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.

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

DOI: 10.18097/PBMC20247006469

INTRODUCTION

The development of novel, more safe and effective drugs requires high costs and long time that should be spent on the discovery of new pharmaceutical agents [1]. Therefore, development of computational approaches is in high demand to decrease time and costs spent for research in drug discovery. Computational methods utilize experimental data published in scientific articles and stored in databases of biologically active compounds. The rapid growth of scientific research in the field of drug discovery and development lead to the accumulation of a large number of publications. The scientific publications are the primary source of information about biological activities of chemical compounds and disease pathogenesis. Therefore, knowledge extraction from the main text of the publication and its abstract can be helpful for enriching the information about biological activity of chemical compounds, including mechanisms, their pharmacological and side effects, toxicity, drug indication, drug-drug interactions and drug repurposing. In addition, the extraction of knowledge about molecular mechanisms of diseases and disorders can be profitable for the development of novel therapeutic strategies and discovery of new drugs. Such information is particularly important for the development of novel, effective, and safe antivirals in a short timeframe in the event of new threats arising from the rapid spread of a new or rare virus.

Previously we developed algorithms aimed at named entity recognition in the scientific texts [2, 3] and for extracting relationships between the recognized entities [3].

The aim of this study was to extract information about the interactions between the virus and the host (human) organism, as well as the mechanisms and effects of potential antiviral compounds, based on algorithms of automated text analysis, that we developed earlier. Also, we performed the search of extracted data in the databases of biologically active compounds and protein sequences. Based on the extracted knowledge, we have created a freely available knowledge base [4] that provides information on the interaction between viruses and their hosts, and on chemical compounds interacting with viral and host targets with the cross-links on the external database (ChEMBL v.34).

METHODS

The approach for extracting information of antiviral chemical compounds is based on the automated selection of relevant publications using keywords, followed by the recognition of named entities and the extraction of associations between them. The proposed computational method for chemical and biological named entity recognition is based on a combination of the methods: conditional random fields [3], a self-developed naïve Bayesian approach [2], and the Hunflair method [5]. We assumed that a given entity belonged to the name of a chemical compound, biological macromolecule or disease, if it was recognized successfully by at least two of the three machine learning algorithms mentioned above. We used a combination of machine learning methods, followed by the use of semi-automatically prepared dictionaries containing synonyms and annotations of chemical compounds and biological molecules. These dictionaries are compiled with automated queries to the ChEMBL and UniProt databases. Such a strategy provides reasonable accuracy of named entity recognition in five-fold cross-validation procedure (Table 1). Additionally, we have prepared and used the set of filters to remove false positive named entities that take into account matching the type of named entity (i.e. biological or chemical molecule) using the developed dictionaries, and by removing the incorrect characters included in the named entity, such as “@”, “!”, “<”, etc. We used regular expressions to detect miRNAs, single nucleotide polymorphisms and amino acid substitutions.

The performance of named entity recognition using our method is, in general, comparable with the most recent and powerful method, HunFlair, with lower performance in diseases and disorders recognition. The comparison results might have some bias due to usage of different training and test corpora. Enrichment of dictionaries and training based on several corpora may be helpful in increase the performance of recognition.

The associations and relationships between chemical and biological objects (viral and host proteins, genes, miRNAs) are extracted using three strategies (Fig. 1). First, named entities are extracted from the set of publications strictly relevant to the scientific topic [6]. This method allows the identification of general concepts of the scientific topic or specific molecular mechanisms of the disease under study, that are described in the articles. Using such approach, we investigated possible types of virus-host interactions that may be shared between SARS-CoV-2 and several other viruses, and found some molecular mechanisms that might be shared between HIV-1 and dengue virus infections [6]. The second approach is based on the usage of the so-called template phrases in the texts of abstracts and was evaluated in the case studies of extracting key target proteins involved in the mechanisms of HIV infection progression [7] and the Hedgehog-pathway associated mechanisms of cancer [3]. Third, we use co-occurrences for extracting associations between entities.

