Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw499
Advance Access Publication Date: 4 August 2016
Applications Note

Genome analysis

FARAO: the flexible all-round annotation organizer

Rickard Hammarén[1,2], Chandan Pal[1,3] and Johan Bengtsson-Palme[1,3,*]

[1]Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Guldhedsgatan 10, SE-413 46, Gothenburg, Sweden, [2]Science for Life Laboratory, Department of Medical Epidemiology and Biostatistics, Karolinska Institute, SE-171 21 Solna, Stockholm, Sweden and [3]Centre for Antibiotic Resistance Research (CARe) at University of Gothenburg, Gothenburg, Sweden

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on May 20, 2016; revised on July 21, 2016; accepted on July 22, 2016

Abstract

Summary: With decreasing costs of generating DNA sequence data, genome and metagenome projects have become accessible to a wider scientific community. However, to extract meaningful information and visualize the data remain challenging. We here introduce FARAO, a highly scalable software for organization, visualization and integration of annotation and read coverage data that can also combine output data from several bioinformatics tools. The capabilities of FARAO can greatly aid analyses of genomic and metagenomic datasets.

Availability and Implementation: FARAO is implemented in Perl and is supported under Unix-like operative systems, including Linux and macOS. The Perl source code is freely available for download under the MIT License from http://microbiology.se/software/farao/.

Contact: johan.bengtsson-palme@microbiology.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Rapid advancements in DNA sequencing technology have quickly shifted the bottleneck for biological knowledge generation from restricted availability of sequence data to limited ability to analyze and mine the datasets generated (Heather and Chain, 2016). Thus, the analysis of nucleotide sequences from recent genomics projects requires tools that cannot only handle massive amounts of sequence data, but can also filter it and present the most relevant information to the user in an efficient manner. In this process, visualization of DNA sequence data and its associated annotations is key. However, annotations are often generated using a range of tools and databases, each with its own output formats and restrictions (Oulas *et al.*, 2015). Hence the ability to combine annotations from different sources is instrumental to form a complete picture of the sequences under scrutiny. Furthermore, with the advent of shotgun metagenomics, which can produce highly fragmented assemblies with thousands or even millions of contigs from a single sample (Bengtsson-Palme *et al.*, 2014), scalability becomes an important issue to consider. Metagenomics and RNA sequencing also introduce the ability to quantify the abundance of contigs and features in the data based on read coverage, which therefore also becomes an important part to incorporate in analysis and visualization frameworks.

A multitude of tools for visualization of annotated sequence features exist, but most of them either cannot integrate annotation and read coverage data (e.g. Kumar *et al.*, 2011; Lee *et al.*, 2009; Pan *et al.*, 2010; Wilkinson *et al.*, 2002), focus mainly on visualizing coverage information (Carver *et al.*, 2012; Woźniak *et al.*, 2011), lack scalability beyond a few hundred sequences (Fiume *et al.*, 2010), cannot efficiently filter the data to the needs of the user (Hou *et al.*, 2010; Peterson *et al.*, 2012), or pose restrictions on what tools and file formats that can be used for input of annotated features and/or coverage information (Milne *et al.*, 2010; Okonechnikov *et al.*, 2012). There also exist a range of web-based tools that can

Table 1. Brief description of the FARAO command-line tools

Name of tool	Description
setup_annotation_db	Creates a new FARAO database from a FASTA file, for annotation or coverage information
add_annotation	Adds annotations from a bioinformatics software to an annotation database
get_annotations	Queries an annotation database for information, e.g. entries matching reference proteins in a certain database with >90% identity
remove_annotations	Deletes specific annotation information from an annotation database
add_mapped_reads	Adds coverage information from a read mapping software to a coverage database
get_coverage	Queries a coverage database for information on mapped reads to each database sequence
remove_mapped_reads	Deletes coverage information for a specific library from a coverage database
estimate_coverage	Used to estimate the coverage of specific features from an annotation database, based on the information in a coverage database
annotation_db_to_mysql	Convert annotation databases to MySQL format
coverage_db_to_mysql	Convert coverage databases to MySQL format

visualize high-throughput data, generally connected to automated pipelines for sequence annotation (e.g. Cantor *et al.*, 2015; Meyer *et al.*, 2008; Sharma *et al.*, 2015). However, none of these tools present a scalable, multi-purpose framework for integrating annotation and coverage information from several different database sources into a database that can be queried at the request of the user.

Consequently, there is a need for highly flexible and adaptable software tools that can bridge disparate annotation and quantification approaches. Such a toolkit should be agnostic to what kind of bioinformatic tools that have been used to annotate sequences, and must be able to adapt to novel annotation utilities as they are introduced, for example by providing extension modules to the software that can be tapped into by users skilled in programming. Given the challenges associated with filtering the data, the software needs to be able to process it in a way that is scalable up to millions of sequences with hundreds of features each. Finally, it is desirable if the software is able to produce publication quality visual output that can be further processed by users. Cantor *et al.* (2015) recently introduced a web server tool with many of these capabilities, which, however, requires the user to upload the sequence and annotation data. This might be prohibitive for users with limited Internet connection, and while web based tools are convenient for many users, they do not lend themselves to integration with Unix and/or Linux based workflows, which typically rely on command-line tools.

