Биомедицинская химия, 2024 том 70, вып. 6, с. 469–474.

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ О ВЗАИМОДЕЙСТВИИ ВИРУСОВ С ОРГАНИЗМОМ ЧЕЛОВЕКА И О ПРОТИВОВИРУСНЫХ СОЕДИНЕНИЯХ НА ОСНОВЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА БОЛЬШИХ КОЛЛЕКЦИЙ ТЕКСТОВ

О.А. Тарасова*, Н.Ю. Бизюкова, Е.А. Столбова, Л.А. Столбов, Р.Р. Такташов, Д.А. Карасев, Н.С. Ионов, С.М. Иванов, А.В. Дмитриев, А.В. Рудик, Д.С. Дружиловский, Б.Н. Соболев, Д.А. Филимонов, В.В. Поройков

Научно-исследовательский институт биомедицинской химии им. В.Н. Ореховича,
119121, Москва, Погодинская ул., 10; *эл. почта: olga.a.tarasova@gmail.com

Разработка эффективных противовирусных препаратов имеет огромное значение в связи с возможностью быстрого распространения вирусных инфекций. Накопление сведений в базах данных биологически активных соединений и в научных публикациях даёт возможность извлекать информацию о взаимодействии между мишенями организма человека и фармакологически активными веществами. Эта информация может быть использована для получения знаний о потенциальной фармакологической активности химических соединений, их побочных эффектах и токсичности. Цель нашего исследования — извлечение информации о взаимодействии вируса и организма человека, а также о потенциальных противовирусных соединениях на основе автоматического анализа большого массива научных публикаций. С помощью разработанных ранее и усовершенствованных в рамках настоящей работы методов мы извлекли информацию о взаимодействии вируса и хозяина и наименованиях соединений, которые взаимодействуют с белками вирусов или организма человека. Мы собрали данные о взаимодействии нескольких вирусов, включая вирусы гепатита B и C, SARS-CoV-2, гриппа A и B и вируса простого герпеса, с (1) организмом человека, (2) потенциальными противовирусными соединениями, а также информацию о взаимодействии между потенциальными противовирусными соединениями и белками хозяина. На основе проведённого анализа данных создана свободно доступная база знаний о взаимодействии химических веществ с белками вирусов, белками организма человека, включая лекарственные соединения, их взаимодействие и применение в терапии вирусных инфекций и других заболеваний.

Ключевые слова: взаимодействие “вирус-организм человека”; химические соединения; биологическая активность; противовирусные соединения; интеллектуальный анализ текстов

DOI: 10.18097/PBMC20247006469

ВВЕДЕНИЕ

Разработка новых, более безопасных и эффективных лекарств требует высоких затрат и длительного времени, необходимого для создания новых фармацевтических препаратов [1]. Поэтому, разработка вычислительных методов исследования особенностей химических соединений с целью повышения эффективности поиска перспективных соединений является крайне важной задачей. Основой применения вычислительных методов являются экспериментальные данные, которые опубликованы в научных статьях и содержатся в базах данных биологически активных соединений. Быстрое развитие исследований в области изучения и разработки лекарственных препаратов приводит к накоплению большого количества публикаций, которые являются основным источником информации о биологической активности химических соединений и патогенезе заболеваний. Поэтому извлечение знаний из текста публикации и её аннотации весьма полезно для обогащения информации о биологической активности химических соединений, включая механизмы, их фармакологические и побочные эффекты, токсичность, показания к применению лекарств, лекарственные взаимодействия и совместное применение лекарственных препаратов. Кроме того, извлечение знаний о молекулярных механизмах заболеваний и расстройств будет полезным при разработке новых терапевтических стратегий и поиске новых лекарств. Такая информация особенно важна для разработки оригинальных, эффективных и безопасных противовирусных препаратов, которые должны быть созданы в короткие сроки в случае возникновения новых угроз, вызванных быстрым распространением нового или малоизученного вируса.

Ранее мы разработали алгоритмы, предназначенные для распознавания наименований биологических и химических объектов в текстах научных публикаций [2, 3] и поиска семантических взаимосвязей между ними [3].

