XXX Symposium on Bioinformatics
and Computer-Aided Drug Discovery

ИБМХ

IBMC
Institute of Biomedical Chemistry

ANALYSIS OF CHEMICALS-VIRUS-HOST INTERACTIONS BASED ON LARGE-SCALE BIOMEDICAL TEXT AND DATA MINING

Olga Tarasova, PhD

Laboratory of Structure-Function Based Drug Design, Senior Scientist
Laboratory of Big Data Analysis for Digital Pharmacology, Head

Institute of Biomedical Chemistry, Moscow, Russia

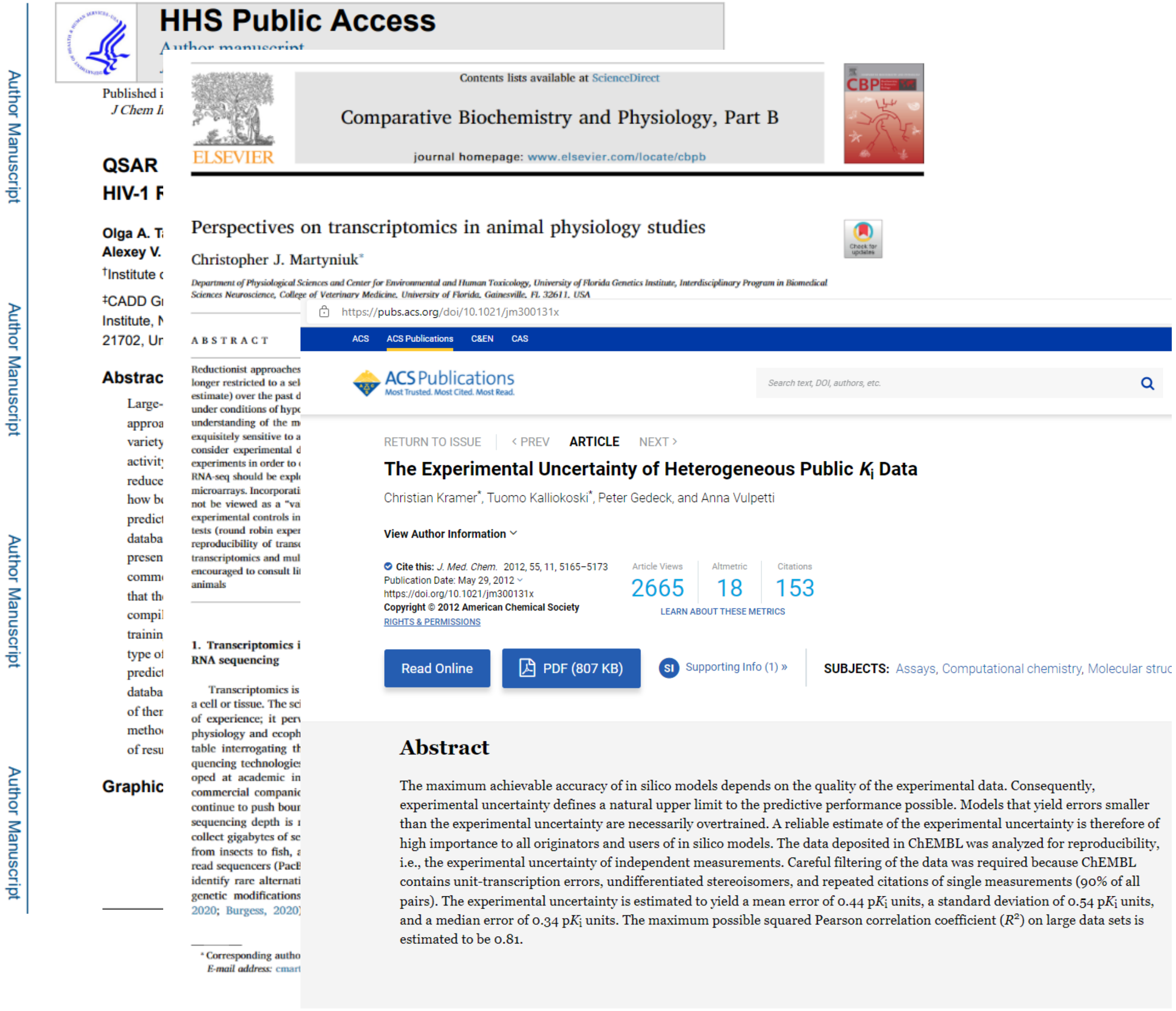
Large-scale biomedical data are available in the databases and scientific publications

DRUGBANK Online

BindingDB

ChEMBL

PubChem



Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Published i
J Chem I

QSAR
HIV-1 F

Olga A. T
Alexey V.
†Institute c
‡CADD G
Institute, N
21702, Ur

Abstrac

Large-
approa
variety
activit
reduce
how b
predict
databa
presen
comm
that th
compil
trainin
type o
predict
databa
of ther
metho
of resu

Graphic

ELSEVIER

Contents lists available at ScienceDirect

Comparative Biochemistry and Physiology, Part B

journal homepage: www.elsevier.com/locate/cbpb

CBP

Perspectives on transcriptomics in animal physiology studies

Christopher J. Martyniuk*

Department of Physiological Sciences and Center for Environmental and Human Toxicology, University of Florida Genetics Institute, Interdisciplinary Program in Biomedical Sciences Neuroscience, College of Veterinary Medicine, University of Florida, Gainesville, FL 32611, USA

Check for updates

ABSTRACT

Reductionist approaches
longer restricted to a sel
estimate) over the past d
under conditions of hypo
understanding of the m
exquisitely sensitive to a
consider experimental d
experiments in order to
RNA-seq should be expl
microarrays. Incorporati
not be viewed as a "va
experimental controls in
tests (round robin exper
reproducibility of trans
transcriptomics and mul
encouraged to consult li
animals

1. Transcriptomics i
RNA sequencing

Transcriptomics is
a cell or tissue. The sci
of experience; it perv
physiology and ecoph
table interrogating th
quencing technologies
oped at academic in
commercial compani
continue to push bour
sequencing depth is
collect gigabytes of se
from insects to fish,
read sequencers (PacE
identify rare alternati
genetic modifications
2020; Burgess, 2020

* Corresponding autho
E-mail address: cmart

https://pubs.acs.org/doi/10.1021/jm300131x

ACS Publications
Most Trusted. Most Cited. Most Read.

The Experimental Uncertainty of Heterogeneous Public K_i Data

Christian Kramer*, Tuomo Kalliokoski*, Peter Gedeck, and Anna Vulpetti

View Author Information

Cite this: *J. Med. Chem.* 2012, 55, 11, 5165–5173
Publication Date: May 29, 2012
https://doi.org/10.1021/jm300131x
Copyright © 2012 American Chemical Society
RIGHTS & PERMISSIONS

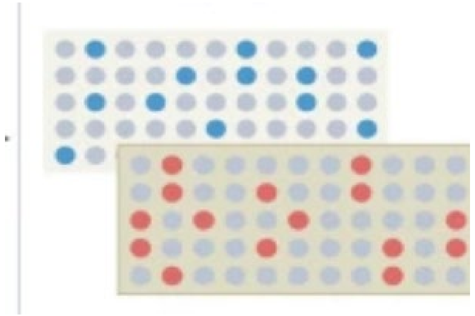
Article Views	Altmetric	Citations
2665	18	153

LEARN ABOUT THESE METRICS

SUBJECTS: Assays, Computational chemistry, Molecular struc

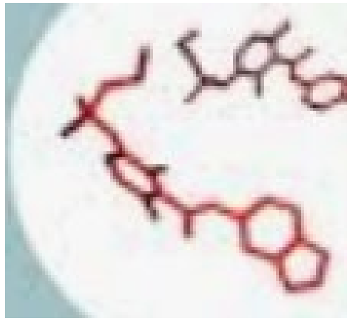
Abstract

The maximum achievable accuracy of in silico models depends on the quality of the experimental data. Consequently, experimental uncertainty defines a natural upper limit to the predictive performance possible. Models that yield errors smaller than the experimental uncertainty are necessarily overtrained. A reliable estimate of the experimental uncertainty is therefore of high importance to all originators and users of in silico models. The data deposited in ChEMBL was analyzed for reproducibility, i.e., the experimental uncertainty of independent measurements. Careful filtering of the data was required because ChEMBL contains unit-transcription errors, undifferentiated stereoisomers, and repeated citations of single measurements (90% of all pairs). The experimental uncertainty is estimated to yield a mean error of 0.44 pK_i units, a standard deviation of 0.54 pK_i units, and a median error of 0.34 pK_i units. The maximum possible squared Pearson correlation coefficient (R^2) on large data sets is estimated to be 0.81.

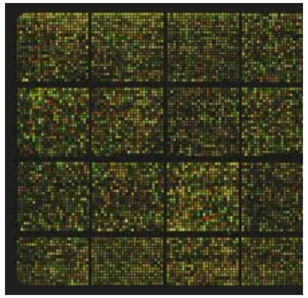


Genome analysis/ genome-wide association studies

Epigenome-wide/ Methylome-wide association studies

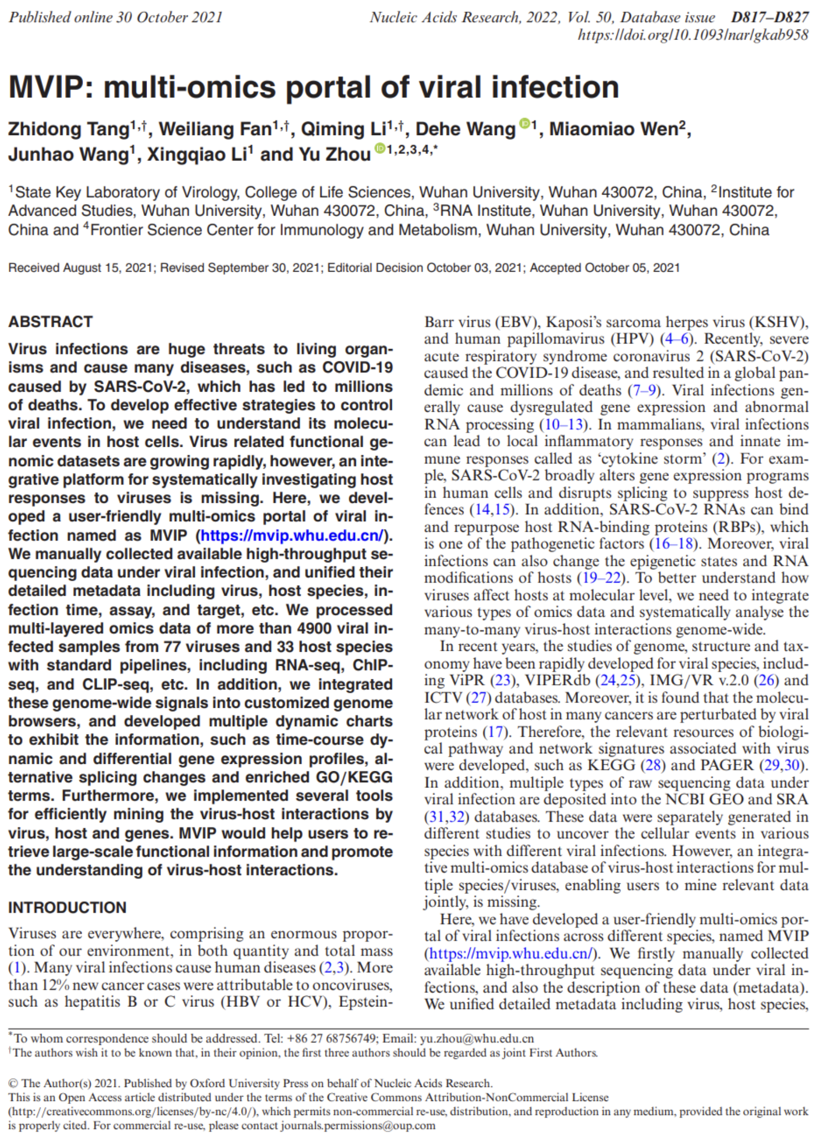


Metabolome-wide association studies



Transcriptome data analysis/ transcriptome wide association studies

Proteome-wide association studies



Published online 30 October 2021 *Nucleic Acids Research, 2022, Vol. 50, Database issue* ***D817–D827***
https://doi.org/10.1093/nar/gkab958

MVIP: multi-omics portal of viral infection

Zhidong Tang[1,†], Weiliang Fan[1,†], Qiming Li[1,†], Dehe Wang[1], Miaomiao Wen[2], Junhao Wang[1], Xingqiao Li[1] and Yu Zhou[1,2,3,4,*]

[1]State Key Laboratory of Virology, College of Life Sciences, Wuhan University, Wuhan 430072, China, [2]Institute for Advanced Studies, Wuhan University, Wuhan 430072, China, [3]RNA Institute, Wuhan University, Wuhan 430072, China and [4]Frontier Science Center for Immunology and Metabolism, Wuhan University, Wuhan 430072, China

Received August 15, 2021; Revised September 30, 2021; Editorial Decision October 03, 2021; Accepted October 05, 2021

ABSTRACT

Virus infections are huge threats to living organisms and cause many diseases, such as COVID-19 caused by SARS-CoV-2, which has led to millions of deaths. To develop effective strategies to control viral infection, we need to understand its molecular events in host cells. Virus related functional genomic datasets are growing rapidly, however, an integrative platform for systematically investigating host responses to viruses is missing. Here, we developed a user-friendly multi-omics portal of viral infection named as MVIP (https://mvip.whu.edu.cn/). We manually collected available high-throughput sequencing data under viral infection, and unified their detailed metadata including virus, host species, infection time, assay, and target, etc. We processed multi-layered omics data of more than 4900 viral infected samples from 77 viruses and 33 host species with standard pipelines, including RNA-seq, ChIP-seq, and CLIP-seq, etc. In addition, we integrated these genome-wide signals into customized genome browsers, and developed multiple dynamic charts to exhibit the information, such as time-course dynamic and differential gene expression profiles, alternative splicing changes and enriched GO/KEGG terms. Furthermore, we implemented several tools for efficiently mining the virus-host interactions by virus, host and genes. MVIP would help users to retrieve large-scale functional information and promote the understanding of virus-host interactions.

INTRODUCTION

Viruses are everywhere, comprising an enormous proportion of our environment, in both quantity and total mass (1). Many viral infections cause human diseases (2,3). More than 12% new cancer cases were attributable to oncoviruses, such as hepatitis B or C virus (HBV or HCV), Epstein-Barr virus (EBV), Kaposi's sarcoma herpes virus (KSHV), and human papillomavirus (HPV) (4–6). Recently, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) caused the COVID-19 disease, and resulted in a global pandemic and millions of deaths (7–9). Viral infections generally cause dysregulated gene expression and abnormal RNA processing (10–13). In mammalians, viral infections can lead to local inflammatory responses and innate immune responses called as 'cytokine storm' (2). For example, SARS-CoV-2 broadly alters gene expression programs in human cells and disrupts splicing to suppress host defences (14,15). In addition, SARS-CoV-2 RNAs can bind and repurpose host RNA-binding proteins (RBPs), which is one of the pathogenetic factors (16–18). Moreover, viral infections can also change the epigenetic states and RNA modifications of hosts (19–22). To better understand how viruses affect hosts at molecular level, we need to integrate various types of omics data and systematically analyse the many-to-many virus-host interactions genome-wide.

