Genomics Proteomics Bioinformatics 13 (2015) 69–72

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

RESOURCE REVIEW

Web Resources for Microbial Data

Qinglan Sun [#,a], Li Liu [#,b], Linhuan Wu [c], Wei Li [d], Quanhe Liu [e], Jianyuan Zhang [f], Di Liu [g], Juncai Ma [*,h]

Information Network Center, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

Received 9 January 2015; revised 22 January 2015; accepted 27 January 2015
Available online 23 February 2015

Handled by Fangqing Zhao

KEYWORDS

Microorganism;
Web resource;
Bioinformatics tools;
Rating

Abstract There are multitudes of web resources that are quite useful for the microbial scientific research community. Here, we provide a brief introduction on some of the most notable microbial web resources and an evaluation of them based upon our own user experience.

Introduction

Microorganisms are found virtually everywhere in the environment and serve as important biological resources. Some microorganisms have become widely utilized in industrial production. With the rapid development of high-throughput sequencing technology in recent years, it is considerably easier to obtain whole genome sequence of microorganisms. To date, thousands of microbes have been sequenced. Due to the difficulties of isolating single microbial species, high-throughput sequencing technology has been applied to mixtures of environmental microbes to obtain metagenome sequencing data. One metagenomics endeavor is the Human Microbiome Project (HMP), which aims to obtain metagenomic sequencing data from a large number of human subjects to enhance our understanding of the relationship between the microbiome and human health. In this review, we provide a brief introduction on some of the most useful microbial web resources (Table 1), including information on collection of microbial cultures, species identification, literature, patent resources, microbial genomics, and metagenomics, as well as tools for analyzing genomic and metagenomic data. In addition, we have evaluated their function and provided ratings for each resource based on our own user experience.

Integrated Microbial Genomes

The Integrated Microbial Genomes (IMG, http://img.jgi.doe.gov/) system [1] is a combined web resource of microbial genome datasets, including genome sequences and gene annotations. The aim of the IMG is to provide free microbial genomic data, together with integrated annotations and comparative analysis services, to scientists worldwide. At the end

* Corresponding author.
E-mail: ma@im.ac.cn (Ma J).
# Equal contribution.
[a] ORCID: 0000-0002-8451-760X.
[b] ORCID: 0000-0001-6977-1004.
[c] ORCID: 0000-0002-5255-1846.
[d] ORCID: 0000-0003-3364-8690.
[e] ORCID: 0000-0001-8263-9446.
[f] ORCID: 0000-0003-4215-5214.
[g] ORCID: 0000-0003-3693-2726.
[h] ORCID: 0000-0001-6382-8014.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2015.01.008

Table 1 Important web resources for microbial research

Name	Link	Main features	Rating	Refs.
IMG	http://img.jgi.doe.gov/	The integrated microbial genomes and annotation	★★★★★	[1,2]
RAST	http://img.jgi.doe.gov/	Gene prediction and annotation tool for microbial genome	★★★★★	[3,4]
MG-RAST	http://metagenomics.anl.gov/	Analysis and annotation tool for metagenome	★★★★★	[5]
HMP-DACC	http://metagenomics.anl.gov/	Data center for the Human Microbiome Project	★★★★★	[6]
PATRIC	http://atricbrc.org/portal/portal/patric/Home	The Integrated pathogen genomes and annotation	★★★★★	[7]
WDCM-CCINFO	http://www.wfcc.info/ccinfo	Worldwide culture collection for microorganisms	★★★★☆	
GCM	http://gcm.wfcc.info	Global catalog of microorganism resource and storage information	★★★★☆	[8]
ABC	http://www.wfcc.info/abc	Papers and patents of microorganisms	★★★★☆	
Website for avian flu information	http://www.avian-flu.info	Data center of avian flu	★★★★☆	[9]
GMS-DBD	http://www.boldmirror.net	DNA barcode data for species identification	★★★★☆	[10]

Note: More stars in the rating column indicate higher importance and usefulness.

of 2014, there were over 10,000 microbial genomes in IMG. Approximately 9000 of them were from GenBank, with the remaining ones submitted by registered users or institutes. Genome sequencing data stored in IMG are annotated using several functional references, such as the Clusters of Orthologous Groups (COG), Pfam, KEGG, TIGRfam, MetaCyc, and the Gene Ontology (GO), providing valuable information for registered users. In addition to integrating microbial genome data, IMG also provides the tool for analyzing publicly-available environmental microbial samples and metagenomes, which is named IMG/M [2].