Целью данного исследования стало извлечение информации о взаимодействиях между вирусом и организмом хозяина (человека), а также механизмах и эффектах потенциальных противовирусных соединений, на основе ранее разработанных алгоритмов интеллектуального анализа текстов. Также произведена интеграция полученных сведений с представленной в базах данных информацией о биологически активных соединениях и наименованиях белков. На основе извлечённой из литературы информации мы разработали свободно-доступную базу знаний [4], которая предоставляет информацию о взаимодействиях между

вирусами и организмом человека, и химических соединениях, которые могут воздействовать на соответствующие белки-мишени, с указанием перекрёстных ссылок на базу данных ChEMBL v.34.

МЕТОДИКА

Алгоритм извлечения информации о противовирусных соединениях основан на автоматическом поиске и загрузке релевантных публикаций с использованием ключевых слов. Затем в сформированной коллекции текстов производится распознавание наименований биологических и химических объектов и поиск ассоциаций между ними. Используемый в рамках работы метод распознавания наименований объектов является комбинацией трёх алгоритмов машинного обучения: условных случайных полей (Conditional Random Fields, CRF) [2], наивного байесовского классификатора [3] и метода HunFlair [5]. В настоящем исследовании мы предполагали, что конкретная сущность принадлежит к наименованию химического соединения, биологической макромолекуле или заболеванию, если она была распознана соответствующим образом при применении хотя бы двух из трёх алгоритмов машинного обучения. Мы применяли комбинацию методов машинного обучения с последующим использованием словарей наименований химических соединений и активных биологических молекул и их синонимов. Словарь был составлен с помощью автоматизированных запросов к базам данных ChEMBL и UniProt. Такой подход обеспечивает приемлемую точность распознавания при пятикратной кросс-валидации (табл. 1). Дополнительно мы использовали набор правил для автоматического исключения ложно распознанных наименований химических и биологических объектов. Для этого мы провели сопоставление типов и строк самих наименований с разработанными словарями, а также исключили наименования, которые включали в себя нестандартные символы (“@”, “!”, “<” и др.). В числе наименований биологических объектов мы также произвели распознавания наименований миРНК, однонуклеотидных полиморфизмов и аминокислотных замен с применением подхода на основе правил, который проводит сопоставление подстрок из текста заданному шаблону.

Точность распознавания наименований объектов с помощью нашего метода в целом сопоставима с наиболее современным методом HunFlair для наименований химических и биологических объектов, но при этом несколько ниже — в распознавании наименований болезней. Результаты приведённого сравнения могут иметь некоторую погрешность из-за использования различных обучающих и тестовых наборов данных (различных корпусов). Обогащение словарей и обучение на основе нескольких корпусов позволит повысить эффективность распознавания.

Ассоциации между химическими и биологическими объектами (белки вируса и хозяина, гены, миРНК) были извлечены с применением трёх подходов (рис. 1).

Таблица 1. Точность распознавания наименований химических и биологических объектов на основе подхода с применением методов машинного обучения с последующим сопоставлением с терминами словаря

Тип наименования	Разработанный нами метод распознавания			Метод HunFlair [5]		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Химические соединения	0,89	0,83	0,86	0,88	0,88	0,88
Белки и гены	0,87	0,84	0,85	0,84	0,86	0,85
Заболевания	0,84	0,79	0,81	0,86	0,87	0,86

