Bioinformatics, 36(22-23), 2021, 5507–5513
doi: 10.1093/bioinformatics/btaa1066
Advance Access Publication Date: 26 December 2020
Original Paper

Databases and ontologies

Virxicon: a lexicon of viral sequences

Mateusz Kudla[1,2,†], Kaja Gutowska[1,3,†], Jaroslaw Synak[1], Mirko Weber[2], Katrin Sophie Bohnsack [2], Piotr Lukasiak[1,3], Thomas Villmann[2], Jacek Blazewicz[1,3] and Marta Szachniuk [1,3,*]

[1]Institute of Computing Science and European Centre for Bioinformatics and Genomics, Poznan University of Technology, Poznan 60-965, Poland, [2]Saxon Institute for Computational Intelligence and Machine Learning, University of Applied Sciences Mittweida, Mittweida 09648, Germany and [3]Department of Structural Bioinformatics, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan 61-704, Poland

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Zhiyong Lu

Received on August 31, 2020; revised on November 18, 2020; editorial decision on December 9, 2020; accepted on December 11, 2020

Abstract

Motivation: Viruses are the most abundant biological entities and constitute a large reservoir of genetic diversity. In recent years, knowledge about them has increased significantly as a result of dynamic development in life sciences and rapid technological progress. This knowledge is scattered across various data repositories, making a comprehensive analysis of viral data difficult.

Results: In response to the need for gathering a comprehensive knowledge of viruses and viral sequences, we developed Virxicon, a lexicon of all experimentally acquired sequences for RNA and DNA viruses. The ability to quickly obtain data for entire viral groups, searching sequences by levels of taxonomic hierarchy—according to the Baltimore classification and ICTV taxonomy—and tracking the distribution of viral data and its growth over time are unique features of our database compared to the other tools.

Availabilityand implementation: Virxicon is a publicly available resource, updated weekly. It has an intuitive web interface and can be freely accessed at http://virxicon.cs.put.poznan.pl/.

Contact: mszachniuk@cs.put.poznan.pl

1 Introduction

Viruses constitute the most abundant biological entities and a large reservoir of genetic diversity (de Cárcer *et al.*, 2015; Suttle, 2005). Understanding their nature turns out crucial in public health policy—outbreaks of viral diseases with varying range tend to occur in the human population every few years (Baize *et al.*, 2014; Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020; Cotten *et al.*, 2013; Gire *et al.*, 2014; Haagmans *et al.*, 2014). Reducing the spread of the virus and determining effective treatment methods (Parrish *et al.*, 2008) require rapid access to up-to-date data and their multi-faceted analysis. Meanwhile, knowledge about viruses develops dynamically, influencing changes in their taxonomy. This is primarily due to the wide application of innovative technologies, such as Next-Generation Sequencing (NGS), which provide huge amounts of new genomic data (Datta, 2015; Marston *et al.*, 2013). Exploration of these data and their phylogenetic analysis allow us to broaden and refine the knowledge about the increasing number of virus species (Bao *et al.*, 2004). In this respect, it is important to develop bioinformatics tools that allow the collection of ever-increasing viral data, including the latest ones, their quick taxonomy-related search and effective analysis.

Currently, there exist several virus-related databases. Some specialize in specific virus types such as DPVweb for plant and fungal virus genomes (Adams and Antoniw, 2006) or GISAID for influenza and coronavirus data (Elbe and Buckland Merrett, 2017). NCBI Viral Genomes (Brister *et al.*, 2015) collects all publicly available virus genome sequences, which undergo a curatorial process. The database offers a variety of filtering options such as taxonomy or accession number. The information about the sequence itself and the supplementary data are available separately (the sequence is given in FASTA format and the information corresponding to the search criteria as CSV or XML) and must be further processed if both types of information are to be combined. The Virus Pathogen Database and Analysis Resource (ViPR) (Pickett *et al.*, 2012) is a virus database with a focus on integrated bioinformatics analysis tools such as multiple sequence alignments or phylogenetic tree construction. It applies advanced filtering methods within a virus family, however, the query-specific information and the sequences themselves are downloaded as separate data. ViralZone (Hulo *et al.*, 2011; Masson et al., 2012) contains molecular biological knowledge about viruses as well as references to genomic and proteomic sequence information. This resource does not store sequences. Instead, it contains references to external databases, such as NCBI or UniProtKB (UniProt Consortium, 2019), where the data (including sequences) can be found and downloaded. In response to the recent pandemic caused by SARS-Cov-2, other useful tools are also being developed, including specialized databases, like the repository of validated PDB models of CoV-2 proteins (Wlodawer et al., 2020).

