BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 19 2014, pages 2820–2821
doi:10.1093/bioinformatics/btu394

Sequence analysis Advance Access publication June 17, 2014

DAFGA: diversity analysis of functional gene amplicons

Yongkyu Kim* and Werner Liesack
Department of Biogeochemistry, Max Planck Institute for Terrestrial Microbiology. 35043 Marburg, Germany

Associate Editor: John Hancock

ABSTRACT

Summary: Diversity analysis of functional marker genes provides physiological insights into microbial guilds that perform an ecologically relevant process. However, it is challenging to group functional gene sequences to valid taxonomic units, primarily because of differences in the evolutionary rates of individual genes and possible horizontal gene transfer events. We developed a python script package named DAFGA, which estimates the evolutionary rate of a particular functional gene in a standardized manner by relating its sequence divergence to that of the 16S rRNA gene. As a result, DAFGA provides gene-specific parameter sets for operational taxonomic unit clustering and taxonomic assignment at desired rank, and it can be implemented into the diversity measurements offered by QIIME.

Availability and implementation: DAFGA is freely available with a manual and test data from https://github.com/outbig/DAFGA.

Contact: yongkyu.kim@mpi-marburg.mpg.de

Received on April 14, 2014; revised on May 22, 2014; accepted on June 11, 2014

1 INTRODUCTION

Microbes are the most abundant and most diverse organisms on earth playing crucial roles in ecosystem functioning. Massively parallel sequencing of 16S rRNA genes allows us to explore the phylogenetic diversity and taxonomic composition of microbial communities in any environment (e.g. Serkebaeva *et al.*, 2013). Contrary to the 16S rRNA gene, genes encoding key catalytic enzymes define particular functional guilds that perform ecologically relevant functions, e.g. methane-oxidizing bacteria (*pmoA*) and nitrogen-fixing bacteria (*nifH*). Exploring the prevalence and diversity of these protein-coding genes can help identify novel lineages of particular functional guilds as well as physiologically differentiate closely related phylotypes. Similarity of tree topology with that of the 16S rRNA gene should reveal the potential of a functional gene to be used as a phylogenetic marker (Case *et al.*, 2007). Protein-coding genes, however, evolve faster than the 16S rRNA gene, and at varying rates. There are considerable variations in the evolutionary rates (ERs), even among individual genes present in the same genome (Du *et al.*, 2013). Use of arbitrary threshold values for clustering functional gene amplicons into operational taxonomic units (OTUs) could lead to incorrect estimates of microbial diversity within functional guilds. Therefore, computational tools for functional gene amplicon analysis have to take into account the ER of a targeted gene. Because variations in ER result primarily from changes in the protein sequence itself rather than from external factors such as changing environment or lifestyle (Bedford *et al.*, 2008), the ER of individual genes can be measured in relation to that of 16S rRNA genes (Degelmann *et al.*, 2010). DAFGA compares the sequence divergence of functional genes with that of 16S rRNA genes obtained from the same source organisms. It then extrapolates the sequence identity thresholds that define different taxonomic ranks and use them for OTU clustering and taxonomic assignment to the lowest rank possible.

