346 записей по структуре и активности ингибиторов интегразы, протеазы и обратной транскриптазы ВИЧ-1, по которым сформирована 21 обучающая и тестовая выборка общим объемом 100 706 структур; 3) из БД PostEra Moonshot, COVID-19 Open Data Portal и Coronavirus Antiviral and Resistance Database получено и проанализировано

8 962 записей по структуре и активности ингибиторов протеазы 3CLpro, протеазы PLpro и полимеразы RdRp SARS-CoV-2, по которым сформировано 3 обучающих выборки общим объемом 8 962 структуры. Всего в семи базах данных было найдено и проанализировано 54 314 записей по структуре и 44 видам активности химических соединений, по которым сформировано 62 обучающие и тестовые выборки общим объемом 259 118 структур.

Следует отметить высокий уровень математической подготовки автора. Разрабатываемые методы самосогласованной логистической SCLC и экстремальной SCEC классификации содержательно излагаются в терминах векторной и матричной алгебры. При необходимости автор корректно использует соответствующие методы многомерной статистики.

Отдельно необходимо отметить высокий профессиональный уровень автора как программиста, умелое использование современных языков и методов программирования и специализированных программ. Алгоритмы логистического SCLC и экстремального SCEC классификаторов реализованы автором самостоятельно на языке C++ для непосредственного применения через команды языка R посредством Rcpp в качестве промежуточной среды. Для предобработки данных автором самостоятельно разработаны процедуры на языке Python, включающие применение специализированной хемоинформатической библиотеки RDKit.

Автор квалифицированно применяет для сравнительных исследований такие хорошо зарекомендовавших себя программы, как PASS (байесовский классификатор) и GUSAR (самосогласованная регрессия), а также пакеты R ‘caret’ (метод опорных векторов) и R ‘neuralnet’ (искусственные нейронные сети).

Необходимо особо отметить, что на основе (Q)SAR моделей, полученных с применением разработанной методология, создан свободно доступный специализированный Web-сервис AntiHIV Pred.

Специально следует подчеркнуть, что работа выполнена при поддержке двух грантов: гранта РФФИ и гранта РНФ. Таким образом, результаты диссертационного исследования неоднократно проходили экспертизу высокопрофессиональных специалистов этих фондов.

Большой объем данных, полученных из надежных источников, их глубокий анализ, тщательная предварительная обработка и последующая профессиональная структуризация, аргументированное построение алгоритмов разрабатываемой методологии, создание специализированных программных модулей, корректная вычислительная обработка с помощью этих программ структурно весьма разнородных и несбалансированных данных, демонстрация высокой прогнозной эффективности новой методологии и создание на ее основе профильного Web-сервиса – все это позволяет квалифицировать результаты диссертационной работы как обоснованные и достоверные, а основные положения, выносимые на защиту, выводы и практические рекомендации, как вполне обоснованные.

Научная новизна и практическая ценность диссертации

Исследование носит фундаментально-прикладной характер. Бесспорны научная новизна и практическая значимость полученных результатов.

Впервые на основе теории самосогласования созданы оригинальные методы логистической SCLC и экстремальной SCEC классификации: разработаны математический аппарат и реализующие его алгоритмы и компьютерные программы. Показана высокая эффективность виртуального скрининга противовирусных соединений, выполненного с применением указанных методов, в условиях гетерогенных структурно разнородных и несбалансированных массивов данных.

Впервые с использованием методов SCLC и SCEC построены *in silico* модели для прогноза ингибирующей активности в отношении шести релевантных антивирусных белков-мишеней: протеазы, обратной транскриптазы и интегразы ВИЧ-1; протеазы 3CLpro, протеазы PLpro и полимеразы RdRp SARS-CoV-2. Все модели показали высокую точность и предсказательную способность.

Как итог, впервые создана новая методология виртуального скрининга антивирусных соединений, применимая в условиях гетерогенности данных. Следует подчеркнуть, что разработанная автором методология носит универсальный характер и может быть использована в поиске соединений с другими видами биологической и фармакологической активности.

Таким образом, диссертационная работа Столбова Л.А. имеет высокую степень научной новизны, а полученные при ее выполнении результаты характеризуются высокой практической значимостью.

Теоретическая и научно-практическая значимость

В диссертационном исследовании разработаны новые методы анализа взаимозависимостей «химическая структура – биологическая активность», которые могут быть применены для направленного поиска новых антивирусных соединений, активных в отношении ВИЧ-1 и SARS-CoV-2.