Here, we present Virxicon—a lexicon of all experimentally determined viral sequences. It serves as a comprehensive knowledge base that gathers in one place the information about viruses and viral sequences, annotated due to current virological taxonomy. According to the International Committee on Taxonomy of Viruses (ICTV) (Lefkowitz et al., 2018), viruses fall into ten groups, which include seven Baltimore classes (Baltimore, 1971): I—dsDNA, II—ssDNA, III—dsRNA, IV—ssRNA(+), V—ssRNA(-), VI—ssRNA-RT, VII—dsDNA-RT; and three additional classes defined by ICTV (Walker et al., 2019): ssDNA/dsDNA, ssRNA(±)/ssRNA(-) and ssRNA (viroids). The Virxicon collection is automatically updated once a week with sequences coming from the NCBI Viral Genome database (Brister et al., 2015) and GenBank (Benson et al., 2012). Compared with other existing databases and tools, Virxicon allows users to conveniently retrieve the information about viruses and viral sequences according to the level of the taxonomic hierarchy and tracking the updated distribution of the viral data over time, which are sought-after and desirable functions. These facilities are unique features of the presented tool. Note that although ViralZone can also filter viral data according to the Baltimore classes, it does not allow users to download the data directly. This fact was one of the motivations behind the development of Virxicon.

2 Materials and methods

2.1 Data acquisition into the database

The weekly update of the Virxicon database is carried out according to the following scheme. The first stage is to retrieve information about currently used classification groups, then the list of all available viral genomes is downloaded from NCBI (Brister et al., 2015). It holds NCBI and GenBank IDs (accession numbers) of all relevant records. Note that different accession numbers in GenBank may correspond to the same accession number in NCBI since GenBank comprises information of different strains. From the accession list, the Virxicon importing module selects accession numbers of the sequences, which are not yet present in the system database or were recently updated in the external resources, to upload or modify them. It also identifies previously imported sequences, which were removed from the accession list, to discard them from Virxicon. Sequences to be placed in the database are assigned to classification groups. Complete records corresponding to their identifiers are downloaded in the XML file. The file is parsed to extract data and create a database entry. A reference to related sequences is obtained via GenBank ID—in the case of the NCBI entry—or via the corresponding NCBI reference sequence ID—in the case of the GenBank entry. Additionally, genes and coding DNA sequences link to the UniProt ID as the reference to the functional information deposited in UniProt Database (UniProt Consortium, 2019). The complete ICTV taxonomy of viruses is downloaded from https://talk.ictvonline.org/files/master-species-lists/. A general scheme of dataflow in the Virxicon system is presented in Figure 1.

2.2 Virus classification

The current taxonomy of viruses applied in Virxicon (Fig. 2)—including the division into particular groups—follows the up-to-date guidelines of the International Committee on Taxonomy of Viruses (ICTV) (ICTV, 2020; Lefkowitz et al., 2018; Walker et al., 2019). It is periodically updated as the knowledge about viruses develops.

The systematics of viruses is based on the molecular type (DNA, RNA), the nucleic acid structure (single- or double-stranded, circular or linear), the strategy of genomic replication (negative-sense, positive-sense, ambisense polarities), the structure of the virus capsid and whether or not it is enveloped (Baltimore, 1971; Lefkowitz, 2015; Simmonds, 2015). Following these criteria, we distinguish DNA and RNA viruses, and RNA/DNA viruses with gene encoding reverse transcriptase in the genome. DNA viruses fall into three groups having single- or double-stranded DNA: dsDNA, ssDNA, ssDNA/dsDNA. RNA viruses divide into four groups according to single- or double-stranded structure and positive (+), negative (-) or ambisense (±) polarities: dsRNA, ssRNA(+), ssRNA(-) and ssRNA(±)/ssRNA(-). Both RNA and DNA viruses have also been distinguished with reverse transcriptase viruses: ssRNA-RT and dsDNA-RT. Besides, a group of viroids is included and annotated as ssRNA (viroids). Altogether there are seven independent classes according to Baltimore classification (Baltimore, 1971) and three additional groups according to current virus taxonomy proposed by ICTV (Walker et al., 2019). The latter three groups include RNA viruses with ambisense strand polarity, viruses covering various types of DNA genomes (examples are viruses from Pleolipoviridae family) and viroids.

2.3 Virxicon system implementation

Virxicon in build based on PostgreSQL relational database using the .NET Core framework backend application, which ensures high efficiency, and allows to run the system on multiple operating systems (Windows, Linux and macOS). All the data are stored locally, which enables their quick access and processing. The Virxicon content is updated using Hangfire. It allows performing time-consuming operations in the background without blocking the main application thread. Access to the database entries is possible via API that uses the GraphQL engine, in which we used advanced mechanisms such as data loaders that significantly reduce the number of the database queries and the response time. We also provide web application created with the React framework, which allows accessing the database using a graphical user interface that works with most of the available web browsers. The system is hosted and maintained by the Institute of Computing Science, Poznan University of Technology, Poland.

3 Results

3.1 Database content

As of August 2, 2020, Virxicon holds over 231 000 viral sequences. Given the taxonomic hierarchy, they are organized into 4 Realms, 9 Kingdoms, 54 Orders, 6 Suborders, 154 Families, 88 Subfamilies, 1162 Genera, 57 Subgenera. The following information is stored for every sequence: original ID (from GenBank or NCBI), sequence definition (short description), complete taxonomy, molecule type (RNA, DNA), complete nucleotide sequence, coordinates of the genes and coding regions, gene and CDS sequences, protein sequences with NCBI IDs, gene ID and protein ID linked to the UniProt ID as the reference to functional information, collection date, update date, place of data collection (country and city—if available), host. The molecule type and coordinates of the coding area are of great importance—they specify the necessary transformations that bring sequences into the same orientation to counteract the differences of individual raw sequence entries.