2 FEATURES AND METHODS

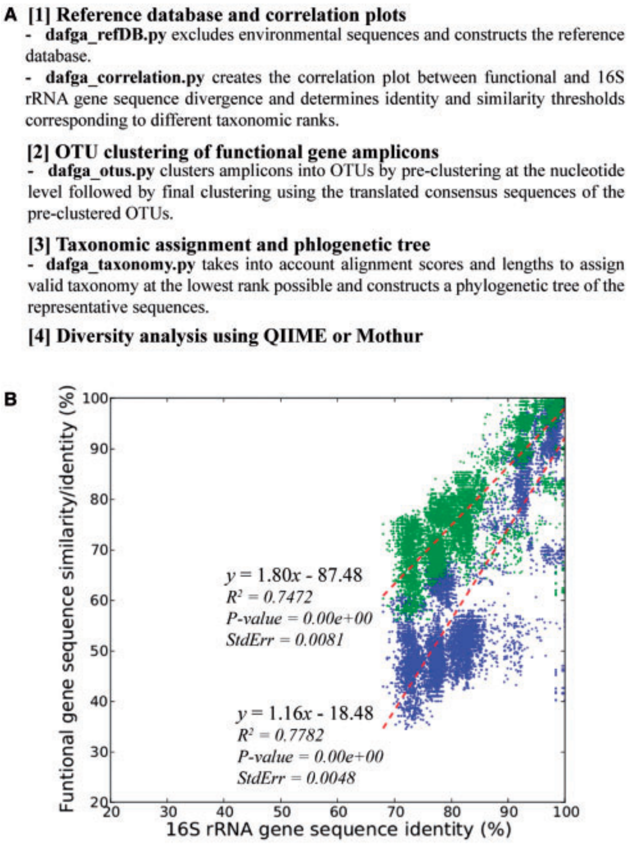
2.1 Reference database and correlation plots

DAFGA parses the publicly available sequences of a functional gene, which are retrieved from the NCBI protein database in genpept format. It excludes all the environmental sequences lacking taxonomic information and subsequently constructs the reference database to be used for taxonomic assignment. DAFGA retrieves the full taxonomic lineage of each reference sequence from the NCBI taxon ID, and fetches near full-length 16S rRNA gene sequences (1200–1600 bp in length) of the source organisms from the NCBI nucleotide database, but only from those organisms that were identified at strain level. The pairs of functional and 16S rRNA gene sequences are used to compute the ER of the functional gene in relation to that of its 16S rRNA gene by pairwise alignments between all the strain-level source organisms using the EMBOSS Smith–Waterman alignment tool (Goujon *et al.*, 2010). The identity thresholds of a functional gene that correspond to different taxonomic ranks are deduced from a linear regression curve of identity scores in sequence alignments and used for OTU clustering and taxonomic assignment. Alternatively, similarity scores can be used to taxonomically assign OTUs (Fig. 1).

2.2 OTU clustering of functional gene amplicons

Insertion and deletion errors in next-generation sequencing (e.g. 454 pyrosequencing and Ion Torrent semiconductor sequencing) can often cause a shift in reading frame during translation, thereby inflating diversity of functional gene sequences. To overcome this problem, DAFGA uses a two-step procedure for OTU clustering. Amplicon reads, which passed user-defined quality-filtering, are preclustered into OTUs defined by high nucleotide sequence identity (>97%), using USEARCH (Edgar, 2010). A consensus sequence of each OTU is generated from multiple sequence alignment and selected to represent the preclustered OTU. The translated consensus sequences are subjected to final OTU clustering using the gene-dependent sequence identity threshold that corresponds to the desired taxonomic rank.

*To whom correspondence should be addressed.



A [1] Reference database and correlation plots

- dafga_refDB.py excludes environmental sequences and constructs the reference database.
- dafga_correlation.py creates the correlation plot between functional and 16S rRNA gene sequence divergence and determines identity and similarity thresholds corresponding to different taxonomic ranks.

[2] OTU clustering of functional gene amplicons

- dafga_otus.py clusters amplicons into OTUs by pre-clustering at the nucleotide level followed by final clustering using the translated consensus sequences of the pre-clustered OTUs.

[3] Taxonomic assignment and phlogenetic tree

- dafga_taxonomy.py takes into account alignment scores and lengths to assign valid taxonomy at the lowest rank possible and constructs a phylogenetic tree of the representative sequences.