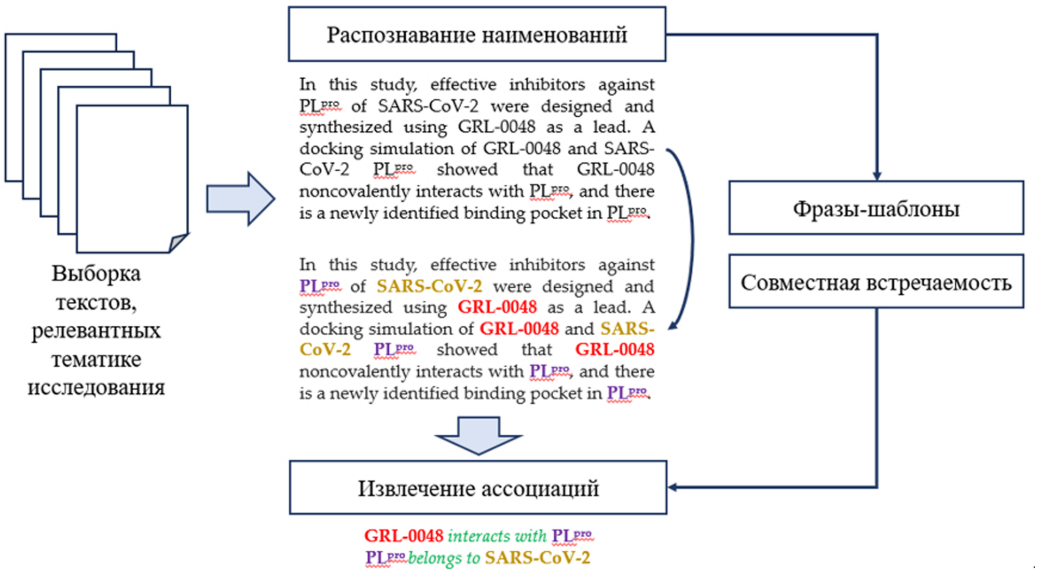

Рисунок 1. Подходы к извлечению ассоциаций между наименованиями объектов из текстов.

При первом подходе наименования объектов извлекаются из выборки публикаций, которые строго релевантны заданной теме; то есть ассоциация устанавливается между распознанным наименованием объекта и, например, описанным в выборке текстов явлением [6]. Такой подход позволяет определить общие понятия в рамках заданной области исследования или обозначить молекулярные механизмы заболевания, которые описаны в текстах научных публикаций. Используя данный подход, мы исследовали возможные типы взаимодействий "вирус-хозяин", общих для SARS-CoV-2 и других вирусов, и таким образом обнаружили общие для ВИЧ-1 и вируса денге молекулярные механизмы патогенеза [6]. Второй подход основан на применении так называемых фраз-шаблонов — фрагментов текста, которые указывают на наличие семантической взаимосвязи между объектами внутри предложений. Этот подход был применён в задаче поиска белков организма человека, обуславливающих различную скорость прогрессии ВИЧ-инфекции [7], а также для идентификации молекулярных механизмов, ассоциированных с путём Hedgehog, в патогенезе неопластических процессов [4]. Третий подход основан на оценке совместной встречаемости определённых терминов в текстах.

В рамках первого подхода используется набор текстов, строго соответствующих интересующей нас теме. При этом мы рассматриваем все именованные сущности, соответствующие биологическим молекулам (белкам, генам и т. д.), распознанные в текстах публикаций, строго относящихся к определённой тематике (т. е. заболеванию, патологическому состоянию и т. д.). Отбор релевантных публикаций обычно осуществляется либо с помощью набора ключевых слов, терминов MeSH, либо на основе моделей машинного обучения (метода опорных векторов, случайного леса, искусственных нейронных сетей), направленных на отбор публикаций, имеющих отношение к ингибированию репликации SARS-CoV-2. Мы протестировали способ отбора релевантных публикаций с применением искусственных нейронных сетей долгой краткосрочной памяти (LSTM). Эффективность распознавания релевантных публикаций (F_1-score) оценивалась путём разделения всей выборки на две равные подвыборки (по 50% от всей выборки на обучающую и тестовую выборку, соответственно). F_1-score распознавания релевантных публикаций составил 0,84 для классификации документов, имеющих отношение к экспериментальному тестированию ингибиторов SARS-CoV-2 с применением клеточных линий и выделенного и очищенного фермента.

Во втором подходе используется набор правил, реализованных с помощью так называемых шаблонных фраз, которые описывают определённые взаимосвязи между объектами, соответствующими идентифицированным сущностям в тексте. Нами показано, что точность распознавания взаимосвязей с применением такого подхода, F_1-score, равна 0,84, что соответствует точности большинства различных методов выявления ассоциаций.

