Latin American Data in Science, v.3, n.2
datainscience.com.br
ISSN: 2763-9290

Development and Characterization of an *In Silico* Database of *In Vitro* Tested Antiviral Compounds

Recebido:30/03/2023 |Aceito: 18/09/2023 | Publicado: 13/10/2023
https://doi.org/10.53805/lads.v3i2.61

Nicolas P. Félix[1,2]; Caio Felipe de A. R. Cheohen[3]; Maria Eduarda A. Esteves[1,4]; Manuela L. da Silva[1,3,4]

ABSTRACT

Antivirals are substances that inhibit viral infection or virus replication, acting as drugs to treat viral diseases. However, due the diversity of pathogens, it is important to seek new antivirals. Among the options, repositioning already approved drugs is a cheaper and faster strategy when compared to classical research and development methods. Since there is a lack of compiled and standardized data on these drugs, this work aims to build a database of *in vitro* antiviral tests. Thus, the compounds and their information were obtained through publications of *in vitro* methods for testing antiviral drugs, we created six databases covering SMILES, MOL, SDF 2D, MOL2, PDB and PDBQT extensions, classified by the presence and/or absence of antiviral activity. Each file contains its IUPAC name and structural data in up to three dimensions.

Keywords: Antiviral compounds; Database; Antiviral database; *In silico*; *In vitro*.

DATA IMPORTANCE

- Provides six databases in different formats containing compounds that were tested *in silico* for antiviral activity;
- Allows for the diversity of formats for various analyses of action *in silico*, facilitating the understanding and use of compounds;
- Decreases the construction time of the three-dimensional structure in several formats with corrections related to pH;
- Facilitates the search for general information about tested compounds and their references.

[1] Instituto Nacional de Metrologia, Qualidade e Tecnologia – Inmetro, Duque de Caxias, Brasil. nicolasportofelix@gmail.com
[2] Universidade Católica de Petrópolis, Petrópolis, Brasil.
[3] Universidade Federal do Rio de Janeiro-UFRJ, Macaé, Brasil.
[4] Instituto Oswaldo Cruz – IOC, Rio de Janeiro, Brasil.

MATERIALS AND METHODS

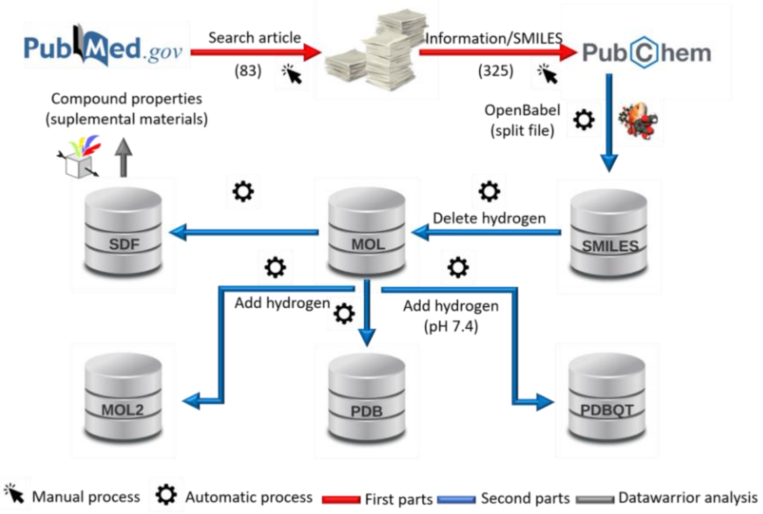
Data extraction

Data were manually extracted from published articles searched in the PubMed database (https://pubmed.ncbi.nlm.nih.gov/). We used the method Boolean search mode, matching one "keyword" at a time to viruses and the term "*in vitro*". Were combined the keywords "antiviral molecule(s)"; "antiviral compound(s)"; "antiviral agent(s)"; "viral inhibitor(s)"; "drug-like molecule(s)"; "viral inhibition(s)" linked by the "AND" operator with "Chikungunya"; "Dengue"; "Ebola"; "Yellow Fever"; "Zika";" Hepatitis"; "Human Coronavirus"; "MERS coronavirus"; "SARS human coronavirus"; "SARS human coronavirus-2"; "Human herpes"; "Human immunodeficiency"; "Human papilloma"; "Norwalk"; "Rotavirus"; "Influenza virus". The third term: "*in vitro*" was also used together with the operator "AND" completing the Boolean filters step (RUCHAWAPOL et al., 2021).

