The Journal of Microbiology (2011) Vol. 49, No. 4, pp. 689-691
Copyright ⓒ 2011, The Microbiological Society of Korea

DOI 10.1007/s12275-011-1213-z

NOTE

rRNASelector: A Computer Program for Selecting Ribosomal RNA Encoding Sequences from Metagenomic and Metatranscriptomic Shotgun Libraries

Jae-Hak Lee[1], Hana Yi[2], and Jongsik Chun[1,2,3]*

[1]Interdisciplinary Graduate Program in Bioinformatics, [2]Institute of Molecular Biology and Genetics, [3]School of Biological Sciences, Seoul National University, Seoul 151-742, Republic of Korea

(Received April 25, 2011 / Accepted August 9, 2011)

Metagenomic and metatranscriptomic shotgun sequencing techniques are gaining popularity as more cost-effective next-generation sequencing technologies become commercially available. The initial stage of bioinformatic analysis generally involves the identification of phylogenetic markers such as ribosomal RNA genes. The sequencing reads that do not code for rRNA can then be used for protein-based analysis. Hidden Markov model is a well-known method for pattern recognition. Hidden Markov models that are trained on well-curated rRNA sequence databases have been successfully used to identify DNA sequence coding for rRNAs in prokaryotes. Here, we introduce rRNASelector, which is a computer program for selecting rRNA genes from massive metagenomic and metatranscriptomic sequences using hidden Markov models. The program successfully identified prokaryotic 5S, 26S, and 23S rRNA genes from Roche 454 FLX Titanium-based metagenomic and metatranscriptomic libraries. The rRNASelector program is available at http://sw.ezbiocloud.net/rrnaselector.

Keywords: rRNASelector, metagenomics, metatranscriptomics, HMMER, rRNA, computer program

Our understanding of microorganisms has advanced over the last two decades through the introduction of metagenomics (Schmidt *et al*., 1991; Handelsman *et al*., 1998; Rondon *et al*., 2000). Metagenomics is the culture-independent study of a collective set of genetic materials that are extracted directly from microbial communities, and thus, it is a powerful tool to unravel natural biodiversity without potential biases resulting from culturing or isolation. Several major technical limitations have been elucidated in the application of metagenomics in routine laboratories. Because PCR-based gene amplification may introduce bias (von Wintzingerode *et al*., 1997; Polz and Cavanaugh, 1998; Handelsman, 2004), massive sequencing of metagenomic shotgun libraries or metatranscriptomic libraries is believed to be the best method to dissect the genetic makeup of samples (Venter *et al*., 2004; Qin *et al*., 2010). The affordability of this approach has improved as more efficient DNA sequencing technologies become commercially available. These technologies are known as next-generation sequencing (NGS) (Margulies *et al*., 2005; Mardis, 2008; Valouev *et al*., 2008).

Given the accessibility of NGS technologies in many microbiological disciplines, one of the major limitations of metagenomics or metatranscriptomics at the level of DNA sequencing is the lack of adequate bioinformatic software to handle massive NGS data. Even though there are several computer programs and web servers available (Huson *et al*., 2007; Glass *et al*., 2010), there is a great need for efficient and easy-to-use software.

In general, metagenomic DNA fragments are assembled or used individually for homology-based searches against protein databases. Ribosomal RNA (rRNA)-coding genes that are important phylogenetic markers should be removed before searching for protein-based homology. This step is particularly critical in metatranscriptomic studies because the majority of fragments do not encode proteins but rRNA.

The presence of the rRNA-coding region in DNA sequences is readily identified using algorithms that are based on hidden Markov models. The RNAmmer (Lagesen *et al*., 2007) program is widely used in prokaryotic genome annotation; it identifies 5S, 16S, and 23S using the HMMer version 2 package (Eddy, 1998), and it is only applicable for large contigs and genome sequences. The Meta-RNA (Huang *et al*., 2009) is a Python program that identifies rRNA genes from metagenomic fragments and is used to select or remove rRNA genes prior to protein-based homology searches. However, this program is based on a script language and lacks graphical user-interfaces. Unlike the aforementioned computer programs, MG-RAST is not a standalone program but a web-based service that provides comprehensive bioinformatic analysis for metagenomic shotgun sequencing libraries. In MG-RAST, metagenomic fragments coding for rRNA are identified using the BLASTN program (Meyer *et al*., 2008). In this study, we introduced a JAVA-based computer program, which was named rRNASelector, for one-step selection of fragments containing rRNA genes using the HMMer package version 3.

To construct hidden Markov models (HMMs) of 5S, 16S, and 23S rRNAs, multiple sequence alignments (MSAs) of bacterial and archaeal rRNA gene sequences were generated

* For correspondence. E-mail: jchun@snu.ac.kr; Tel.: +82-2-880-8153; Fax: +82-2-874-8153

Table 1. Test datasets and performance of rRNASelector and Meta-RNA

Dataset	NCBI SRA accession	Type (Instrument)	Number of reads	Program	Number of 5S rRNA	Number of 16S rRNA	Number of 23S RNA	Accuracy[a]
Puerto Rico rainforest soil	SRR034257	Metagenome (454 FLX Titanium)	233,836	rRNASelector	7	68	111	100
				Meta-RNA	7	68	113	100
Windshield splatter	SRR014856	Metagenome (454 FLX Titanium)	93,380	rRNASelector	9	78	165	100
				Meta-RNA	9	74	151	100
Tidal salt marsh (water)	SRR013513	Metatranscriptome (454 FLX Titanium)	237,675	rRNASelector	10	8,930	119,615	99.9
				Meta-RNA	9	8,892	119,570	99.7
Mushroom Spring (water)	SRR106861	Metatranscriptome (454 FLX Titanium)	112,968	rRNASelector	6	28,812	71,277	100
				Meta-RNA	6	28,800	71,048	99.8

[a] Accuracy was based on the BLASTN search of selected 16S rRNA fragments against the EzTaxon-e database (http://eztaxon-e.ezbiocloud.net/).

by the following procedures. The MSAs of 5S, 16S, and 23S rRNA were retrieved from the 5S Ribosomal RNA Database (Szymanski *et al.*, 2002), EzTaxon-e database (Chun *et al.*, 2007), and NCBI genome databases. All sequences were aligned and verified using the secondary structural model. The sequences were used to generate HMM profiles using the hmmbuild program, which is part of the HMMer package version 3 (Eddy, 2009). The JAVA programming language was used to generate the rRNASelector program, which was tested for running on Microsoft Windows, Linux and Mac OS X.

