

Извлечение знаний о биологической активности лекарственных препаратов: исследование на примере соединений с противовирусными свойствами

Бизюкова Н.*, Соболев Б., Карасев Д., Ионов Н., Сухачёв В., Такташов Р., Рудик А., Иванов С., Тарасова О.

Институт биомедицинской химии им. В.Н. Ореховича, Москва, Россия

* nad.smol@gmail.com

Ключевые слова: биологическая активность; интеллектуальный анализ текстов; противовирусные препараты; распознавание наименований объектов; извлечение ассоциаций

Мотивация и цель: Важным источником знаний, необходимых во многих областях биомедицинской науки, являются базы данных. Пополнение баз данных, как правило, производится вручную путем экспертной оценки первоисточников информации – текстов научных публикаций. Алгоритмы интеллектуального анализа текстов позволяют существенно уменьшить финансовые и временные затраты для пополнения баз данных по биомедицинской информации. В большинстве алгоритмов интеллектуального анализа текстов можно выделить следующие этапы: (1) отбор релевантных текстов (статьи, патенты), (2) распознавание наименований целевых объектов (биологических и/или химических), (3) поиск ассоциаций между распознанными объектами, и (4) обработка и представление извлеченной информации для дальнейшего использования. Большинство из разработанных и свободно доступных методов реализуют отдельные из перечисленных этапов. Немногие из интегральных алгоритмов зачастую фокусируются на решении конкретной проблемы (например, молекулярных механизмах [1] или репозиционировании лекарственных препаратов для терапии болезни Паркинсона [2]). Поэтому целью данной работы стала разработка подхода, позволяющего извлекать всесторонние сведения о биологической активности химических соединений из текстов научных публикаций, и его апробация на примере противовирусных препаратов.

Методы и алгоритмы: Реализация разрабатываемого метода произведена на скриптовом языке Python (версия 3.11) в виде отдельных функций. Отбор препаратов для дальнейшего поиска сведений о биологической активности проводился следующим образом: были сформированы автоматические запросы к базе данных ChEMBL на основе раздела Drug Indications, среди которых упоминается термин “viral diseases” – для противовирусных препаратов. Дополнительно из БД ChEMBL нами были извлечены идентификаторы, известные мишени и возможные синонимы найденных лекарственных препаратов. С целью поиска текстов для анализа использована база данных библиографической информации PubMed, а в качестве единицы текста – аннотации научных публикаций. В основу запроса в БД PubMed легли наименования лекарственных препаратов и их синонимы, а также MeSH-термины (1) “Animals” – для исследований *in vivo*, (2) “Cells” – для исследований *in vitro*. Дополнительно мы собрали коллекцию текстов, содержащих клинические данные; с этой целью была

использована фильтрация по типам публикаций. Литературные обзоры были исключены на этапе запросов. Распознавание наименований реализовано с применением ранее разработанного алгоритма на основе метода условных случайных полей (CRF) [3–6] и библиотеки flair [7], а также регулярных выражений и словарей. В результате нами были извлечены наименования следующих объектов: химические соединения, белки/гены, виды, клеточные линии, заболевания, мРНК и однонуклеотидные полиморфизмы. С целью извлечения ассоциаций между перечисленными объектами разработан список правил, которые основаны на использовании семантической связи в текстах, например, «inhibits», «regulate», «ameliorate». Мы выбрали типы ассоциаций между объектами, значимые, на наш взгляд, для понимания биологической активности лекарственных препаратов, такие как: (1) химическое соединение-химическое соединение (метаболизм, межлекарственные взаимодействия), (2) химическое соединение-белок (терапевтические мишени, изменение клинических показателей), (3) химическое соединение-заболевание (применение в терапии, лекарственно-индуцированные патологии), (4) белок-белок (молекулярные механизмы, комплексы биомаркеров заболеваний), (5) химическое соединение-вид (индуцирующее/ингибирующее воздействие на микроорганизм, экспериментальные условия тестирования), (6) заболевание-заболевание (сопутствующие патологические состояния). С целью разграничения механизмов и эффектов, причин и следствий воздействия одного объекта на другой, семантические правила были разделены на категории и приведены к одному, наиболее общему понятию. К примеру, семантические связи «ameliorate», «alleviate», «correct» для ассоциации «химическое соединение-заболевание» отражают некоторое положительное воздействие лекарственного препарата на патогенез. Для исключения дубликатов объектов (например, вследствие, использования синонимов) реализованы запросы по наименованиям их объектов к фактографическим базам данных и онтологиям (ChEMBL [8], UniProt [9], PubChem [10], NCBI Taxonomy [11], Human Disease Ontology [12]). Извлеченные ассоциации были представлены в табличном виде, где каждая ассоциация имеет формат: <объект 1><обобщенная семантическая связка><объект 2> (напр., «imidapril inhibit angiotensin-converting enzyme»). На основании полученных таблиц сформированы файлы, формат которых приемлем для графического представления сетей взаимосвязей с применением платформы CytoScape [13].

