BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 14 2012, pages 1947–1948
doi:10.1093/bioinformatics/bts273

Databases and ontologies

Advance Access publication May 9, 2012

MicrobeDB: a locally maintainable database of microbial genomic sequences

Morgan G. I. Langille[1,*], Matthew R. Laird[2], William W. L. Hsiao[3], Terry A. Chiu[2], Jonathan A. Eisen[4] and Fiona S. L. Brinkman[2]

[1]Department of Biochemistry & Molecular Biology, Dalhousie University, Halifax, Nova Scotia, [2]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, [3]BCCDC Public Health Microbiology & Reference Laboratory, Vancouver, British Columbia, Canada and [4]Genome Center, University of California Davis, Davis, California, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Analysis of microbial genomes often requires the general organization and comparison of tens to thousands of genomes both from public repositories and unpublished sources. MicrobeDB provides a foundation for such projects by the automation of downloading published, completed bacterial and archaeal genomes from key sources, parsing annotations of all genomes (both public and private) into a local database, and allowing interaction with the database through an easy to use programming interface. MicrobeDB creates a simple to use, easy to maintain, centralized local resource for various large-scale comparative genomic analyses and a back-end for future microbial application design.

Availability: MicrobeDB is freely available under the GNU-GPL at: http://github.com/mlangill/microbedb/

Contact: morgan.g.i.langille@gmail.com

Received on January 11, 2012; revised on April 11, 2012; accepted on May 1, 2012

1 INTRODUCTION

The study of bacterial and archaeal genomes has rapidly progressed from the analysis of single genomes to comparisons between hundreds and thousands. Any type of biological analyses or development of novel bioinformatic methods that uses more than a handful of genomes requires a basic but non-trivial method for obtaining, organizing and storing this genomic information. In the past, this has been a problem primarily limited to large scale data providers such as IMG (Markowitz *et al*., 2012), NCBI (Sayers *et al*., 2011), GOLD (Pagani *et al*., 2011) and CMR (Davidsen *et al*., 2010). Although many of these centers provide genomic data in a variety of static formats such as Genbank and Fasta, these are often inadequate for complex queries. To carry out these analyses efficiently, a relational database such as MySQL (http://mysql.com) can be used to allow rapid querying across many genomes at once. Some existing data providers such as CMR allow downloading of their database files directly, but these databases are designed for large web-based infrastructures and contain numerous tables that demand a steep learning curve. Also, addition of unpublished genomes to these databases is often not supported. A well known and widely used system is the Generic Model Organism Database (GMOD) project (http://gmod.org). GMOD is an open-source project that provides a common platform for building model organism databases such as FlyBase (McQuilton *et al*., 2011) and WormBase (Yook *et al*., 2011). GMOD supports a variety of options such as GBrowse (Stein *et al*., 2002) and a variety of database choices including Chado (Mungall and Emmert, 2007) and BioSQL (http://biosql.org). GMOD provides a comprehensive system, but for many researchers such a complex system is not needed. For example, Chado and the simpler BioSQL schemas have over 130 and 20 database tables, respectively. We propose a minimalistic system that is easy to set up, requires minimal administration for automatic updates, focusing on a lab based setting where unpublished genomes can be easily added, and allowing individual users to work with an unchanging snapshot of genomes from a given download date. To fulfill these goals, we created MicrobeDB, an open-source project that has been used in several comparative genome projects (Ho Sui *et al*., 2009; Winstanley *et al*., 2009) and as a backend for previously developed applications (Langille and Brinkman, 2009; Yu *et al*., 2010).

2 FEATURES

MicrobeDB offers an easy to access, manageable and centralized database for microbial genomes. The main features of MicrobeDB are automated downloading of archaeal and bacterial genomes from NCBI, organized storage of the flat files, annotations and genomic metadata stored in a MySQL database, and a Perl API database for interacting with the data. A single script (that can be scheduled to run weekly, monthly, etc.) looks after downloading and storing new genomes, parsing and loading the data into the MySQL database, and cleaning up any old 'versions' that have not been saved by individual users.

2.1 Genome data source

By default all genomes available in the NCBI RefSeq database (Pruitt *et al.*, 2011) are downloaded using the Aspera downloader (Beloslyudtsev, 2010). Users can optionally choose to include incomplete genomes and/or limit to a subset of genomes at the genera or species level of their choice. In addition, users may download the data in several formats beyond the standard gbk format required by MicrobeDB such as fna, faa, gff, etc. After download, all genomes are uncompressed into their original flat files, and stored under a date stamped central directory.

*To whom correspondence should be addressed.