[4] Diversity analysis using QIIME or Mothur

Fig. 1. Workflow of DAFGA (**A**) and correlation plot of *pmoA* and 16S rRNA gene sequence divergence (**B**). The *pmoA* gene was selected as a representative example. *pmoA* reference sequences deposited in FunGene (http://fungene.cme.msu.edu/) were obtained in genpept format from NCBI non-redundant protein database. Similarity and identity values of protein sequence alignments were plotted in green and blue, respectively. Dashed red line is a linear regression curve of, respectively, similarity and identity of *pmoA* over 16S rRNA gene sequence identity

Centroid sequences are selected to represent each final OTU and used for both taxonomic assignment and phylogenetic tree construction. The OTU mapping file is generated by merging preclustering and clustering output. When examined with *pmoA* test amplicon data (Lüke and Frenzel, 2011), this two-step clustering reduced the number of OTUs as compared with a single-step approach based on translated amino acid sequences. In particular, singleton OTUs significantly declined.

2.3 Taxonomic assignment and phylogenetic tree

The number of functional gene sequences that are available for the construction of reference databases is limited and biased toward taxa characterized by sequenced genomes. For instance, dependent on the functional marker gene, the number of sequences deposited in FunGene ranges from 100 to maximum 77 876 (including environmental sequences), while >3 million 16S rRNA sequences are available (Fish *et al.*, 2013). In similarity-based taxonomic classification of functional gene sequences, lack of homologues at low taxonomic ranks, such as genus or species, can often lead to incorrect classification due to high sequence divergence with the most homologous sequence. For instance, a statistically significant alignment (*E*-value < 1e-10) shows 40% amino acid identity between environmental sequence and the reference sequence of a particular species. This explicitly indicates that the environmental sequence is derived not from the same species but from another species, so that the relationship is validly recognized only on a higher taxonomic level such as, for example, family or order level. Therefore, DAFGA takes into account the query coverage and the absolute identity or similarity scores in alignment with the best homologues, and performs assignment at the most likely taxonomic rank by referring to the correlation plot between the functional and 16S rRNA gene sequences. It creates an OTU table with taxonomic information and a phylogenetic tree of representative sequences. These outputs can be used as input files for various diversity measurements in QIIME (Caporaso *et al.*, 2010) or Mothur (Schloss *et al.*, 2009).

3 CONCLUSION

The most critical step in the diversity analysis of functional marker genes and rRNA genes is to cluster taxonomically homogeneous sequences into valid OTUs. DAFGA offers a standardized procedure to determine the ER of functional genes in relation to that of the 16S rRNA gene, and it provides sequence identity thresholds that correspond to different taxonomic ranks. As such, DAFGA allows for the automated cluster analysis and taxonomic categorization of environmental sequence data and provides a computational means to evaluate the potential of a functional gene as a phylogenetic marker.

Funding: Yongkyu Kim received a postdoctoral fellowship of the Max Planck Society. The project was supported by the Deutsche Forschungsgemeinschaft (collaborative research center SFB 987).

Conflict of Interest: None declared.

REFERENCES

Bedford,T. *et al.* (2008) Overdispersion of the molecular clock varies between yeast, *Drosophila* and mammals. *Genetics*, **179**, 977–984.

Caporaso,J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Case,R.J. *et al.* (2007) Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.*, **73**, 278–288.

Lüke,C. and Frenzel,P. (2011) Potential of *pmoA* amplicon pyrosequencing for methanotroph diversity studies. *Appl. Environ. Microbiol.*, **77**, 6305–6309.

Degelmann,D.M. *et al.* (2010) Different atmospheric methane-oxidizing communities in European beech and Norway spruce soils. *Appl. Environ. Microbiol.*, **76**, 3228–3235.

Du,X. *et al.* (2013) Why does a protein's evolutionary rate vary over time? *Genome Biol. Evol.*, **5**, 494–503.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Fish,J.A. *et al.* (2013) FunGene: the functional gene pipeline and repository. *Front. Microbiol.*, **4**, 1–14.

Goujon,M. *et al.* (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, **38**, W695–W699.

Serkebaeva,Y.M. *et al.* (2013) Pyrosequencing-based assessment of the *Bacteria* diversity in surface and subsurface peat layers of a northern wetland, with focus on poorly studied phyla and candidate divisions. *PLoS One*, **8**, e63994.

Schloss,P.D. *et al.* (2009) Introducing mothur: open-source, platform-independent, community–supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.