BIOINFORMATICS APPLICATIONS NOTE

2013, pages 1–3
doi:10.1093/bioinformatics/btt123

Sequence analysis

Advance Access publication March 19, 2013

FunFrame: functional gene ecological analysis pipeline

David Weisman, Michie Yasuda and Jennifer L. Bowen*

Department of Biology, University of Massachusetts Boston, Boston, MA 02125, USA

Associate Editor: Inanc Birol

ABSTRACT

Summary: Pyrosequencing of 16S rDNA is widely used to study microbial communities, and a rich set of software tools support this analysis. Pyrosequencing of protein-coding genes, which can help elucidate functional differences among microbial communities, significantly lags behind 16S rDNA in availability of sequence analysis software. In both settings, frequent homopolymer read errors inflate the estimation of microbial diversity, and de-noising is required to reduce that bias. Here we describe FunFrame, an R-based data-analysis pipeline that uses recently described algorithms to de-noise functional gene pyrosequences and performs ecological analysis on de-noised sequence data. The novelty of this pipeline is that it provides users a unified set of tools, adapted from disparate sources and designed for different applications, that can be used to examine a particular protein coding gene of interest. We evaluated FunFrame on functional genes from four PCR-amplified clones with sequence depths ranging from 9084 to 14 494 sequences. FunFrame produced from one to nine Operational Taxonomic Units for each clone, resulting in an error rate ranging from 0 to 0.18%. Importantly, FunFrame reduced spurious diversity while retaining more sequences than a commonly used de-noising method that discards sequences with frameshift errors.

Availability: Software, documentation and a complete set of sample data files are available at http://faculty.www.umb.edu/jennifer.bowen/software/FunFrame.zip.

Contact: Jennifer.Bowen@umb.edu

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

Received on November 9, 2012; revised on March 5, 2013; accepted on March 6, 2013

1 INTRODUCTION

Pyrosequencing of 16S rDNA is commonly used to study microbial community structure, and existing bioinformatics pipelines are primarily designed for the analysis of 16S rDNA (Caporaso *et al*., 2010b; Schloss *et al*., 2009). Noise from DNA sequencing errors upwardly biases the estimates of microbial diversity (Huse *et al*., 2007; Kunin *et al*., 2010). Substantial progress has been made to remove noise artifacts from 16S rDNA sequence datasets (Huse *et al*., 2010; Quince *et al*., 2009, 2011).

Targeted metagenomics of protein-coding genes offers the possibility of focusing on the diversity, abundance and expression of specific genes, particularly those genes that encode enzymes critical to biogeochemical cycling. Protein-coding sequences fundamentally differ from non-coding 16S rDNA in that the genetic code implicitly constrains the sequences to the 61 amino-coding triplets. This distinction is particularly relevant with pyrosequencing, as the technology is prone to misread the lengths of long homopolymers, thereby creating the appearance of frameshift mutations. To reduce the inflated diversity bias with protein-coding genes, a commonly used approach discards sequences containing unexpected stop codons (Jones *et al*., 2008; Iwai *et al*., 2010; Rozera *et al*., 2009). A more nuanced error detection algorithm, HMM-FRAME, was recently introduced (Zhang and Sun, 2011). In HMM-FRAME, a hidden Markov model (HMM) of the target protein, combined with a probabilistic model of homopolymer errors, detects and corrects the frameshifts caused by homopolymer read errors. In addition, the algorithm reports HMM alignment scores, which can be used in downstream quality filtering.

To facilitate targeted metagenomics using PCR-amplified protein-coding genes, we produced FunFrame, a complete bioinformatics pipeline that uses HMM-FRAME followed by chimera detection, Operational Taxonomic Unit (OTU) clustering, rarefaction and diversity estimation (Supplementary Fig. S1). Additionally, FunFrame performs clustering and ordination using UniFrac and Bray–Curtis metrics (Supplementary Figs S2 and S3).

