Databases and ontologies

AciDB 1.0: A database of acidophilic organisms, their genomic information and associated metadata

Gonzalo Neira[1], Diego Cortez[1], Joaquin Jil[1] and David S. Holmes[1,2,3,*]

[1]Center for Bioinformatics and Genome Biology, Fundación Ciencia & Vida, Santiago, Chile, [2]Universidad San Sebastián, Santiago, Chile, [3]Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago, Chile

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: There are about 600 available genome sequences of acidophilic organisms (grow at a pH<5) from the three domains of the Tree of Life. Information about acidophiles is scattered over many heterogeneous sites making it extraordinarily difficult to link physiological traits with genomic data. We were motivated to generate a curated, searchable database to address this problem.

Results: AciDB 1.0 is a curated database of sequenced acidophiles that enables researchers to execute complex queries linking genomic features to growth data, environmental descriptions and taxonomic information.

Availability: AciDB 1.0 is freely available online at: http://AciDB.cl. The source code is released under an MIT license at: https://gitlab.com/Hawkline451/acidb/

Contact: dsholmes2000@yahoo.com

1 Introduction

Acidophiles are organisms that grow optimally in environments with a pH lower than 5. They include Bacteria, Archaea and Eukarya and are widely distributed in many branches of the Tree of Life (Cardenas et al., 2016). Acidophiles are found in many environments such as acidic sulfate soils, hot springs, volcanic areas, deep sea vents, acid mine and rock drainages and acid salt brines (Baker-Austin and Dopson, 2007). They play major roles in several biotechnological applications including copper recovery (bioleaching) and environmental remediation projects (Sharma et al., 2016). They participate significantly in metal and nutrient recycling in natural environments and have recently become important in studies of origin of life, metabolic evolution and astrobiology.

The publication of about 600 acidophile genomes has promised to revolutionize our understanding of this important group of microorganisms. However, associating genomic data with other relevant metadata has proved to be a major task. Information about pH, temperature and other growth parameters resides in disjointed, unstructured repositories impeding discoverability. Phylogenetic classification can also be difficult to link to genome data and other metadata, stifling progress in omics applications.

In this work, we present AciDB 1.0, a curated database of metadata of acidophiles and their genomes. AciDB 1.0 solves the data accessibility issue by providing researchers with sorted and searchable metadata linked to genome information that allow the user to explore the database and to personalize datasets.

2 Methods

2.1 Database content collection

An extensive literature and metagenomic metadata search was performed in order to obtain all publicly available genome sequences for acidophiles (growth pH < 5) from National Center for Biotechnology Information (NCBI) Genbank or RefSeq databases (O' Leary et al., 2016; Benson et al., 2012) and the Joint Genome Institute (JGI) (Chen et al., 2019) as of July 2019. If optimal growth values were not available, the optimum was estimated as the midpoint of the growth range or environmental values whenever available. For metagenome assembled genomes (MAGs), sampling data is provided. Genomes with pH ≤5 were considered. Organisms without pH information were included if they had ≥95% average nucleotide identity (ANI) with organisms that had a mean growth pH ≤5. Taxonomic information is included, ranging from domain to species/strain plus the taxid of each genome. Genomic features such as assembly level, sequencing date, genome size, GC percentage, number of open reading frames (ORFs) and direct FTP link to the respective genome

repository are provided. Genome completeness and genome contamination was calculated using checkM v1.2.2 (Parks et al., 2015) in order to allow genome quality filters. References to the genome sequencing publication and metadata are provided.

2.2 Web application architecture

AciDB 1.0 uses a modern frontend-backend architecture with three main components: a frontend server, backend server and a relational database. The database was built using PostgreSQL. The backend was implemented in Python using Django REST server as a framework and the front-end is a SPA (Single Page Application) built using React, Material UI and several other libraries. This architecture has the following advantages: modularity of each server, scalability in case one component needs more resources and flexibility when choosing frameworks or libraries for the servers or the database.

2.3 Database design and backend server

The database contains five tables, two non-normalized tables and three normalized tables, and uses a design similar to the star schema used in data warehouse systems. The two non-normalized tables are the taxonomy table and the organism table respectively. In the taxonomy table, any Domain has *n* Phyla, which have *m* Classes and so on for each attribute. This means that in the taxonomy table, each taxon of a given taxonomic rank (Domain, Family... Species) can have a variable number of taxa of the next smaller taxonomic rank (e.g Archaea (Domain) has 6 Phyla, Firmicutes (Phyla) has 2 Classes). The organisms table contains attributes that include growth condition, genomic data and accession links.