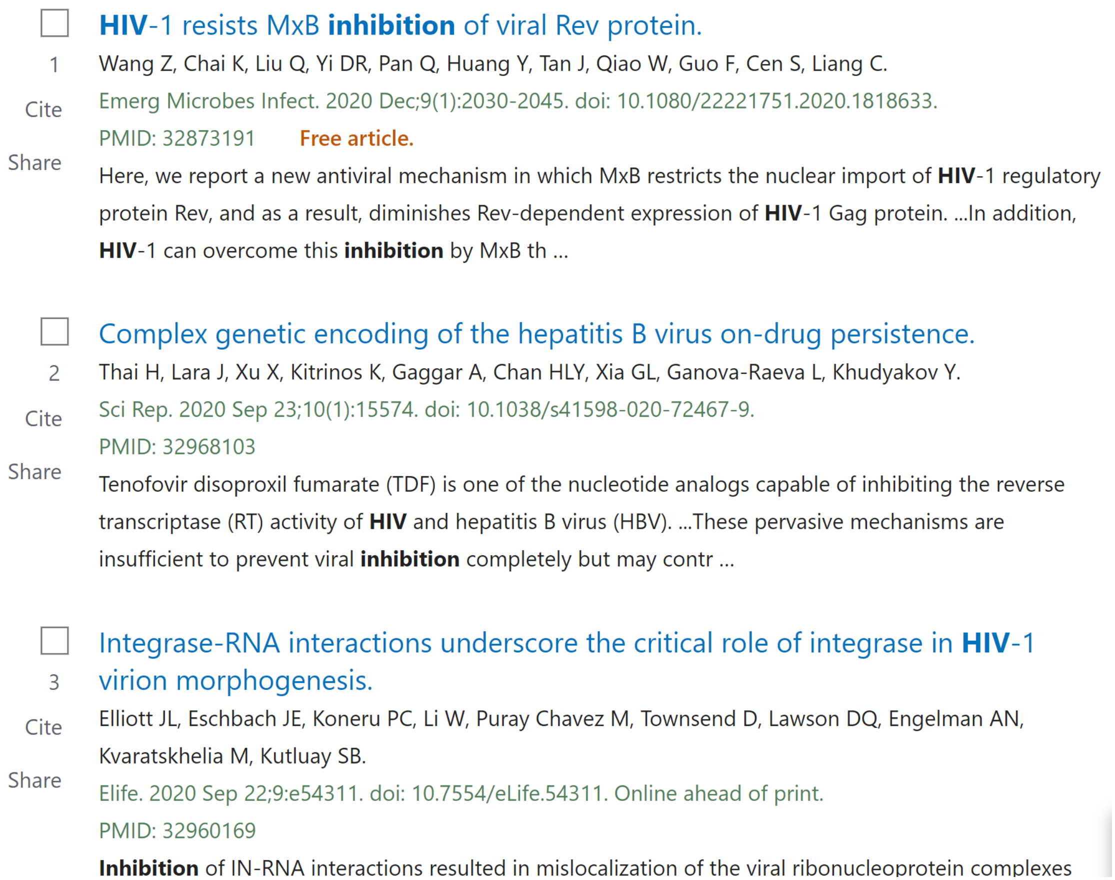
In recent years, the studies of genome, structure and taxonomy have been rapidly developed for viral species, including ViPR (23), VIPERdb (24,25), IMG/VR v.2.0 (26) and ICTV (27) databases. Moreover, it is found that the molecular network of host in many cancers are perturbated by viral proteins (17). Therefore, the relevant resources of biological pathway and network signatures associated with virus were developed, such as KEGG (28) and PAGER (29,30). In addition, multiple types of raw sequencing data under viral infection are deposited into the NCBI GEO and SRA (31,32) databases. These data were separately generated in different studies to uncover the cellular events in various species with different viral infections. However, an integrative multi-omics database of virus-host interactions for multiple species/viruses, enabling users to mine relevant data jointly, is missing.

Here, we have developed a user-friendly multi-omics portal of viral infections across different species, named MVIP (https://mvip.whu.edu.cn/). We firstly manually collected available high-throughput sequencing data under viral infections, and also the description of these data (metadata). We unified detailed metadata including virus, host species,

[*] To whom correspondence should be addressed. Tel: +86 27 68756749; Email: yu.zhou@whu.edu.cn
[†] The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Integration of data obtained in the specific E/G/M/T/PWAS may be helpful for the comprehensive understanding of particular diseases mechanisms, and the methods for meta-analysis were proposed and described [Nan Wang, Shuilin Jin; Methods Mol. Biol., 2022].

Text and Data Mining Tools Can Help Researchers



□ 1 Cite Share

HIV-1 resists MxB **inhibition** of viral Rev protein.

Wang Z, Chai K, Liu Q, Yi DR, Pan Q, Huang Y, Tan J, Qiao W, Guo F, Cen S, Liang C.

Emerg Microbes Infect. 2020 Dec;9(1):2030-2045. doi: 10.1080/22221751.2020.1818633.

PMID: 32873191 Free article.

Here, we report a new antiviral mechanism in which MxB restricts the nuclear import of **HIV**-1 regulatory protein Rev, and as a result, diminishes Rev-dependent expression of **HIV**-1 Gag protein. ...In addition, **HIV**-1 can overcome this **inhibition** by MxB th ...

□ 2 Cite Share

Complex genetic encoding of the hepatitis B virus on-drug persistence.

Thai H, Lara J, Xu X, Kitrinos K, Gaggar A, Chan HLY, Xia GL, Ganova-Raeva L, Khudyakov Y.

Sci Rep. 2020 Sep 23;10(1):15574. doi: 10.1038/s41598-020-72467-9.

PMID: 32968103

Tenofovir disoproxil fumarate (TDF) is one of the nucleotide analogs capable of inhibiting the reverse transcriptase (RT) activity of **HIV** and hepatitis B virus (HBV). ...These pervasive mechanisms are insufficient to prevent viral **inhibition** completely but may contr ...

□ 3 Cite Share

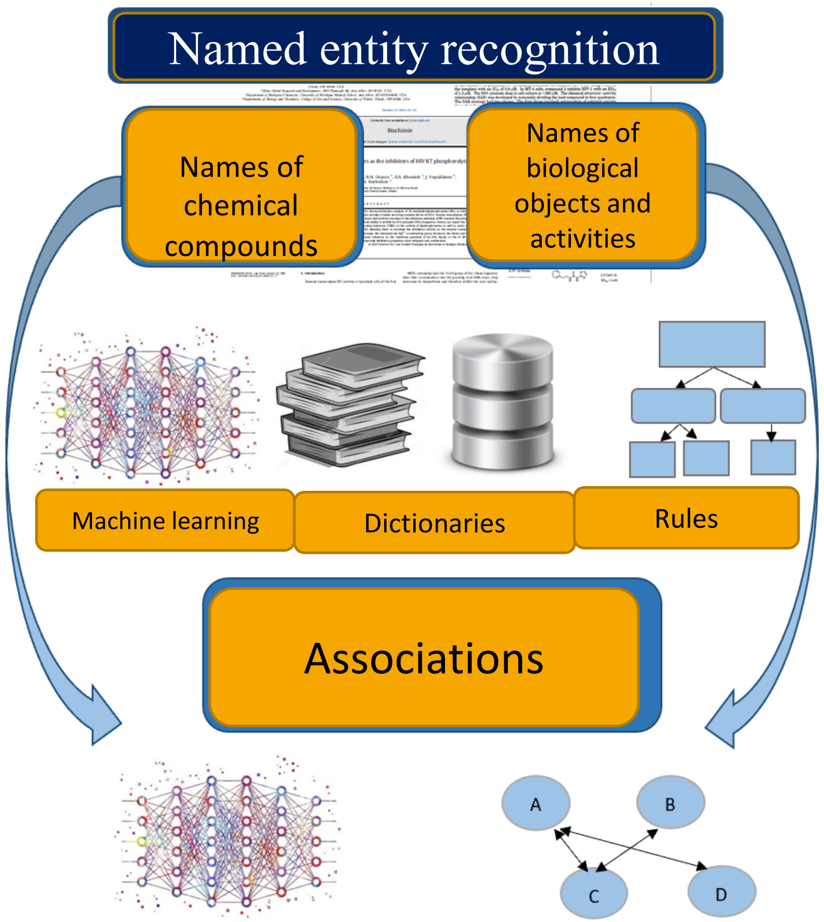
Integrase-RNA interactions underscore the critical role of integrase in **HIV**-1 virion morphogenesis.

Elliott JL, Eschbach JE, Koneru PC, Li W, Puray Chavez M, Townsend D, Lawson DQ, Engelman AN, Kvaratskhelia M, Kutluay SB.

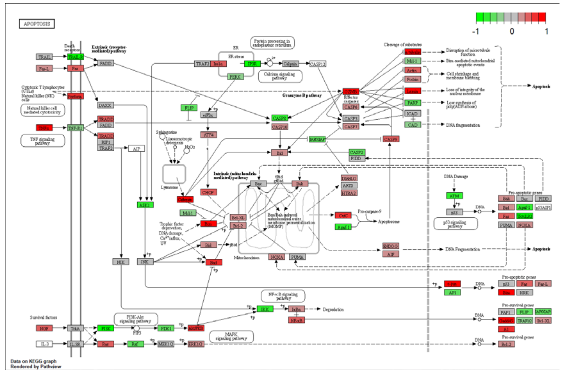
Elife. 2020 Sep 22;9:e54311. doi: 10.7554/eLife.54311. Online ahead of print.

PMID: 32960169

Inhibition of IN-RNA interactions resulted in mislocalization of the viral ribonucleoprotein complexes

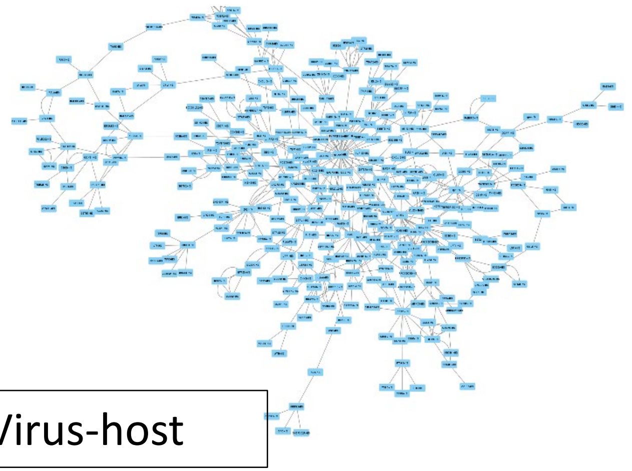


Textome - is a comprehensive set of biological literature that contains useful information and provides retrieval new knowledge using bioinformatics, ML and AI.

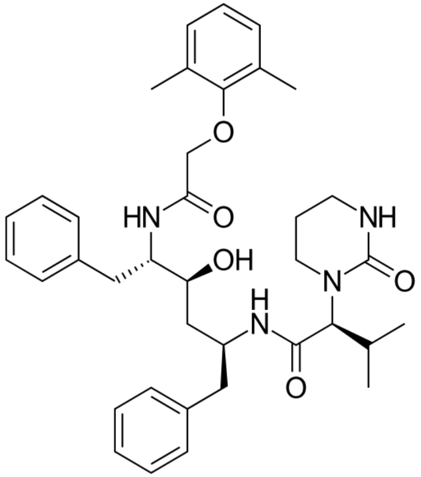


Textome
Transcriptome

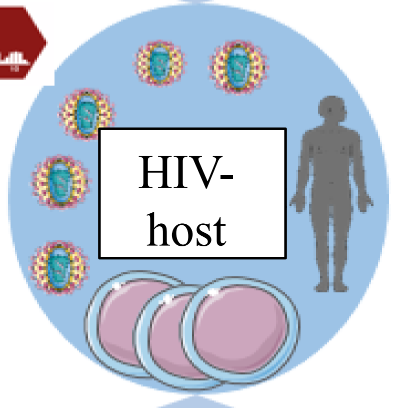
An analysis of HIV infection progression velocity



Virus-host

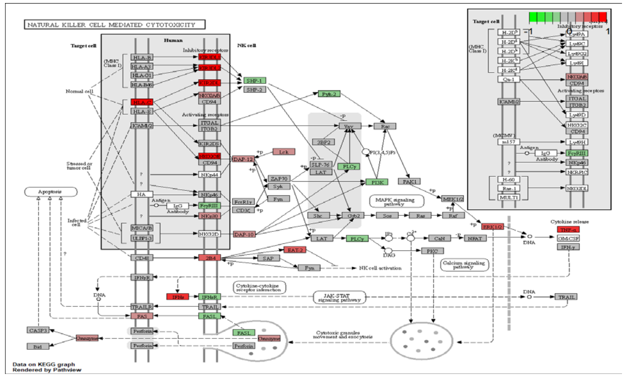


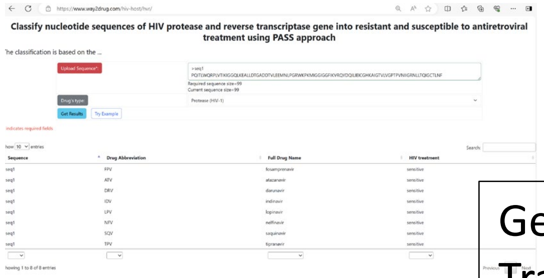
An analysis of HIV drug resistance and efficacy of ARVT



HIV-host

Search for master regulators of the long-term nonprogression





https://www.way2drug.com/hiv-host/hvr/

Classify nucleotide sequences of HIV protease and reverse transcriptase gene into resistant and susceptible to antiretroviral treatment using PASS approach

'he classification is based on the ...

Upload Sequence*: >seq1
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Required sequence size=99
Current sequence size=99

Drug's type: Protease (HIV-1)

Get Results | Try Example

indicates required fields

how 10 entries

Search:

Sequence	Drug Abbreviation	Full Drug Name	HIV treatment
seq1	FPV	fosamprenavir	sensitive
seq1	ATV	atazanavir	sensitive
seq1	DRV	darunavir	sensitive
seq1	IDV	indinavir	sensitive
seq1	LPV	lopinavir	sensitive
seq1	NFV	nelfinavir	sensitive
seq1	SQV	saquinavir	sensitive
seq1	TPV	tipranavir	sensitive

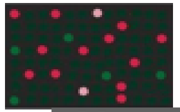
howing 1 to 8 of 8 entries

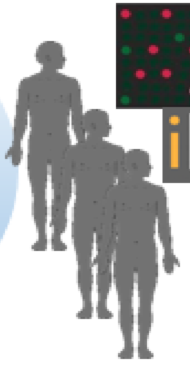
Previous Next

Virus-drug
Drug-host

Genome (virus)
Transcriptome (host)