Contrasting with the 16S rDNA pipelines QIIME (Caporaso *et al*., 2010b) and Mothur (Schloss *et al*., 2009), FunFrame centers around the R Project for Statistical Computing (R Core Team, 2012), which facilitates analysis with a rich set of ecological, statistical and visualization tools (Borcard *et al*., 2011; Oksanen *et al*., 2011).

2 METHODS

The FunFrame pipeline begins with HMM-FRAME (Zhang and Sun, 2011) for pyrosequencing error analysis; UCHIME (Edgar *et al*., 2011) for chimera detection (running in *de novo* mode without a reference database) and ESPRIT-Tree (Cai and Sun, 2011) for OTU clustering. FunFrame performs ecological analyses on the resulting OTU table. Sub-sampled diversity estimation is computed with QIIME (Caporaso *et al*., 2010b). Bray–Curtis distances of Hellinger-transformed counts are computed in Vegan (Oksanen *et al*., 2011). To compute unweighted and weighted UniFrac metrics (Hamady *et al*., 2010), representative OTUs are aligned with PyNAST (Caporaso *et al*., 2010a), a phylogeny is inferred using FastTree (Price *et al*., 2010) and UniFrac metrics are computed from the tree. Principal coordinates analysis and hierarchical clustering are performed in R on the Bray–Curtis and UniFrac metrics. Rarefaction curves and alpha diversity estimates are computed in Vegan. Redundancy analysis and constrained correspondence analysis are computed with user-supplied environmental variables and displayed as biplots using Vegan.

FunFrame programs are written in R and Python. A bash script runs the full pipeline; alternatively, users can run pipeline stages individually.

*To whom correspondence should be addressed.

A user-customizable configuration file specifies all parameter settings. Installation and operating instructions are provided with the software distribution, as are sample data with expected outputs.

Environmental clones containing the gene *nirS*, a gene in the microbial denitrification pathway (Zumft, 1997), were prepared (GenBank accessions KC203032–KC203035) and sequenced along with amplicon libraries (Supplementary Table S1) of environmental samples taken from sediments of the Great Sippewissett Salt Marsh, Cape Cod, MA, USA (Supplementary Methods). Sequences mismatching the 5′ primer or having ambiguous bases were removed, and remaining sequences were trimmed to 432 bp. With these inputs, we ran FunFrame using the cytochrome D1 HMM from Pfam (accession PF02239.10), and retained sequences with HMM scores >85. OTUs were defined as sequences within a 0.05 divergence. To compare FunFrame against stop-codon filtering, FunFrame was modified to replace HMM-FRAME with logic that translates sequences in three reading frames, and retains those with a full reading frame.

3 RESULTS AND DISCUSSION

In evaluating HMM-FRAME over the stop-codon filtering approach, we sought two objectives. First, we wanted to minimize the upward diversity estimation bias due to homopolymer errors, and we measured this property by analyzing the OTU counts of four clone libraries, measured at sequence depths of 9084, 5780, 11 382 and 14 494. Of these four libraries, FunFrame reported 9, 3, 3 and 1 OTUs per clone, whereas the stop-codon filter approach reported 11, 4, 3 and 2 OTUs per clone, respectively (Supplementary Table S2). Figure 1 shows that FunFrame produces lower OTU counts than stop-codon filtering.

Our second objective was to maximize the number of non-noisy sequences retained, which effectively increases the likelihood of detecting rare species. Starting with 119 663 total reads, the HMM filtering approach retained 117 659 (~98%) reads, whereas stop-codon filtering retained 104 347 (~87%) reads. After subsequent chimera detection, FunFrame retained 108 603 (~91%) versus 95 902 (~80%) reads for stop-codon filtering (Supplementary Table S3). This improvement is reflected in the greater number of sequences shown in the solid, compared with the dotted, lines in Figure 1. Analysis of these data demonstrates the pipeline's capacity to uncover ecologically meaningful patterns in environmental sequences (Supplementary Figs S2 and S3).