Характер взаимосвязи определяется фразами-шаблонами, которые в свою очередь соотнесены с группами, сформированным на основе их семантического значения. Взаимосвязь считается извлечённой из текста, если в предложении встречались два наименования различных объектов и фраза-шаблон, которая соответствуют их типам. Например, наличие фразы-шаблона "interacts" в предложении совместно с наименованиями химического соединения C и белка P позволяет извлечь ассоциацию вида "C interacts with P". В некоторых случаях взаимосвязи между объектами в предложениях не могут быть извлечены с применением фраз-шаблонов, однако такие ассоциации могут быть предположены исходя из контекста (например, наименованиями белков и вирусов, которым они принадлежат), поэтому в таких ситуациях мы использовали подход на основе совместной встречаемости терминов.

Третий подход к распознаванию ассоциаций основан на обнаружении в одном фрагменте текста двух объектов. Более частые совместно найденные объекты соответствуют хорошо изученным ассоциациям, а редкие совпадения могут представлять собой вновь или недавно найденные ассоциации. По этой причине мы не анализируем отдельно ассоциации, имеющие статистически значимые различия в частоте встречаемости, и рассматриваем каждую найденную ассоциацию как имеющую значение для соответствующей области исследования.

С применением сочетания описанных подходов к извлечению взаимосвязей и ассоциаций была извлечена информация о молекулярных механизмах патогенеза вирусных заболеваний и химических соединениях, активность которых исследуется в отношении этих процессов. Полученные сведения стали частью разрабатываемой свободно-доступной базы знаний [4].

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Методы интеллектуального анализа текстов были применены с целью автоматического извлечения имеющихся в литературе сведений, касающихся ряда вирусов, включая вирусы гепатитов B и C, SARS-CoV-2, вирусы гриппа A и B, простого герпеса, поскольку они имеют большое влияние на здоровье человека и всесторонне изучаются, что привело к накоплению значительного объёма литературных источников. Мы извлекли информацию о взаимодействии этих вирусов с (1) хозяином (организмом человека) и (2) потенциальными противовирусными соединениями. Подобные сведения о взаимодействиях мишеней организма человека и вируса могут быть представлены в литературе как результаты разнообразных *in vitro* исследований, как в клеточных системах, так и на рекомбинантных белках: эта информация была извлечена с применением разработанного алгоритма интеллектуального анализа текстов. Для того, чтобы извлечь соответствующие сведения, мы сформировали коллекцию текстов, релевантных *in vitro* и *in vivo* исследованиям

(1) молекулярных взаимодействий между вирусом и хозяином и (2) активности 77 известных противовирусных лекарственных препаратов, извлечённых из базы данных ChEMBL (v.34). Список известных противовирусных лекарственных препаратов был использован для формирования запросов в базу данных PubMed наряду с MeSH-терминами, ключевыми словами и типами публикаций. Таким образом удалось загрузить 171735 текстов аннотаций и заголовков публикаций, находящихся в свободном доступе, которые стали объектом дальнейших исследований. Мы извлекли из текстов более 83000 наименований химических соединений, и 11878 наименований белков и генов; 8848 и 7544 были идентифицированы в базах данных, соответственно (табл. 2).

На основе извлечённой информации мы создали свободно-доступную базу знаний [4], включающую в себя сведения о взаимодействиях химических соединений с вирусными белками и их мишенями в организме хозяина. Текущая на момент написания публикации версия базы знаний [4] включает в себя 5683 записи об извлечённых из текстов научных публикаций ассоциациях и взаимосвязях между наименованиями химических и биологических объектов. Представленная база знаний содержит идентификаторы объектов, полученные при сопоставлении извлечённых из текстов наименований записям ChEMBL, а также список возможных синонимов объектов, с указанием кросс-ссылок на источник. В базу знаний мы включили тип ассоциации на основе фразы-шаблона, которая использовалась при её извлечении из текста. Например, ассоциации между химическим соединением и заболеванием могут быть описаны фразами “may cause” (“может приводить к” — для токсических и побочных эффектов), “used in therapy” (“используется в терапии”), “has positive effect on” (“имеет благоприятный эффект на” — для исследуемых в отношении заболевания биологически активных соединений), “increases risk of” (“увеличивает риск возникновения” — для токсических и побочных эффектов), и т. д. База знаний доступна в сети Интернет [4]. Среди включённых в базу знаний — 5683 ассоциаций, для 253 — конкретный тип взаимосвязи не известен.

