Latin American Data in Science, v.5, n.1
datainscience.com.br
ISSN: 2763-9290

PseudoResistance DB: A new Database of antibiotics related to *Pseudomonas aeruginosa* antibiotic resistance

Recebido: 19/11/24|Aceito: 02/07/25 | Publicado: 15/09/25
https://doi.org/10.53805/lads.v5i1.69

Caio B. Nunes[1], Vinnícius M. S. Gomes[*2], Caio Felipe de A. R. Cheohen[3], Manuela L. da Silva[1,2,3]

ABSTRACT

Antibiotic resistance represents a significant challenge to global health, compromising the effectiveness of treatments. Pseudomonas aeruginosa, an opportunistic Gram-negative bacterium, is a priority pathogenic agent for research, being categorized as critical level by the World Health Organization (WHO) for the development of new therapeutic strategies. The antibiotics listed were screened through text mining correlated with P. aeruginosa of the Scielo database. The final database contains 98 antibiotics, containing textual information, Protein Data Bank, Partial Charge & Atom Type (PDBQT), Simplified Molecular Input Line Entry System (SMI), IUPAC International Chemical Identifier (INCHI), Molecular Design Limited Molfile (MOL2), Structure-Data File (SDF), Chemical Markup Language (CML), Cartesian Coordinates File (XYZ), Scalable Vector Graphics (SVG), Molecular File (MOL) and Protein Data Bank (PDB), converted using OpenBabel.

Keywords: Antibiotic Resistance; Pseudomonas aeruginosa; In silico; Text mining.

DATA IMPORTANCE

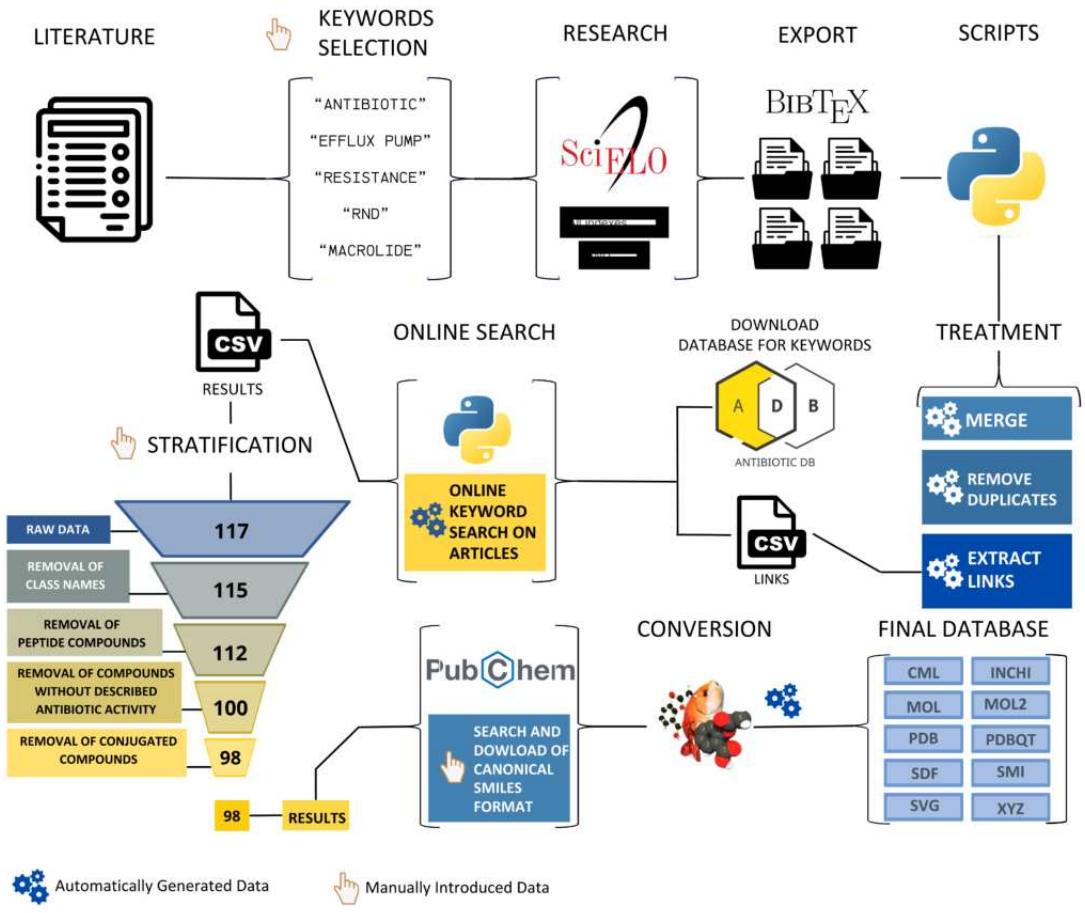
- The database aids in identifying antibiotics specifically associated with Pseudomonas aeruginosa, which is critical for developing effective treatment strategies against its resistant strains;
- Provides 98 antibiotics related to Pseudomonas aeruginosa, critical for addressing antibiotic resistance research;
- Includes molecular models in formats CML, INCHI, MOL, MOL2, PDB, PDBQT, SDF, SMI, SVG, XYZ ready for in silico analysis, like molecular docking and dynamics in physiological pH.