Решающие правила, сформированные в условиях мало сбалансированных выборок, с использованием методов самосогласованной логистической классификации SCLC и самосогласованной экстремальной классификации SCEC, показали высокую распознающую и прогностическую способность, достигающую BA=88.9% и AUC_{ROC}=95.4% и сопоставимую, а во многих случаях превосходящую, аналогичные характеристики моделей, созданных с применением других методов.

Свободно доступный Web-сервис AntiHIV Pred, созданный на основе прогностических моделей, полученных с применением разработанных методов, позволяет выполнять виртуальный скрининг и проводить направленное конструирование новых соединений с высокой анти-ВИЧ-1 активностью.

Особую ценность данной работе придает то, что она выполнена комплексно и логически последовательно: 1) сформирован большой массив данных по токсическим эффектам и антивирусной активности в отношении ВИЧ-1 и SARS-CoV-2 известных химических соединений; 2) выполнен детальный анализ, предварительная обработка и последующая структуризация этих данных; 3) сформирован математический аппарат и построены алгоритмы разрабатываемых методов; 4) созданы специализированные программные средства, реализующие новые методы; 5) с помощью этих программ проведено построение решающих правил как результат вычислительной обработки структурно весьма разнородных и несбалансированных данных; 6) показана высокая прогностическая способность построенных решающих правил, подтверждающая

эффективность новой методологии; 7) создан на основе валидированных прогнозных зависимостей специальный свободно доступный Web-сервис.

Следует отдельно подчеркнуть, что в результате проделанной работы автором создана новая методология общего характера и предложен реальный рабочий инструмент для виртуального поиска новых биологически и фармакологически активных соединений.

Два гранта, при поддержке которых выполнена работа: грант РФФИ и грант РНФ, успешно завершены, результаты опубликованы в открытой печати.

Таким образом, диссертационное исследование Столбова Л.А. имеет высокую теоретическую и практическую значимость.

Соответствие диссертации паспорту специальности

Поставленные цели, задачи, область исследования, предмет исследования, примененные современные и обоснованные методы, полученные выводы и рекомендации являются весьма актуальными для биоинформатики. Считаю, что представленная диссертации полностью соответствует паспорту специальности 1.5.8. Математическая биология, биоинформатика по следующим направлениям.

п.4. Математическое и компьютерное моделирование биологического действия ксенобиотиков. Компьютерное конструирование лекарств. Анализ взаимосвязей «структура-активность». Компьютерная фармакология и токсикология.

п.12. Разработка и применение новых вычислительных алгоритмов для анализа экспериментальных данных в биологии и медицине.

п.14. Математические модели, численные методы, алгоритмы и программные средства применительно к процессам получения, накопления, обработки и систематизации биологических и медицинских данных и знаний.

п.16. Разработка и применение методов машинного обучения и искусственного интеллекта для анализа и прогнозирования свойств биологических объектов на основе анализа больших биомедицинских данных.

Полнота освещения результатов диссертации в печати

По теме исследования опубликована 21 работа, из них 6 статей в журналах, рекомендованных ВАК Минобрнауки РФ и индексируемых в базах данных Scopus и Web of Science. Основные положения диссертации докладывались на 9 международных конференциях. Получено свидетельство о государственной регистрации программы для ЭВМ.

Структура и оформление работы

Диссертация оформлена в классическом стиле, в соответствии с существующими требованиями, изложена на 132 страницах, включает 35 рисунков, 13 таблиц в основной части и 3 таблицы в приложении. Работа состоит из Списка сокращений; Введения; главы 1 «Обзор литературы», включающей три раздела; главы 2 «Материалы и методы», включающей шесть разделов; главы 3 «Результаты и обсуждение», включающей пять разделов самостоятельно полученных данных; Заключения; Выводов; Списка работ, опубликованных по теме диссертации; Финансирования работы; Благодарностей; Списка литературы; Приложений А-Д.

Список литературы включает в себя 169 источников и охватывает период в 47 лет с 1977 по 2023 гг. Необходимо специально отметить значительный объем и временну́ю широту литературного поиска.

В первой главе проведен обзор и анализ литературных данных по теме диссертации. Представлены основные методы компьютерного конструирования лекарств, освещены методы SBDD и методы LBDD. Изложены особенности применения SAR и QSAR анализа для виртуального скрининга биологически активных веществ, в том числе, описаны методы дескрипторного представления химической структуры, способы построения (Q)SAR моделей и оценки их качества. Отражена актуальность использования методов *in silico* для поиска ингибиторов мишеней ВИЧ-1 и SARS-CoV-2, в том числе проведен детальный обзор этих мишеней.