Количество ассоциаций, извлечённых из текстов, между конкретными типами объектов представлено в таблице 3.

В частности, количество связей между сущностями отражает специфику тематики коллекции научных текстов, сформированной для анализа. Поскольку большинство публикаций были получены в соответствии с запросом, который включал наименование противовирусного препарата, большинство найденных ассоциаций — это ассоциации между химическим соединением и заболеванием, или двумя химическими соединениями. В то же время, некоторые публикации касались взаимодействия вируса и хозяина и тестирования активности химических соединений *in vitro* или *in vivo*, поэтому мы также смогли извлечь более 1200 взаимосвязей между химическими веществами и белками/генами, которые отражают конкретный молекулярный механизм действия противовирусного химического соединения или конкретного вирусного заболевания.

Типы ассоциаций между двумя химическими соединениями представлены на рисунке 2. Эта информация может быть особенно полезной для выбора наиболее перспективных фармацевтических субстанций и лекарственных препаратов с противовирусным эффектом, а также их комбинаций.

Таблица 2. Количество наименований биологических и химических объектов, извлечённых из текстов коллекции, посвящённых исследованию биологической активности противовирусных лекарственных препаратов

Тип наименования	Количество распознанных наименований в текстах аннотаций	Количество наименований, идентифицированных в базах данных
Химические соединения	83571	8848
Белки/гены	11878	7544
миРНК	5106	—
Заболевания	55080	51028
Аминокислотные замены и однонуклеотидные полиморфизмы	111224	10953

Таблица 3. Извлечённые ассоциации и взаимосвязи между объектами — химическими и биологическими молекулами, заболеваниями, полиморфизмами (аминокислотными заменами)

Типы объектов	Количество ассоциаций между указанными объектами, которые были идентифицированы в базах данных
Химическое соединение-Заболевание	1673
Химическое соединение-Химическое соединение	1227
Химическое соединение-Ген/белок	1204
Химическое соединение-Аминокислотная замена/Однонуклеотидный полиморфизм	397

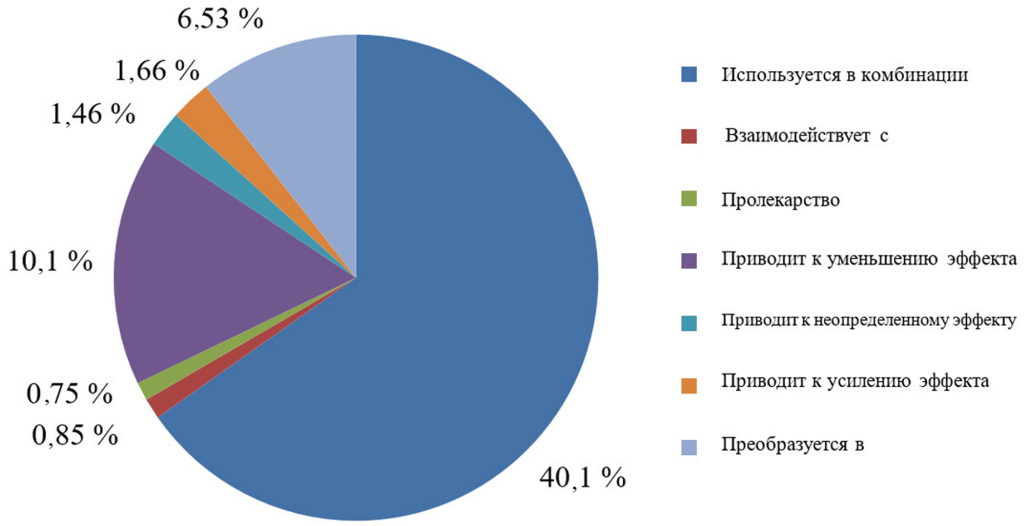

Рисунок 2. Типы ассоциаций между наименованиями химических соединений, извлечённых из текстов научных публикаций и включённых в базу знаний. Цветной вариант рисунка доступен в электронной версии статьи на сайте журнала.