The rRNASelector program identifies metagenomic fragments coding for prokaryotic rRNA genes if they meet the following two conditions: (i) a sequence fragment shows an overlap (>60 bp) with a rRNA HMM profile and (ii) the E-value is below 10^{-5}. Fragments satisfying these conditions are selected, and rRNA-coding regions are extracted. The unselected fragments are stored for subsequent protein-based analyses.

The performance of rRNASelector was evaluated using publically available NGS data. Two metagenomic sequence libraries (NCBI SRA accessions SRR034257 and SRR014856) and two metatranscriptomic libraries (SRR013513 and SRR-106861) were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/sra) and used as sample datasets. The overall performance of identification of rRNA genes is summarized in Table 1. Because no standard annotation is available for these datasets, we evaluated the accuracy of our program using BLASTN-based identification of selected genes against the reference 16S rRNA database. Selected 16S rRNA fragments were searched against the EzTaxon-e database (http://eztaxon-e.ezbiocloud.net/), which contains >30,000 representative prokaryotic 16S rRNA sequences with a minimal E-value of 10^{-5}. The average accuracy using four metagenomic or metatranscriptomic datasets was 99.999% (Table 1). The rRNA Selector program was useful in metatranscriptomics, and the majority of RNA fragments were identified as rRNA-coding genes (54 and 89% in two transcriptomic datasets; Table 1).

We compared the outcome of rRNASelector with the Meta-RNA (Huang *et al.*, 2009) using the same datasets (Table 2). The two programs generated similar results with few discrepancies. Differences may be attributed to differences in the HMM profiles and detection parameters in the two programs.

In conclusion, rRNASelector is a valuable tool to identify or filter out rRNA genes for non-rRNA-based studies from massive metagenomic or metatranscriptomic sequencing libraries. This program was implemented in Java and runs on Linux, Microsoft Windows, and Mac OS. The program, sample datasets, and user guide are available online at http://sw.ezbiocloud.net/rrnaselector.

This work was supported by a National Research Foundation grant (#2011-0016498) of Korea.

References

Chun, J., J.H. Lee, Y. Jung, M. Kim, S. Kim, B.K. Kim, and Y.W. Lim. 2007. Eztaxon: A web-based tool for the identification of prokaryotes based on 16S ribosomal rna gene sequences. *Int. J. Syst. Evol. Microbiol.* 57, 2259-2261.

Eddy, S.R. 1998. Profile hidden markov models. *Bioinformatics* 14, 755-763.

Eddy, S.R. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* 23, 205-211.

Glass, E.M., J. Wilkening, A. Wilke, D. Antonopoulos, and F. Meyer. 2010. Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes. *Cold Spring Harb. Protoc.* 2010, pdb prot5368.

Handelsman, J. 2004. Metagenomics: Application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669-685.

Handelsman, J., M.R. Rondon, S.F. Brady, J. Clardy, and R.M. Goodman. 1998. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem. Biol.* 5, R245-249.

Huang, Y., P. Gilna, and W. Li. 2009. Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* 25, 1338-1340.

Huson, D.H., A.F. Auch, J. Qi, and S.C. Schuster. 2007. Megan analysis of metagenomic data. *Genome Res.* 17, 377-386.

Lagesen, K., P. Hallin, E.A. Rodland, H.H. Staerfeldt, T. Rognes, and D.W. Ussery. 2007. Rnammer: Consistent and rapid annotation of ribosomal rna genes. *Nucleic Acids Res.* 35, 3100-3108.

Mardis, E.R. 2008. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387-402.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, and *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376-380.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E.M. Glass, M. Kubal, T. Paczian, and *et al.* 2008. The metagenomics rast server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9, 386.

Polz, M.F. and C.M. Cavanaugh. 1998. Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* 64, 3724-3730.

Qin, J., R. Li, J. Raes, M. Arumugam, K.S. Burgdorf, C. Manichanh, T. Nielsen, and *et al.* 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59-65.

Rondon, M.R., P.R. August, A.D. Bettermann, S.F. Brady, T.H. Grossman, M.R. Liles, K.A. Loiacono, and *et al*. 2000. Cloning the soil metagenome: A strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* 66, 2541-2547.

Schmidt, T.M., E.F. DeLong, and N.R. Pace. 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* 173, 4371-4378.

Szymanski, M., M.Z. Barciszewska, V.A. Erdmann, and J. Barciszewski. 2002. 5S ribosomal RNA database. *Nucleic Acids Res.* 30, 176-178.

Valouev, A., J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, and *et al*. 2008. A high-resolution, nucleosome position map of c. Elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18, 1051-1063.

Venter, J.C., K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, and *et al*. 2004. Environmental genome shotgun sequencing of the sargasso sea. *Science* 304, 66-74.

von Wintzingerode, F., U.B. Gobel, and E. Stackebrandt. 1997. Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis. *FEMS Microbiol. Rev.* 21, 213-229.