The database also provides statistical data in the form of tables (such as Table 1), pie charts with data distribution and column charts with data growth. Pie charts visualize data distribution by molecule type and virus group (see Fig. 3), topology and molecular type, topology and virus group, resource and molecular type, resource and virus group. Column charts show the monthly growth of the total number of sequences and the number of sequences for individual groups, molecular types, topologies and resources. All statistics are updated automatically along with the database contents.

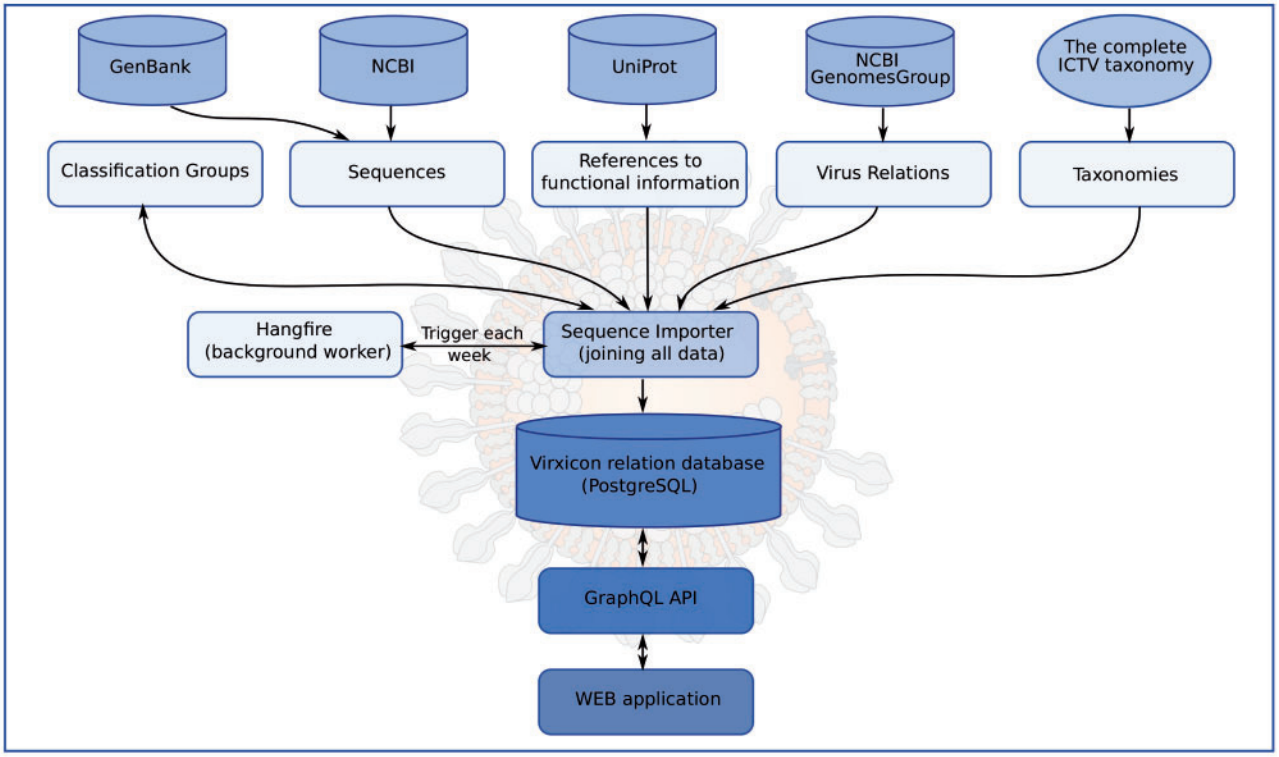

GenBank
NCBI
UniProt
NCBI GenomesGroup
The complete ICTV taxonomy
Classification Groups
Sequences
References to functional information
Virus Relations
Taxonomies
Hangfire (background worker)
Trigger each week
Sequence Importer (joining all data)
Virxicon relation database (PostgreSQL)
GraphQL API
WEB application