[1] Universidade Federal do Rio de Janeiro, Macaé, Brasil.
[2] Programa de Pós-Graduação em Biologia Computacional e Sistemas. vinniciusmschelk@gmail.com
[3] Programa de Pós-Graduação Multicêntrico em Ciências Fisiológicas.

MATERIAL AND METHODS

The database construction on antibiotic interactions with RND resistance systems in Pseudomonas aeruginosa involved several stages, as schematically illustrated in Figure 1. Initially, keywords were manually selected based on intrinsic components of antibiotic resistance in Pseudomonas aeruginosa, drawing from comprehensive literature reviews, including those by Amaral et al (2014), Blair et al (2014) and Colclough et al (2020). These keywords were used to conduct searches in the Scientific Electronic Library Online (SciELO) database, selected for its extensive coverage of relevant scientific literature in Latin America and its automated access to published works.

Figure 1.- Flowchart represents the hierarchical process of data input and processing. The first stage involves the manual entry of data, which includes selecting keywords, searching the SciELO database, exporting results in BibTeX format, and processing the data through scripts for merging, removing duplicates, and extracting links. The second stage, generated automatically, encompasses online searches for articles using scripts, downloading antibiotic databases, and converting chemical structures to the SMILES format via PubChem. The final result is a consolidated database containing multiple molecular formats.



Keyword Selection for Guiding Literature Searches

The following searches were conducted to collect textual material from the SciELO using boolean operators (AND) and (ALL INDEXES).

- "efflux pump" AND "antibiotic"
- "MexAB-OprM" AND "antibiotic"
- "Mex*" AND "antibiotic"
- "Opr*" AND "antibiotic"
- "Pseudomonas aeruginosa" AND "antibiotic"
- "Pseudomonas aeruginosa" AND "antibiotic resistance"

- “Pseudomonas aeruginosa” AND “efflux pump”
- “RND” AND “antibiotic”

The online search results were exported in the BibTeX format, a widely used bibliographic flat-file database file format for managing references.

Automatically Data Processing

Python 3.8.10 version was applied to merge the eight exported files. The first command, ’open’, was employed to access both the input files (each containing a set of references in BibTeX format) and the output file, where all references would be consolidated. The ‘read’ command was used to read the content of each input file, ensuring that all bibliographic information was accurately processed. Subsequently, the ‘write’ command was applied to record the read content into the main output file, which would compile all references into a single document.

To remove duplicate references from the complete BibTeX file, the previously described commands were used to open the input and output files and analyze the text. A list named ‘entries’ was constructed to store unique bibliographic entries, creating an output free of duplicate references. This approach ensured the comprehensiveness of the bibliographic database while avoiding redundancies.