The experimental validation of the developed approach

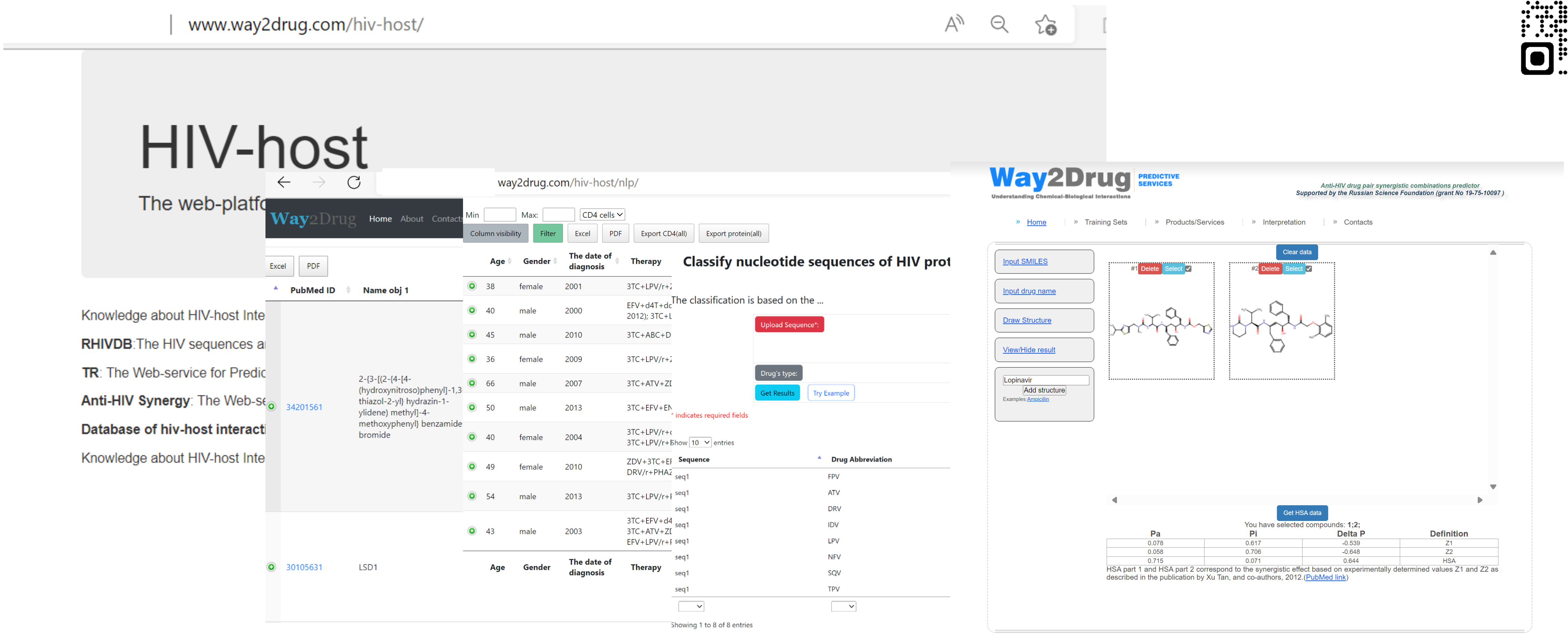




Virus-host
Transcriptome (host)

A principal scheme of HIV-host interaction analysis

HIV-host web resource



www.way2drug.com/hiv-host/

HIV-host

The web-platf

Knowledge about HIV-host Inte

RHIVDB:The HIV sequences a

TR: The Web-service for Predic

Anti-HIV Synergy: The Web-se

Database of hiv-host interacti

Knowledge about HIV-host Inte

way2drug.com/hiv-host/nlp/

PubMed ID	Name obj 1
34201561	2-{3-[(2-{4-[4-(hydroxynitroso)phenyl]-1,3-thiazol-2-yl} hydrazin-1-ylidene) methyl]-4-methoxyphenyl} benzamide bromide
30105631	LSD1

Min Max: CD4 cells

Age	Gender	The date of diagnosis	Therapy
38	female	2001	3TC+LPV/r+Z
40	male	2000	EFV+d4T+dc 2012); 3TC+L
45	male	2010	3TC+ABC+D
36	female	2009	3TC+LPV/r+Z
66	male	2007	3TC+ATV+ZI
50	male	2013	3TC+EFV+EN
40	female	2004	3TC+LPV/r+ 3TC+LPV/r+I
49	female	2010	ZDV+3TC+EI DRV/r+PHAZ
54	male	2013	3TC+LPV/r+I
43	male	2003	3TC+EFV+d4 3TC+ATV+ZI EFV+LPV/r+I
Age	Gender	The date of diagnosis	Therapy

Classify nucleotide sequences of HIV prot

The classification is based on the ...

Upload Sequence*:

Drug's type:

* indicates required fields

Show 10 entries

Sequence	Drug Abbreviation
seq1	FPV
seq1	ATV
seq1	DRV
seq1	IDV
seq1	LPV
seq1	NFV
seq1	SQV
seq1	TPV

Showing 1 to 8 of 8 entries

Anti-HIV drug pair synergistic combinations predictor

Supported by the Russian Science Foundation (grant No 19-75-10097)

Lopinavir

Examples:Ampicillin

You have selected compounds: **1;2;**

Pa	Pi	Delta P	Definition
0.078	0.617	-0.539	Z1
0.058	0.706	-0.648	Z2
0.715	0.071	0.644	HSA

HSA part 1 and HSA part 2 correspond to the synergistic effect based on experimentally determined values Z1 and Z2 as described in the publication by Xu Tan, and co-authors, 2012.(PubMed link)

Hiv-host includes: (1) information on extracted names of interacting HIV and human macromolecules; (2) models for predicting the efficacy of antiretroviral therapy; (3) a web service for predicting HIV drug resistance; (4) a specialised database on HIV sequences and viral load dynamics, immunograms of HIV-infected patients on specific antiretroviral therapy regimens; (5) a web resource for predicting drug synergies in inhibiting HIV replication.

http://www.way2drug.com/hiv-host/

Big data analysis of medical virology to find new effective and safe antiviral compounds and optimise therapy for infectious diseases

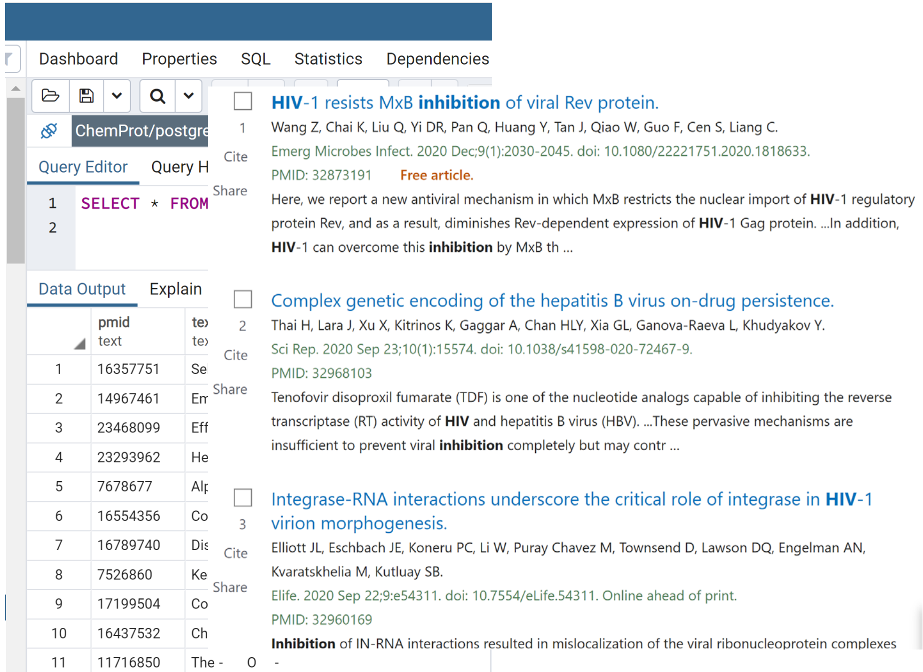
The aim of our study is to develop an *in silico* approach for the extracting knowledge about viruses and the host (the human body), and potential antiviral agents based on the mining of massive amounts of scientific publications

DrugProt, CHEMDNER

Annotated corpora

```
22301815  T  11   23   carbohydrate   FAMILY
22301815  T  65   73   cortisol       TRIVIAL
22301815  T  154  164  endosulfan     TRIVIAL
22301815  A  15   25   endosulfan     TRIVIAL
22301815  A  30   44   organochlorine FAMILY
```

Publication databases



```
1  SELECT * FROM
2
```

	pmid text	te te
1	16357751	Se
2	14967461	Em
3	23468099	Eff
4	23293962	He
5	7678677	Alp
6	16554356	Co
7	16789740	Dis
8	7526860	Ke
9	17199504	Co
10	16437532	Ch
11	11716850	The -

□ 1 Cite Share

HIV-1 resists MxB **inhibition** of viral Rev protein.

Wang Z, Chai K, Liu Q, Yi DR, Pan Q, Huang Y, Tan J, Qiao W, Guo F, Cen S, Liang C.

Emerg Microbes Infect. 2020 Dec;9(1):2030-2045. doi: 10.1080/22221751.2020.1818633.

PMID: 32873191 Free article.

Here, we report a new antiviral mechanism in which MxB restricts the nuclear import of **HIV**-1 regulatory protein Rev, and as a result, diminishes Rev-dependent expression of **HIV**-1 Gag protein. ...In addition, **HIV**-1 can overcome this **inhibition** by MxB th ...

□ 2 Cite Share

Complex genetic encoding of the hepatitis B virus on-drug persistence.

Thai H, Lara J, Xu X, Kitrinos K, Gaggar A, Chan HLY, Xia GL, Ganova-Raeva L, Khudyakov Y.

Sci Rep. 2020 Sep 23;10(1):15574. doi: 10.1038/s41598-020-72467-9.

PMID: 32968103

Tenofovir disoproxil fumarate (TDF) is one of the nucleotide analogs capable of inhibiting the reverse transcriptase (RT) activity of **HIV** and hepatitis B virus (HBV). ...These pervasive mechanisms are insufficient to prevent viral **inhibition** completely but may contr ...

□ 3 Cite Share

Integrase-RNA interactions underscore the critical role of integrase in **HIV**-1 virion morphogenesis.

Elliott JL, Eschbach JE, Koneru PC, Li W, Puray Chavez M, Townsend D, Lawson DQ, Engelman AN, Kvaratskhelia M, Kutluay SB.

Elife. 2020 Sep 22;9:e54311. doi: 10.7554/eLife.54311. Online ahead of print.

PMID: 32960169

Inhibition of IN-RNA interactions resulted in mislocalization of the viral ribonucleoprotein complexes

O -

Machine learning

Regular expressions

Conditional random fields, *J. Lafferty et al., 2001*	Naïve Bayes, *O.Tarasova et al., 2022*

BioBERT,
L. *Weber et al., 2021*

5-Alkyl-2-[(methylthiomethyl)thio]-6-(benzyl)-pyrimidin-4-(1H)-ones as potent non-nucleoside reverse transcriptase inhibitors of S-DABO series.

5-Alkyl-2-[(methylthiomethyl)thio]-6-(benzyl)-pyrimidin-4-(1H)-ones as potent non-nucleoside **reverse transcriptase** inhibitors of S-DABO series.

Effects of SKF 108922, an HIV-1 protease inhibitor, on retrovirus replication in mice.

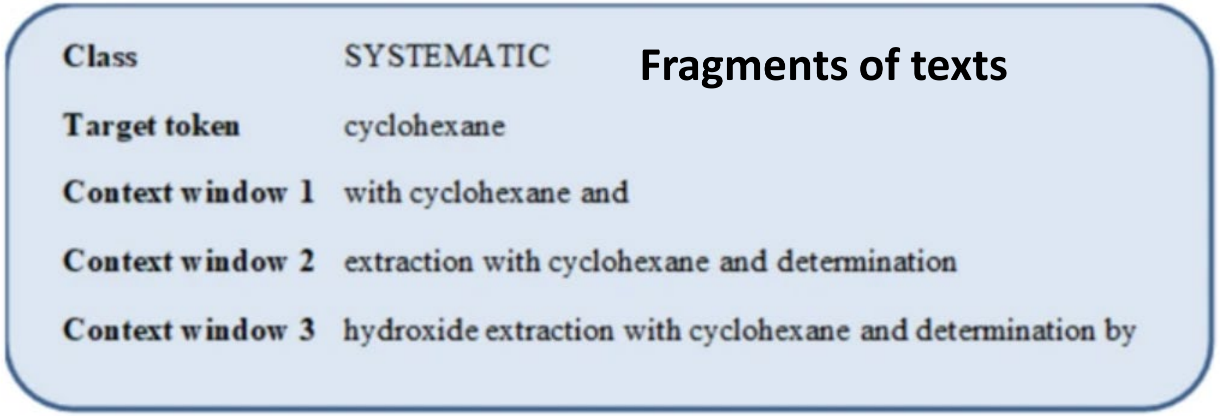
Effects of **SKF 108922**, an HIV-1 **protease** inhibitor, on retrovirus replication in mice.

Pyrrolyl aryl sulfones (PASs) have been recently reported as a new class of human immunodeficiency virus type 1 (HIV-1) reverse transcriptase (RT) inhibitors acting at the non-nucleoside binding site of this enzyme.

Pyrrolyl aryl sulfones (**PASs**) have been recently reported as a new class of human immunodeficiency | virus type 1 (HIV-1) **reverse transcriptase** (**RT**) inhibitors acting at the non-nucleoside binding site of this enzyme.

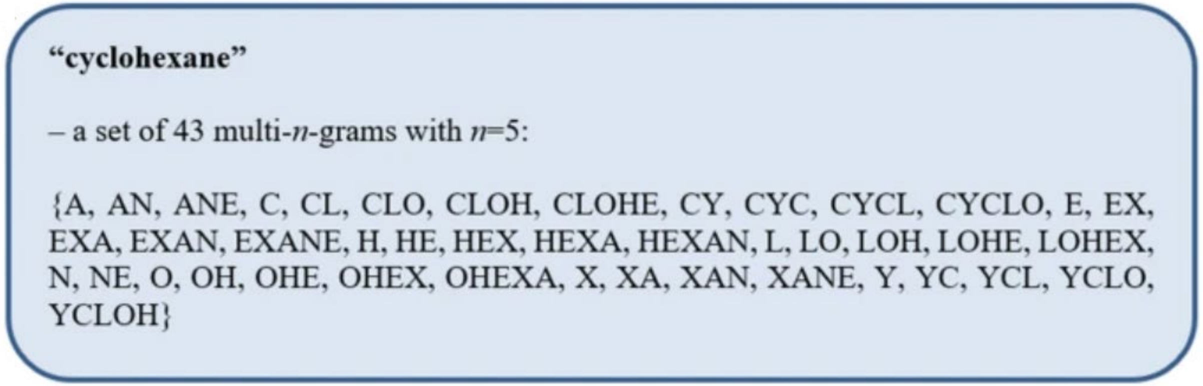
Associations and relations extraction

Naïve Bayes approach for chemical and biological NER



Fragments of texts

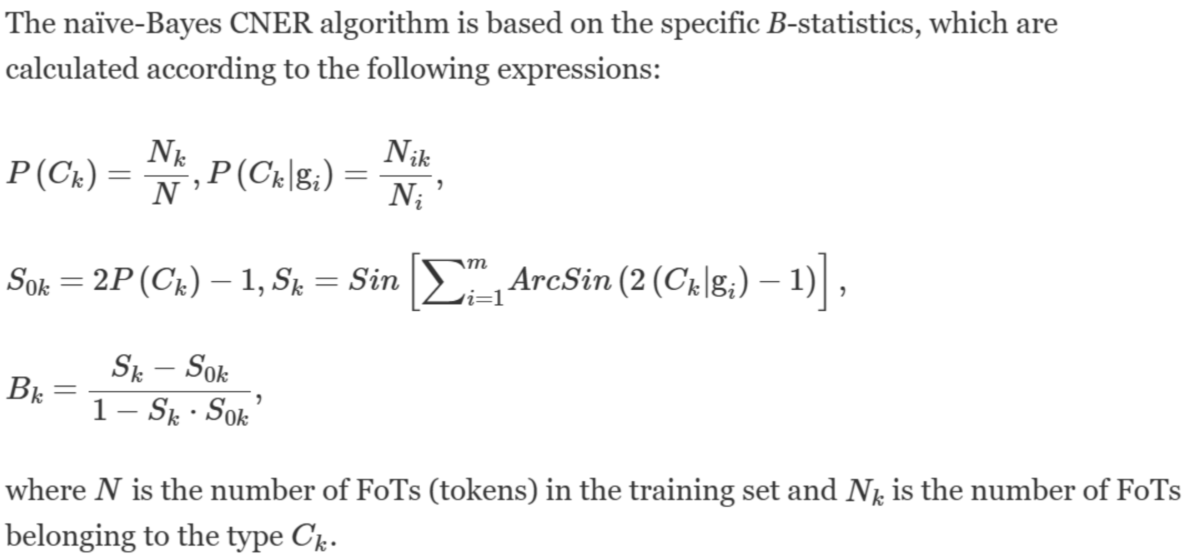
Class	SYSTEMATIC
Target token	cyclohexane
Context window 1	with cyclohexane and
Context window 2	extraction with cyclohexane and determination
Context window 3	hydroxide extraction with cyclohexane and determination by



“cyclohexane”

– a set of 43 multi-*n*-grams with *n*=5:

{A, AN, ANE, C, CL, CLO, CLOH, CLOHE, CY, CYC, CYCL, CYCLO, E, EX, EXA, EXAN, EXANE, H, HE, HEX, HEXA, HEXAN, L, LO, LOH, LOHE, LOHEX, N, NE, O, OH, OHE, OHEX, OHEXA, X, XA, XAN, XANE, Y, YC, YCL, YCLO, YCLOH}



The naïve-Bayes CNER algorithm is based on the specific *B*-statistics, which are calculated according to the following expressions:

$$P(C_k) = \frac{N_k}{N}, P(C_k|g_i) = \frac{N_{ik}}{N_i},$$

$$S_{0k} = 2P(C_k) - 1, S_k = Sin\left[\sum_{i=1}^{m} ArcSin(2(C_k|g_i) - 1)\right],$$

$$B_k = \frac{S_k - S_{0k}}{1 - S_k \cdot S_{0k}},$$

where N is the number of FoTs (tokens) in the training set and N_k is the number of FoTs belonging to the type C_k.