Результаты: После запросов к ChEMBL с термином «viral diseases» среди назначений препаратов было отобрано 111 низкомолекулярных соединений. Среди противовирусных агентов мы отобрали только соединения с прямым противовирусным действием, против ВИЧ, вирусов гепатита С и В, SARS-CoV 2, гриппа и простого герпеса. Для указанных вирусов в БД ChEMBL содержится более 60 % соединений с противовирусной активностью. Итого, после фильтрации по организму, в список для анализа вошли 77 противовирусных препаратов. Запросы к PubMed привели к значительным пересечениям текстов по отдельным противовирусным препаратам, что неудивительно, поскольку, зачастую, они назначаются в комбинации. После фильтрации повторяющихся текстов, в коллекцию вошли более 150 тысяч аннотаций, релевантных изучению биологической активности противовирусных препаратов. Средняя точность распознавания наименований биологических и химических объектов с применением комбинации методов достигла 0.87. Для поиска ассоциаций нами

сформулировано более 500 правил, которые позволяют идентифицировать взаимосвязи «объект–активность–объект» со средней точностью, равной 0.78. Сети извлеченных взаимосвязей визуализированы с применением программы Cytoscape. При ретроспективной верификации результатов обнаружены как сведения о препаратах, которые были тщательно изучены ранее, так и менее представленные в литературе материалы. В частности, найдены результаты исследований, проводимых в течение нескольких лет, о взаимосвязи длительного приема абакавира и проявлений атеросклероза и ремоделирования стенки артерий у молодых людей, живущих с ВИЧ-инфекцией [14–16]. Изучаются возможности репозиционирования противогрибкового препарата позаконазола в качестве противовирусного средства [17]. Дополнительно мы обнаружили, что в комбинации с ганцикловиром наблюдается синергетический эффект в подавлении цитомегаловирусной инфекции [17]. Для цидофовира – препарата, основным показанием к применению которого являются заболевания, вызванные цитомегаловирусом – при сочетанном применении с антидепрессантами – флувоксамином, венлафаксином, дулоксетином и бупропионом, и нормотимическими средствами – препаратами лития, карбамазепином, ламотриджином, показано влияние на замедление процесса их экскреции [18]. Была продемонстрирована активность адефовира (адефовир дипивоксил) в отношении клеточной линии OCI/AML-2 острого миелоидного лейкоза, а в комбинации с цитарабином (цитостатический препарат) исследователи наблюдали усиление эффекта [19].

Выводы: Разработанный интегральный подход применен к извлечению знаний о биологической активности низкомолекулярных соединений на примере противовирусных препаратов. Ретроспективная верификация результатов по данным литературы показала, что подход позволяет быстро и эффективно извлекать как широко изученные, так и менее известные эффекты и механизмы лекарственных препаратов.

Финансирование: Работа выполнена в рамках Программы фундаментальных научных исследований в Российской Федерации на долгосрочный период (2021–2030 гг.) (№ 124050800018-9).

Application of text mining methods to extract comprehensive information about the biological activity of drugs: case-study for antiviral compounds

Biziukova N.*, Sobolev B., Karasev D., Ionov N., Sukhachev V., Taktashov R., Rudik A., Ivanov S., Tarasova O.

Institute of Biomedical Chemistry, Moscow, Russia

* nad.smol@gmail.com

Key words: biological activity; text-mining; antiviral drugs; antihypertensive drugs; named entity recognition; relation extraction

Motivation and Aim: Databases are the main source of knowledge required in many areas of biomedical science. As a rule, databases are replenished manually by expert evaluation of information primary sources – scientific publications. Text mining

algorithms allow to significantly reduce time costs for replenishment of biomedical information databases. In the majority of text-mining algorithms the following stages can be distinguished: (1) selection of relevant texts (articles, patents), (2) named entity recognition (biological and/or chemical objects), (3) search for associations between the recognized entities, and (4) processing and representation of the extracted information for further use. Most of the developed and freely available methods are implementations of individual of the above listed steps; few of the integral algorithms often focus on solving a specific problem (e.g., molecular mechanisms [1], repositioning drugs for Parkinson's disease therapy [2]). Therefore, the aim of this work was to develop an approach that allows to extract comprehensive information about the biological activity of chemical compounds from the texts of scientific publications, and its validation within case-study for antiviral drugs.