Fig. 1. Dataflow in the Virxicon system

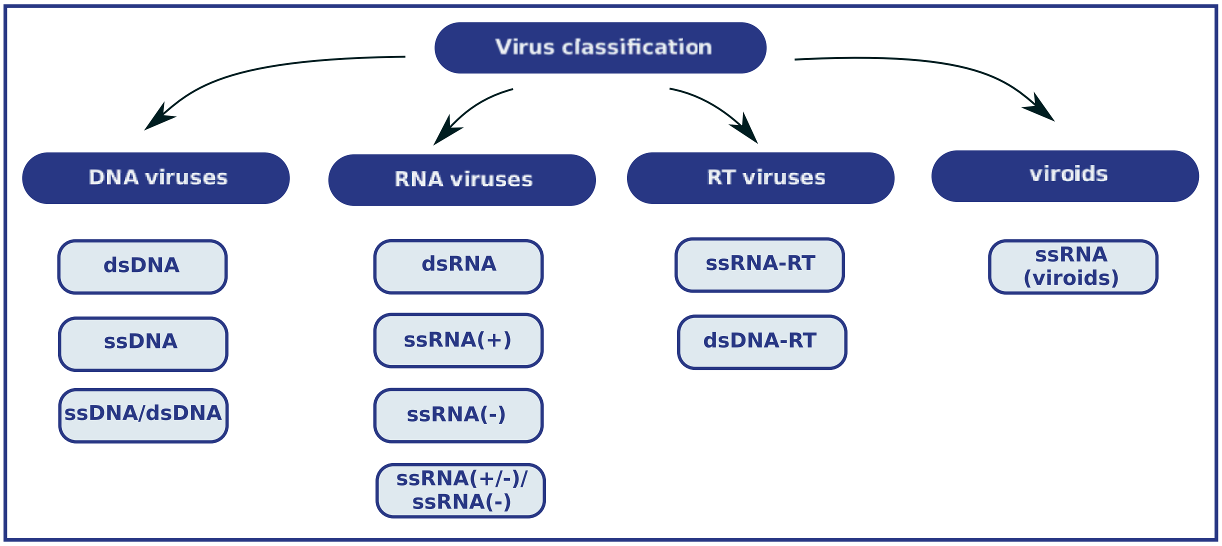


Fig. 2. Viral groups in Virxicon

3.2 User interface

Virxicon can be operated via a web application or API. The former is easier and more convenient to use, while the latter allows using the system more widely.

To run the web application, users open the address http://virxicon.cs.put.poznan.pl/ in their web browsers. The interface consists of 5 pages: Search, News, Statistics, Help and Citations (Fig. 4). The *News* page publishes information about the updates. In *Statistics*, users can see current holdings of the database with charts showing data distribution and data growth—these data can be downloaded in the textual and graphical format. Both pages are updated weekly. *Help* explains all the options of web application and API, using short video tutorials. The *Search* page enables defining the direct or indirect query to search for viral sequences. In direct case, users manually enter the species name or select the level of taxonomic hierarchy from the drop-down list. The indirect option allows selecting virus group(s), molecular type(s), topology(-ies) and resource(s). If users choose several criteria, the system combines them into a single query and looks for entries, which meet their conjunction. If they do not define any criteria and click the *Search* button, Virxicon outputs the list of all records in the database.

The search result is a list of viral sequences. Users can select entries of the interest and download their basic descriptions in the CSV file or their sequences in FASTA format. The whole result table can be additionally filtered based on the user-provided keywords. By clicking on the selected row of the table, users go to the page with

Table 1. The number of sequences in the database by classification, molecular types, topologies and resources (August 2, 2020)

	Virus group		Molecular type				Topology				Resource			
			RNA		DNA		Linear		Circular		NCBI		GenBank	
DNA viruses														
dsDNA	19 896	8.6%	3	0.0%	19 893	8.6%	13 963	6.0%	5933	2.6%	4385	1.9%	15 511	6.7%
ssDNA	22 405	9.7%	9	0.0%	22 396	9.7%	4912	2.1%	17 493	7.6%	1528	0.7%	20 877	9.0%
ssDNA/dsDNA	28	0.0%	0	0.0%	28	0.0%	2	0.0%	26	0.0%	14	0.0%	14	0.0%
Total number	42 329	18.3%	12	0.0%	42 317	18.3%	18 877	8.2%	23 452	10.1%	5927	2.6%	36 402	15.7%
RNA viruses														
dsRNA	87 236	37.7%	86 651	37.4%	585	0.3%	87 236	37.7%	0	0.0%	1379	0.6%	85 857	37.1%
ssRNA(+)	51 650	22.3%	51 427	22.2%	223	0.1%	51 624	22.3%	26	0.0%	1978	0.9%	49 672	21.5%
ssRNA(-)	15 762	6.8%	15 698	6.8%	64	0.0%	15 412	6.7%	350	0.2%	1057	0.5%	14 705	6.4%
ssRNA(±)/ ssRNA(-)	3255	1.4%	3247	1.4%	8	0.0%	3251	1.4%	4	0.0%	245	0.1%	3010	1.3%
Total number	157 903	68.2%	157 023	67.8%	880	0.4%	157 523	68.0%	380	0.2%	4659	2.0%	153 244	66.2%
RT viruses														
ssRNA-RT	9129	3.9%	4074	1.8%	5055	2.2%	9.121	3.9%	8	0.0%	94	0.0%	9035	3.9%
dsDNA-RT	10 901	4.7%	14	0.0%	10 887	4.7%	1542	0.7%	9359	4.0%	112	0.0%	10 789	4.7%
Total number	20 030	8.7%	4088	1.8%	15 942	6.9%	10 663	4.6%	9367	4.0%	206	0.1%	19 824	8.6%
Viroids														
ssRNA(viroids)	11 241	4.9%	11 110	4.8%	131	0.1%	482	0.2%	10 759	4.6%	39	0.0%	11 202	4.8%