Извлечённая из текстов и включённая в базу знаний информация будет полезной для понимания молекулярных механизмов действия известных противовирусных соединений, направленных на разнообразные вирусы, что важно при оценке химических соединений с потенциальной противовирусной активностью в случаях, когда химическое соединение с известной активностью в отношении одного вируса может быть эффективно и в отношении другого. Примером таких исследований может стать цидофовир, показанием к применению которого является инфекция цитомегаловируса [8]. Было показано, что он также активен в отношении вируса осповакцины (Vaccinia virus) [9, 10], инфекции коровьей оспы [9], а также в терапии папиллом, вызванных инфекцией вирусом папилломы человека [11]. Практическое использование разработанной нами базы знаний [4] даёт возможность определить, какие лекарственные препараты или их комбинации могут быть применены для терапии конкретных вирусных инфекций, а также активность каких химических соединений была изучена в отношении вирусных и человеческих мишеней. Более того, представленная база знаний предоставляет возможность более подробно ознакомиться с первоисточником — текстами научных публикаций — из которых были извлечены агрегированные в ней сведения.

ЗАКЛЮЧЕНИЕ

С применением ряда ранее разработанных алгоритмов, предназначенных для интеллектуального анализа текстов и обработки больших массивов данных, мы извлекли сведения об ассоциациях между химическими соединениями, мишенями вирусов и хозяина. Кроме того, мы определили аминокислотные замены и однонуклеотидные полиморфизмы, которые могут быть ассоциированы с лекарственной устойчивостью вирусов или восприимчивостью хозяина к вызываемой вирусом инфекции. Извлечённые знания были агрегированы и представлены в формате свободно-доступной базы знаний [4], которая включает информацию о двунаправленных ассоциациях между химическими соединениями, мишенями вирусов и хозяина, заболеваниями. Информация о молекулярных взаимодействиях между вирусом и организмом-хозяином крайне важна для всестороннего понимания механизмов вирусных инфекций и разработки новых терапевтических стратегий.

ФИНАНСИРОВАНИЕ

Исследование выполнено при поддержке Программы фундаментальных научных исследований в Российской Федерации на долгосрочный период (2021–2030 гг.) (№ 124050800018-9).

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая статья не содержит каких-либо исследований с участием людей или с использованием животных в качестве объектов.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

ЛИТЕРАТУРА

1. *Catacutan D.B., Alexander J., Arnold A., Stokes J.* (2024) Machine learning in preclinical drug discovery. Nat. Chem. Biol., **20**(8), 960–973. DOI: 10.1038/s41589-024-01679-1
2. *Tarasova O.A., Rudik A.V., Biziukova N.Y., Filimonov D.A., Poroikov V.V.* (2022) Chemical named entity recognition in the texts of scientific publications using the naïve Bayes classifier approach. J. Cheminformatics, **14**(1), 55. DOI: 10.1186/s13321-022-00633-4