Rapid Annotation using Subsystem Technology

The Rapid Annotation using Subsystem Technology (RAST, http://rast.nmpdr.org/) [3,4] is an automated online tool for analyzing microbial genomes. RAST screens the genome sequence submitted by the registered users and predicts gene-coding sequences, rRNA, and tRNA using glimmer2, search_for_rnas, and tRNAscan-SE, respectively. RAST then automatically produces two classes of functional annotations for the predicted gene sequences. One is a subsystem-based functional annotation that depends on recognition of functional variants of subsystems. The other is a nonsubsystem-based functional annotation, which considers combined evidence from a number of tools. RAST integrates these two annotations and provides exceptionally strong gene annotation results. RAST also uses the gene annotation results for metabolic pathway reconstruction, which makes the resource useful for comprehensive annotation efforts. RAST also excels at making collections of functionally related protein families (FIGfams). When a new genome is submitted to RAST and made public by the submitter, annotated genes are compared and added to the FIGfams collection. The expanding FIGfams collection has proven to be a robust and scalable solution to microbial genome annotation efforts. To date, over 12,000 users have registered in RAST and have submitted over 60,000 distinct microbial genomes for annotation.

Metagenomics RAST

The Metagenomics RAST (MG-RAST, http://metagenomics.anl.gov/) server [5] is an open online system for management and comparative analysis of metagenomic data. Registered users can submit raw sequencing data in FASTA format, along with sampling information into the system. The uploaded sequences are subsequently processed and analyzed. Summaries of the raw data as well as the analyzed results are generated automatically. The MG-RAST server can manage many different types of data including phylogenetic and metabolic data. In addition to analyzing single-sample datasets, the MG-RAST can also provide comparative analysis for multiple metagenomes and genomes. The metagenome annotation and comparative analysis systems are designed such that tools and/or new data can be added or replaced at any stage of the analysis process as needed to accommodate new methods. In order to protect data privacy, user access is controlled. Registered users retain full control of their data. Nonetheless, collaboration and sharing of data between multiple users are both possible and encouraged within the MG-RAST server.

HMP – Data Analysis and Coordination Center

The Data Analysis and Coordination Center (DACC, http://hmpdacc.org/) is the data center for the HMP [6]. It includes sampling information as well as microbial genome and metagenome sequences. The HMP was launched in 2008 and funded by the National Institutes of Health (NIH). The main objective of this resource is to produce microbial data, which enhances our understanding of the relationship between the human microbiome and human health. The HMP has investigated the microorganisms from multiple locations on the human body, including the gastrointestinal tract, oral cavity, nasal passages, skin, and urogenital tract. The Project has produced a significant amount of genomic and metagenomic data.

Currently, there are more than 1300 human microorganisms that have had their entire genome sequenced. Additionally, there are more than 1200 microbiome samples collected from various organs of hundreds of human subjects that have been sequenced using whole metagenomic shotgun sequencing technology. All of these data are available on the HMP-DACC. In addition to its role as a data source, the HMP-DACC is an information portal for news and research progress related to human microbiome research.

Pathosystems Resource Integration Center

The Pathosystems Resource Integration Center (PATRIC, http://atricbrc.org/portal/portal/patric/Home) [7] is a pathogen information system with rich data and analysis tools that supports studies on bacteria-based infectious diseases. PATRIC integrates a variety of data types, including sequence typing data, genomes, transcriptomes, 3D protein structures, and protein–protein interactions. There are currently more than 10,000 genomes and related transcriptomic data available in PATRIC. The genomes in PATRIC are annotated using the RAST server. Summaries and comparisons of annotations for homologous genes are also available in PATRIC. PATRIC collects and integrates available bacterial genomes and related disease information to minimize time and effort needed for biological analyses. PATRIC provides several tools for researchers to find data quickly and provides a personal workspace to save pertinent data. Users can also upload their own data into this personal workspace. Both public and personal data can be analyzed together with the provided suite of tools.

World Data Center for Microorganisms

The World Data Center for Microorganisms (WDCM, http://www.wfcc.info/ccinfo) was established as a data center of the World Federation for Culture Collections (WFCC). The WFCC collects, authenticates, manages, and distributes cultured microorganisms and cells. The aim of the WFCC is to promote the establishment of connections between microbial culture collections and provide related services to the research community. The WFCC pioneered the development of an international database for cultured microorganisms and cells, which led to the establishment of the WDCM. The WDCM is currently housed in the Institute of Microbiology, Chinese Academy of Sciences (CAS). The WDCM has integrated information from 678 culture collections from 71 countries. The information includes data on organization, cultured microorganisms, management, services, and scientific interests of the collections. The WDCM is a basic and useful information center for microbiological researchers.