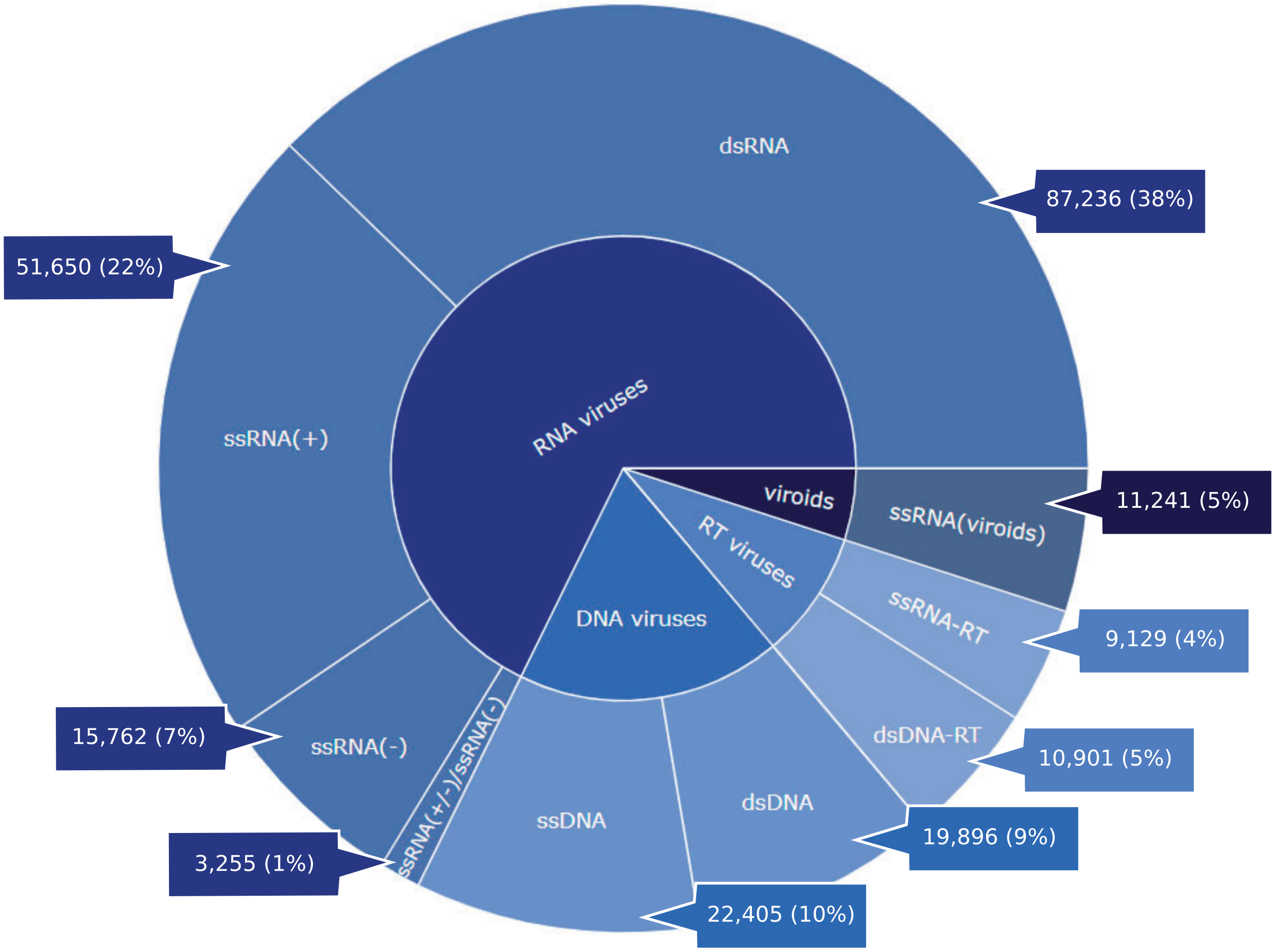
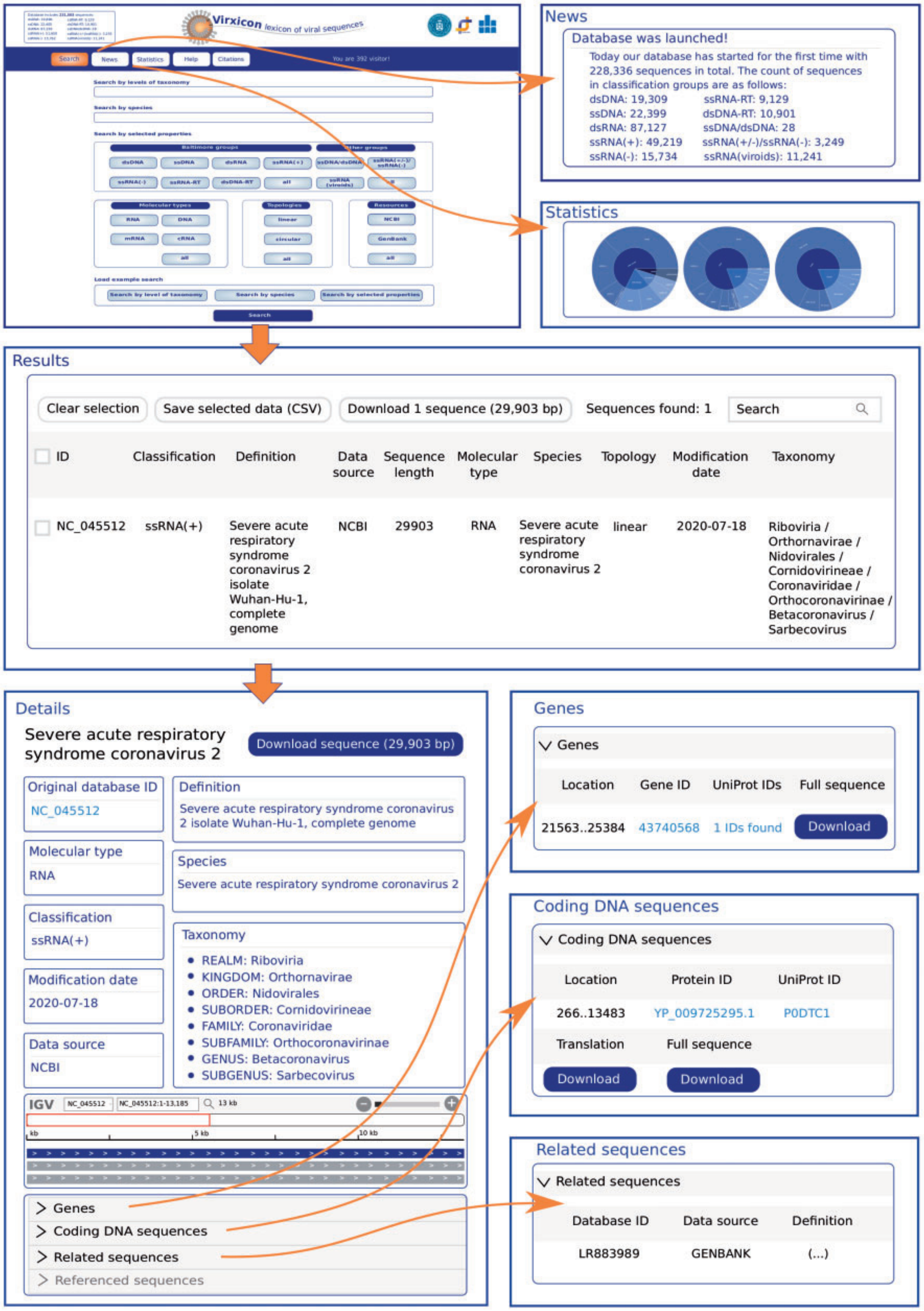