3. *Biziukova N.Y., Ivanov S.M., Tarasova O.A.* (2024) Identification of proteins and genes associated with Hedgehog signaling pathway involved in neoplasm formation using text-mining approach. Big Data Mining Analytics, **7**(1), 107–130. DOI: 10.26599/BDMA.2023.9020007
4. База знаний о взаимодействии химических веществ и вирусов с организмом человека. Доступ получен 28 ноября, 2024, https://www.way2drug.com/viruses/nlp/ [Knowledge base on the interaction of chemicals and viruses with the human body. Retrieved November 28, 2024, from: https://www.way2drug.com/viruses/nlp/]
5. *Weber L., Sänger M., Münchmeyer J., Habibi M., Leser U., Akbik A.* (2021) HunFlair: An easy-to-use tool for state-of-the-art biomedical named entity recognition. Bioinformatics, **37**(17), 2792–2794. DOI: 10.1093/bioinformatics/btab042
6. *Tarasova O., Ivanov S., Filimonov D.A., Poroikov V.* (2020) Data and text mining help identify key proteins involved in the molecular mechanisms shared by SARS-CoV-2 and HIV-1. Molecules, **25**(12), 2944. DOI: 10.3390/molecules25122944
7. *Tarasova O., Biziukova N., Shemshura A., Filimonov D., Kireev D., Pokrovskaya A., Poroikov V.* (2023) Identification of molecular mechanisms involved in viral infection progression based on text mining: Case study for HIV infection. Int. J. Mol. Sci., **24**(2), 1465. DOI: 10.3390/ijms24021465
8. *Lea A.P., Bryson H.M.* (1996) Cidofovir. Drugs, **52**(2), 225–230. DOI: 10.2165/00003495-199652020-00006
9. *Quenelle D.C., Collins D.J., Kern E.R.* (2003) Efficacy of multiple- or single-dose cidofovir against vaccinia and cowpox virus infections in mice. Antimicrob. Agents Chemother., **47**(10), 3275–3280. DOI: 10.1128/AAC.47.10.3275-3280.2003
10. *de Clercq E.* (2001) Vaccinia virus inhibitors as a paradigm for the chemotherapy of poxvirus infections. Clin. Microbiol. Rev., **14**(2), 382–397. DOI: 10.1128/CMR.14.2.382-397.2001
11. *Petersen B.L, Buchwald C., Gerstoft J., Bretlau P., Lindeberg H.* (1998) An aggressive and invasive growth of juvenile papillomas involving the total respiratory tract. J. Laryngol. Otol., **112**(11), 1101–1104. DOI: 10.1017/s0022215100142586

Поступила в редакцию: 15. 10. 2024.
После доработки: 10. 12. 2024.
Принята к печати: 11. 12. 2024.

EXTRACTING INFORMATION ON VIRUS-HUMAN INTERACTIONS AND ON ANTIVIRAL COMPOUNDS BASED ON AUTOMATED ANALYSIS OF LARGE TEXT COLLECTIONS

O.A. Tarasova*, N.Yu. Biziukova, E.A. Stolbova, L.A. Stolbov, R.R. Taktashov, D.A. Karasev, N.S. Ionov, S.M. Ivanov, A.V. Dmitriev, A.V. Rudik, D.S. Druzhilovskiy, B.N. Sobolev, D.A. Filimonov, V.V. Poroikov

Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; *e-mail: olga.a.tarasova@gmail.com

The development of effective antivirals is of great importance due to the threat associated with the rapid spread of viral infections. The accumulation of data in scientific publications and in databases of biologically active compounds provides an opportunity to extract specific information about interactions between chemicals and their viral and host targets. This information can be used for elucidation of knowledge about potential antiviral activity of chemical compounds, their side effects and toxicities. Our study aims to extract knowledge about virus-host interactions and potential antiviral agents based on the mining of massive amounts of scientific publications. With a set of previously developed algorithms, we have extracted comprehensive information on virus-host interactions and chemical compounds that interact with both viral and host targets. We collected data on the interactions of several viruses, including hepatitis B and C viruses, SARS-CoV-2, influenza A and B, and herpes simplex viruses, with (1) the host (human body), (2) potential antiviral agents, and, also extracted information on the interactions between potential antiviral agents and host proteins. Based on the data analysis performed, we created a freely available knowledge base on the interaction of chemical compounds with viral proteins and their host targets, allowing the exploration of both well-studied and recently discovered novel virus-host-chemical-compound interactions.

The whole English version is available at http://pbmc.ibmc.msk.ru.

Key words: virus-host interactions; chemical compounds; biological activity; antivirals; text mining

Funding. This study was supported by the Program for Basic Research in the Russian Federation for a long-term period 2021–2030 (project no. 124050800018-9).

Received: 15.10.2024; revised: 10.12.2024; accepted: 11.12.2024.