IAP (average), LOO CV		
	N-gram = 5	*N-gram* = 6
Context window = 0	0.86	0.96
Context window = 1	0.95	0.96

Tarasova O. et al., *Journal of Chemoinformatics, 2022*

Verification of recognized entities. Dictionaries

Chemical named entities

- CAS common chemistry API
- ChemSpider Web API
- PubChem PUG REST
- Manually

Proteins and genes

UniProt website REST API

Diseases

Human Disease Ontology

- dictionaries can help to filter out some false positives of chemical named entities and improve accuracy of recognition;
- dictionaries can be efficiently used for recognition of diseases and disorders

Accuracy of named entity recognition

Chemicals, ML + dictionary

Validation type	Precision	Recall	F_1-score
5-fold CV	0.89	0.83	0.86
Manual annotation, external test	0.84	0.79	0.81

Proteins, ML + dictionary

5-fold CV	0.87	0.84	0.85
Manual annotation, external test	0.84	0.79	0.81

Diseases and disorders, ML + dictionary

5-fold CV	0.84	0.79	0.81
Manual annotation, external test	0.80	0.76	0.78

Extraction of associations between entities

1. Named entity recognition in the abstracts of relevant publications

ML-based selection of relevant publications; associations with /relations to a set of keywords characterising a set of publications belonging to a particular class

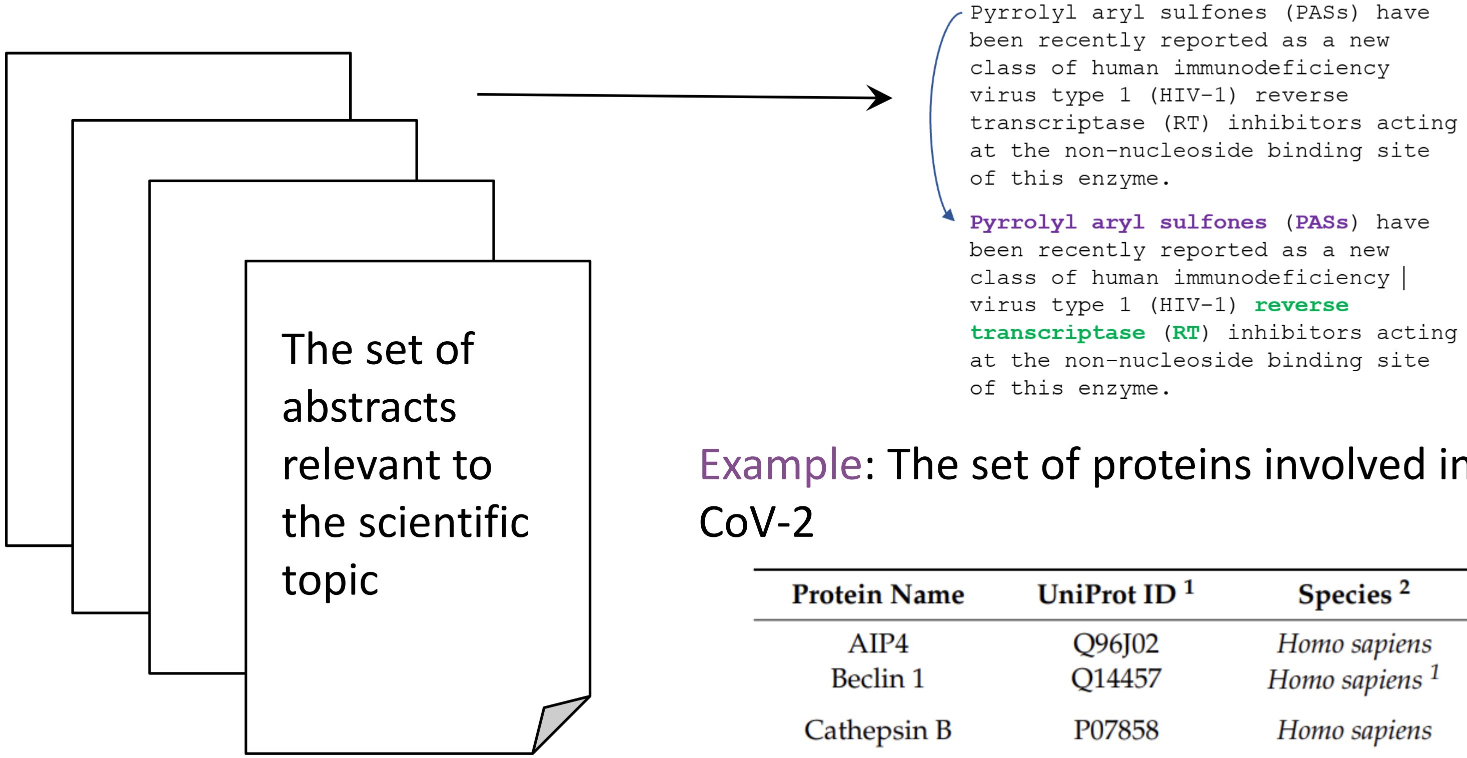
2. Rule-based approach using a set of pattern phrases

Identification of relationships in an abstract/full text or part thereof

3. Co-occurrences

Search for possible relationships that have not yet been investigated or shown in an experiment

Extraction of associations. Named entity recognition in the abstracts of relevant publications

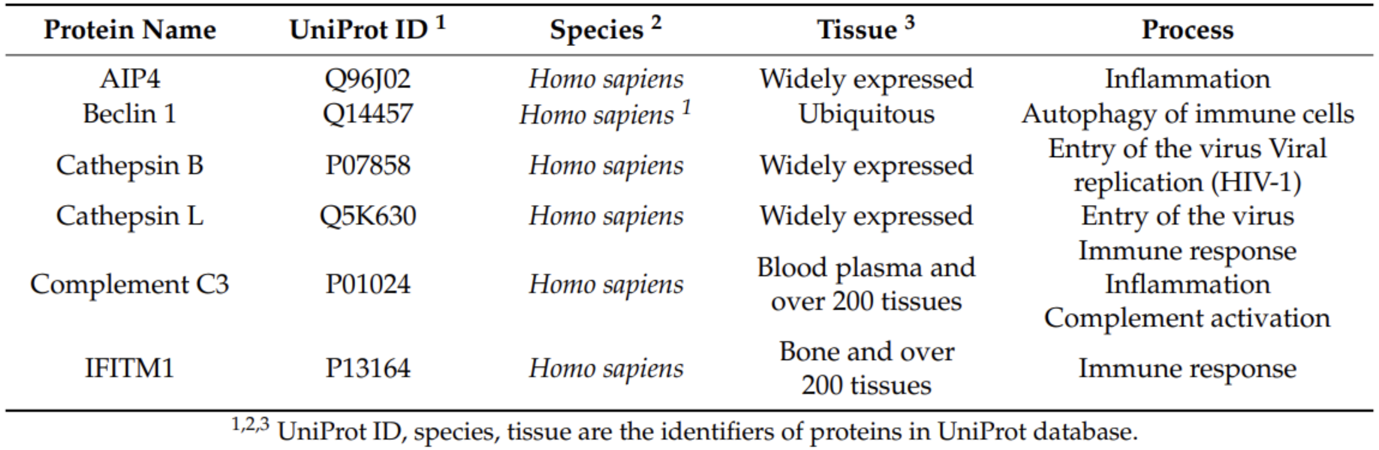


The set of abstracts relevant to the scientific topic

```
Pyrrolyl aryl sulfones (PASs) have
been recently reported as a new
class of human immunodeficiency
virus type 1 (HIV-1) reverse
transcriptase (RT) inhibitors acting
at the non-nucleoside binding site
of this enzyme.
```

```
Pyrrolyl aryl sulfones (PASs) have
been recently reported as a new
class of human immunodeficiency |
virus type 1 (HIV-1) reverse
transcriptase (RT) inhibitors acting
at the non-nucleoside binding site
of this enzyme.
```

Example: The set of proteins involved in antiviral response against HIV-1 and SARS-CoV-2



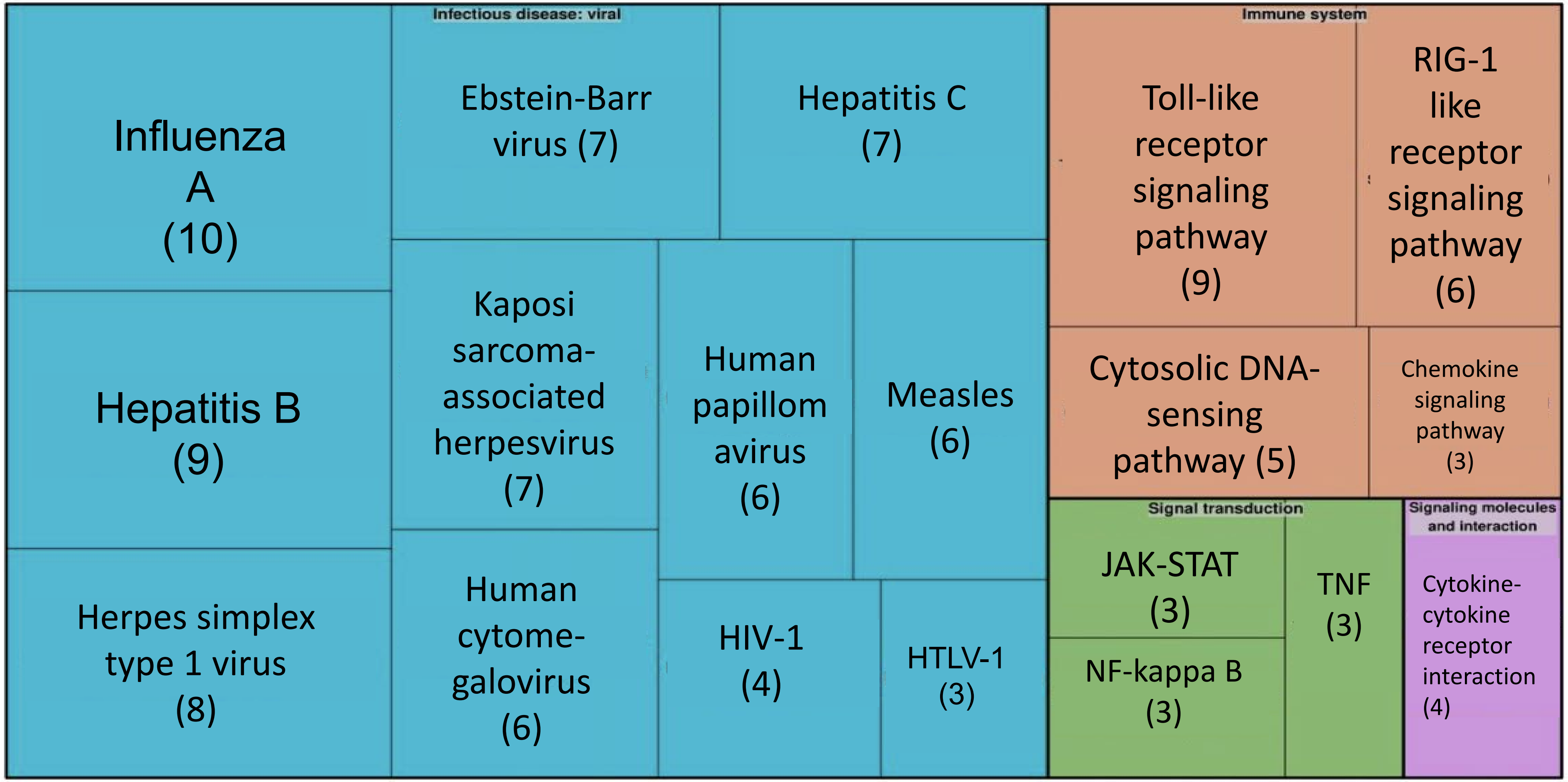
Protein Name	UniProt ID [1]	Species [2]	Tissue [3]	Process
AIP4	Q96J02	*Homo sapiens*	Widely expressed	Inflammation
Beclin 1	Q14457	*Homo sapiens* [1]	Ubiquitous	Autophagy of immune cells
Cathepsin B	P07858	*Homo sapiens*	Widely expressed	Entry of the virus Viral replication (HIV-1)
Cathepsin L	Q5K630	*Homo sapiens*	Widely expressed	Entry of the virus
Complement C3	P01024	*Homo sapiens*	Blood plasma and over 200 tissues	Immune response Inflammation Complement activation
IFITM1	P13164	*Homo sapiens*	Bone and over 200 tissues	Immune response

[1,2,3] UniProt ID, species, tissue are the identifiers of proteins in UniProt database.

O.A. Tarasova et al., *Molecules*, 2020

R.M. Bonotto et al., *Antiviral Res.*, 2023; Hashimoto R. et al., *Mol Ther Nucleic Acids*., 2021

Gene set enrichement analysis based on literature mining results



Infectious disease: viral

Influenza A (10)

Hepatitis B (9)

Herpes simplex type 1 virus (8)

Ebstein-Barr virus (7)

Hepatitis C (7)

Kaposi sarcoma-associated herpesvirus (7)

Human cytome-galovirus (6)

Human papillomavirus (6)

Measles (6)

HIV-1 (4)

HTLV-1 (3)

Immune system

Toll-like receptor signaling pathway (9)

RIG-1 like receptor signaling pathway (6)

Cytosolic DNA-sensing pathway (5)

Chemokine signaling pathway (3)

Signal transduction

JAK-STAT (3)

NF-kappa B (3)

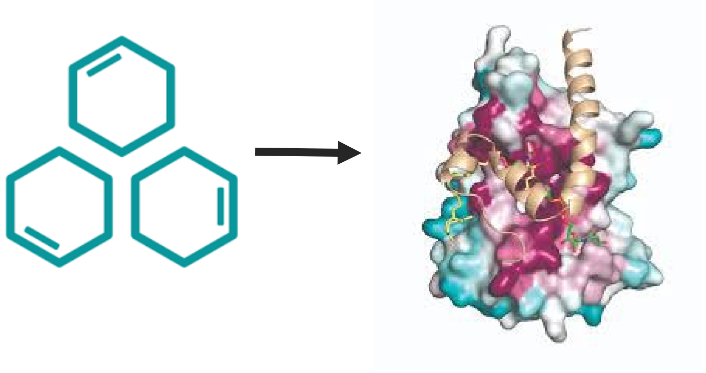
TNF (3)

Signaling molecules and interaction

Cytokine-cytokine receptor interaction (4)

KEGG pathways enriched in the genes associated with human proteins involved in both SARS-CoV-2–host and Dengue–host interactions. Each color represents an individual pathway. The size of each box reflects the number of proteins involved in that particular pathway. The number of proteins involved in each pathway is given in brackets.