Methods and Algorithms: The developed method was implemented using Python (version 3.11) in the form of separate functions. The selection of drugs for further search of information on biological activity was performed as follows: we generated automatic queries to the ChEMBL database based on the Drug Indications section, among which the term “viral diseases” was mentioned. Additionally, we extracted identifiers, known targets and possible synonyms of the found drugs from the ChEMBL database. For the purpose of text retrieval, the PubMed bibliographic information database was used for analysis, and abstracts of scientific publications were chosen as the text unit. The PubMed database query was based on drug names and their synonyms, as well as MeSH terms (1) “Animals” for in vivo studies, (2) “Cells” for in vitro studies. Additionally, we assembled a collection of texts containing clinical data; for this purpose, filtering by publication type was used. Literature reviews were excluded with the query. Named entity recognition was implemented using a previously developed algorithm based on the conditional random field (CRF) method [3–6] and the flair library [7], as well as regular expressions and dictionaries. As a result, we extracted the names of the following entities: chemical compounds, proteins/genes, species, cell lines, diseases, miRNAs and single nucleotide polymorphisms. In order to extract associations between the listed objects, we developed a set of rules that are based on the use of semantic links in texts (e.g., “inhibits”, “regulate”, “ameliorate”). We selected the types of associations between objects that we believe are significant for understanding the biological activity of drugs such as: (1) chemical compound-chemical compound (metabolism, drug-drug interactions), (2) chemical compound-protein (therapeutic targets, changes in clinical parameters), (3) chemical compound-disease (therapeutic applications, drug-induced pathologies), (4) protein-protein (molecular mechanisms, disease biomarker complexes), (5) chemical compound-species (inducing/inhibitory effects on microorganism, experimental testing conditions), (6) disease-disease (associated pathologic conditions). In order to distinguish between mechanisms and effects, causes and consequences of the influence of one object on another, semantic rules were categorized and brought to one, the most general concept. For example, the semantic relations “ameliorate”, “alleviate”, “correct” for the association “chemical compound-disease” reflect some positive effect of the drug on pathogenesis. To avoid duplicate entities (e.g., due to, for example, the use of synonyms), queries on the names of their entities to databases and ontologies (ChEMBL [8], UniProt [9], PubChem [10], NCBI Taxonomy [11], Human Disease Ontology [12]) were implemented. The extracted associations were tabulated, where each association has the format: <object 1><generalized semantic conjunction><object 2> (e.g., “imidapril inhibit angiotensin-

converting enzyme”). Based on the obtained tables, files were generated, the format of which is acceptable for graphical representation of interaction networks using the CytoScape platform [13].

Results: After queries to ChEMBL for the designation “viral diseases”, 111 low molecular weight compounds were selected. Among antiviral agents, we selected only compounds with direct antiviral activity against HIV, hepatitis C and B, SARS-CoV-2, influenza and herpes simplex. For the above viruses, the ChEMBL database contained more than 60 % of compounds with antiviral activity. In total, after filtering by organism, 77 antiviral drugs were included in the list for analysis. Queries to PubMed resulted in significant overlap of texts for both individual antiviral drugs, which is not surprising since, often, they are prescribed in combination. After filtering out repetitive texts, the collection included more than 150,000 abstracts relevant to the study of the biological activity of antiviral drugs. The average accuracy of named entity recognition for chemical and biological objects using the combination of methods reached 0.87. To retrieve associations, we formulated more than 500 rules that allow us to identify object–acti-vity–object chains with an average accuracy of 0.78. The extracted association networks were visualized using CytoScape software. Retrospective verification of the results revealed both information on drugs that had been thoroughly studied earlier and less represented in the literature. In particular, we found the results of studies conducted over several years on the relationship between long-term administration of abacavir and manifestations of atherosclerosis and arterial wall remodeling in young people living with HIV infection [14–16]. The repositioning of the antifungal drug posaconazole as an antiviral agent is being investigated [17]. Additionally, we found that in combination with ganciclovir, a synergistic effect was observed in suppressing cytomegalovirus infection [17]. For cidofovir, a drug whose main indication for use is diseases caused by cytomegalovirus, when used in combination with antidepressants – fluvoxamine, venlafaxine, duloxetine and bupropion, and normotimics – lithium, carbamazepine, lamotrigine, an effect on slowing down their excretion was shown [18]. The activity of adefovir (adefovir dipivoxil) was demonstrated against the OCI/AML-2 cell line of acute myeloid leukemia, and in combination with cytarabine (a cytostatic drug), researchers observed an enhanced effect [19].