The three normalized tables are the strain, reference and growth detail tables respectively. The strain and reference tables have a one-to-many relationship with the organisms table. The growth detail table has a one-to-one relationship with the organisms table and stores additional growth conditions of each organism.

The use of non-normalized tables in combination with small normalized tables facilitates rapid querying that can be a problem when using normalized tables by themselves.

The backend provides an ORM (Object-Relational mapping) tool that facilitates interaction with the database. The server processes the requests from the frontend and serves the data through different endpoints in the REST API (Application Programming Interface) and returns a JSON file.

2.4 Frontend server

The web interface was built using React 16.8 and Material UI composed of a set of components that work asynchronously, providing a responsive and intuitive graphical interface. The basic components, such as buttons, lists and text fields, were developed using Material UI, while more complex components were developed with the help of external libraries. The taxonomy browser uses the material-ui-tree library. The database browser uses react-tree, while the charts were implemented using react-vis.

The frontend mounts the web components when loading a page for the first time or updates an existing component in the interface using the data from the REST API. Each component uses a different endpoint for the request, allowing the interface to render multiple components at the same time. This feature means that the entire page does not have to be recomposed on every request, significantly expediting the process

3 Functionality

3.1 Taxonomy Browser

AciDB 1.0 allows users to navigate through a tree structure by using NCBI taxonomy information. The tree includes hierarchical information from domain to strain level, it is recommended for the search of a single organism and allows the user to access the complete metadata available in the database.

3.2 Database Browser

The database browser can be used to obtain information of multiple organisms at once. The available filter search allows the user to make simple and fast queries with conditional statements using several parameters at the same time. Growth condition, genome metadata and taxonomy are all available to use in the query. The displayed information can be then downloaded in a TSV file for downstream analyses.

3.3 Scatter plot

The scatter plot tool allows the user to generate customizable plots permitting a fast visualization of the information in the database. Each axis can be changed to different values including growth conditions and genomic metadata. Genomes can be filtered by quality to visualize only high-quality draft and complete genomes. Filters for the genome source (isolated or non-isolated) and assembly level (complete or draft) are also available. Genomes can be selected in the scatter plot and summaries of their corresponding metadata, including genome ftp links, can be obtained as a TSV file.

3.4 Advanced search

The advanced search function gives the user the option to make complex queries to obtain specific datasets from the database. The search can be based on taxonomy information (including all NCBI taxonomy levels), growth ranges (pH and temperature) and complete genome metadata. The advanced search retrieves all metadata available for each genome as a TSV file.

4 Conclusions

AciDB 1.0 is an extensive, searchable database of metadata of acidophilic microorganisms and their genomes that facilitates queries and allows information to be downloaded from a highly curated dataset. This promotes opportunities to (i) improve our knowledge of the mechanisms of biological acid resistance, (ii) understand the role of acidophiles in biogeochemical cycles and biomining processes and (iii) leverage information relevant to origin-of-life studies and astrobiology. The database is planned to be updated twice a year with new genomes and new functions such as integration with commonly used services e.g. BLAST.

Acknowledgements

The authors would like to thank Carolina Gonzalez, Eva Vergara and Katelyn Boase for testing the website.

Funding

This work was supported by FONDECYT 1181717 and Programa de Apoyo a Centros con Financiamiento Basal AFB170004 to Fundación Ciencia & Vida.

Conflict of Interest: none declared.

References

Baker-Austin, C., & Dopson, M. (2007). Life in acid: pH homeostasis in acidophiles. Trends in microbiology, 15(4), 165-171.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. Nucleic acids research, 41(D1), D36-D42.

Cárdenas, J. P., Quatrini, R., & Holmes, D. S. (2016). Genomic and metagenomic challenges and opportunities for bioleaching: a mini-review. Research in microbiology, 167(7), 529-538.

Chen, I. M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., ... & Smirnova, T. (2019). IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic acids research, 47(D1), D666-D677.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., ... & Astashyn, A. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research, 44(D1), D733-D745.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research, 25(7), 1043-1055.

Sharma, A., Parashar, D., and Satyanarayana, T. (2016). “Acidophilic microbes: biology and applications,” in *Biotechnology of Extremophiles: Advances and Challenges*, ed. P.H. Rampelotto (Porto Alegre: Springer),215–241.