Table 1. Accuracy of named entity recognition based on machine learning followed by terms search in dictionaries and the comparison with earlier developed method

Type of named entity	Our method			HunFlair [5]		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Chemicals	0.89	0.83	0.86	0.88	0.88	0.88
Proteins and genes	0.87	0.84	0.85	0.84	0.86	0.85
Diseases and disorders	0.84	0.79	0.81	0.86	0.87	0.86

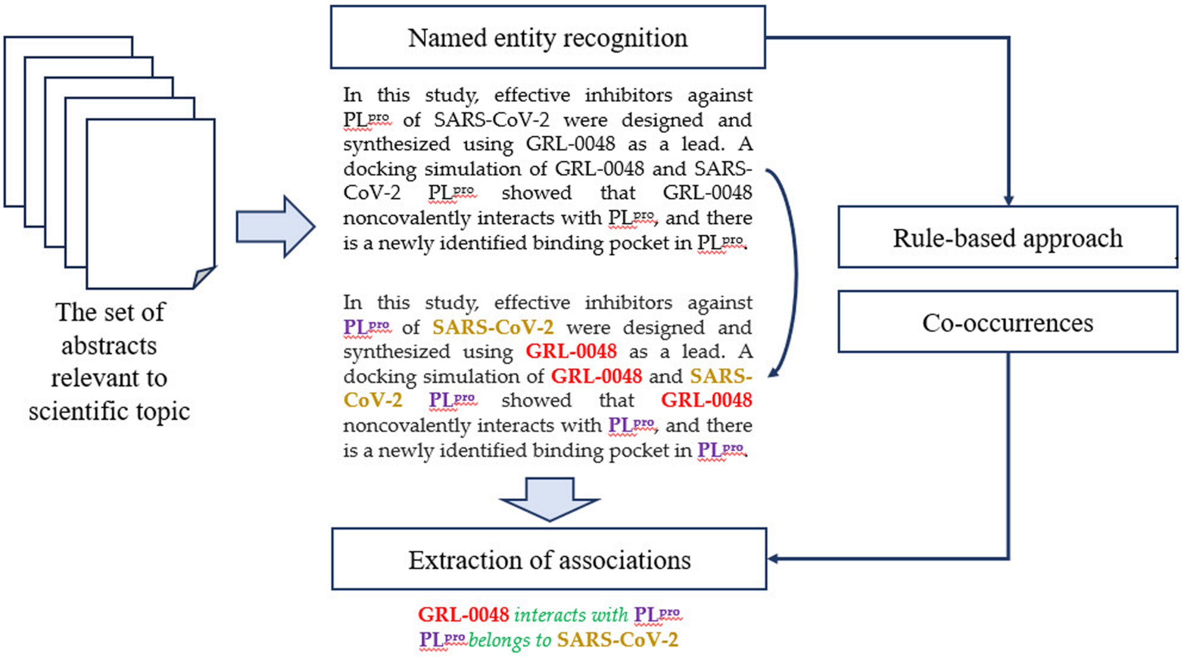
The set of abstracts relevant to scientific topic
Named entity recognition
In this study, effective inhibitors against PLpro of SARS-CoV-2 were designed and synthesized using GRL-0048 as a lead. A docking simulation of GRL-0048 and SARS-CoV-2 PLpro showed that GRL-0048 noncovalently interacts with PLpro, and there is a newly identified binding pocket in PLpro.
In this study, effective inhibitors against PLpro of SARS-CoV-2 were designed and synthesized using GRL-0048 as a lead. A docking simulation of GRL-0048 and SARS-CoV-2 PLpro showed that GRL-0048 noncovalently interacts with PLpro, and there is a newly identified binding pocket in PLpro.
Rule-based approach
Co-occurrences
Extraction of associations
GRL-0048 interacts with PLpro
PLpro belongs to SARS-CoV-2

Figure 1. Principles of extracting associations and relationships between entities in text.

In the first strategy, we collect a set of texts strictly relevant to the topic of the interest. We consider all named entities corresponding to biological molecules (proteins, genes, etc.) recognized in the texts of publications strictly related to the specific subject (i.e. disease, pathological condition, etc.). The selection of relevant publications is typically done either using a set of keywords, MeSH terms, or based on machine learning models (support vector machines, random forest, artificial neural networks) aimed at selecting publications relevant to the inhibition of SARS-CoV-2 replication. We tested this strategy in the selection of relevant publications using Long Short-Term Memory (LSTM) recurrent neural networks (RNN). The performance of relevant publications recognition (F_1-score) was evaluated by dividing the whole set in two equal sets (by 50% of the whole set into training and test sets respectively). F_1-score of relevant publication recognition was 0.84 for classification of publication sets with publications relevant to experimental testing of SARS-CoV-2 inhibitors in cell-based and cell-free assays.