Afterwards, a script was developed to extract links from a text file and save them in a CSV file. The ‘open’ command was used to open and read the text file, while the ‘read’ command retrieved the entire content of the file. This allowed a regular expression, created with the ‘re.compile’ command, to be applied for identifying and extracting links. The ‘re.findall’ command was utilized to locate all links present in the text, returning them in list format. The extracted links were then organized into a DataFrame using the ‘pd.DataFrame’ command from the Pandas library. To ensure that the links were in the correct format, the ‘str.strip’ command was employed to remove brackets and commas from the BibTeX typography. Finally, the ‘to_csv’ command was used to save the resulting DataFrame in a Comma-Separated Values (CSV) file, enabling an organized export of the links. The online search script involved utilizing the Antibiotic DB to obtain names of antibiotics, automating the search for these terms in online scientific articles through the extracted links. The code employs the ‘requests’ library to send requests to the web pages and ‘BeautifulSoup’ to process and extract relevant text from the HTML tags.

A pre-processing step was necessary for the list of antibiotics extracted from the Antibiotic DB. Terms containing quotation marks in their nomenclature were removed to ensure the terms were correctly recognized as strings, thus preventing syntax errors in the code. These cleaned terms were stored in the variable ‘palavras_chave.’ The code then iterated over a list of links to scientific articles, accessing each one to extract the full text. Subsequently, the script checked for the presence of the listed antibiotics within the extracted content by utilizing the function ‘encontrar_palavras_em_documento_online.’ For each keyword, the function assessed its presence in the provided text, disregarding case sensitivity, by employing the expression ‘palavra.lower() in texto.lower()’, which compared the keywords against the article text. When any keyword was identified in the text, the article link and the name of the detected antibiotic were stored in a results list. This list was in due course organized into a DataFrame using the ‘pandas’ library to facilitate data manipulation. Finally, the results were saved in both CSV and Excel formats, allowing for the storage and analysis of antibiotic occurrences in the accessed scientific articles.

Manual Data Processing

Text mining yielded a total of 2.099 occurrences (hits) across the analyzed articles,

encompassing 117 distinct terms. To refine the data for future computational research, a filtering step was implemented. During this phase, terms associated with classes were discarded, as they did not refer directly to specific molecules. Additionally, terms related to combination treatments were excluded due to the presence of multiple ligands within a single file. Antimicrobial peptides were also eliminated from consideration due to their high molecular weight and structural complexity, which necessitated advanced computational capacity often unavailable due to technical limitations. Furthermore, molecules lacking documented antibacterial activity in the PubChem database were removed.

After applying these exclusion criteria, 98 molecules were selected and categorized into 26 distinct classes.

To publish the database, the molecules were initially researched manually in PubChem to obtain their textual representations in the Simplified Molecular Input Line Entry System (SMILES) format. This information served as the foundation for converting the molecules into various other formats. A pipeline developed in Google Colab, supported by the Open Babel software (as described in the supplementary material), facilitated the conversion of the molecules into multiple formats: Protein Data Bank, Partial Charge & Atom Type (PDBQT), Simplified Molecular Input Line Entry System (SMI), IUPAC International Chemical Identifier (INCHI), Molecular Design Limited Molfile (MOL2), Structure-Data File (SDF), Chemical Markup Language (CML), Cartesian Coordinates File (XYZ), Scalable Vector Graphics (SVG), Molecular File (MOL) and Protein Data Bank (PDB).

DATA DESCRIPTION

The conversion into different formats ensures the flexibility and applicability of the data across various scientific contexts. For instance, the PDBQT, MOL2, and SDF formats are extensively utilized in molecular simulation and docking studies, as they record not only the 3D coordinates of atoms but also additional information, such as partial charges and atom types, which are essential for molecular flexibility analyses (AGU et al., 2023). Moreover, the converted molecules are associated with a table that includes detailed information such as the molecule's identifier in PubChem, the compound's name in English, and the SMILES notation. The entire dataset, comprising files mentioned before, along with their metadata, has been made available in the Mendeley Data repository, enabling public access and scientific reproducibility.