The values/results of the compounds contained in extracts, without isolation of compounds were excluded. Counting 83 articles with publication age from 1985 to 2022, containing Poliovirus, Rift valley fever, Punta toro phlebovirus, and Sandfly fever Naples phlebovirus as viruses cited that were not used as keywords present in accounted articles.

After the first step, the compounds were searched on the PubChem website (https://pubchem.ncbi.nlm.nih.gov/) using their names described in the articles, collecting information such as the IUPAC name and the canonical SMILES, where the information of 325 compounds was determined and thus inserted into the database (DB) (Fig. 1 and Fig. 2). The use of canonical smiles is due to the impossibility of accurately collecting all the isotopic and stereochemical changes (contained in isomeric SMILES) of the molecules used in the articles in relation to the SMILES contained in Pubmed, used for collection and information on the compounds. Four structures: r-chlorcyclizine, s-chlorcyclizine, catechin, and epicatechin, had their isomeric structures collected because they did not show differences between them in their canonical SMILES and because their isomers were used in the articles.

Figure 1: Workflow representing the six databases creation process. This process was divided into three parts: the search (red arrows), the conversion (blue arrows), and physical-chemical property analysis by the Datawarrior program for addition of supplementary material (gray arrow). The first process was divided into two stages: searching for articles containing the compounds tested as antivirals *in vitro* and collecting information on the compounds found. The second process was divided into six changing steps varying in dimensionality and protonation for each file type and preparation need. The third process is the determination of physical-chemical properties by DataWarrior from the SDF database, calculating 9 properties for each of the 325 molecules.



Data conversion and calculate properties:

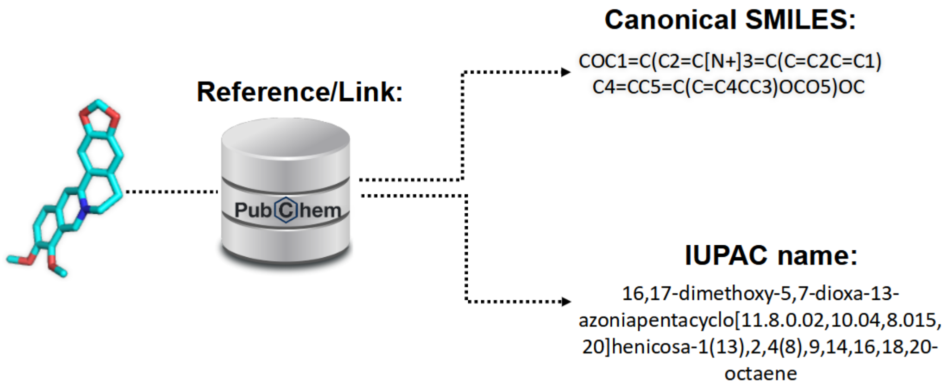
From the canonic SMILES obtained by PubChem (https://pubchem.ncbi.nlm.nih.gov/), the 325 compounds were separated from a single file into individual files by the OpenBabel software via the Linux terminal command “obabel *.smi –O *.smi –m”, being "-m" the option responsible for multiple files generation (OPENBABEL, 2011), generating the first one-dimensional database, called “antiviral_smi”. After splitting into multiple files, the SMI had hydrogen added to some structures (CHEOHEN, 2022; O’BOYLE, 2011).

The second database, called “antiviral_mol” originated by the OpenBabel software using the command “obabel *.smi -O *.mol -d --gen2d”, contains files with two-dimensional structures of the molecules based on information from the SMILES present in the “antiviral_smi” with the option “-d” to remove hydrogens from compounds that were added after splitting into several files and the option “--gen2d” for creating 2D coordinates (OPENBABEL, 2011). The “antiviral_mol” DB was used as a template for all other two-dimensional and three-dimensional antiviral DBs (Fig. 1).

The third database contains two-dimensional files in SDF format, called “antiviral_sdf2d” generated from the OpenBabel program, via the Linux terminal command “obabel *.mol -O *.sdf”. The fourth DB, called “antiviral_pdb” contains files in three dimensions in PDB format, formed by the OpenBabel program using the command “obabel *.mol -O *.pdb --gen3d”, with the option “--gen3d” for creating 3D coordinates (OPENBABEL, 2011). The calculations of physicochemical and pharmacokinetic properties were originated by the DataWarrior program using the "antiviral_sdf2d" database, which contains 9 properties calculated and described for each compound available in the supplementary material (SANDER, 2015).