Conclusion: The developed integral approach was applied to the extraction of knowledge about the biological activity of low molecular weight compounds using antiviral drugs as an example. Retrospective verification of the results against literature data showed that the approach allows for fast and efficient extraction of both widely studied and less known effects and mechanisms of drugs.

Funding: The work was performed within the framework of the Program for Basic Research in the Russian Federation for a long-term period (2021–2030) (No.124050800018-9).

Список литературы/References

1. Bachman J.A., Gyori B.M., Sorger P.K. Automated assembly of molecular mechanisms at scale from text mining and curated databases. *Mol Syst Biol.* 2023;19(5):e11325. doi 10.15252/msb.202211325
2. Tandra G. et al. Literature-Based Discovery Predicts Antihistamines Are a Promising Repurposed Adjuvant Therapy for Parkinson's Disease. *Int J Mol Sci.* 2023;24(15):12339. doi 10.3390/ijms241512339
3. Tarasova O. et al. Identification of Molecular Mechanisms Involved in Viral Infection Progression Based on Text Mining: Case Study for HIV Infection. *Int J Mol Sci.* 2023;24(2):1465. doi 10.3390/ijms24021465
4. Tarasova O. et al. Extraction of Data on Parent Compounds and Their Metabolites from Texts of Scientific Abstracts. *J Chem Inf Model.* 2021;61(4):1683-1690. doi 10.1021/acs.jcim.0c01054

5. Biziukova N. et al. Automated Extraction of Information From Texts of Scientific Publications: Insights Into HIV Treatment Strategies. *Front Genet.* 2020;11:618862. doi 10.3389/fgene.2020.618862
6. Biziukova N., Ivanov S., Tarasova O. Identification of Proteins and Genes Associated with Hedgehog Signaling Pathway Involved in Neoplasm Formation Using Text-Mining Approach. *Big Data Mining Analytics.* 2024;7(1):107-130. doi 10.26599/BDMA.2023.9020007
7. Weber L. et al. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics.* 2021;37(17):2792-2794. doi 10.1093/bioinformatics/btab042
8. Davies M. et al. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* 2015;43(W1):W612-W620. doi 10.1093/nar/gkv352
9. Apweiler R. et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 2004;32:D115-D119. doi 10.1093/nar/gkh131
10. Kim S. et al. PubChem 2023 update. *Nucleic Acids Res.* 2023;51(D1):D1373-D1380. doi 10.1093/nar/gkac956
11. Federhen S. The NCBI Taxonomy database. *Nucleic Acids Res.* 2012;40:D136-D143. doi 10.1093/nar/gkr1178
12. Schriml L.M. et al. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* 2012;40:D940-D946. doi 10.1093/nar/gkr972
13. Shannon P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498-2504. doi 10.1101/gr.1239303
14. Dirajlal-Fargo S. et al. Longitudinal Changes in Subclinical Vascular Disease in Ugandan Youth With Human Immunodeficiency Virus. *Clin Infect Dis.* 2023;76(3):e599-e606. doi 10.1093/cid/ciac686
15. Pocock M.O. et al. Pathophysiology of ischaemic heart disease. *Curr Opin HIV AIDS.* 2017;12(6):548-553. doi 10.1097/COH.0000000000000411
16. Hsue P.Y. et al. Association of abacavir and impaired endothelial function in treated and suppressed HIV-infected patients. *AIDS.* 2009;23(15):2021-7. doi 10.1097/QAD.0b013e32832e7140
17. Mansuri Z. et al. Drug Interactions Between Psychotropic Medications and Treatment of Mpox (tecovirimat and cidofovir). *Prim Care Companion CNS Disord.* 2024;26(1):23lr03608. doi 10.4088/PCC.23lr03608
18. Mercorelli B. et al. The Clinically Approved Antifungal Drug Posaconazole Inhibits Human Cytomegalovirus Replication. *Antimicrob Agents Chemother.* 2020;64(10):e00056-20. doi 10.1128/AAC.00056-20
19. Khoury H. et al. Octadecyloxyethyl Adefovir Exhibits Potent *in vitro* and *in vivo* Cytotoxic Activity and Has Synergistic Effects with Ara-C in Acute Myeloid Leukemia. *Chemotherapy.* 2018;63(4):225-237. doi 10.1159/000491705