OXFORD

Briefings in Bioinformatics, 2018, 1–3

doi: 10.1093/bib/bby013
Editorial

Editorial

Recent developments of software and database in microbial genomics and functional genomics

With the rapid progress in next-generation sequencing technologies and the exponential increase in available microbial genomes, mining biological knowledge from these genomes represents a challenge to the whole scientific community. In this themed issue, we present the invited review articles by the scientists who have made significant contributions in this field to review the state of the art of their influential work and discuss the upcoming challenges for the further development.

This issue aims to provide a comprehensive overview of recent developments of software and database in microbial genomics and functional genomics, including those for replication origin (Ori-Finder system and DoriC database), prophage (PHAST and PHASTER Web servers), operon (DOOR database), secondary metabolite biosynthetic gene clusters (antiSMASH software), antimicrobial resistance (AMR) (PATRIC database), orthologous genes/proteins (COG database), pathway/genome (MicroScope and BioCyc database), genome visualization (CGView software family), metagenomic classification, assembly and analysis (Kraken, Centrifuge, VALET, MG-RAST, etc.) and multiple sequence alignment (MSA) (MAFFT online service). We will introduce the articles briefly in the order of their appearance to give a quick reference guide for the readers.

For the purpose of phylogenetic classification of proteins from complete genomes, a collection of the COGs (Clusters of Orthologous Groups of proteins, re-branded as Clusters of Orthologous Genes later) was constructed 20 years ago. Since then, the COG approach has become one of the foundations of comparative and evolutionary genomics. At present, this pioneer work is still extremely popular as an essential tool in microbial genomics and a solid platform for comparative genomic analysis. Galperin *et al.* [1] review the key principles and the applications of the COG approach, and also discuss the unresolved problems in the COG approach and the possible solutions (Recommended in F1000 Prime).

Toward a full understanding of the newly sequenced genomes, the MicroScope platform has been developed as an integrated environment to perform comparative genomic and metabolic analyses of microbial genomes. Furthermore, the MicroScope platform also provides a collaborative environment to share and improve knowledge on these genomes. From the end users' point of view, Médigue *et al.* [2] present a comprehensive description of MicroScope services to help the microbiologists extend their analyses of species of interest.

To provide a reference on the thousands of sequenced microbial genomes and their metabolic pathways, the BioCyc collection has been generated based on the prediction by computer programs, the information integrated from other bioinformatics databases and the curation from the biomedical literature by biologist curators. Karp *et al.* [3] outline the recent improvements to BioCyc, including the expansion of BioCyc database content as well as the related bioinformatics tools for query, visualization and analysis.

To facilitate the research on AMR, PATRIC (Pathosystems Resource Integration Center) includes AMR information at both the genome and gene level, curates the AMR-specific functional roles manually and builds the machine learning-based classifiers to predict the AMR phenotypes and their genetic determinants for the submitted genomes. Consequently, researchers can easily explore AMR data and design experiments based on whole genomes or individual genes, and they can also quickly obtain these data in their private genomes and compare with the PATRIC collection [4].

Toward a rapid and reliable identification of all the potential secondary metabolite biosynthetic gene clusters in the newly sequenced genomes, the genome mining platform antiSMASH (antibiotics and Secondary Metabolite Analysis Shell) has been released to combine and extend the functionality of the previous tools with a user-friendly Web interface. Blin *et al.* discuss the principles underlying the predictions of antiSMASH and other computational tools, and provide practical advice for their applications. In addition, Blin *et al.* [5] point out the important caveats that should be taken into consideration when designing and interpreting genome mining studies.

To perform the prophage and cryptic prophage identification in bacterial genomes, PHAST (PHAge Search Tool) and PHASTER (PHAge Search Tool-Enhanced Release) have been presented in succession, which have already become two of the most widely used Web servers in this field. Arndt *et al.* [6] review the main capabilities of PHAST and PHASTER, provide some practical guidance regarding their applications and discuss possible future directions for the development of PHASTEST, a successor to PHASTER.

To identify the operons and transcriptional units accurately, an operon database DOOR (Database of prOkaryotic OpeRons) for genome analyses and functional inference has been developed and updated continually. Cao *et al.* review the information stored in DOOR database as well as the utility tools in support of operon-based analyses and information discovery. In addition, Cao *et al.* [7] explain how to computationally derive such data and demonstrate how to facilitate systems-level studies based on them.

For the purpose of large-scale identification and characterization of replication origins in microbial genomes, Ori-Finder system and DoriC database have been developed and maintained based on the Z-curve method. Luo *et al.* introduce the main methodology of Ori-Finder system used to predict the replication origins for bacteria (Ori-Finder Web server) or archaea (Ori-Finder 2 Web server), and review the development of DoriC (Database of *ori*Cs in bacterial and archaeal genomes) and its application, including the large-scale analyses of *ori*Cs and strand bias in prokaryotic genomes. In addition, Luo *et al.* [8] also present some future directions and aspects for extending the application of Ori-Finder and DoriC.