The fifth and sixth databases in MOL2 and PDBQT formats respectively are named “antiviral_mol2” and “antiviral_pdbqt”. These two DBs contain three-dimensional files with the addition of hydrogen, being in the DB “antiviral_pdbqt” with the addition of hydrogens at physiological pH (7.4). Using the OpenBabel program via the Linux terminal commands “obabel *.mol -O *. mol2 –h --gen3d” and “obabel *.mol -O *. pdbqt -p 7.4 --gen3d” were made databases “antiviral_mol” and “antiviral_pdbqt” respectively, with the option “-p” together with the value of “7.4” to generate hydrogens at pH 7.4 and the option “-h” to add the hydrogens in the compounds (OPENBABEL, 2011).

Figure. 2: Visual representation of the information contained for each compound in the supplementary material. Containing information on the molecules, taken from the PubChem database and added in the supplementary material their SMILES and IUPAC name with links to the page where the information on each molecule was taken.


Reference/Link:
PubChem
Canonical SMILES:
COC1=C(C2=C[N+]3=C(C=C2C=C1)C4=CC5=C(C=C4CC3)OCO5)OC
IUPAC name:
16,17-dimethoxy-5,7-dioxa-13-azoniapentacyclo[11.8.0.02,10.04,8.015,20]henicosa-1(13),2,4(8),9,14,16,18,20-octaene

The file names in the databases were manually modified to the same mentioned in the articles, without the presence of spaces and symbols, to facilitate their use. Having the withdrawal of associated molecules as components/excipients for the standardization of programs that recognize only one compound as a ligand (MORRIS, 2008).

In addition to the structures contained in the database, a table is available with informative data and references of the compound, containing the International Union of Pure and Applied Chemistry (IUPAC) name, the indication of the presence of antiviral activity, and a link referencing the origin of the collected information (Fig. 2). The classification of activity and inactivity present in the supplementary material was performed according to the quantification method used by the article, such as the use of effective concentration (EC), inhibitory concentration (IC), selectivity (SI) and/or therapeutic index (TI) as a dose-response or by viral quantification methods, such as plaque assay and Real Time Quantitative Polymerase Chain Reaction (qPCR). In dose-response methods, the absence of EC values characterizes structure without antiviral activity and quantification by concentration as EC indicates compounds with antiviral activity. In viral quantification, there is no change in the infective percentage in relation to the negative control are considered compounds without activity, so the presence of changes compared to the control is classified as antiviral activity. The aim of this classification is to standardize the variations in methodologies, compound concentrations, cell cultures, and cutoff values for IC50, EC50, SI, and TI, as well as to indicate the potential or lack thereof for utilizing the compound in the context of the study from which it was investigated.

DATASET

The "antiviral_smi" DB contains one-dimensional structures where hydrogens were presented by the separation of SMILES by OpenBabel without structural alteration of the compounds, allowing its use in pharmacokinetic structure prediction programs.

The two-dimensional DBs are "antiviral_mol" and "antiviral_sdf" where the group of mol files was removed from the hydrogens contained in the SMILES files to be the point of origin of the three-dimensional DBs, and the DB with SDF files does not contain hydrogens added directly and allowing future addition by the user (CHEOHEN, 2022; O'BOYLE, 2011; IRWIN, 2005).

The three-dimensional databases "antiviral_pdb", "antiviral_mol2" and "antiviral_pdbqt" were derived from the two-dimensional database in MOL. The structures in pdb format contain hydrogens in explicit form without direct addition by OpenBabel, unlike the structures present in the mol2 and PDBQT DBs contain explicit hydrogens added directly by OpenBabel. In addition, the DB files "antiviral_pdbqt" were adjusted to a specific pH of 7.4, which approximates physiological pH, making them suitable for docking simulations.

- The complete database available on MendeleyData (https://data.mendeley.com/datasets/jxgf4n3y6s/1) can be accessed through DOI:10.17632/jxgf4n3y6s.1. It contains 157 synthetic antivirals, seven semisynthetic and 157 natural antivirals. Additional information is available in the supplementary material and the excel table contains the tabs "INTRO", wich is the initial panel for the tabs containing specific data of DB, the "*IN VITRO* RESEARCH", "IUPAC" and "PROPERTIES" tabs. *IN VITRO* RESEARCH: This tab refers to the research of molecules tested as antivirals for different viruses. It's classified in "No antiviral activity" and "antiviral activity" to determine the activity. The tab is divided into columns associated with the 22 viruses, grouped into families or genera and the reference article is available as an annotation.

