BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 10 2009, pages 1338–1340
doi:10.1093/bioinformatics/btp161

Sequence analysis

Identification of ribosomal RNA genes in metagenomic fragments

Ying Huang, Paul Gilna and Weizhong Li*

California Institute for Telecommunications and Information Technology, University of California, La Jolla, San Diego, California, USA

Received on November 3, 2008; revised on March 16, 2009; accepted on March 17, 2009

Advance Access publication April 3, 2009

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Identification of genes coding for ribosomal RNA (rRNA) is considered an important goal in the analysis of data from metagenomics projects. Here, we report the development of a software program designed for the identification of rRNA genes from metagenomic fragments based on hidden Markov models (HMMs). This program provides rRNA gene predictions with high sensitivity and specificity on artificially fragmented genomic DNAs.

Availability: Supplementary files, scripts and sample data are available at http://tools.camera.calit2.net/camera/meta_rna.

Contact: liwz@sdsc.edu

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

1 INTRODUCTION

The emerging field of metagenomics promises a more comprehensive and complete understanding of the microbial world. Many projects have been reported with metagenomic approaches to study microbes and microbial communities that live in many different environmental conditions (Tringe and Rubin, 2005). Analyzing the sequence data generated by these projects is far from easy and requires accessible and user-friendly tools (Raes *et al.*, 2007). An essential step in any metagenomics project is the identification of genes encoding for ribosomal RNAs (rRNAs), which are widely used for phylogenetic analysis and quantification of microbial diversity. Several methods haven been proposed for predicting non-coding RNA genes (Meyer, 2007), but a recent benchmark study by Freyhult *et al.* (2007) indicated that the most commonly used methods yield less than encouraging results. Lagesen *et al.* (2007) proposed RNAmmer, a program based on hidden Markov models (HMMs) for annotation of rRNA genes. Their algorithm predicts rRNAs in complete genomics sequences with high accuracy. However, a major concern for their predictions is the inability to deal with fragments of rRNAs. Compared with assembled genomic sequences from single species, the raw sequence reads from a typical metagenomic study often remain unassembled due to insufficient coverage. For a typical metagenome dataset, the length of sequence read is ∼100–450 bp using 454 pyrosequencing, or ∼700 bp long if using Sanger sequencing. Meanwhile, the full lengths of most of 16S and 23S rRNAs are >1200 bp. Therefore, most of rRNA genes in metagenomic sequencing reads are fragmentary, and will be overlooked by RNAmmer that focus on full length rRNAs. To overcome this limitation, we used HMMs that can discover incomplete rRNA gene fragments for predictions. In this article, we apply our algorithm on simulated sets of sequence reads of various lengths. Our method provides rRNA predictions with high-sensitivities and specificities on the benchmark dataset.

2 ALGORITHM DEVELOPMENT

As an important molecular machine in all living organisms, the ribosome can be broken down into two subunits, the small and the large subunit. In prokaryotes, the large subunit of the ribosome contains 5S and 23S rRNAs, while the small subunit contains 16S rRNAs. Therefore, we will try to build predictors for 5S, 16S and 23S rRNAs. To obtain a reliable multiple sequence alignment (MSA) for HMM building, we retrieved MSAs of 5S rRNAs from the 5S Ribosomal Database (Szymanski *et al.*, 2002), and MSAs of 16S and 23S rRNAs from the European rRNA database (Wuyts *et al.*, 2004). These databases provide high-quality alignment that combine sequence and structural information. The MSAs were then divided into bacterial and archaeal domains. All sequences with more than five ambiguous nucleotides in either end were removed from the alignment, and then sequences were further clustered at 98% identity threshold to reduce bias. We then used software package HMMER (Eddy, 1998) version 2.3.2 to create HMMs from these alignments. We used 'fs' mode in HMMER package for HMM building instead of 'ls' mode implemented in RNAmmer. In HMMER package, 'ls' mode is suitable for identification of a complete sequence domain, while 'fs' mode is capable of finding domain fragments and maybe useful to detect incomplete rRNA genes. In addition, domain information for sequences is not available in metagenomic projects, so HMMs from bacterial and archaeal rRNA alignments were both used to search input sequences. Each sequence was classified to the domain that reported the most significant *E*-value, and results obtained from corresponding HMMs were used as final result.

3 EVALUATION

Performance of our rRNA prediction algorithm was evaluated using artificial DNA fragments generated from fully sequenced archaea and bacteria genomes. GenBank files for all fully sequenced genomes were retrieved from the ENTREZ Genome Project

*To whom correspondence should be addressed.

Table 1. Prediction sensitivities for different fragment lengths

Prediction method	hmm_fs			BLASTN		
Length of reads	5S	16S	23S	5S	16S	23S
100	91.9	98.2	96.2	79.4	89.9	94.8
200	95.8	97.9	98.6	85.7	96.7	97.8
300	96.8	99.3	99.0	88.3	99.0	98.2
400	97.6	98.3	99.2	89.1	97.5	98.5
500	98.2	99.2	99.1	89.2	99.2	98.4
600	98.0	98.8	99.1	89.5	98.4	98.5
700	98.7	99.5	99.3	90.3	99.5	98.7
800	98.2	99.2	99.6	90.8	99.2	99.1

Here, hmm_fs represents our algorithm. Sensitivities are represented in percentage (%).