To visualize and compare circular genomes by graphical genome maps, the CGView (Circular Genome Viewer) software family has been designed to generate the visually impressive graphical genome maps for bacteria, organelles and viruses. Stothard *et al.* [9] describe the capabilities of the original CGView program and subsequent companion applications, such as the CGView Server and the CGView Comparison Tool, and also discuss the newer GView program, a rewrite of CGView with support for linear maps and interactive editing, and its companion Web application (the GView Server), which provides the analysis pipelines for comparative genomics particularly.

To assist in the analysis of metagenomics data sets, a great variety of computational tools have been developed. Breitwieser *et al.* [10] review the methods and databases for classification and assembly of metagenomics data, and also discuss the challenges presented by inconsistencies in microbial taxonomy as well as contamination in the genome resources. This review has successfully attracted the readers' attention, and once became the most read paper of *Briefings in Bioinformatics*.

Despite recent advances in metagenomic assembly, validation is still the key for moving forward because of the imperfect assembled contigs. From a computational and technological perspective, Olson *et al.* highlight the recent developments in metagenomic assembly, summarize the key approaches for genomic and metagenomic assembly validation and demonstrate the insights derived from assemblies through the lens of validation. Olson *et al.* [11] also discuss the potential impact of long-read technologies on metagenomics, and future challenges and opportunities in the field of metagenomic assembly and validation.

To provide the automatic phylogenetic and functional analysis of metagenomes, MG-RAST (Metagenomics RAST server) has been designed as a hosted, open-source and open-submission platform. Meyer *et al.* introduce the implementation, backend components, current workflow of MG-RAST version 4, which has increased throughput dramatically with the same amount of resources and brings a new user interface fully relying on the API (Application Programmers Interface) for data access and the analysis reuse. Meyer *et al.* also present the lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis [12].

To meet the challenge of the MSA in the era of big data, MAFFT (Multiple Alignment based on Fast Fourier Transformation), a popular MSA program, has significantly improved in performance and usability recently. Katoh *et al.* [13] describe in detail the Web interface for newly added options for large data, as well as interactive usage to refine sequence data sets and MSAs, which will facilitate the MAFFT users in handling big data and extracting biological information more quickly and effectively (Recommended in F1000 Prime).

In conclusion, this special issue will be helpful to the scientific community in the fields of microbiology, bioinformatics/computational biology, genomics, etc. Personally, I have benefited enormously from these outstanding works, whether studying as a graduate student a dozen years ago, or working as a researcher up to now. Here, I would like to take this opportunity to thank the authors for their excellent contributions to this issue.

Feng Gao[1,2,3]
[1]*Department of Physics, School of Science, Tianjin University, Tianjin, China*
[2]*Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin, China*
[3]*SynBio Research Platform, Collaborative Innovation Center of Chemical Science and Engineering, Tianjin, China.*
E-mail: *fgao@tju.edu.cn*

Acknowledgements

The author would like to thank Professor Chun-Ting Zhang for the invaluable assistance and inspiring discussions. This work was funded by National Natural Science Foundation of China (grant numbers 31571358, 21621004 and 31171238) and the National High-Tech Research and Development Program (863) of China (grant number 2015AA020101).

References

1. Galperin MY, Kristensen DM, Makarova KS, *et al.* Microbial genome analysis: the COG approach. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx117.
2. Médigue C, Calteau A, Cruveiller S, *et al.* MicroScope—an integrated resource for community expertise of gene functions and comparative analysis of microbial genomic and metabolic data. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx113.
3. Karp PD, Billington R, Caspi R, *et al.* The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx085.
4. Antonopoulos DA, Assaf R, Aziz RK, *et al.* PATRIC as a unique resource for studying antimicrobial resistance. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx083.
5. Blin K, Kim HU, Medema MH, *et al.* Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx146.
6. Arndt D, Marcu A, Liang Y, *et al.* PHAST, PHASTER and PHASTEST: tools for finding prophage in bacterial genomes. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx121.
7. Cao H, Ma Q, Chen X, *et al.* DOOR: a prokaryotic operon database for genome analyses and functional inference. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx088.
8. Luo H, Quan CL, Peng C, *et al.* Recent development of Ori-Finder system and DoriC database for microbial replication origins. *Brief Bioinform* 2018, doi: 10.1093/bib/bbx174.
9. Stothard P, Grant JR, Van Domselaar G. Visualizing and comparing circular genomes using the CGView family of tools. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx081.

10. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx120.
11. Olson ND, Treangen TJ, Hill CM, *et al*. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx098.
12. Meyer F, Bagchi S, Chaterji S, *et al*. MG-RAST version 4—lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx105.
13. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform* 2017, doi: 10.1093/bib/bbx108.