The script used to convert the molecules is attached as as supplementary file, it enables the conversion of molecular files into various formats using the Open Babel software, integrated with an interactive interface in Google Colab. It supports multiple output formats (such as pdbqt, smi, inchi, and others), adjusts protonation based on a user-specified pH, and performs 2D or 3D coordinate generation. After the conversion, the resulting files can be downloaded individually or grouped into ZIP files by format.

The process begins by requesting the user to upload molecular files, input formats supported include those compatible with Open Babel, such as mol, sdf, and xyz. Once the files are uploaded, users can configure several options, including output formats, coordinate generation (2D, 3D, or none), and hydrogen manipulation (adding, deleting, or retaining hydrogens), in this last the user can specify the desired pH value.

It processes, each of the uploaded files, applying transformations designated by the user. A log file in .txt format is generated for each input file, capturing details of the Open Babel process, including errors or warnings. The converted files are organized by output format and compressed into ZIP files. Each ZIP file contains all the converted files of a specific format, facilitating bulk downloads. For example,

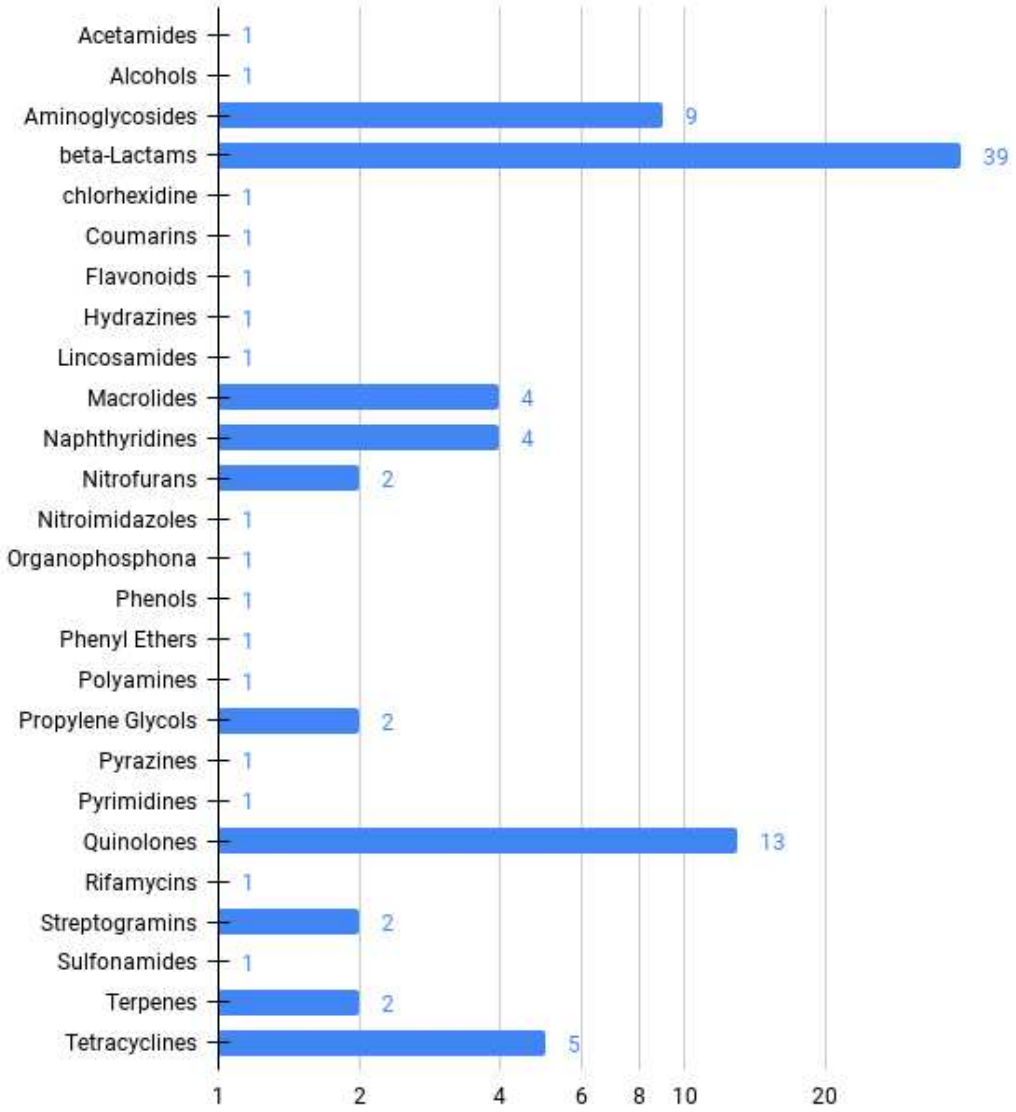
files converted to the SMILES format (.smi) are grouped into a smi.zip archive.