Table 2. Prediction specificities for different fragment lengths

Prediction method	hmm_fs			BLASTN		
Length of reads	5S	16S	23S	5S	16S	23S
100	88.6	92.7	94.5	92.8	91.5	94.8
200	90.4	91.2	94.0	93.0	88.1	94.6
300	91.7	93.5	94.4	94.9	86.9	94.8
400	92.3	95.4	94.3	94.2	88.6	94.9
500	93.7	91.9	93.3	95.0	84.4	94.1
600	92.0	91.4	94.2	94.1	86.5	94.6
700	93.9	91.0	94.9	95.6	85.5	95.6
800	92.6	89.6	94.5	94.1	82.3	94.9

Here, hmm_fs represents our algorithm. Specificities are represented in percentage (%).

(downloaded on September 30, 2008). To reduce the impact of sequence redundancy, we removed species related to training set (see Supplementary Tables for remaining species used for evaluation). To simulate the current sequencing techniques, fragments of the lengths 100–800 bp (in intervals of 100 bp) were randomly sampled from each genome to 1 × genome coverage for each length. These fragments were used to investigate prediction performance of both our method and RNAmmer, they were also analyzed by BLASTN against 5S Ribosomal Database and SILVA database (Pruesse *et al.*, 2007) to identify rRNA genes (with *E*-value of 10^{-5} or less). In current analysis, sampling of fragments was done without considering the sequencing errors, therefore estimated performances are optimistic. The annotation information of rRNA genes was also retrieved from GenBank files. Sequence fragments that had an overlap (>40 nt) with a known rRNA gene in the same strand were considered as a positive sample. The ratios of true-positives relative to all annotated fragments (sensitivity) and to all predicted fragments (specificity) were used as a performance measure. Both exactly matching predictions and partially matching predictions with correct strand were counted as true-positives.

Tables 1 and 2 show the prediction sensitivities and specificities for all fragment lengths. The result for RNAmmer is shown in Supplementary Table S5. The sequence length of most 16S and 23S rRNA genes substantially exceeds 800 bp, therefore can not be detected by a full domain model like RNAmmer. It can be shown that our algorithm can predict sequence reads with rRNAs with a high sensitivity and specificity (>90% in almost all configurations). More important, the prediction performance does not vary much on different read lengths. One commonly used method for predicting rRNAs in metagenomic projects is based on BLAST (Altschul *et al.*, 1997, Frias-Lopez *et al.*, 2008). However, Lagesen *et al.* (2007) indicated that results based on BLAST can be problematic due to its inconsistency. Compared with BLASTN, our algorithm achieves much better sensitivities (average 10.2% improvement) while the specificities are around 2.3% less for 5S RNA. The performances for 23S rRNA are almost the same for our algorithm and BLASTN. The biggest improvement comes from 16S rRNA prediction, it demonstrates that our algorithm improves the specificities significantly and keeps the sensitivities slightly better.

The average running time of our algorithm was 744 ms per 800 bp read, and 145 ms per 200 bp read for a single 2.33G Xeon® CPU. The running time for BLASTN was 239 ms per 800 bp read, and 123 ms per 200 bp read. Additional analyses were performed on Sargasso Sea metagenomic project (Venter *et al.*, 2004) consisted of 811 372 entries totaling over 800 Mbp. On this set the search speed was 1088 s per Mbp, and our algorithm identified 660 5S, 1337 16S and 2300 23S rRNA genes or fragments of genes.

4 CONCLUSION

With the continued growth of metagenomic sequencing projects, identification of rRNA genes within sequence fragments from these projects continues to be a very important task. Here, we reported a HMM based algorithm to detect rRNA genes in short metagenomic fragments with high accuracies. Our algorithm is written in Python, and runs well on Linux/Unix and Windows XP systems with the installation of Python and HMMER package. The scripts, sample dataset and usage instruction are available online at http://tools.camera.calit2.net/camera/meta_rna as a downloadable application.

Funding: Gordon and Betty Moore Foundation (CAMERA project, http://camera.calit2.net).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Freyhult,E.K. *et al.* (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.*, **17**, 117–125.

Frias-Lopez,J. *et al.* (2008) Microbial community gene expression in ocean surface waters. *Proc. Natl Acad. Sci. USA*, **105**, 3805–3810.

Lagesen,K. *et al.* (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.

Meyer,I.M. (2007) A practical guide to the art of RNA gene prediction. *Brief. Bioinform.*, **8**, 396–414.

Raes,J. *et al.* (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr. Opin. Microbiol.*, **10**, 490–498.

Pruesse,E. *et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–7196.

Szymanski,M. *et al.* (2002) 5S Ribosomal RNA database. *Nucleic Acids Res.*, **30**, 176–178.

Tringe,S.G. and Rubin,E.M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.*, **6**, 805–814.

Venter,J.C. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.

Wuyts,J. *et al.* (2004) The European ribosomal RNA database. *Nucleic Acids Res.*, **32**, D101–D103.