Fig. 3. Data distribution in viral groups with the division into RNA, DNA, RT viruses and viroids

the entry details: original ID linked to the source repository, definition, molecular type, species name, classification, modification date, taxonomy, data sources, information about genes, CDS, related and reference sequences. Information about genes contains the location and provides the ability to download sequences for individual genes. For the CDS, users can see the location, and protein ID



News

Database was launched!

Today our database has started for the first time with 228,336 sequences in total. The count of sequences in classification groups are as follows:

dsDNA: 19,309	ssRNA-RT: 9,129
ssDNA: 22,399	dsDNA-RT: 10,901
dsRNA: 87,127	ssDNA/dsDNA: 28
ssRNA(+): 49,219	ssRNA(+/-)/ssRNA(-): 3,249
ssRNA(-): 15,734	ssRNA(viroids): 11,241

Results

□ ID	Classification	Definition	Data source	Sequence length	Molecular type	Species	Topology	Modification date	Taxonomy
□ NC_045512	ssRNA(+)	Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome	NCBI	29903	RNA	Severe acute respiratory syndrome coronavirus 2	linear	2020-07-18	Riboviria / Orthornavirae / Nidovirales / Cornidovirineae / Coronaviridae / Orthocoronavirinae / Betacoronavirus / Sarbecovirus

Details

Severe acute respiratory syndrome coronavirus 2

Original database ID: NC_045512

Definition: Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

Molecular type: RNA

Species: Severe acute respiratory syndrome coronavirus 2

Classification: ssRNA(+)

Taxonomy
- REALM: Riboviria
- KINGDOM: Orthornavirae
- ORDER: Nidovirales
- SUBORDER: Cornidovirineae
- FAMILY: Coronaviridae
- SUBFAMILY: Orthocoronavirinae
- GENUS: Betacoronavirus
- SUBGENUS: Sarbecovirus

Modification date: 2020-07-18

Data source: NCBI

Genes

Location	Gene ID	UniProt IDs	Full sequence
21563..25384	43740568	1 IDs found	Download

Coding DNA sequences

Location	Protein ID	UniProt ID
266..13483	YP_009725295.1	P0DTC1
Translation	Full sequence	
Download	Download	

Related sequences

Database ID	Data source	Definition
LR883989	GENBANK	(...)

Fig. 4. The web-based interface of Virxicon

linking to the reference repository, and download the amino acid and nucleotide sequence. The related and reference sequences are described by the database ID, data source name and definition. The built-in IGV genome browser provides a dynamic and interactive graphical representation of viral genomes. It allows zooming in to single base pairs or zooming out to browse entire genes and CDS. Clicking on a region of gene or CDS reveals detailed information about the corresponding gene/CDS.

API provides an alternative way of using Virxicon. It allows downloading all detailed information about a given virus sequence (in JSON format) with relaxed limitations for the size of downloaded data. Compared the the web application, it offers additional filtering options: classification groups—fetch available classification groups, molTypes—fetch available molecular types, topologies—fetch available sequence topologies, taxonomies—fetch available taxonomy levels, sequence—fetch information about a single sequence determined by the sequence identifier, sequences—fetch information about sequences including filters such as sequence identifiers, classification groups, data source (NCBI, GenBank), mol types, taxonomies, topologies, organisms. API also enables metadata queries: databaseInfo—retrieves information about the current state of the database, databaseHistoricalInfo—gives the historical state of the database, news—retrieves information about news.