Global Catalogue of Microorganisms

The Global Catalogue of Microorganisms (GCM, http://gcm.wfcc.info/) [8] is a reliable and robust microorganism information system that helps culture collections manage and share data. The GCM was designed and constructed by the WDCM. Currently, the GCM contains strain information for more than 273,000 strains, covering 43,436 microbial species from 67 collections from 34 countries and regions. After users submit cultured microorganism catalogue information, the GCM extracts the species names and strain numbers from various sources, including GenBank, SwissProt, and PubMed. The GCM then links the species names and strain numbers to the corresponding genomic resource, including nucleotide sequences, protein sequences, and any published reference articles. The data are processed automatically with subsequent manual validation to ensure an accurate microbial resource database. The GCM provides services for all academic and industrial users. A number of online tools, including search functions and statistical analysis tools, are integrated into the GCM to facilitate ease of use.

Analyzer of Bio-resource Citations

The Analyzer of Bio-resource Citations (ABC, http://www.wfcc.info/abc/) is a microbial information system that connects microbial species names and strain numbers to the corresponding published reference articles. Through this system, researchers can easily identify published articles that have studied a particular microorganism. This is also useful for cultural collections, which can use the citation information when cataloging the microorganisms in their collection. According to the WDCM, in 2010, there were about 600 registered culture collections in nearly 70 countries. Behind this platform, there is a full-text file system with a bio-resource term indexing and mining engine, which automatically explores the citation information. ABC also provides an interface for users to access full-text articles (in flash format) and check the citation information automatically extracted by the mining engine. Another useful function for researchers is the statistics results generated by the statistics module, which accurately demonstrates the state of bio-resource citation information (automatic or curated results). There is currently citation information for more than 120,000 articles that have been published in 50,307 journals since January of 1953. The publications include information about 63,000 microbial strains, which belong to 131 culture collections in 50 countries.

Website for Avian Flu Information

The Website for Avian Flu Information (http://www.avian-flu.info) [9] was designed and created by the Institute of Microbiology, CAS in 2004. The website has been maintained and routinely updated for more than 10 years in support of public concerns and research interest that resulted from pathogenic influenza virus outbreaks. Information on influenza infection cases as well as sequencing data have increased significantly. The website provides information on outbreak reports, clinical diagnoses, prevention policy, scientific publications, medicines, and vaccines. In addition, the website has an integrated influenza virus sequence database and bioinformatics tools, to facilitate the analysis of various influenza viral sequences.

Global Mirror System of DNA Barcode Data

The Global Mirror System of DNA Barcode Data (GMS-DBD, http://www.boldmirror.net/) [10] is a web-based system that was designed and built by the Institute of Microbiology,

CAS. The system distributes DNA barcode sequences that are produced by the International Barcode of Life (iBOL) project. iBOL is the largest cooperation for studying biodiversity genomics. It aims to produce DNA barcode records from 5 million specimens from 500,000 species, including animals, plants, and fungi, to aid in species identification. The vast amount of DNA barcoding data that have been generated by collaborative research requires well-organized information system capabilities. In addition to the Barcode of Life Data system (BOLD) established in Canada, the GMS-DBD also plays an important role in the iBOL project. The GMS-DBD has been established in seven countries and aids the distribution and sharing of DNA barcode information worldwide.

Competing interests

The authors declared that there are no competing interests.

Acknowledgements

This research was supported by the National High-tech R&D Program of China (863 Program, Grant No. 2014AA021501) and the National Scientific-Basic Special Fund from the Ministry of Science and Technology of China (Grant No. 2014FY110500).

References

[1] Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Pillay M, et al. IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res 2014;42:D560–7.

[2] Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Pillay M, et al. IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Res 2014;42:D568–73.

[3] Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST server: rapid annotations using subsystems technology. BMC Genomics 2008;9:75.

[4] Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res 2014;42:D206–14.

[5] Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 2008;9:386.

[6] Wortman J, Giglio M, Creasy H, Chen A, Liolios K. A data analysis and coordination center for the human microbiome project. Genome Res 2010;11:O13.

[7] Snyder EE, Kampanya N, Lu J, Nordberg EK, Karur HR, Shukla M, et al. PATRIC: the VBI pathosystems resource integration center. Nucleic Acids Res 2007;35:D401–6.

[8] Wu L, Sun Q, Sugawara H, Yang S, Zhou Y, McCluskey K, et al. Global catalogue of microorganisms (gcm): a comprehensive database and information retrieval, analysis, and visualization system for microbial resources. BMC Genomics 2013;14:933.

[9] Liu D, Liu QH, Wu LH, Liu B, Wu J, Lao YM, et al. Website for avian flu information and bioinformatics. Sci China C Lift Sci 2009;52:470–3.

[10] Liu D, Liu L, Guo G, Wang W, Sun Q, Parani M, et al. BOLDMirror: a global mirror system of DNA barcode data. Mol Ecol Resour 2013;13:991–5.