The second approach uses a set of rules, which were implemented with the so-called pattern phrases that described particular relations between the objects corresponding to the identified entities in the text. We have demonstrated that the accuracy of association recognition using this approach, F_1-score = 0.84, is consistent with the accuracy of most different association detection methods [3]. We define a relation as a link between two objects in a form of pattern phrase defining the semantic of their co-existence in the framework of a sentence (text fragment) in a way that corresponds with their previously recognized entity. For example, the indication in the text that a certain substance S interacts with a particular protein E can help to determine the relation "S interacts with E" based on the analysis of the pattern phrases and their synonyms. If the named entities are found in the text of a scientific publication based on co-occurrence, but the particular relationship is not defined, we consider it as association. In some cases, relations between entities in sentences cannot be extracted using phrase patterns, but such associations can be assumed based on context (e.g., names of proteins and viruses to which they belong). In such cases we used a co-occurrence of terms as an approach for extracting associations.

The third strategy is based on the detection of two objects in the same abstract text that are relevant to a particular scientific topic. More frequent co-occurrences correspond to the well-studied associations, and rare co-occurrences may represent newly discovered associations, so we do not distinguish statistically significant differences in occurrence frequencies and consider every association found as a result of co-occurrence identification.

We extracted information about the interactions of chemicals with viruses and host targets and to build a freely available knowledge base [4].

RESULTS AND DISCUSSION

We collected data for a number of viruses, including hepatitis B and C viruses, SARS-CoV-2, influenza A and B, and herpes simplex virus as having the high impact on the human health and analyzed in multiple publications. We extracted data on the interactions of these viruses with (1) the host (human body), (2) potential antiviral drugs. The data on interaction with both host and viral targets can be represented in literature as the results of various *in vitro* experiments in both cell-based and cell-free assays: these data were extracted using the developed approach. We also extracted the associations on interactions between known antiviral drugs and the host (human body), and the information on interactions between potential antiviral drugs and host proteins. To extract associations and relationships between named entities associated with the viruses, the host and potential antiviral chemical compounds, we collected texts relevant to (1) virus-host interactions and (2) the *in vitro* and *in vivo* studies of 77 known drugs and their synonyms retrieved from the ChEMBL database (v.34), which were used in the queries based on MeSH terms and keywords. A total of 171735 texts were collected and used for information retrieval. We extracted over 83000 unique chemical named entities, genes, and 11878 proteins; 8848 and 7544 of them were found in the databases, respectively (Table 2).

Based on the data analysis performed, we have created a freely available knowledge base on the interaction of chemical compounds with viral proteins and their host targets. The current version

Table 2. Extracted data on chemical named entities, proteins, genes, and diseases

Entity type	Number of entities extracted from abstracts of publications	Number of entities found in the databases
Chemicals	83571	8848
Proteins/ Genes	11878	7544
miRNAs	5106	—
Disease	55080	51028
Amino acid substitutions and SNPs	111224	10953

of the knowledge base contains 5683 records on entities identified in the publications and the relations and/or associations between them. The knowledge base includes the names and identifiers with the cross-links to the ChEMBL database of the objects corresponding to named entity in the texts of publications and the types of relations identified based on the texts of publications. For instance, the relations described the semantic links between a chemical and a disease may include “may cause”, “used in therapy of”, “has positive effect on”, “increases risk of”, etc. Knowledge base is available on the Internet [4]. The knowledge base includes the 5430 relations between the identified entities and 253 associations, for which exact type of relations is unknown.

Number of relations and associations between particular entities is provided in Table 3.

In particular, the number of relations between entities reflects the specific topic of the scientific texts retrieved for analysis. Since most of the publications were retrieved by the query that included or was associated with the name of an antiviral drug, the majority of the entity relations were found to be between chemical and disease and between a pair of chemicals. At the same time, some publications were relevant to virus-host interactions and *in vitro* or *in vivo* testing of the activity of chemical compounds, so we also extracted over 1200 relations between chemicals and proteins/genes that reflect the particular molecular mechanism of an antiviral chemical compound or a particular viral disease.

The types of relations for chemicals-chemicals are provided in the Figure 2. They can be particularly useful for selection of the most promising pharmaceutical agents and drugs with antiviral effect and their combinations.

