BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 7 2012, pages 1033–1034
doi:10.1093/bioinformatics/bts079

Sequence analysis

Advance Access publication February 12, 2012

Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2

Martin Wu[1,*] and Alexandra J. Scott*
Department of Biology, University of Virginia, Charlottesville, VA 22904, USA

Associate Editor: David Posada

ABSTRACT

Summary: With the explosive growth of bacterial and archaeal sequence data, large-scale phylogenetic analyses present both opportunities and challenges. Here we describe AMPHORA2, an automated phylogenomic inference tool that can be used for high-throughput, high-quality genome tree reconstruction and metagenomic phylotyping. Compared with its predecessor, AMPHORA2 has several major enhancements and new functions: it has a greatly expanded phylogenetic marker database and can analyze both bacterial and archaeal sequences; it incorporates probability-based sequence alignment masks that improve the phylogenetic accuracy; it can analyze DNA as well as protein sequences and is more sensitive in marker identification; finally, it is over 100× faster in metagenomic phylotyping.

Availability: http://wolbachia.biology.virginia.edu/WuLab/Software.html.

Contact: mw4yv@virginia.edu

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

Received on November 18, 2011; revised on February 3, 2012; accepted on February 7, 2012

1 INTRODUCTION

Although the small subunit rRNA (SSU rRNA) is the 'gold standard' in microbial diversity studies, it has its limitations. For one, it has been shown that the GC compositional bias in rRNA could cause misleading phylogenetic artifacts (Hasegawa and Hashimoto, 1993). For another, rRNA makes up only a tiny fraction (~0.1%) of any given genome. Therefore, the organismal phylogeny based on rRNA alone needs to be corroborated with the analyses of other phylogenetic markers. Thirdly, there are large variations in the rRNA copy number in different bacterial and archaeal species, which prevents the use of rRNA for accurate estimation of the relative abundance of species in a population.

Protein-coding genes such as *EF-Tu*, *rpoB*, *recA* and *HSP70* have been used as alternative phylogenetic markers to complement rRNA-based analyses (Ludwig and Klenk, 2000). Because their sequences are conserved at the amino acid level, protein-coding genes offer at least two advantages over rRNA: (i) their protein sequences are less sensitive to GC compositional bias (Lockhart *et al.*, 1992; Loomis and Smith, 1990). (ii) Their fast-evolving nucleotide sequences have more power to resolve the relationships between closely related organisms. In the past, the use of protein-coding genes in phylogenetic analyses has been hindered largely by the lack of available protein sequences. Recent explosive growth of microbial genomic sequences, however, has transformed the way we study microbial diversity. For example, now many genes can be combined to make robust 'genome trees', which could serve as a framework for genome-based bacterial and archaeal classification (Klenk and Goker, 2010; Wu and Eisen, 2008). The use of dozens of protein marker genes in addition to rRNA also greatly increases the power of metagenomic phylotyping. Furthermore, many of the protein marker genes are single-copy gene in the genome. Therefore, the species composition estimated from these markers should be more accurate than that estimated based on rRNA (Venter *et al.*, 2004; Wu and Eisen, 2008).

Table 1. Feature comparison between AMPHORA and AMPHORA2.

Feature	AMPHORA	AMPHORA2
Marker genes	Bacteria only	Bacteria + archaea
Alignment masking	Manual for the seed alignments	Fully automated with Zorro
Input sequences	Protein only	Protein + DNA
Phylotyping	Parsimony only Serial, slow	Parsimony + ML Parallel, fast
Search engine	HMMER2 + BLASTP	HMMER3

The promise of phylogenomics comes with new challenges. First, we need to select a set of robust marker genes that are universally distributed and are relatively recalcitrant to lateral gene transfers. Second, we need an efficient way of identifying their orthologs (and not just homologs) from genomic-scale sequences and of rapidly aligning them with the reference sequences. Third, to make high-quality trees, we need to filter out the unreliable alignment regions prior to the phylogenetic inference. The original AMPHORA package was designed to overcome these bottlenecks for automated, high-throughput, high-quality phylogenomic analyses (Wu and Eisen, 2008). However, several issues remain (Table 1). To address these issues, we redesigned AMPHORA, made significant enhancements and incorporated new algorithms and functions. We named it AMPHORA2 and describe it below.

2 APPLICATION DESCRIPTION

2.1 Adding archaeal phylogenetic markers

AMPHORA relies on a core phylogenetic marker database to identify a set of marker genes from the input sequences. It is limited

*To whom correspondence should be addressed.

to handling bacterial sequences because it only contains 31 bacterial markers in the database (Wu and Eisen, 2008). In AMPHORA2, we greatly expand the phylogenetic database by including marker genes from the archaeal domain. To limit potential complications from paralogy and lateral gene transfers, we selected archaeal marker genes using two criteria: (i) they are 'universally' distributed. (ii) They are single-copy genes in the genome. To identify these genes, an all-against-all BLASTP search was performed between 112 064 proteins from 50 representative archaeal genomes. Proteins were then clustered using the Markov Cluster Algorithm (MCL) (*e*-value cutoff = 1e-15) (Enright *et al.*, 2002). The 108 clusters with an average of 1.00 ± 0.02 genes per organism and present in at least 45 organisms were chosen as candidate marker genes. Maximum likelihood trees were made for each marker gene and markers with non-species-specific gene duplications were removed from the list. In the end, 104 archaeal genes were selected and added to the AMPHORA2 phylogenetic marker database (Supplementary Material 1).