Table 1. Annotations stored in the MicrobeDB database

Table/object[a]	Field descriptions[a]	Example
Genome project	Organism name	Pseudomonas aeruginosa LESB58
	NCBI taxon ID	557722
	Genome size (Mb)	6.6
	Pathogenic in	Human
	GC %	66.3
	Oxygen requirements	Aerobic
Replicon	Replicon type	Chromosome
	Accession (RefSeq)	NC_011770
	Replicon size (bp)	6601757
	Number of genes	6027
	Replicon sequence	TTTAAAGAG…
Gene	Gene type	CDS
	Locus ID	PLES_00001
	Start position	483
	End position	2027
	Gene name	dnaA
	Product	chromosomal replication initiation
	DNA sequence	GTGTCCGT…
	Protein sequence	MSVELWQQ…
Version	Download date	2011-12-17
	Flat file directory	/share/genomes/2011-12-17/
	Used by	Morgan, Matthew

[a]Not all fields and tables in MicrobeDB are listed.

2.2 Annotation extraction and storage

The second step of each update parses annotations and metadata for each genome and stores the information in a locally installed MySQL database. Information is split into different levels of 'objects', including *Gene* (e.g. accession, start position, end position, product, name, etc.), *Replicon* (e.g. size, number of genes, replicon type, etc.) and *Genome Project* (NCBI taxon id, NCBI genome project id, GC%, habitat, pathogen, etc.) (Table 1). This information is obtained from the Genbank formatted files for each genome, from metadata tables from NCBI, or derived computationally (e.g. gene counts, GC%, etc.) (Table 1). Additionally, a simplified version of the NCBI taxonomy is stored for each genome and is associated with each Genome Project object. The MicrobeDB schema is easily extended so that users can add their own custom data fields if needed (e.g. SNP positions, regulatory elements, etc.). The MySQL database can be accessed using any MySQL client or through the MicrobeDB Perl API that is supplied with MicrobeDB. The MicrobeDB Perl API provides simple querying and retrieval of information in the MySQL database from within the user's own applications without having to write actual SQL queries. In addition there are many free graphical interfaces for interacting with MySQL databases that do not require programming skills including web based such as phpMyAdmin (http://phpmyadmin.net), and local desktop clients such as MySQL Work Bench (http://www.mysql.com/products/workbench/).

2.3 Unpublished genomes

Unpublished genomes (those not in NCBI) can be loaded into MicrobeDB by placing their Genbank formatted files into a directory and running a single script. MicrobeDB does not support genome annotation or create Genbank files, but many programs are available for production of these files such as RAST (Aziz *et al*., 2008) or ARTEMIS (Carver *et al*., 2008). NCBI-specific metadata that is not available for unpublished genomes is simply left as blank fields in MicrobeDB without affecting functionality.

2.4 Stable versions of genomes

MicrobeDB keeps each update as a separate 'version'. This allows users to save and work on a particular snapshot of genomes knowing that the underlying dataset remains consistent. Each MicrobeDB version has an associated download date and users can save a version until their research is complete. Old unsaved versions that are no longer needed will be automatically removed after each update is completed to save storage space.

Overall, MicrobeDB provides support for researchers that require a manageable local organization of bacterial and archaeal genomes for either large comparative genome projects or for constructing new bioinformatic applications.

Funding: This work was supported by the Canadian Institutes of Health Research, Michael Smith Foundation for Health Research, Genome Canada, and Gordon and Betty Moore Foundation.

Conflict of Interest: none declared.

REFERENCES

Aziz,R.K. *et al*. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, **9**, 75.

Beloslyudtsev,D. (2010) Aspera Transfer Guide. In: *SRA Handbook*. National Center for Biotechnology, Bethesda, MD.

Carver,T. *et al*. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, **24**, 2672–2676.

Davidsen,T. *et al*. (2010) The comprehensive microbial resource. *NAR*, **38**, D340–D345.

Ho Sui,S.J. *et al*. (2009) The association of virulence factors with genomic islands. *PloS ONE*, **4**, e8094.

Langille,M.G.I. and Brinkman,F.S.L. (2009) IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics*, **25**, 664–665.

Markowitz,V.M. *et al*. (2012) IMG: the integrated microbial genomes database and comparative analysis system. *NAR*, **40**, D115–D22.

McQuilton,P. *et al*. (2011) FlyBase 101 - the basics of navigating FlyBase. *NAR*, **40**, D706–D714.

Mungall,C.J. and Emmert,D.B. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.

Pagani,I. *et al*. (2011) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *NAR*, **40**, D571–D579.

Pruitt,K.D. *et al*. (2011) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *NAR*, **40**, D130–D135.

Sayers,E.W. *et al*. (2011) Database resources of the National Center for Biotechnology Information. *NAR*, **40**, D13–D25.

Stein,L.D. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

Winstanley,C. *et al*. (2009) Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of Pseudomonas aeruginosa. *Genome Res.*, **19**, 12–23.

Yook,K. *et al*. (2011) WormBase 2012: more genomes, more data, new website. *NAR*, **40**, D735–D741.

Yu,N.Y. *et al*. (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.