Dataset

This pipeline provides flexibility and standardization in molecular file conversion, making it a valuable tool for researchers in computational chemistry, drug design, and related fields. It allows seamless integration of Open Babel functionalities with an easy-to-use interface for efficient processing and download of molecular data.

The PseudoResistance DB was generated from the textual research and organized into several classes of antibiotics. Each class contains antibiotics categorized according to their classification in the AntibioticDB (Farrell et al. 2018). Figure 2 shows a detailed description of the organization and classification of the antibiotics contained into PseudoResistance DB. The results highlight that the most of the cataloged antibiotics are concentrated in classes commonly used in clinical practice, such as beta-lactams, aminoglycosides, and quinolones (KIM; HOOPER, 2014).

Figure 2.- Bar chart illustrating the distribution of 98 antibiotics across 26 distinct classes. The classes include beta-lactams (39), aminoglycosides (9), and quinolones (13), whereas classes such as acetamides, alcohols, and phenols each contain only 1 antibiotic.



SUPPLEMENTARY MATERIALS

Repository name: Mendeley Data
DOI: 10.17632/bxdn3p33z2.1
Data access link: https://data.mendeley.com/datasets/bxdn3p33z2/1

ACKNOWLEDGEMENTS

C.B.N, V.M.S.G., and C.F.A.R.C. thank CAPES for the scholarship funding. We would also like to thank CAPES, CNPq, and FAPERJ for their support.

REFERENCES

AGU, P. C. et al. Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. Scientific Reports, v. 13, n. 1, 17 ago. 2023. DOI: 10.1038/s41598-023-40160-2

AMARAL, L. et al. Efflux pumps of Gram-negative bacteria: what they do, how they do it, with what and how to deal with them. Frontiers in Pharmacology, v. 4, 2014. DOI: 10.3389/fphar.2013.00168

BLAIR, J. M.; RICHMOND, G. E.; PIDDOCK, L. J. Multidrug efflux pumps in Gram-negative bacteria and their role in antibiotic resistance. Future Microbiology, v. 9, n. 10, p. 1165–1177, out. 2014. DOI: 10.2217/fmb.14.66

COLCLOUGH, A. L. et al. RND efflux pumps in Gram-negative bacteria; regulation, structure and role in antibiotic resistance. Future Microbiology, v. 15, n. 2, p. 143–157, jan. 2020. DOI: 10.2217/fmb-2019-0235

FARRELL, L. J. et al. Revitalizing the drug pipeline: AntibioticDB, an open access database to aid antibacterial research and development. The Journal of Antimicrobial Chemotherapy, v. 73, n. 9, p. 2284–2297, 1 Sep. 2018.

KIM, E. S.; HOOPER, D. C. Clinical Importance and Epidemiology of Quinolone Resistance. Infection & Chemotherapy, v. 46, n. 4, p. 226, 2014. DOI: 10.3947/ic.2014.46.4.226