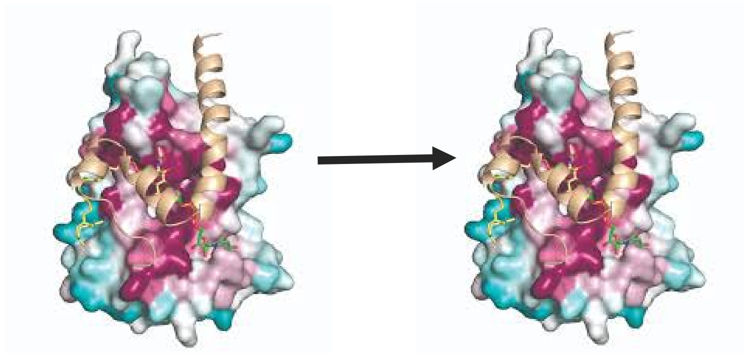
Extraction of associations. Pattern phrases



Chemical compound

Protein

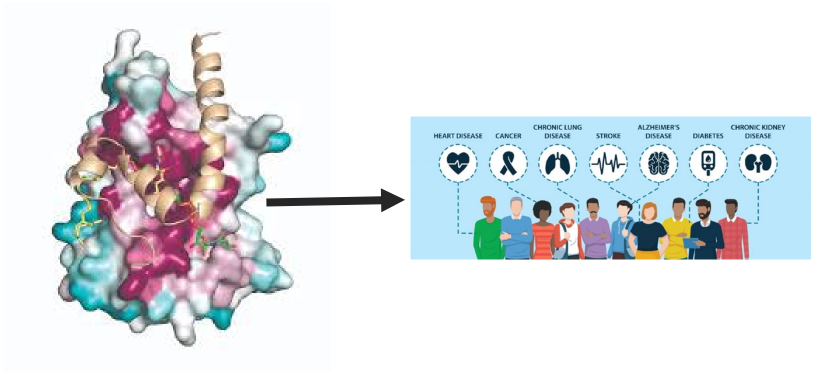
Set of pattern phrases



Protein/gene

Protein/gene

Set of pattern phrases



Protein/gene

Disease

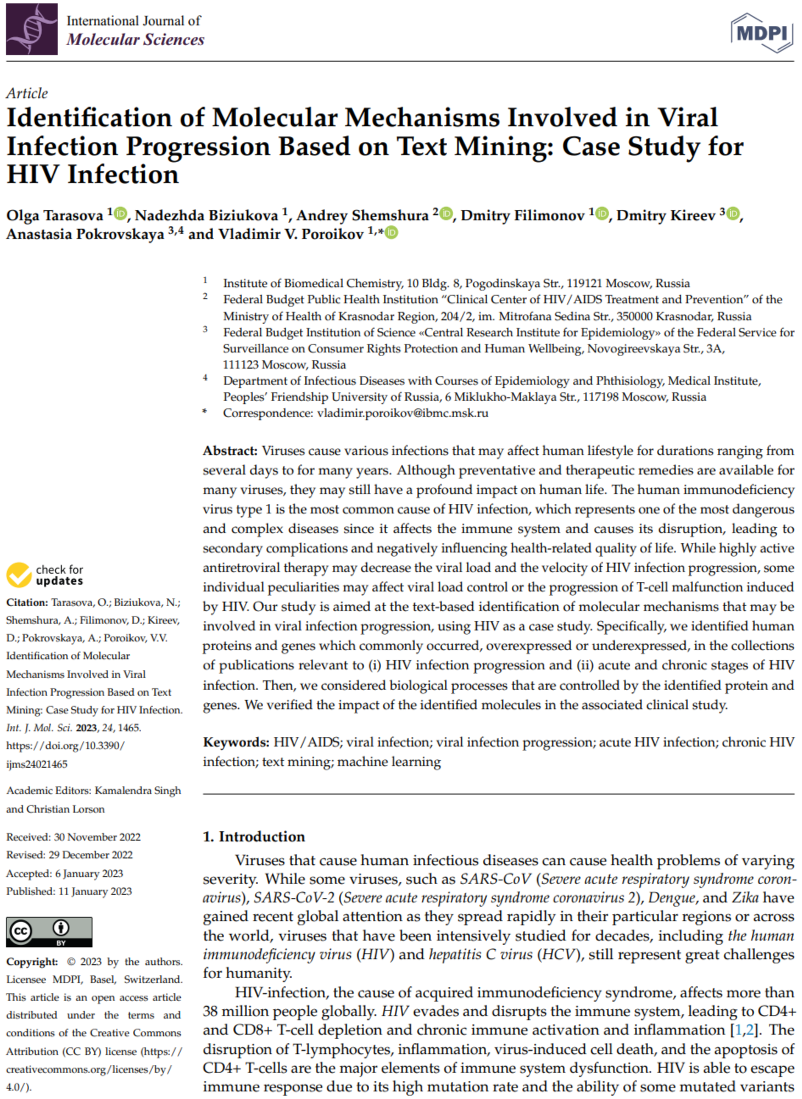
Main term	Pattern	Example
interact	Interaction of P1 P2	Interaction of Vpr with other proteins such as EF2
regulate	regulation of P1 P2	Regulation of IDO by HIV-1 Tat protein
inhibit	C1 inhibits P1	RT1t49 inhibits recombinant RT

Accuracy of associations extraction: precision is 0.78, recall is 0.91, F_1-score is 0.84

Recognition for two names (protein and disease) in one abstract selected by relevance by the particular diseases or disorders

Example: Hedgehog pathway and Cancer

Chemical named entity recognition and relation extraction



International Journal of
Molecular Sciences

MDPI

Article

Identification of Molecular Mechanisms Involved in Viral Infection Progression Based on Text Mining: Case Study for HIV Infection

Olga Tarasova [1], Nadezhda Biziukova [1], Andrey Shemshura [2], Dmitry Filimonov [1], Dmitry Kireev [3], Anastasia Pokrovskaya [3,4] and Vladimir V. Poroikov [1,*]

[1] Institute of Biomedical Chemistry, 10 Bldg. 8, Pogodinskaya Str., 119121 Moscow, Russia
[2] Federal Budget Public Health Institution "Clinical Center of HIV/AIDS Treatment and Prevention" of the Ministry of Health of Krasnodar Region, 204/2, im. Mitrofana Sedina Str., 350000 Krasnodar, Russia
[3] Federal Budget Institution of Science «Central Research Institute for Epidemiology» of the Federal Service for Surveillance on Consumer Rights Protection and Human Wellbeing, Novogireevskaya Str., 3A, 111123 Moscow, Russia
[4] Department of Infectious Diseases with Courses of Epidemiology and Phthisiology, Medical Institute, Peoples' Friendship University of Russia, 6 Miklukho-Maklaya Str., 117198 Moscow, Russia
[*] Correspondence: vladimir.poroikov@ibmc.msk.ru

Abstract: Viruses cause various infections that may affect human lifestyle for durations ranging from several days to for many years. Although preventative and therapeutic remedies are available for many viruses, they may still have a profound impact on human life. The human immunodeficiency virus type 1 is the most common cause of HIV infection, which represents one of the most dangerous and complex diseases since it affects the immune system and causes its disruption, leading to secondary complications and negatively influencing health-related quality of life. While highly active antiretroviral therapy may decrease the viral load and the velocity of HIV infection progression, some individual peculiarities may affect viral load control or the progression of T-cell malfunction induced by HIV. Our study is aimed at the text-based identification of molecular mechanisms that may be involved in viral infection progression, using HIV as a case study. Specifically, we identified human proteins and genes which commonly occurred, overexpressed or underexpressed, in the collections of publications relevant to (i) HIV infection progression and (ii) acute and chronic stages of HIV infection. Then, we considered biological processes that are controlled by the identified protein and genes. We verified the impact of the identified molecules in the associated clinical study.

Keywords: HIV/AIDS; viral infection; viral infection progression; acute HIV infection; chronic HIV infection; text mining; machine learning

check for updates

Citation: Tarasova, O.; Biziukova, N.; Shemshura, A.; Filimonov, D.; Kireev, D.; Pokrovskaya, A.; Poroikov, V.V. Identification of Molecular Mechanisms Involved in Viral Infection Progression Based on Text Mining: Case Study for HIV Infection. *Int. J. Mol. Sci.* **2023**, *24*, 1465. https://doi.org/10.3390/ijms24021465

Academic Editors: Kamalendra Singh and Christian Lorson

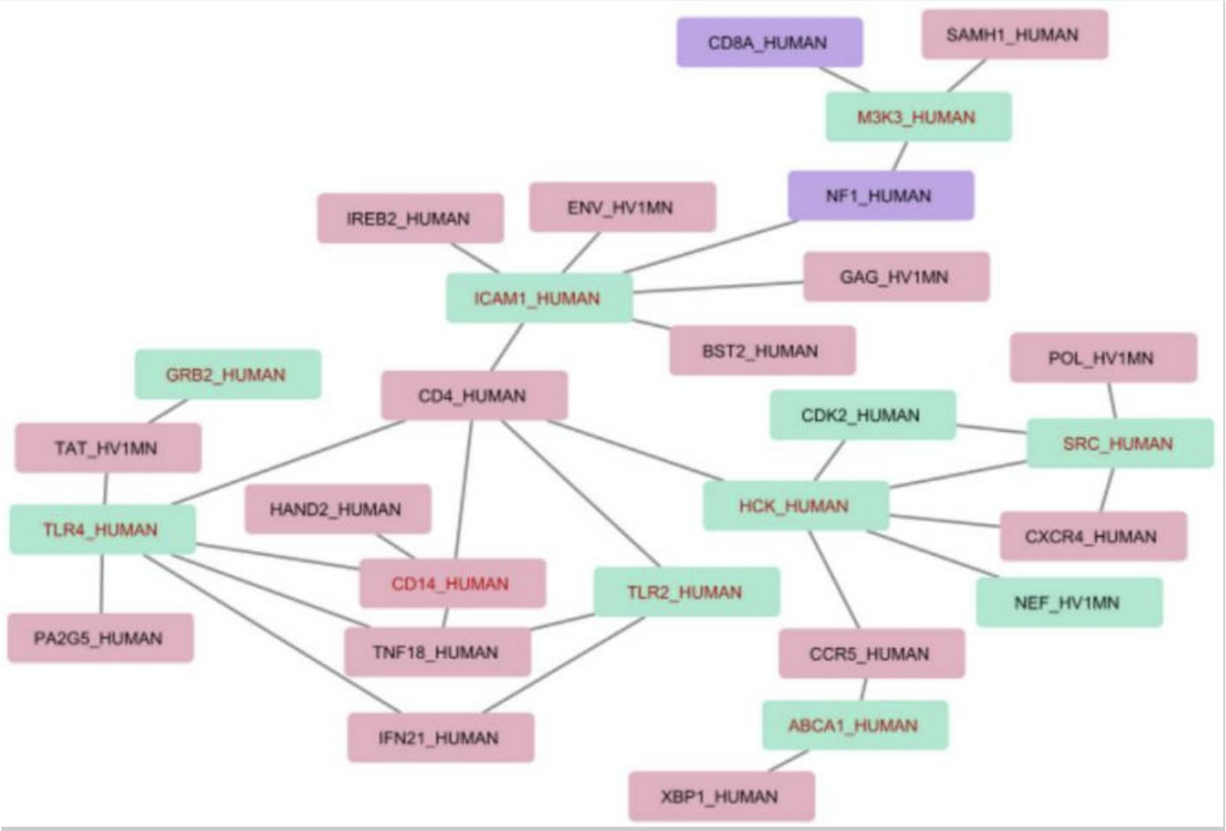
Received: 30 November 2022
Revised: 29 December 2022
Accepted: 6 January 2023
Published: 11 January 2023

1. Introduction

Viruses that cause human infectious diseases can cause health problems of varying severity. While some viruses, such as *SARS-CoV* (*Severe acute respiratory syndrome coronavirus*), *SARS-CoV-2* (*Severe acute respiratory syndrome coronavirus 2*), *Dengue*, and *Zika* have gained recent global attention as they spread rapidly in their particular regions or across the world, viruses that have been intensively studied for decades, including *the human immunodeficiency virus* (*HIV*) and *hepatitis C virus* (*HCV*), still represent great challenges for humanity.

HIV-infection, the cause of acquired immunodeficiency syndrome, affects more than 38 million people globally. *HIV* evades and disrupts the immune system, leading to CD4+ and CD8+ T-cell depletion and chronic immune activation and inflammation [1,2]. The disruption of T-lymphocytes, inflammation, virus-induced cell death, and the apoptosis of CD4+ T-cells are the major elements of immune system dysfunction. HIV is able to escape immune response due to its high mutation rate and the ability of some mutated variants

For a variery of human genes, information about changes in expression levels during the course of HIV infection has been shown in the experiment



Red font - differentially expressed genes for which the differences are reliable and confirmed in experiment

Experimental validation of the results based on the prospective clinical study

11 patients before the start of antitertoviral therapy, (ART), peripheral blood mononulcear cells (PBMC)

9 patients after 24 weeks of ART, PBMC

Patients with HIV-infection over 1 year/ less than 1 year (5 patients/ 6 patients)

- Central Research Institute of Epidemiology, Moscow, Russia
- Krasnodar clinical center of HIV profilaxys and cure

RNASeq, HiSeq1500 (Illumina)

Differences in expression level:

- 606 genes (p < 0,1);

- 183 genes (p<0.05)

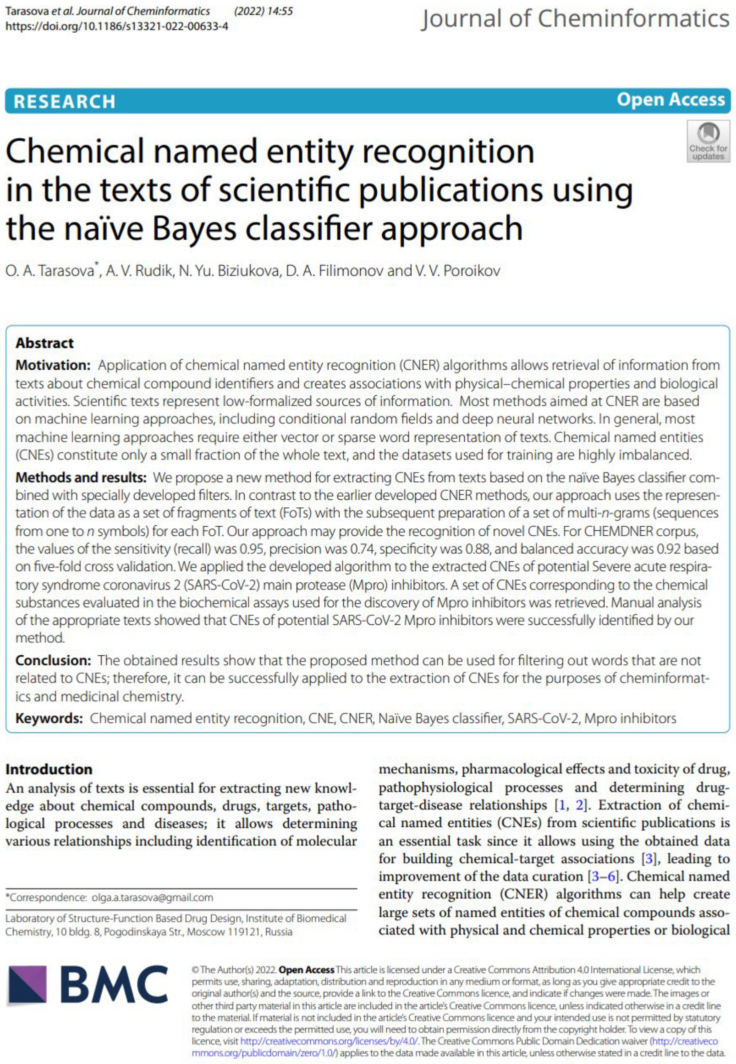
Differences in expression level, 24 weeks after HAART: 165 genes (p < 0.05) compared to before treatment

16 genes (p< 0.05) in two groups based on the immunological effectiveness

Genes that identified in the text mining that were found to be differentially expressed in the experiment

Gene	Name	Process	PMIDs	Log$_2$FoldChange	P$_{adj}$
CLEC5A ↑	C-type lectin domain family 5 member A	Immune response; **negative regulation of apoptotic process; negative regulation of myeloid cell apoptotic** process	31867016	1.67	0.0006
TLR2 ↑	Toll-like receptor 2, CD282	Immune response Inflammatory response **apoptotic process** Regulation of gene expression	32093694 32516401 28730622	1.4	0.019
CD14 ↑	Monocyte differentiation antigen CD14	Immune response **apoptotic process** Inflammatory response	34211989 33487130	1.14	0.04
CD86 ↑	T-lymphocyte activation antigen CD86	Immune response Negative regulation of T cell proliferation	34630420	0.89	0.046
NAMPT ↑	Nicotinamide phosphoribosyltransferase	Autophagy	-	2.0	0.03

Named entity recognition and relation extraction for solving various biological tasks



Tarasova *et al. Journal of Cheminformatics* *(2022) 14:55*
https://doi.org/10.1186/s13321-022-00633-4

Journal of Cheminformatics

RESEARCH Open Access

Chemical named entity recognition in the texts of scientific publications using the naïve Bayes classifier approach

O. A. Tarasova*, A. V. Rudik, N. Yu. Biziukova, D. A. Filimonov and V. V. Poroikov

Abstract

Motivation: Application of chemical named entity recognition (CNER) algorithms allows retrieval of information from texts about chemical compound identifiers and creates associations with physical–chemical properties and biological activities. Scientific texts represent low-formalized sources of information. Most methods aimed at CNER are based on machine learning approaches, including conditional random fields and deep neural networks. In general, most machine learning approaches require either vector or sparse word representation of texts. Chemical named entities (CNEs) constitute only a small fraction of the whole text, and the datasets used for training are highly imbalanced.