Extracted data implemented in the knowledge base can be useful for understanding the molecular mechanisms of known antiviral agents against different viruses, which may be particularly useful for evaluating candidates against different viral infections in cases where a chemical compound or combinations is active against one virus is found to be active against another virus. An example of such cases can be studies of cidofovir, which is active against human cytomegalovirus infection [8], in its use against vaccinia virus [9, 10], cowpox infection [9], and in treatment of papillomas associated with human papillomavirus infection [11]. We believe that the developed knowledge base can also be useful in understanding which drugs or drug combinations can be used to treat specific viral infections, which chemical compounds have been evaluated against specific viral and host targets, and provides the ability to look up the source of the information retrieved.

Table 3. Extracted data on chemicals-virus-host interactions

Entity type	Number of relations found in literature between entities identified in the databases
Chemical-disease	1673
Chemical-chemical	1227
Chemical-gene/protein	1204
Chemical-SNP	397

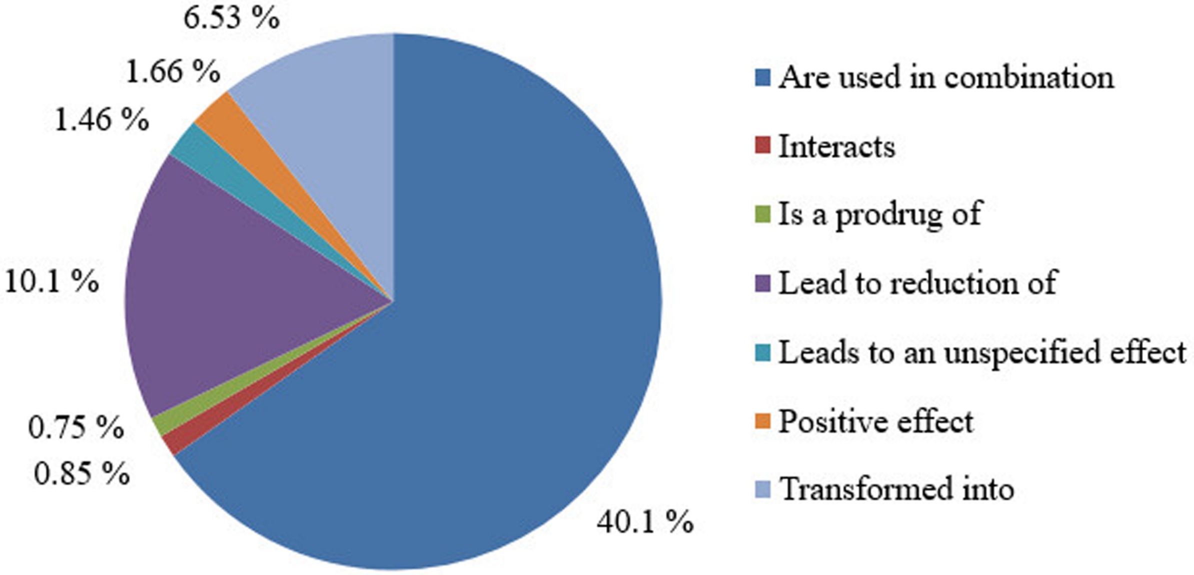

Figure 2. Types of relations between chemicals extracted from the literature and those included in the knowledge base. The color version of the figure is available in the electronic version of the article.

CONCLUSIONS

By application of the set of earlier developed algorithms of text and data mining, we extracted the knowledge about interactions between chemicals, virus, and host targets. Additionally, we extracted information about amino acid substitutions that could be associated with viral drug resistance, SNPs that could be associated with susceptibility to a particular drug or viral infection. The extracted knowledge was shaped as a freely available knowledge base [4] containing information about pairwise relations between chemicals, viral and host targets, diseases. Information on virus-host interactions is useful for a comprehensive understanding of viral infection mechanisms and the development of new therapeutic strategies.

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).

COMPLIANCE WITH ETHICAL STANDARDS

This article does not contain any research involving humans or the use of animals as objects.

CONFLICT OF INTEREST

The authors declare no conflicts of interests.

REFERENCES

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. 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

Received: 15. 10. 2024.
Revised: 10. 12. 2024.
Accepted: 11. 12. 2024.

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.

Полный текст статьи на русском языке доступен на сайте журнала (http://pbmc.ibmc.msk.ru).

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

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

Поступила в редакцию: 15.10.2024; после доработки: 10.12.2024; принята к печати: 11.12.2024.