With these limitations in mind, we have developed FARAO – the Flexible All-Round Annotation Organizer – that functions as a highly configurable set of command-line tools that: (i) integrate annotation and coverage information for the same sequence set; (ii) are scalable to millions of sequences and features; (iii) can filter out sequences with annotations satisfying criteria given by the user; (iv) can handle annotations produced by a range of bioinformatics tools; and (v) provide a flexible interface for writing custom parsers for virtually any format not supported out of the box.

2 Implementation

FARAO is a set of command-line tools implemented in Perl 5 and should be functional under any version of Unix or Linux, including macOS. It uses the DBI and DBD modules for MySQL compatibility, the PostScript-Simple module for EPS graphics creation, and the GD library for PNG image generation. None of these libraries are required to install FARAO and use its basic functionality, although installing them greatly extends the capabilities of the software.

FARAO consists of ten programs, each with specific functions for interacting with annotation and coverage data (Table 1). In addition, FARAO comes with a bundled set of parsers for various bioinformatics software (see the manual; Supplementary Item 1). The set of parsers can easily be expanded upon with additional parsers written by the user or downloaded from the library located at http://microbiology.se/software/farao/parsers. Detailed instructions on how to create custom parsers are available in the FARAO manual. The runtimes and memory requirements of FARAO are highly dependent on the use cases. For example, for a database consisting of 1.7 million contigs (encompassing 2.4 Gbp of sequence data) with annotations from 8 different sources, occupying 1.4 Gb of disk space, running a typical request for retrieving all annotations of a certain kind took 1 minute and 31 s, using 1.7 Gb of RAM. In contrast, looking up all annotations for a specific entry only took 15 s and used less than 32 Mb of RAM.

3 Conclusions

We introduce FARAO, an annotation and read coverage visualization software that enables integration of annotation and coverage information, is highly scalable, can combine output data from several bioinformatics tools and produces high-quality EPS output for figures. In addition, FARAO eases annotation of genomes and metagenomes by enabling filtering of sequence data by the features they contain.

Acknowledgements

The authors would like to thank Dr. Anna Johnning, Dr. Fredrik Boulund and Dr. Carl-Fredrik Flach for valuable input on FARAO features and visualization.

Conflict of Interest: none declared.

References

Bengtsson-Palme,J. *et al.* (2014) Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. *Front. Microbiol.*, **5**, 648.

Cantor,M. *et al.* (2015) Elviz – exploration of metagenome assemblies with an interactive visualization tool. *BMC Bioinformatics*, **16**, 130.

Carver,T. *et al.* (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, **28**, 464–469.

Fiume,M. *et al*. (2010) Savant: genome browser for high-throughput sequencing data. *Bioinformatics*, **26**, 1938–1944.

Heather,J.M. and Chain,B. (2016) The sequence of sequencers: the history of sequencing DNA. *Genomics*, **107**, 1–8.

Hou,H. *et al*. (2010) MagicViewer: integrated solution for next-generation sequencing data visualization and genetic variation detection and annotation. *Nucleic Acids Res.*, **38**, W732–W736.

Kumar,K. *et al*. (2011) AGeS: a software system for microbial genome sequence annotation. *PLoS ONE*, **6**, e17469.

Lee,D. *et al*. (2009) WeGAS: a web-based microbial genome annotation system. *Biosci. Biotechnol. Biochem.*, **73**, 213–216.

Meyer,F. *et al*. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Milne,I. *et al*. (2010) Tablet–next generation sequence assembly visualization. *Bioinformatics*, **26**, 401–402.

Okonechnikov,K. *et al*. (2012) Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, **28**, 1166–1167.

Oulas,A. *et al*. (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinf. Biol. Insights*, **9**, 75–88.

Pan,X. *et al*. (2010) Domain view: a web tool for protein domain visualization and analysis. *J. Struct. Funct. Genomics*, **11**, 241–245.

Peterson,E.S. *et al*. (2012) VESPA: software to facilitate genomic annotation of prokaryotic organisms through integration of proteomic and transcriptomic data. *BMC Genomics*, **13**, 131.

Sharma,A.K. *et al*. (2015) Woods: A fast and accurate functional annotator and classifier of genomic and metagenomic sequences. *Genomics*, **106**, 1–6.

Wilkinson,M.D. *et al*. (2002) Genquire: genome annotation browser/editor. *Bioinformatics*, **18**, 1398–1399.

Woźniak,M. *et al*. (2011) CAMBerVis: visualization software to support comparative analysis of multiple bacterial strains. *Bioinformatics*, **27**, 3313–3314.