Methods and results: We propose a new method for extracting CNEs from texts based on the naïve Bayes classifier combined with specially developed filters. In contrast to the earlier developed CNER methods, our approach uses the representation of the data as a set of fragments of text (FoTs) with the subsequent preparation of a set of multi-*n*-grams (sequences from one to *n* symbols) for each FoT. Our approach may provide the recognition of novel CNEs. For CHEMDNER corpus, the values of the sensitivity (recall) was 0.95, precision was 0.74, specificity was 0.88, and balanced accuracy was 0.92 based on five-fold cross validation. We applied the developed algorithm to the extracted CNEs of potential Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) main protease (Mpro) inhibitors. A set of CNEs corresponding to the chemical substances evaluated in the biochemical assays used for the discovery of Mpro inhibitors was retrieved. Manual analysis of the appropriate texts showed that CNEs of potential SARS-CoV-2 Mpro inhibitors were successfully identified by our method.

Conclusion: The obtained results show that the proposed method can be used for filtering out words that are not related to CNEs; therefore, it can be successfully applied to the extraction of CNEs for the purposes of cheminformatics and medicinal chemistry.

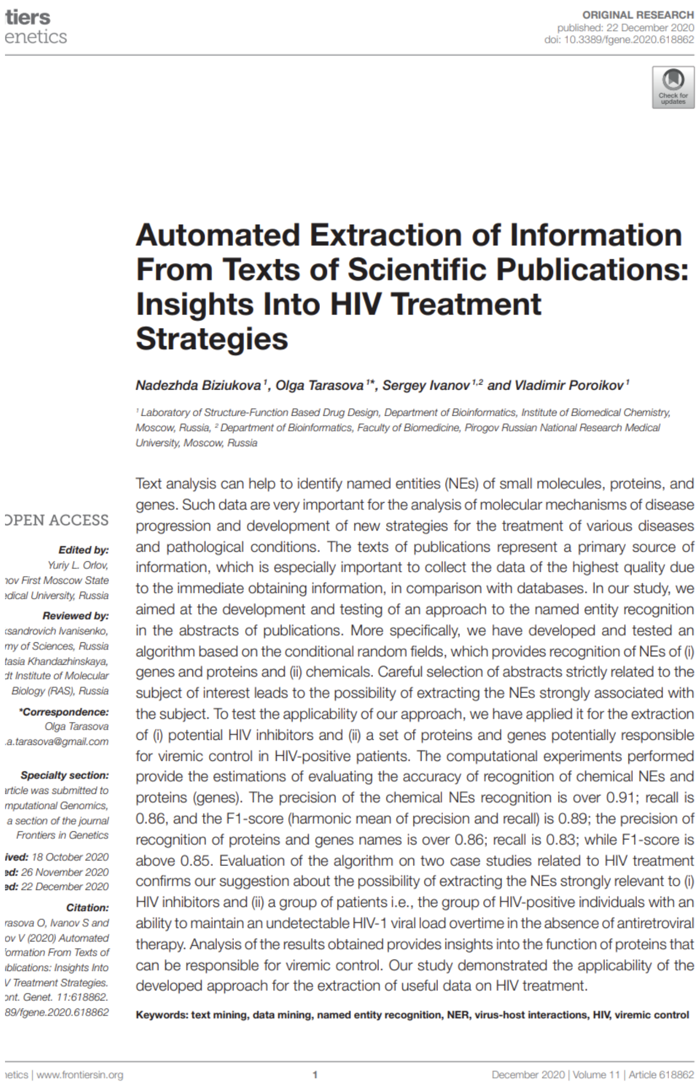
Keywords: Chemical named entity recognition, CNE, CNER, Naïve Bayes classifier, SARS-CoV-2, Mpro inhibitors

Introduction

An analysis of texts is essential for extracting new knowledge about chemical compounds, drugs, targets, pathological processes and diseases; it allows determining various relationships including identification of molecular mechanisms, pharmacological effects and toxicity of drug, pathophysiological processes and determining drug-target-disease relationships [1, 2]. Extraction of chemical named entities (CNEs) from scientific publications is an essential task since it allows using the obtained data for building chemical-target associations [3], leading to improvement of the data curation [3–6]. Chemical named entity recognition (CNER) algorithms can help create large sets of named entities of chemical compounds associated with physical and chemical properties or biological

*Correspondence: olga.a.tarasova@gmail.com

Laboratory of Structure-Function Based Drug Design, Institute of Biomedical Chemistry, 10 bldg. 8, Pogodinskaya Str., Moscow 119121, Russia



tiers
enetics

ORIGINAL RESEARCH
published: 22 December 2020
doi: 10.3389/fgene.2020.618862

Automated Extraction of Information From Texts of Scientific Publications: Insights Into HIV Treatment Strategies

Nadezhda Biziukova [1], Olga Tarasova [1], Sergey Ivanov [1,2] and Vladimir Poroikov [1]*

[1] Laboratory of Structure-Function Based Drug Design, Department of Bioinformatics, Institute of Biomedical Chemistry, Moscow, Russia, [2] Department of Bioinformatics, Faculty of Biomedicine, Pirogov Russian National Research Medical University, Moscow, Russia

PEN ACCESS

Edited by:
Yuriy L. Orlov,
ıov First Moscow State
ːdical University, Russia

Reviewed by:
ksandrovich Ivanisenko,
my of Sciences, Russia
tasia Khandazhinskaya,
dt Institute of Molecular
Biology (RAS), Russia

****Correspondence:***
Olga Tarasova
.a.tarasova@gmail.com

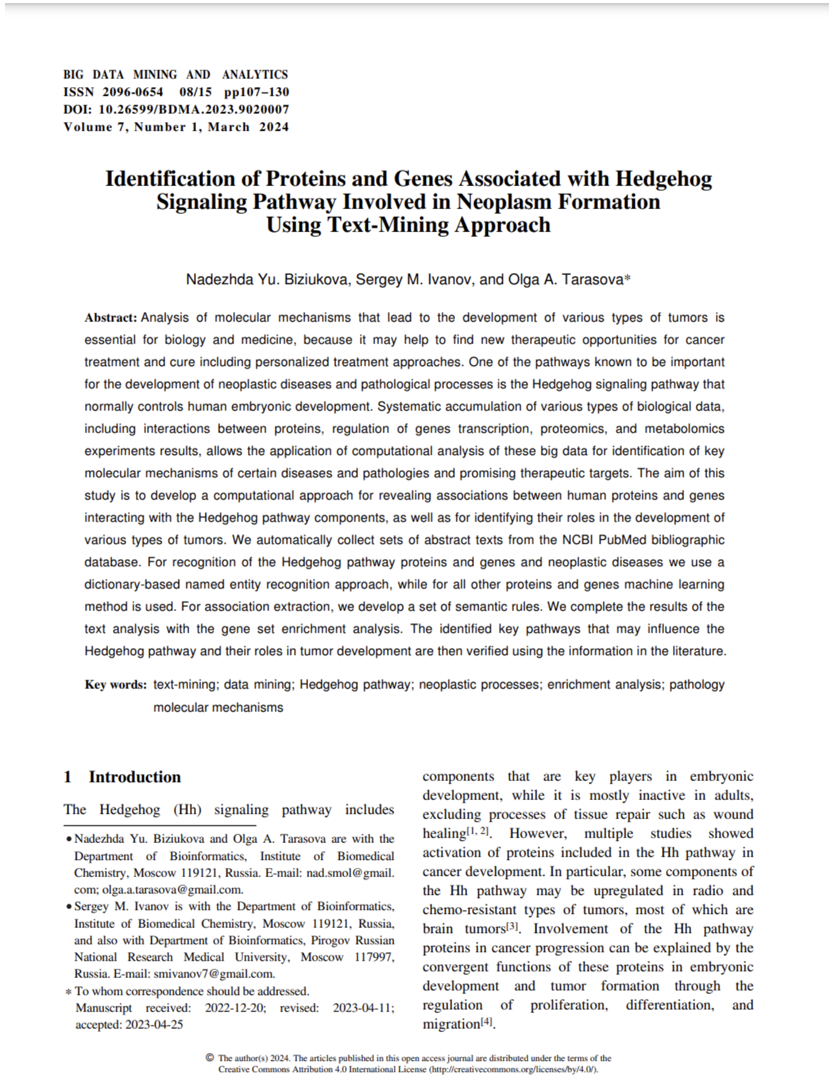
Specialty section:
ırticle was submitted to
mputational Genomics,
a section of the journal
Frontiers in Genetics

ived: *18 October 2020*
ed: *26 November 2020*
ed: *22 December 2020*

Citation:
rasova O, Ivanov S and
ov V (2020) Automated
ormation From Texts of
ıblications: Insights Into
V Treatment Strategies.
ont. Genet. 11:618862.
89/fgene.2020.618862

Text analysis can help to identify named entities (NEs) of small molecules, proteins, and genes. Such data are very important for the analysis of molecular mechanisms of disease progression and development of new strategies for the treatment of various diseases and pathological conditions. The texts of publications represent a primary source of information, which is especially important to collect the data of the highest quality due to the immediate obtaining information, in comparison with databases. In our study, we aimed at the development and testing of an approach to the named entity recognition in the abstracts of publications. More specifically, we have developed and tested an algorithm based on the conditional random fields, which provides recognition of NEs of (i) genes and proteins and (ii) chemicals. Careful selection of abstracts strictly related to the subject of interest leads to the possibility of extracting the NEs strongly associated with the subject. To test the applicability of our approach, we have applied it for the extraction of (i) potential HIV inhibitors and (ii) a set of proteins and genes potentially responsible for viremic control in HIV-positive patients. The computational experiments performed provide the estimations of evaluating the accuracy of recognition of chemical NEs and proteins (genes). The precision of the chemical NEs recognition is over 0.91; recall is 0.86, and the F1-score (harmonic mean of precision and recall) is 0.89; the precision of recognition of proteins and genes names is over 0.86; recall is 0.83; while F1-score is above 0.85. Evaluation of the algorithm on two case studies related to HIV treatment confirms our suggestion about the possibility of extracting the NEs strongly relevant to (i) HIV inhibitors and (ii) a group of patients i.e., the group of HIV-positive individuals with an ability to maintain an undetectable HIV-1 viral load overtime in the absence of antiretroviral therapy. Analysis of the results obtained provides insights into the function of proteins that can be responsible for viremic control. Our study demonstrated the applicability of the developed approach for the extraction of useful data on HIV treatment.

Keywords: text mining, data mining, named entity recognition, NER, virus-host interactions, HIV, viremic control



BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 08/15 pp107–130
DOI: 10.26599/BDMA.2023.9020007
Volume 7, Number 1, March 2024

Identification of Proteins and Genes Associated with Hedgehog Signaling Pathway Involved in Neoplasm Formation Using Text-Mining Approach

Nadezhda Yu. Biziukova, Sergey M. Ivanov, and Olga A. Tarasova*

Abstract: Analysis of molecular mechanisms that lead to the development of various types of tumors is essential for biology and medicine, because it may help to find new therapeutic opportunities for cancer treatment and cure including personalized treatment approaches. One of the pathways known to be important for the development of neoplastic diseases and pathological processes is the Hedgehog signaling pathway that normally controls human embryonic development. Systematic accumulation of various types of biological data, including interactions between proteins, regulation of genes transcription, proteomics, and metabolomics experiments results, allows the application of computational analysis of these big data for identification of key molecular mechanisms of certain diseases and pathologies and promising therapeutic targets. The aim of this study is to develop a computational approach for revealing associations between human proteins and genes interacting with the Hedgehog pathway components, as well as for identifying their roles in the development of various types of tumors. We automatically collect sets of abstract texts from the NCBI PubMed bibliographic database. For recognition of the Hedgehog pathway proteins and genes and neoplastic diseases we use a dictionary-based named entity recognition approach, while for all other proteins and genes machine learning method is used. For association extraction, we develop a set of semantic rules. We complete the results of the text analysis with the gene set enrichment analysis. The identified key pathways that may influence the Hedgehog pathway and their roles in tumor development are then verified using the information in the literature.

Key words: text-mining; data mining; Hedgehog pathway; neoplastic processes; enrichment analysis; pathology molecular mechanisms

1 Introduction

The Hedgehog (Hh) signaling pathway includes components that are key players in embryonic development, while it is mostly inactive in adults, excluding processes of tissue repair such as wound healing[1, 2]. However, multiple studies showed activation of proteins included in the Hh pathway in cancer development. In particular, some components of the Hh pathway may be upregulated in radio and chemo-resistant types of tumors, most of which are brain tumors[3]. Involvement of the Hh pathway proteins in cancer progression can be explained by the convergent functions of these proteins in embryonic development and tumor formation through the regulation of proliferation, differentiation, and migration[4].