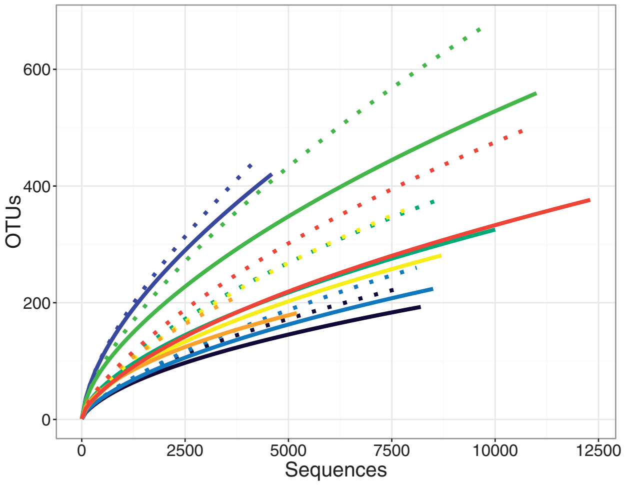

Fig. 1. Rarefaction curves based on FunFrame compared with stop-codon filtering. Colors represent biological samples; solid and dotted lines produced by FunFrame and stop-codon filtering, respectively. FunFrame tends to retain more sequences and produce less OTU inflation

In both criteria, FunFrame performed better than the stop-codon approach. FunFrame discards sequences with HMM scores below a configurable threshold, and adjusting this threshold can trade-off sequencing depth for OTU inflation. The results indicate that this parameter can be set such that both objectives exceed the results from the stop-codon filtering approach.

Microbial ecological analysis based on functional genes is an enormously powerful paradigm, which we believe will become widely used as DNA sequencing costs continue to decline. We offer FunFrame to the community with the hope that it will contribute to the development of this important area.

ACKNOWLEDGEMENTS

Hilary Morrison performed the pyrosequencing at the Marine Biological Laboratory. Sarah Hallin and Chris Jones kindly contributed their exceptionally well-curated *nirS* alignment. We thank an anonymous reviewer for making substantial and valuable technical suggestions.

Funding: Research funding came from the National Science Foundation (DEB-0717155 to J.E.H. and H.G.M., and award DBI-0400819 to J.L.B.). University of Massachusetts of Boston funding to J.L.B. provided additional salary support for M.Y.

Conflict of Interest: none declared.

REFERENCES

Borcard,D. *et al.* (2011) *Numerical Ecology with R*. Springer, New York.

Cai,Y. and Sun,Y. (2011) Esprit-tree: hierarchical clustering analysis of millions of 16s rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res.*, **39**, e95.

Caporaso,J.G. *et al.* (2010a) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, **26**, 266–267.

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

Edgar,R.C. *et al.* (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Hamady,M. *et al.* (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and phylochip data. *ISME J.*, **4**, 17–27.

Huse,S.M. *et al.* (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.*, **8**, R143.

Huse,S.M. *et al.* (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.*, **12**, 1889–1898.

Iwai,S. *et al.* (2010) Gene-targeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment. *ISME J.*, **4**, 279–285.

Jones,C.M. *et al.* (2008) Phylogenetic analysis of nitrite, nitric oxide, and nitrous oxide respiratory enzymes reveal a complex evolutionary history for denitrification. *Mol. Biol. Evol.*, **25**, 1955–1966.

Kunin,V. *et al.* (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.*, **12**, 118–123.

Oksanen,J. *et al.* (2011) *vegan: Community Ecology Package*. Version 2.0-2.

Price,M.N. *et al.* (2010) Fasttree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Quince,C. *et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods*, **6**, 639–641.

Quince,C. *et al.* (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.

R Core Team. (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rozera,G. *et al.* (2009) Massively parallel pyrosequencing highlights minority variants in the HIV-1 ENV quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology*, **6**, 15.

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.

Zhang,Y. and Sun,Y. (2011) Hmm-frame: accurate protein domain classification for metagenomic sequences containing frameshift errors. *BMC Bioinformatics*, **12**, 198.

Zumft,W.G. (1997) Cell biology and molecular basis of denitrification. *Microbiol. Mol. Biol. Rev.*, **61**, 533–616.