3.3 Virxicon versus other viral databases

Four existing virus databases, NCBI virus, ViPR, ViralZone and Virxicon, collect various information, and provide mechanisms that allow users to search, visualize and download it. In Table 2, we have gathered the essential features of these resources to help users in choosing which one is best suited to their needs. The features fall into four groups: database contents, criteria to use in data filtering, data available for download, other facilities. In the *database contents*, we distinguished sequences, classification into virus groups, other viral data and functional annotation. Virxicon, as the only resource in the analysed set, provides all of these data; in the latter case, for each gene and coding RNA sequence, it gives a reference to the UniProt database. The second group of features, *data filtering criteria*, lists seven selected types of search criteria that the user interface can offer. Virxicon can filter by species, taxonomy level, specified virus groups (one or more) according to Baltimore and ICTV classification, molecule type and topology (linear or circular). Such a wealth of criteria is a huge advantage of our database. On the other hand, it does not allow to search by sequence or sequential homology—these options are implemented in NCBI and ViPR, both apply BLAST (Altschul *et al.*, 1990) to enable homology-based searching. In the third group, *data available for download*, we placed information about whether the tool allows users to download single sequences, whole groups of sequences belonging to the selected viral class (all selected sequences are downloaded in one FASTA file), and other data presented as a result of querying the database. A unique feature implemented in Virxicon is the possibility to download whole groups of viruses according to the classification; this option is not present in other databases. Finally, facilities mentioned in the fourth group indicate whether the database allows users to submit new data, obtain the interactive visualization of the genomic data (genome browser), access the resource via API and get the statistical information about the database contents along with the graphical visualization. NCBI virus and ViPR are the primary databases that allow the users to submit their data, while ViralZone and Virxicon make use of the data stored in the other resources. A useful feature of Virxicon is to enable access via API. It significantly extends the use of data from the database and allows users to perform complex searches that are not supported by the web interface.

Table 2. Selected features of viral databases

	NCBI virus	ViPR	ViralZone	Virxicon
I Database contents				
Sequences	✓	✓		✓
Virus groups			✓	✓
Other viral data	✓	✓	✓	✓
Functional annotations			✓	✓
II Data filtering criteria				
Sequence	✓			
Sequence homology	✓	✓		
Species	✓	✓	✓	✓
Taxonomy	✓	✓	✓	✓
Virus group			✓	✓
Molecular type				✓
Topology				✓
III Data available for download				
Single sequences	✓	✓		✓
Group-wide sequences				✓
Search results	✓	✓		✓
IV Other facilities				
Submission of user data	✓	✓		
Genome browser	✓	✓		✓
Additional access via API				✓
Visualized statistics				✓

4 Conclusion

Several databases have been created to date that store information about viral sequences (Adams and Antoniw, 2006; Brister *et al.*, 2015; Elbe and Buckland Merrett, 2017; Goodacre *et al.*, 2018; Hulo *et al.*, 2011; Lefkowitz *et al.*, 2008; Masson *et al.*, 2012; Mihara *et al.*, 2016; Pickett *et al.*, 2012; Sharma *et al.*, 2015; UniProt Consortium, 2019). Their limitations—like inability to obtain the information about all viruses from the group of interest (Wasik *et al.*, 2019)—made us design Virxicon, the resource aiming to collect viral data, present their distribution across different viral classes, search them based on a variety of criteria and download entire groups of viruses under the current taxonomy.

The uniqueness of Virxicon lies in accessing the data for individual groups of viruses (according to the Baltimore classification and ICTV taxonomy) by one-click search, and retrieving all sequences that meet the combination of multiple search criteria. Therefore, the database is particularly suitable for data extraction and their high-throughput analysis. The usability of its alpha version has been already proven in two projects: in the first one, we collected Coronavirus sequences to investigate SARS-CoV-2 viruses in terms of their subtype spreading based on alignment-free methods for RNA sequence comparison (Kaden *et al.*, 2020); in the second we used Virxicon contents to refine and evaluate RNAComposer-predicted 3 D models of 5’UTR and 3’UTR regions of SARS-CoV-2 (Antczak *et al.*, 2016; Lukasiak *et al.*, 2015; Szachniuk, 2019). Further applications of Virxicon are planned, for example a study on sequence propensity to form quadruplexes of various categories (Popenda *et al.*, 2020; Zok *et al.*, 2020), and modelling of viral infections (Wasik *et al.*, 2013, 2014).

Virxicon is constantly updated—after each update, it shows how many new virus sequences have arrived in each category—and developed. Future works include responding to any changes in virological systematics via ICTV supervision, expanding the system functionality by adding the ability to search for user-provided sequences, homology-based search using BLAST (Altschul *et al.*, 1990) or HMMER (Durbin, 1998) algorithms, improving data visualization and attaching new data or links to them—for example, secondary and tertiary structures, literature references.

Funding

This project was supported by the National Science Centre, Poland [2016/23/B/ST6/03931, 2019/35/B/ST6/03074], the statutory funds of Poznan University of Technology and the grant of the European Social Fund (ESF). Funding for open access charge: Institute of Bioorganic Chemistry, Polish Academy of Sciences.