2.2 Automated sequence alignment filtering

Sequence alignment masking and filtering improve the accuracy of phylogenetic analysis (Eisen, 1998). One great advantage of using AMPHORA is that it provides automated high-quality alignment masking and filtering. Although the process itself is automated, the 'stencil' masks that AMPHORA uses to mask the alignments were created manually, which can be laborious and subjective. We recently developed a probability-based alignment masking program named Zorro (Wu *et al.*, 2012) that assigns a confidence score to each column in the sequence alignment. The confidence score can be used to mask and filter the alignments (e.g. by removing any column with a confidence score less than a cutoff) or weigh the columns (e.g. by using RAxML's column weight option). The large-scale, high-quality capacity of Zorro makes it practical to quickly expand the phylogenetic marker database to include hundreds of marker genes. Incorporating Zorro masks within AMPHORA2 also makes it much easier for users to add markers of their own choice and to build their personalized phylogenetic marker database. The stencil masks in the AMPHORA2 marker database were all created using the Zorro program.

2.3 Increasing the speed and accuracy of metagenomic phylotyping

Phylotyping (i.e. mapping sequence reads to taxa) is a key component of metagenomic studies. It helps researchers address the two central questions: who is there and what they are doing (e.g. by anchoring functional genes to a phylogenetic marker)? AMPHORA uses a tree-based algorithm for phylotyping. AMPHORA first inserts the query sequence into a reference tree by parsimony and then classifies it based on its position in the tree. It does this one sequence at a time. New placement algorithms make it possible to place hundreds of sequences into a reference tree simultaneously and thus have the potential to dramatically speed up the phylotyping process (Berger *et al.*, 2011; Matsen *et al.*, 2010). In addition, fast and more accurate likelihood-based placement algorithms are now available and they provide matrices to directly assess the confidence of the phylotyping (Berger *et al.*, 2011; Matsen *et al.*, 2010). AMPHORA2 takes advantage of RAxML's and pplacer's new placement algorithms and now can perform either parsimony or likelihood-based phylotyping. Our benchmark tests showed that implementing these new algorithms resulted in more than 100× speed gain in phylotyping. AMPHORA2 places sequences into the NCBI taxonomic hierarchy and assigns a confidence score at each rank of taxonomic classification (see Supplementary Material 2). The phylotype results can be imported into spreadsheet programs. Unlike its predecessor, AMPHORA2 can phylotype metagenomic sequences from a mixed population of bacteria and archaea.

2.4 Other enhancement and new functions

AMPHORA2 now supports the analyses of DNA sequences, which means that users can apply AMPHORA2 directly to metagenomic reads without the need to first annotate the sequences. AMPHORA2 will make the open reading frame (ORF) calls and identify the markers from the translated peptide sequences. Sequence similarity search using HMMER (version 2.3) is a rate-limiting step in AMPHORA. To speed things up, sequences are first searched and filtered with BLASTP to reduce the workload of HMMER. AMPHORA2 now uses HMMER3 as its sole search engine, which has a speed comparable to BLASTP but is much more sensitive and accurate in detecting homologous sequences.

3 CONCLUSIONS

The second generation of AMPHORA is a much more powerful phylogenomic analysis tool that should be useful for the study of microbial evolution and ecology in the genomic era.

Conflict of Interest: none declared.

REFERENCES

Berger,S.A. *et al.* (2011) Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biol.*, **60**, 291–302.

Eisen,J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Hasegawa,M. and Hashimoto,T. (1993) Ribosomal RNA trees misleading? *Nature*, **361**, 23.

Klenk,H.P. and Goker,M. (2010) En route to a genome-based classification of Archaea and Bacteria? *Syst. Appl. Microbiol.*, **33**, 175–182.

Lockhart,P.J. *et al.* (1992) Substitutional bias confounds inference of cyanelle origins from sequence data. *J. Mol. Evol.*, **34**, 153–162.

Loomis,W.F. and Smith,D.W. (1990) Molecular phylogeny of Dictyostelium discoideum by protein sequence comparison. *Proc. Natl Acad. Sci. USA*, **87**, 9093–9097.

Ludwig,W. and Klenk,H.-P. (2000) Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics. In Boone,D.R. *et al.* (eds) *Bergey's Manual of Systematic Bacteriology*. Springer, New York, N.Y., pp. 49–65.

Matsen,F.A. *et al.* (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, **11**, 538.

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

Wu,M. *et al.* (2012) Accounting For Alignment Uncertainty in Phylogenomics. *PLoS One*, **7**, e30288.

Wu,M. and Eisen,J.A. (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.*, **9**, R151.