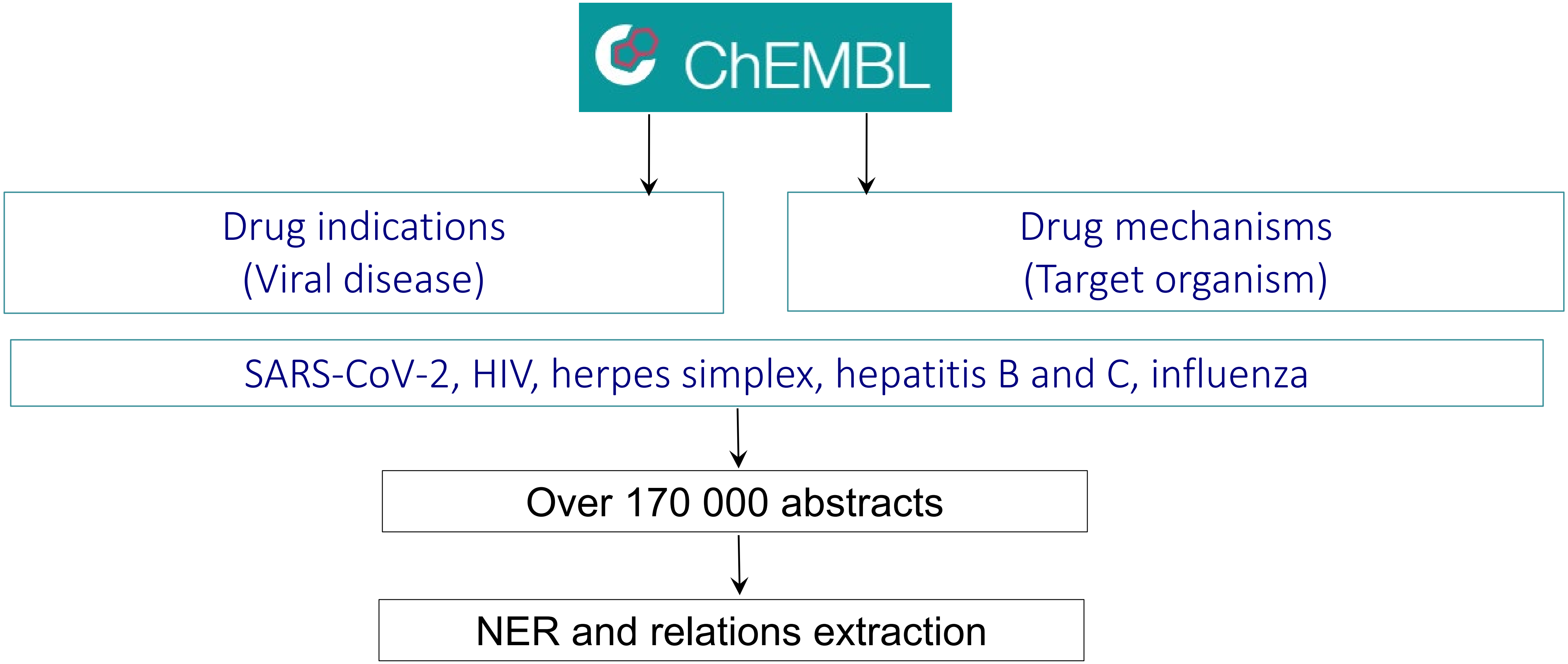
- Nadezhda Yu. Biziukova and Olga A. Tarasova are with the Department of Bioinformatics, Institute of Biomedical Chemistry, Moscow 119121, Russia. E-mail: nad.smol@gmail.com; olga.a.tarasova@gmail.com.
- Sergey M. Ivanov is with the Department of Bioinformatics, Institute of Biomedical Chemistry, Moscow 119121, Russia, and also with Department of Bioinformatics, Pirogov Russian National Research Medical University, Moscow 117997, Russia. E-mail: smivanov7@gmail.com.

∗ To whom correspondence should be addressed.

Manuscript received: 2022-12-20; revised: 2023-04-11; accepted: 2023-04-25

Extracting knowledge about viruses and the host (the human body), and potential antiviral agents based on the mining of massive amounts of scientific publications

Selection of antiviral compounds with direct antiviral activity



ChEMBL

Drug indications
(Viral disease)

Drug mechanisms
(Target organism)

SARS-CoV-2, HIV, herpes simplex, hepatitis B and C, influenza

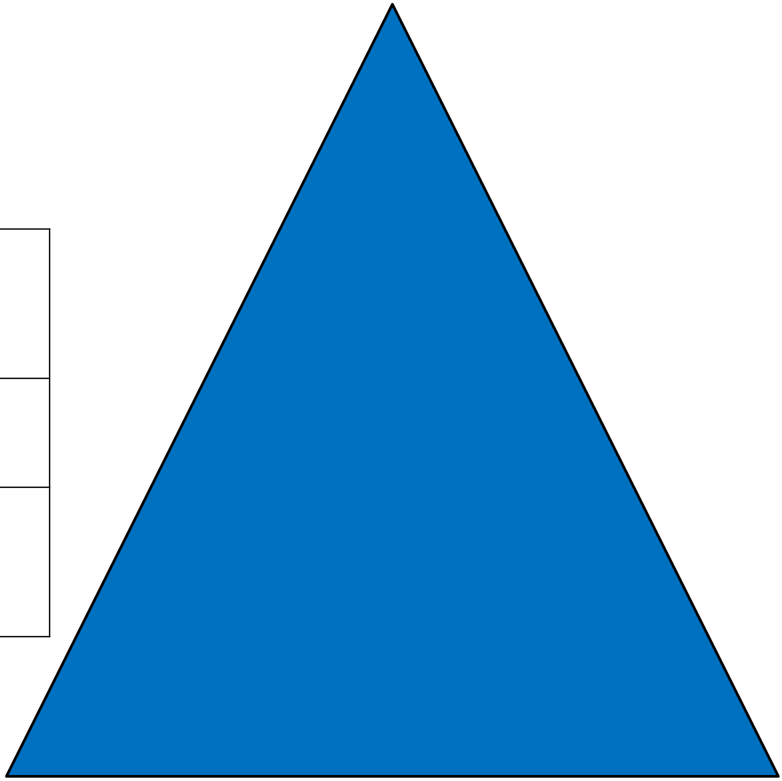
Over 170 000 abstracts

NER and relations extraction

N	Object types	Method of association extraction	Examples of relations
1	Chemical-Chemical	Pattern phrases	Combinations of drugs used in therapy; drug effects on biochemical parameters; drug-drug interactions; metabolism and biotransformation
2	Chemical-Gene	Pattern phrases	Effect of drug on biochemical parameters; effect on protein/gene
3	Chemical-Disease	Pattern phrases	Side effects and toxicity; use in therapy; use in prevention; possible biomarkers of disease
4	Chemical-miRNA	Co-occurrence	Effects of chemical compounds on molecular mechanisms
5	Chemical-SNP	Pattern phrases; Co-occurrence	Relationship between amino acid/nucleotide substitution and drug resistance/susceptibility
6	Chemical-Genotype	Pattern phrases	(Typically associations of HLA genotypes with drug hypersensitivity)
7	Gene-Gene	Pattern phrases	Interactions between proteins (including signaling pathways); part-to-part associations (protein-family)
8	Gene-Disease	Pattern phrases	Possible biomarkers; molecular mechanisms of pathogenesis; use in therapy
9	Gene-miRNA	Co-occurrence	Participation in regulation
10	Gene-SNP	Co-occurrence	Which protein substitutions lead to resistance/susceptibility to the drug
11	Disease-Disease	Pattern phrases	Main disease-symptom; comorbidities; multicomponent diseases
12	Disease-miRNA	Co-occurrence	Involvement in pathogenesis
13	Disease-SNP	Co-occurrence	Associations of changes in proteins/genes with disease manifestation and pathogenesis
14	Disease-Genotype	Pattern phrases; Co-occurrence	Other reactions and pathological processes associated with HLA genotypes

NER and extracting associations or relations

Virus



Object types	Number of recognized (unique)	Found in databases (unique)
Gene	15 844	4 034
miRNA	640	-
Disease	55 080	7 998
SNP	10953	-

Associations / relations	Gene-Disease;
	Gene-gene;
	Gene-miRNA;
	Chemical-Genotype	

Object types	Number of recognized (unique)	Found in databases (unique)
Chemical named entities	83 571	6 972

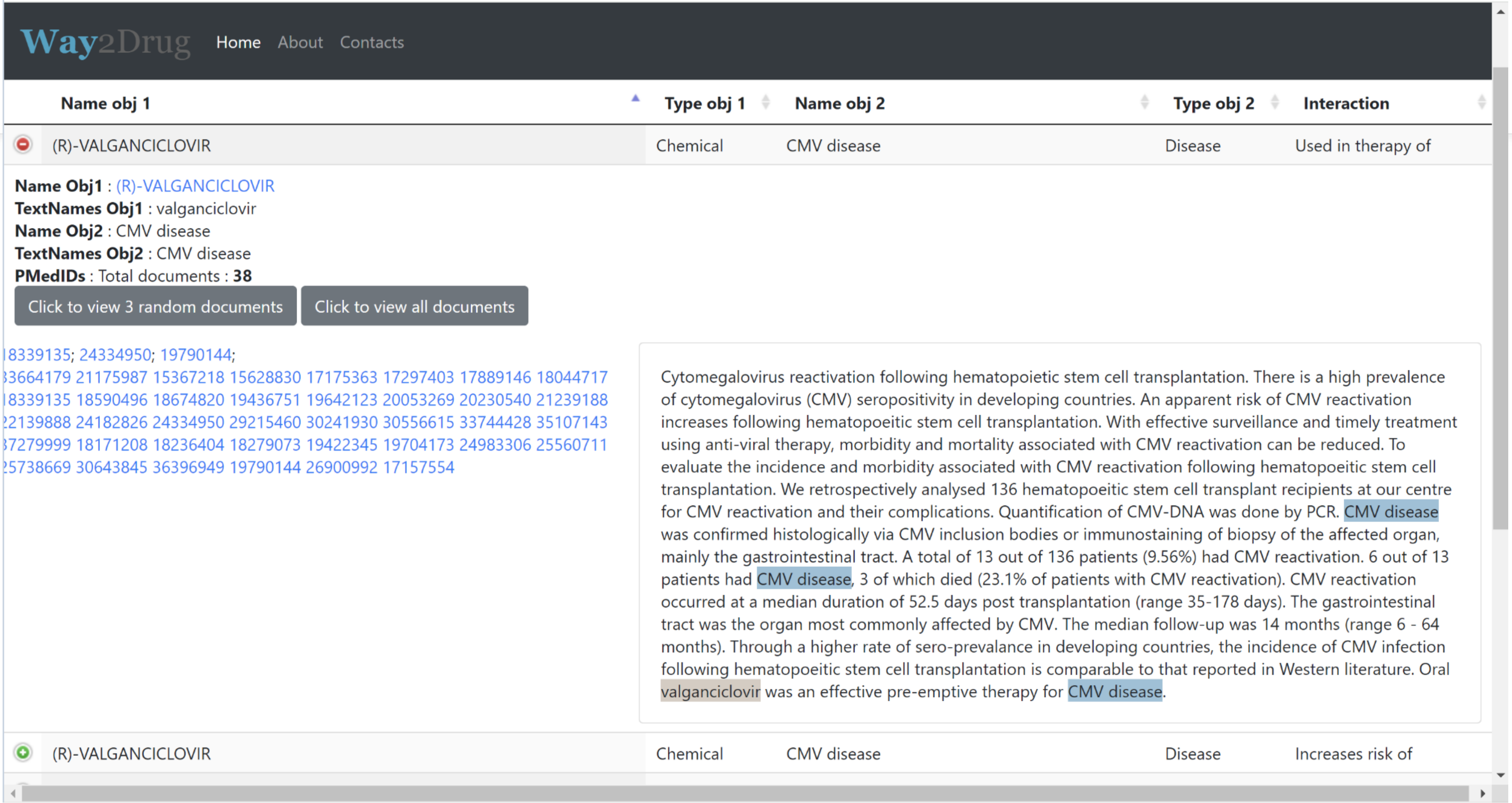
Associations / relations	Chemical-miRNA;
	Chemical-Genotype;
	Chemical-chemical	

Chemicals

Host (homo sapiens)

Associations / relations	Chemical-Disease;
	Chemical-Gene;
	Chemical-miRNA

Virus-host-chemicals interactions based on textome analysis for several viruses. Preliminary results



	Name obj 1	Type obj 1	Name obj 2	Type obj 2	Interaction
⊖	(R)-VALGANCICLOVIR	Chemical	CMV disease	Disease	Used in therapy of

Name Obj1 : (R)-VALGANCICLOVIR
TextNames Obj1 : valganciclovir
Name Obj2 : CMV disease
TextNames Obj2 : CMV disease
PMedIDs : Total documents : **38**

Click to view 3 random documents | Click to view all documents

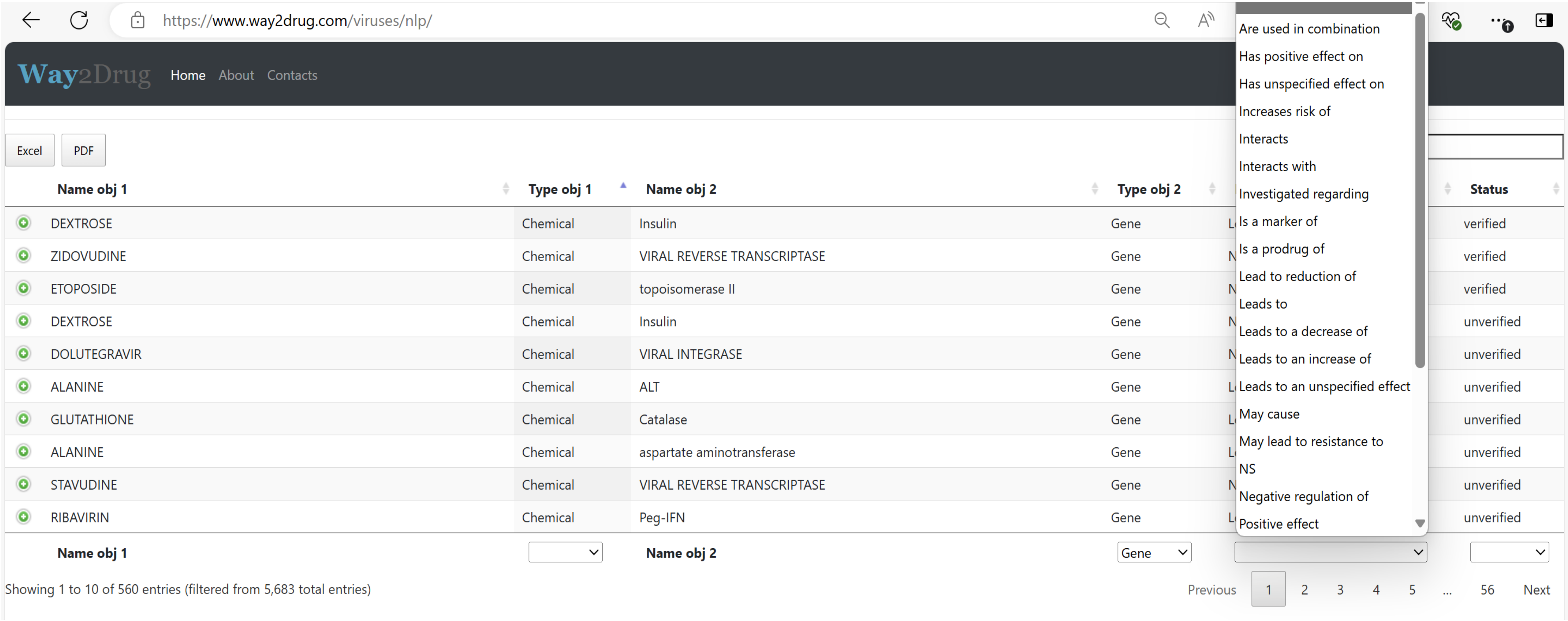
18339135; 24334950; 19790144;
33664179 21175987 15367218 15628830 17175363 17297403 17889146 18044717
18339135 18590496 18674820 19436751 19642123 20053269 20230540 21239188
22139888 24182826 24334950 29215460 30241930 30556615 33744428 35107143
37279999 18171208 18236404 18279073 19422345 19704173 24983306 25560711
25738669 30643845 36396949 19790144 26900992 17157554

Cytomegalovirus reactivation following hematopoietic stem cell transplantation. There is a high prevalence of cytomegalovirus (CMV) seropositivity in developing countries. An apparent risk of CMV reactivation increases following hematopoetic stem cell transplantation. With effective surveillance and timely treatment using anti-viral therapy, morbidity and mortality associated with CMV reactivation can be reduced. To evaluate the incidence and morbidity associated with CMV reactivation following hematopoietic stem cell transplantation. We retrospectively analysed 136 hematopoietic stem cell transplant recipients at our centre for CMV reactivation and their complications. Quantification of CMV-DNA was done by PCR. CMV disease was confirmed histologically via CMV inclusion bodies or immunostaining of biopsy of the affected organ, mainly the gastrointestinal tract. A total of 13 out of 136 patients (9.56%) had CMV reactivation. 6 out of 13 patients had CMV disease, 3 of which died (23.1% of patients with CMV reactivation). CMV reactivation occurred at a median duration of 52.5 days post transplantation (range 35-178 days). The gastrointestinal tract was the organ most commonly affected by CMV. The median follow-up was 14 months (range 6 - 64 months). Through a higher rate of sero-prevalance in developing countries, the incidence of CMV infection following hematopoeitic stem cell transplantation is comparable to that reported in Western literature. Oral valganciclovir was an effective pre-emptive therapy for CMV disease.