Во второй главе подробно описаны материалы и методы проведенных исследований. Следует отметить, что примененные диссертантом методики сбора и обработки данных валидны и соответствуют задачам исследования.

В первом разделе главы 2 описаны использованные в работе данные об активности известных экспериментально изученных химических соединений,

с указанием источников, откуда эти данные получены. Изложены особенности искусственной генерации обучающих выборок. Описаны особенности формирования обучающих и тестовых выборок по данным системы Tox21; по ингибиторам ВИЧ-1, применительно к выбранным для анализа трем релевантным белкам-мишеням; по ингибиторам SARS-CoV-2, также применительно к выбранным для анализа трем релевантным белкам-мишеням. Детально изложена методика предварительной обработки и формирования обучающих и тестовых выборок по структуре и активности ингибиторов ВИЧ-1 и SARS-CoV-2, а также расширенных выборок по структуре ингибиторов ВИЧ-1, содержащих качественные данные об активности.

Во втором разделе главы 2 описаны методы дескрипторного представления химической структуры соединений обучающих и тестовых выборок: MNA дескрипторы и более детально QNA дескрипторы.

В третьем разделе главы 2 детально излагается математический аппарат методов самосогласованной регрессии и классификации, представляющий стержень разрабатываемой методологии. Последовательно приводятся формулы, составляющие логическую последовательность алгоритмов самосогласованной логистической классификации SCLC и самосогласованной экстремальной классификации SCEC. Отдельно излагается методика ортогонализации и итерирования входного пространства дескрипторного описания.

В четвертом разделе главы 2 перечислено программное обеспечение, использованное при выполнении расчетных исследований, в том числе разработанное автором самостоятельно.

В пятом разделе главы 2 перечислены критерии оценки точности и прогностической способности создаваемых моделей.

В шестом разделе главы 2 изложена методика оценки точности создаваемых моделей применительно к процессу виртуального скрининга.

В третьей главе представлены результаты проведенных исследований и их обсуждение.

В первом разделе главы 3 описаны и обсуждены результаты

тестирования новых разработанных классификационных методов SCLC и SCEC на искусственно сгенерированных данных и результаты валидации этих методов с использованием выборок Tox21. Полученные результаты свидетельствуют о возможности применения разработанных методов для построения моделей с использованием несбалансированных выборок.

Во втором разделе главы 3 изложены и обсуждены результаты использования процедуры предварительной обработки структурной информации в выборках и ассоциированных значений активности. Это позволило провести подготовку верифицированных обучающих выборок для ингибиторов ферментов ВИЧ-1 и SARS-CoV-2.

В третьем разделе главы 3 изложены и обсуждены результаты построения методом самосогласованной регрессии SCR количественных моделей ингибирования мишеней ВИЧ-1 с использованием различных выборок. Найденные значения коэффициента детерминации предсказания находятся в диапазоне Q^2=0.714–0.829, что свидетельствует о достаточно хорошей точности SCR подхода.

В четвертом разделе главы 3 приводятся и обсуждаются результаты применения методов SCLC и SCEC при построении зависимостей «структура-активность» для ингибиторов ферментов ВИЧ-1 и SARS-CoV-2.

Проведено сопоставление показателей точности прогноза, полученных на этих выборках с применением двух новых методов, с точностью моделей, полученных с помощью самосогласованной регрессии SCR, метода опорных векторов SVM, байесовского классификатора PASS и искусственных нейронных сетей ANN.

Показано, что классификационные зависимости, полученные с использованием методов SCLC и SCEC, обеспечивают сопоставимую или более высокую точность и предсказательную способность, в сравнении с классификационными зависимостями, полученными с использованием методов SCR, SVM, PASS и ANN. Так, максимально полученное значение сбалансированной точности составило BA_{SCLC}=0.858, BA_{SCEC}=0.866, BA_{SCR}=0.861, BA_{SVM}=0.834, BA_{PASS}=0.861, BA_{ANN}=0.830.

Продемонстрировано преимущество SCLC и SCEC классификаторов в

сравнении с количественными регрессионными моделями.

В пятом разделе главы 3 описан свободно доступный Web-сервис AntiHIV Pred, позволяющий с использованием различных моделей выполнять по структурной формуле соединения прогноз разных видов таргетной анти-ВИЧ-1 активности.