- Additional information: This tab contains eight columns with additional information about each antiviral from the DB;
- compounds: This column presents the name of the molecule cited in their respective reference article;
- Compounds class: Is the classification of the synthetic, semi-synthetic or natural origin of the antivirals presents in the DB;
- SMILES: Column presented with one-dimensional structural information devoid of isotopic and stereochemical information (canonical SMILE);
- IUPAC name: Is the identification of the chemical name of each antiviral based on IUPAC nomenclature;
- link/reference: Is a PubChem reference link where the "SMILES" and "IUPAC" was accessed;
- No antiviral activity: This column indicates which virus the antiviral was tested and did not show activity (according to the reference article);
- Antiviral activity: Indicates which virus the molecule was tested and showed activity (according to the reference article);
- Antiviral activity and not antiviral activity in different articles: Indicates conflicting inhibitory activity between articles in which the antiviral was tested against different viruses. Example: article one recognizes activity against viruses X, Y and not for Z. But article two attested inhibition for Z. Than Z is described in this column.
- Physical chemical properties: This table contains the physicochemical properties of each antiviral. Columns describe properties implemented in pharmacokinetic rules. (EGAN et al., 2001; GHOSE et al., 1999; LIPINSKY et al., 2001; MUEGGE et al., 2001; VEBER et al., 2002).
- Compounds: The tab presentsthe reference name of the antiviral;
- Total-Molweight: Sum of the atomic weight values in a antiviral;
- Logarithm of the partition coefficient (CLoP): Is a measure of the lipophilicity of a antiviral;
- Log solubility coefficient (CLoS): Is a measure of the ability of a substance to dissolve in both, water and octanol;
- H-Acceptors: Is the number of possible hydrogen acceptors;
- H-Donors: Is the number of potential hydrogen donors;
- Total-Surface-Area: Is the sum of all areas of the faces of one antiviral;
- Polar-Surface-Area (PSA): Is a sum of tabulated surface contributions from polar fragments;
- Rotatable-Bonds: Is the number of bonds that can freely rotate around an axis in a molecule;
- Mutagenic: Is related to the ability of an antiviral to have mutagenic effects.

SUPPLEMENTARY MATERIALS

Dataset: Felix et al_Dataset

ACKNOWLEDGEMENTS

Thanks to CNPq, FAPERJ, CAPES and GPBIC group.

REFERENCES

CHEOHEN, C. F. A.; ANDRIOLO, B. V.; DA SILVA, M. L. Database of Active Pharmaceutical Ingredients (APIs) present in the Brazilian Pharmacopeia. Latin American Data in Science, v. 2, n. 1, p. 7-12, 2022. DOI: 10.53805/lads.v2i1.35.

EGAN, W. J.; MERZ, K. M. Jr.; BALDWIN, J. J. Prediction of drug absorption using multivariate statistics. J Med Chem. 2000 Oct 19;43(21):3867-77. DOI: 10.1021/jm000292e.

GHOSE, A. K.; VISWANADLHAN, V. N.; WENDOLOSKII, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. J Comb Chem. 1999 Jan;1(1):55-68. DOI: 10.1021/cc9800071.

IRWIN, J. J.; SHOICHET, B. K. ZINC− a free database of commercially available compounds for virtual screening. Journal of chemical information and modeling, v. 45, n. 1, p. 177-182, 2005. DOI: 10.1021/ci049714+.

LIPINSKI C. A. et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001 Mar 1;46(1-3):3-26. DOI: 10.1016/s0169-409x(00)00129-0.

MORRIS, G. M.; LIM-WILBY, M. Molecular Docking. In: KUKOL, A. (Ed.). Methods in Molecular Biology. Totowa: [s.n.]. v. 443p. 365–382, 2008.DOI: 10.1007/978-1-59745-177-2_19.

MUEGGE I.; HEALD S. L.; BRITTELLI D. Simple selection criteria for drug-like chemical matter. J Med Chem. 2001 Jun 7;44(12):1841-6. DOI: 10.1021/jm015507e.

O'BOYLE, N. M. et al. Open Babel: An open chemi.cal toolbox. Journal of Cheminformatics, v. 3, n. 33, p. 1– 14, 2011.DOI: 10.1186/1758-2946-3-33.

OPENBABEL. Obabel and babel:Convert, Filter and Manipulate Chemical Data documentation.2011.Available:https://openbabel.org/docs/dev/Command-line_tools/babel.html

RUCHAWAPOL, C. et al. Natural Products and Their Derivatives against Human Herpesvirus Infection. Molecules. v. 26, n. 20, 1 out. 2021.DOI: 10.3390/molecules26206290.

SANDER, T. et al. DataWarrior: an open-source program for chemistry aware data visualization and analysis. Journal of chemical information and modeling, v. 55, n. 2, p. 460-473, 2015. DOI: 10.1021/ci500588j.

VEBER, D. F. et al. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem. 2002 Jun 6;45(12):2615-23. DOI: 10.1021/jm020017n.