	Name obj 1	Type obj 1	Name obj 2	Type obj 2	Interaction
⊕	(R)-VALGANCICLOVIR	Chemical	CMV disease	Disease	Increases risk of

SARS-CoV-2, HIV-1, hepatitis C and B, influenza, Zika, Dengue, Western Nile

Examples of relations between identified objects extracted using literature mining

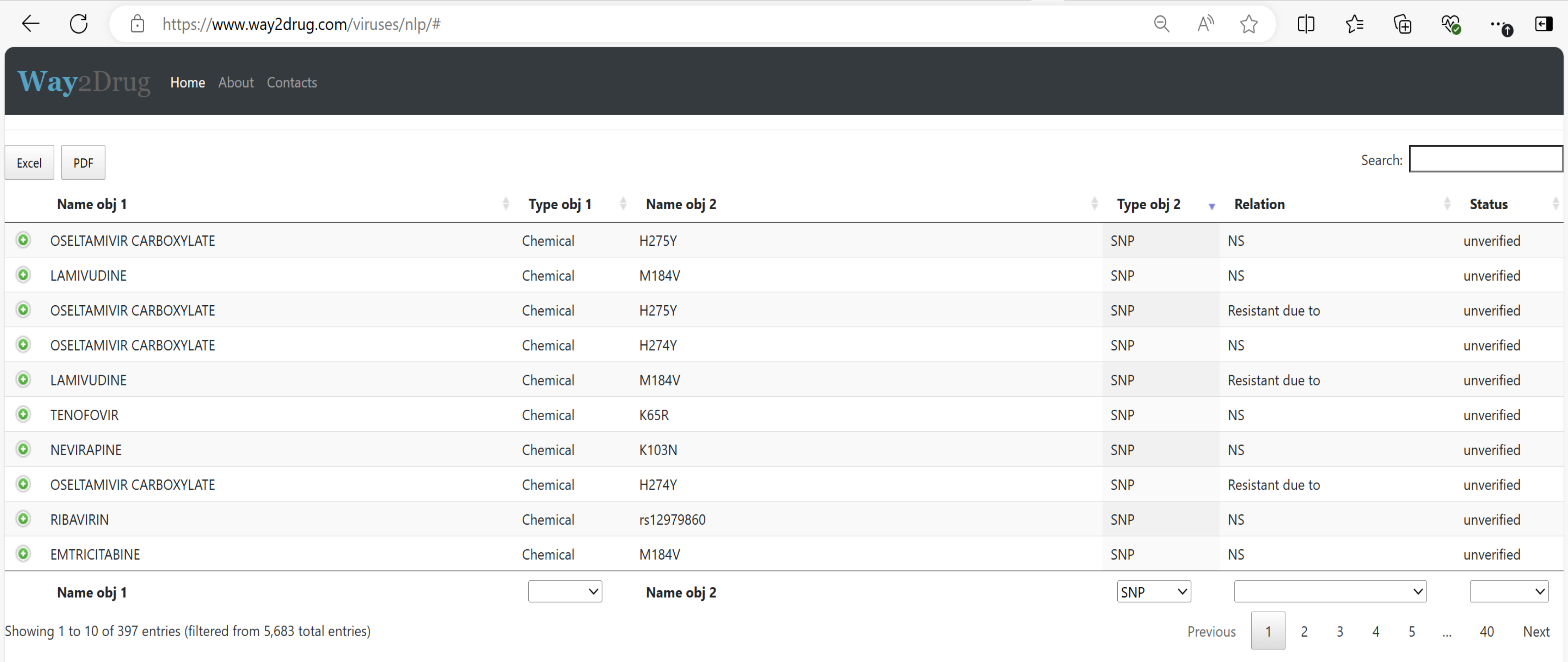


Name obj 1	Type obj 1	Name obj 2	Type obj 2		Status
DEXTROSE	Chemical	Insulin	Gene	L	verified
ZIDOVUDINE	Chemical	VIRAL REVERSE TRANSCRIPTASE	Gene	N	verified
ETOPOSIDE	Chemical	topoisomerase II	Gene	N	verified
DEXTROSE	Chemical	Insulin	Gene	N	unverified
DOLUTEGRAVIR	Chemical	VIRAL INTEGRASE	Gene	N	unverified
ALANINE	Chemical	ALT	Gene	L	unverified
GLUTATHIONE	Chemical	Catalase	Gene	L	unverified
ALANINE	Chemical	aspartate aminotransferase	Gene	L	unverified
STAVUDINE	Chemical	VIRAL REVERSE TRANSCRIPTASE	Gene	N	unverified
RIBAVIRIN	Chemical	Peg-IFN	Gene	L	unverified
Name obj 1		Name obj 2	Gene		

Showing 1 to 10 of 560 entries (filtered from 5,683 total entries)

https://www.way2drug.com/viruses/nlp/

Examples of amino acid substitutions associated with viral drug resistance



Name obj 1	Type obj 1	Name obj 2	Type obj 2	Relation	Status
OSELTAMIVIR CARBOXYLATE	Chemical	H275Y	SNP	NS	unverified
LAMIVUDINE	Chemical	M184V	SNP	NS	unverified
OSELTAMIVIR CARBOXYLATE	Chemical	H275Y	SNP	Resistant due to	unverified
OSELTAMIVIR CARBOXYLATE	Chemical	H274Y	SNP	NS	unverified
LAMIVUDINE	Chemical	M184V	SNP	Resistant due to	unverified
TENOFOVIR	Chemical	K65R	SNP	NS	unverified
NEVIRAPINE	Chemical	K103N	SNP	NS	unverified
OSELTAMIVIR CARBOXYLATE	Chemical	H274Y	SNP	Resistant due to	unverified
RIBAVIRIN	Chemical	rs12979860	SNP	NS	unverified
EMTRICITABINE	Chemical	M184V	SNP	NS	unverified
Name obj 1		Name obj 2	SNP		

Showing 1 to 10 of 397 entries (filtered from 5,683 total entries)

Conclusions

- We developed the approach to extract associations between automatically recognised entities corresponding to (a) chemical named entities; (b) proteins, genes, miRNAs; (c) diseases and disorders.
- The developed approach and algorithms were validated in several studies, including the identification of genes associated with HIV infection progression and therapeutic outcome; the search for proteins and genes involved in neoplasm development and associated with the Hedgehog pathway.
- Based on the developed approach, we created an automated pipeline aimed at extracting knowledge about viruses and the host (the human body) and potential antiviral agents based on the mining of massive amounts of scientific publications.

Selected publications

- **Tarasova, O.**, Biziukova, N., Shemshura, A., Filimonov, D., Kireev, D., Pokrovskaya, A., Poroikov, V. Identification of Molecular Mechanisms Involved in Viral Infection Progression Based on Text Mining: Case Study for HIV Infection. International Journal of Molecular Sciences.; 2023. 24(2), 1465; https://doi.org/10.3390/ijms24021465

- Ivanov SM, **Tarasova OA**, Poroikov VV. Transcriptome-based analysis of human peripheral blood reveals regulators of immune response in different viral infections. Front Immunol. 2023 Sep 19;14:1199482. DOI: 10.3389/fimmu.2023.1199482.

- Rhee, S.-Y., Boehm, M., **Tarasova, O.**, Di Teodoro, G., Abecasis, A.B., Sönnerborg, A., Bailey, A.J., Kireev, D., Zazzi, M., Shafer, R.W. Spectrum of Atazanavir Selected Protease Inhibitor Resistance Mutations.; Pathogens. 2022. 11(5), 546. DOI: 10.3390/pathogens11050546.

- Pikalyova, K., Orlov, A., Lin, A., **Tarasova, O.**, Marcou, M., Horvath, D., Poroikov, V., Varnek, A. HIV-1 drug resistance profiling using amino acid sequence space cartography.; Bioinformatics, 38 (8). 2022. 2307-2314. DOI: 10.1093/bioinformatics/btac090.

- **Tarasova, O.**, Poroikov, V. Machine learning in discovery of new antivirals and optimization of viral infections therapy.; Current Medicinal Chemistry. 2021. 28 (38). 7840-7861. DOI: 10.2174/0929867328666210504114351.

- Khandazhinskaya, A.L., Mercurio, V., Maslova, A.A., Nahui Palomino, R.A., Novikov, M.S., Matyugina, E.S., Paramonova, M.P., Kukhanova, M.K., Fedorova, N.E., Yurlov, K.I., Kushch, A.A., **Tarasova, O.**, Margolis, L., Kochetkov, S.N., Vanpouille, C.; Dual-targeted anti-CMV/anti-HIV-1 heterodimers. Biochimie. 2021. 189. 169 (180). DOI: 10.1016/j.biochi.2021.06.011.

- Tarasova, O., Rudik, A., Kireev, D., Poroikov, V.; RHIVDB: A Freely Accessible Database of HIV Amino Acid Sequences and Clinical Data of Infected Patients. Frontiers in Genetics. 2021. 12. 679029. DOI: 10.3389/fgene.2021.679029.

- **Tarasova, O.A.**, Biziukova, N.Y., Rudik, A.V., Dmitriev, A.V., Filimonov, D.A., Poroikov, V.V.; Extraction of Data on Parent Compounds and Their Metabolites from Texts of Scientific Abstracts. Journal of Chemical Information and Modeling. 2021. 61 (4). 1683-1690. DOI: 10.1021/acs.jcim.0c01054.

- **Tarasova, O.A.**, Rudik, A.V., Ivanov, S.M., Lagunin, A.A., Poroikov, V.V., Filimonov, D.A.; Machine Learning Methods in Antiviral Drug Discovery. Topics in Medicinal Chemistry. 2021. 37, 245-279. DOI: 10.1007/7355_2021_121.

Selected publications

- Ivanov, S., Filimonov, D., **Tarasova, O.**; A computational analysis of transcriptional profiles from CD8(+) T lymphocytes reveals potential mechanisms of HIV/AIDS control and progression. Computational and Structural Biotechnology Journal. 2021. 19. 2447-2459. DOI: 10.1016/j.csbj.2021.04.056.

- Biziukova, N., **Tarasova, O.**, Ivanov, S., Poroikov, V.; Automated Extraction of Information From Texts of Scientific Publications: Insights Into HIV Treatment Strategies. Frontiers in Genetics. 2020. 11, 618862. DOI: 10.3389/fgene.2020.618862.

- Ivanov, S., Lagunin, A., Filimonov, D., **Tarasova, O.**; Network-Based Analysis of OMICs Data to Understand the HIV–Host Interaction. Frontiers in Microbiology. 2020. 11, 1314. DOI: 10.3389/fmicb.2020.01314.

- **Tarasova, O.**, Ivanov, S., Filimonov, D.A., Poroikov, V.; Data and text mining help identify key proteins involved in the molecular mechanisms shared by SARS-CoV-2 and HIV1. Molecules. 2020. 25 (12). 25122944. DOI: 10.3390/molecules25122944.

- **Tarasova, O.**, Biziukova, N., Kireev, D., Lagunin, A., Ivanov, S., Filimonov, D., Poroikov, V.; A computational approach for the prediction of treatment history and the effectiveness or failure of antiretroviral therapy;2020;International Journal of Molecular Sciences. 21(3). 748. DOI: 10.3390/ijms21030748.

- Poroikov, V.V., Filimonov, D.A., Gloriozova, T.A., Lagunin, A.A., Druzhilovskiy, D.S., Rudik, A.V., Stolbov, L.A., Dmitriev, A.V., **Tarasova, O.A.**, Ivanov, S.M., Pogodin, P.V.; Computer-aided prediction of biological activity spectra for organic compounds: the possibilities and limitations. Russian Chemical Bulletin. 2019. 68 (12). 2143-2154. DOI: 10.1007/s11172-019-2683-0.

- **Tarasova, O.A.**, Biziukova, N.Y., Filimonov, D.A., Poroikov, V.V., Nicklaus, M.C. Data Mining Approach for Extraction of Useful Information about Biologically Active Compounds from Publications. Journal of Chemical Information and Modeling. 2019. 59, (9), 3635-3644, DOI: 10.1021/acs.jcim.9b00164

- Demidova, A.V., **Tarasova, O.A.** Application of neural networks to the analysis of the resistance of the human immunodeficiency virus to HIV reverse transcriptase inhibitors. CEUR Workshop Proceedings. 2407.47-52. 2019.

Selected publications

- **Tarasova, O.**, Biziukova, N., Filimonov, D., Poroikov, V.;A computational approach for the prediction of HIV resistance based on amino acid and nucleotide descriptors. Molecules. 2018. 23(11), 2751. DOI: 10.3390/molecules23112751.

- **Tarasova, O.**, Poroikov, V., Veselovsky, A. Molecular Docking Studies of HIV-1 Resistance to Reverse Transcriptase Inhibitors: Mini-Review. 2018. Molecules. 23(5). 1233. DOI: 10.3390/molecules23051233.

- **Tarasova, O.**, Poroikov, V. HIV resistance prediction to reverse transcriptase inhibitors: Focus on open data. 2018. Molecules. 23. 4. 956.

- Dmitriev, A., Rudik, A., Filimonov, D., Lagunin, A., Pogodin, P., Dubovskaja, V., Bezhentsev, V., Ivanov, S., Druzhilovsky, D., **Tarasova, O.**, Poroikov, V. Integral estimation of xenobiotics' toxicity with regard to their metabolism in human organism. 2017. Pure and Applied Chemistry. 89 (10), 1449-1458. DOI: 10.1515/pac-2016-1205.

- **Tarasova, O.**, Filimonov, D., Poroikov, V. PASS-based approach to predict HIV-1 reverse transcriptase resistance. Journal of Bioinformatics and Computational Biology. 2017. DOI: 15(2),1650040. DOI: 10.1142/S0219720016500402.

- **Tarasova, O.A.**, Filimonov, D.A., Poroikov, V.V.;Computational prediction of human immunodeficiency resistance to reverse transcriptase inhibitors. Biomeditsinskaya Khimiya. 2017. 63(5), 457-460. DOI: 10.18097/PBMC20176305457.

Acknowledgements

Nadezhda Biziukova, PhD student, IBMC
Nikita Ionov, junior scientist, IBMC
Sergey Ivanov, PhD, senior researcher, IBMC
Anastassia Rudik, PhD, senior researcher, IBMC
Dmitry Filimonov, PhD, lead researcher, IBMC,
Alexey Lagunin, Dr. Sci., lead researcher, IBMC, head of bioinformatics department, RNRMU
Vladimir Poroikov, Dr. Sci., Corr. Member of RAS
Anastassia Pokrovskaya, Dr. Sci., Central Research Institutre of Epidemiology
Dmitry Kireev, PhD, Central Research Institutre of Epidemiology
Andrey Shemshura, PhD, Clinical Center of HIV treatment and Cure, Krasnodar

Thank you for your attention

Olga.a.Tarasova@gmail.com

@fiolandesky

Funding: The study was supported by RNF grants 17-75-10187, 19-75-10097,
the Program of Basic Scientific Research in the Russian Federation for the long-term period (2021-2030) (№ 124050800018-9)