Conflict of Interest: none declared.

References

Adams,M.J. and Antoniw,J.F. (2006) DPVweb: a comprehensive database of plant and fungal virus genes and genomes. *Nucleic Acids Res.*, **34**, D382–D385.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Antczak,M. *et al.* (2016) New functionality of RNAComposer: application to shape the axis of miR160 precursor structure. *Acta Biochim. Polonica*, **63**, 737–744.

Baize,S. *et al.* (2014) Emergence of Zaire Ebola virus disease in Guine. *N. Engl. J. Med.*, **371**, 1418–1425.

Baltimore,D. (1971) Expression of animal virus genomes. *Bacteriol. Rev.*, **35**, 235–241.

Bao,Y. *et al.* (2004) National Center for Biotechnology Information Viral Genomes Project. *J. Virol.*, **78**, 7291–7298.

Benson,D.A. *et al.* (2012) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.

Brister,J.R. *et al.* (2015) NCBI viral genomes resource. *Nucleic Acids Res.*, **43**, D571–D577.

Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. (2020) The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.*, **5**, 536.

Cotten,M. *et al.* (2013) Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. *Lancet*, **382**, 1993–2002.

Datta,S. (2015) Next-generation sequencing in clinical virology: discovery of new viruses. *World J. Virol.*, **4**, 265.

de Cárcer,D.A. *et al.* (2015) Biodiversity and distribution of polar freshwater DNA viruses. *Sci. Adv.*, **1**, e1400127.

Durbin,R. et al. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Elbe,S. and Buckland Merrett,G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, **1**, 33–46.

Gire,S.K. *et al.* (2014) Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, **345**, 1369–1372.

Goodacre,N. *et al.* (2018) A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. *mSphere*, **3**, e00069-18.

Haagmans,B.L. *et al.* (2014) Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. *Lancet Infect. Dis.*, **14**, 140–145.

Hulo,C. *et al.* (2011) ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res.*, **39**, D576–D582.

International Committee on Taxonomy of Viruses Executive Committee. (2020) The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat. Microbiol.*, **5**, 668.

Kaden,M. *et al.* (2020) Analysis of SARS-CoV-2 RNA-Sequences by Interpretable Machine Learning Models. *bioRxiv*.

Lefkowitz,E.J. *et al.* (2008) Virus databases. In: Mahy, B.W.J. and Van Regenmortel, M.H.V. (eds.) *Encyclopedia of Virology. 3rd edn.* Academic Press, London.

Lefkowitz,E.J. (2015) Taxonomy and classification of viruses. In: Jorgensen, J.H. and Pfaller, M.A. (eds.) *Manual of Clinical Microbiology. 11th edn.* ASM Press, Washington, DC.

Lefkowitz,E.J. *et al.* (2018) Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.*, **46**, D708–D717.

Lukasiak,P. *et al.* (2015) RNAssess a web server for quality assessment of RNA 3D structures. *Nucleic Acids Res.*, **43**, W502–W506.

Marston,D.A. *et al.* (2013) Next generation sequencing of viral RNA genomes. *BMC Genomics*, **14**, 444.

Masson,P. *et al.* (2012) ViralZone: recent updates to the virus knowledge resource. *Nucleic Acids Res.*, **41**, D579–D583.

Mihara,T. *et al.* (2016) Linking virus genomes with host taxonomy. *Viruses*, **8**, 66.

Parrish,C.R. *et al.* (2008) Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol. Mol. Biol. Rev.*, **72**, 457–470.

Pickett,B.E. *et al.* (2012) ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.*, **40**, D593–D598.

Popenda,M. *et al.* (2020) Topology-based classification of tetrads and quadruplex structures. *Bioinformatics*, **36**, 1129–1134.

Sharma,D. *et al.* (2015) Unraveling the web of viroinformatics: computational tools and databases in virus research. *J. Virol.*, **89**, 1489–1501.

Simmonds,P. (2015) Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J. Gen. Virol.*, **96**, 1193–1206.

Suttle,C.A. (2005) Viruses in the sea. *Nature*, **1437**, 356–361.

Szachniuk,M. (2019) RNApolis: computational platform for RNA structure analysis. *Found. Comput. Decis. Sci.*, **44**, 241–257.

UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.

Walker,P.J. *et al.* (2019) Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019). *Arch. Virol.*, **164**, 2417–2429.

Wasik,S. *et al.* (2013) ModeLang: a new approach for experts-friendly viral infections modeling. *Comput. Math. Methods Med.*, **2013**, 1–8.

Wasik,S. *et al.* (2014) Multi-agent model of hepatitis C virus infection. *Artif. Intelli. Med.*, **60**, 123–131.

Wasik,S. *et al.* (2019) Detecting life signatures with RNA sequence similarity measures. *J. Theor. Biol.*, **463**, 110–120.

Wlodawer,A. *et al.* (2020) Ligand centered assessment of SARS CoV2 drug target models in the Protein Data Bank. *FEBS J.*, **287**, 3703–3718.

Zok,T. *et al.* (2020) ElTetrado: a tool for identification and classification of tetrads and quadruplexes. *BMC Bioinformatics*, **21**, 40.