Заключение и выводы завершают диссертацию и включают практические рекомендации, которые логично вытекают из результатов работы, отражают поставленные задачи, аргументированы и имеют научно-практическую значимость.

Соответствие содержания автореферата основным положениям и выводам диссертации

Автореферат адекватно отражает основное содержание диссертационного исследования, полностью соответствует разделам, положениям и выводам диссертационной работы и оформлен в соответствии с предъявляемыми требованиями.

Вопросы и замечания

Принципиальных замечаний и возражений по диссертационной работе нет, однако в процессе ознакомления с ней возникли следующие вопросы, требующие дополнительного уточнения, но не затрагивающие существа работы.

1. В работе в качестве оценки точности прогноза используется метрика сбалансированной точности BA. Между тем, эта оценка является несмещенной оценкой точности Acc только в случае одинакового числа активных и неактивных соединений в обучающей или тестовой выборке. Поясните, какие статистически обоснованные преимущества при оценке валидности решающих правил дает использование BA?

2. В работе в качестве оценки точности прогноза используется также метрика RSMD (рассчитывается по формуле 12, с. 37), которая по своему статистическому содержанию очень близка коэффициенту детерминации R^2 (рассчитывается по формуле 11, с. 37). Вычисленный по объединенным данным таблицы 3 (с. 77) и таблицы 4 (с. 82) непараметрический коэффициент корреляции Спирмена между R^2 и RSMD составил -0.8740, что

соответствует $p<0.0021$. Поясните, какую дополнительную информацию при оценке валидности решающих правил дает использование RSMD, в сравнении с R^2?

3. В таблицах 5-13 в качестве оценки точности классификации приведены значения сбалансированной точности BA. Величина этого показателя в ряде случаев не очень высока, например, 0.703 для SVM при прогнозе ингибирующей RdRp SARS-CoV-2 активности (таблица 10, с. 92). При этом все выборки являются весьма ассиметричными (в данном случае, из 2632 структур только 91 структура активная), но значения Sensitivity и Specificity не приведены. Иллюстративный пример: BA=0.703 при Sensitivity=0.456 (ниже случайного угадывания) и Specificity=0.950. Насколько ассиметричны были полученные оценки Sensitivity и Specificity в случае не очень высоких значений BA, рассчитанных на сильно не симметричных обучающих выборках?

4. В настоящее время известно достаточно много типов искусственных нейронных сетей различной архитектуры, включающих разные виды активационных функций. Число скрытых слоев и число нейронов в них также может сильно варьироваться. Между тем, в работе архитектура ANN нигде не детализируется. Каковы архитектуры нейросетевых классификационных моделей, построенных для прогноза изучаемых видов активности?

Заключение

Диссертация Столбова Леонида Алексеевича на тему «Разработка подходов к виртуальному скринингу антивирусных соединений с учетом гетерогенности информации» является законченной научно-квалификационной работой, в которой содержится решение актуальной задачи биоинформатики по созданию на основе теории самосогласования новой методологии скрининга *in silico* в условиях гетерогенных данных анти-ВИЧ-1 и анти-SARS-CoV-2 активных соединений.

.С учетом актуальности, высокого методического уровня, научной новизны, теоретической и практической значимости исследования считаю, что представленная работа соответствует требованиям п. 9 «Положения о

присуждении ученых степеней», утвержденного постановлением Правительства РФ № 842 от 24.09.2013 (в редакции постановления Правительства РФ № 101 от 26.01.2023), предъявляемым к диссертациям на соискание ученой степени кандидата наук, а ее автор, Столбов Леонид Алексеевич, заслуживает присуждения искомой степени кандидата биологических наук по специальности 1.5.8. – Математическая биология, биоинформатика.

Официальный оппонент:
Заведующий лабораторией
информационных технологий в фармакологии
и компьютерного моделирования лекарств
НЦИЛС ФГБОУ ВО «Волгоградский
государственный медицинский университет»
Минздрава России,
доктор биологических наук Васильев Павел Михайлович

ФГБОУ ВО «Волгоградский государственный медицинский университет» Минздрава России
Адрес: 400131, г. Волгоград, пл. Павших борцов, д. 1
Тел: +7 (8442) 38-50-05
E-mail: post@volgmed.ru
https://www.volgmed.ru/

Подпись тов. Васильева П.М.
УДОСТОВЕРЯЮ
Инспектор по кадрам
Волгоградский государственный медицинский университет
Управление кадров

06.10.2023