Genome Identifier: A Tool for Phylogenetic Analysis of Microbial Genomes

YUKI SHIMOYAMA 1  TOKUMASA HORIJE 2,a)

Received: December 16, 2018, Accepted: February 26, 2019

Abstract: Bacterial whole-genome sequences have recently become widely available via innovative and rapid progress in technologies such as high-throughput sequencing and computing. Genomes of environmental microorganisms have also been sequenced, and their number is expected to increase in the future. Typically, phylogenetic analysis is performed after genome sequencing of such organisms. 16S rRNA is a standard locus for the phylogenetic analysis of prokaryotes. However, 16S rRNA phylogenetic trees are not always reliable because of out-paralogs and horizontal gene transfer. To overcome this problem, multiple genes (or proteins) should be employed. Therefore, we developed “Genome Identifier,” which can be used for constructing a concatenated phylogenetic tree in the form of a species tree by predicting genes from newly sequenced genomic data and collecting homologous sequences from other species.

Keywords: Genome Identifier, genome, phylogenetic tree

1. Introduction

Recent advances such as high-throughput sequencing and bioinformatics have enabled the easy retrieval of whole-genome sequences of bacteria, and about 4,000 prokaryotic genomes have been published thus far. In early genome projects, the subjects for the genome sequencing were mainly model organisms. However, many non-model organisms and phylogenetically unknown organisms are currently being sequenced, and the amount of whole-genome sequence data is expected to increase in the future. Generally, after determining the whole-genome sequence of a strain, the open reading frames (ORFs) are first estimated, and then, functional and taxonomic analyses are performed using the sequence data. There are some pipelines for newly determined genome sequences. For example, xBASE2 [1], RAST [2], and Prokka [3] are used for functional analysis, and AMPHORA2 [4] and PhyloPhAn [5] are used for taxonomic analysis.

16S rRNA and several conserved proteins are widely used for phylogenetic analyses. However, 16S rRNA phylogenetic trees are not reliable if horizontal gene transfer has occurred [6], [7], [8], [9], and some topologies of the phylogenetic trees are contradictory among the proteins resulting from horizontal gene transfer or loss of out-paralogs [10], [11], [12], [13], [14]. Ortholog-sequence-concatenated trees were developed to avoid this contradiction [15], [16], [17], and this became a popular method [17], [18], [19]. To identify the history of the lineage including sample genomes, the construction of the ortholog-sequence-concatenated tree is also more suitable than the construction of trees of individual orthologs. However, no program is available for the phylogenetic analysis of strains using estimated ORFs from the genome. There are two difficult points in automating the construction of an ortholog-sequence-concatenated tree for phylogenetic analysis after a genome sequence is newly available. The first is to choose the species used with the given strain for phylogenetic analysis because homologous sequence data from closely related species are required. Generally, the homologous sequence data for each ORF are detected by a homology search program such as BLAST+ [15]. However, the order of species in the similarity-ranking list for each homolog depends on the query sequence, and paralogs are often included. Therefore, researchers must manually choose the species by comparing the order of species in each similarity-ranking list. The second point is that the calculation time is too long if the number of species is large. To overcome these problems, we developed “Genome Identifier,” which can automatically construct an ortholog-sequence-concatenated tree for an appropriate group of species using signature genes. The appropriate group of species is selected automatically, and the 31 signature genes shared by most bacteria or the 104 signature genes shared by most archaea are detected by AMPHORA2 [4], which is included in the pipeline. Using the signature gene, the calculation time can be shortened, and phylogenetic analysis becomes possible even if the number of species is very large.

2. Methods

Genome Identifier requires BLAST+ [20], Bioperl [21], HMMER [22], Prodigal [23], AMPHORA2 [4], MAFFT [24], trimAL [25], and FastTree 2 [26]. All the required programs other than Bioperl are bundled with Genome Identifier (BLAST+ ver. 2.50, HMMER ver. 3.1b2, Prodigal ver. 2.6.3, AMPHORA2 modified version, MAFFT ver. 7.307, trimAL ver. 1.4, and Fast-
Tree 2 ver. 2.1.9). Genome Identifier operates on Linux operating systems as tested on Ubuntu 16.06 LTS and 18.04 LTS. The program and a user guide are available at https://sites.google.com/view/GenomeIdentifier.

Genome Identifier constructs a phylogenetic tree from the given genome sequence in the following steps (Fig. 1). Input files in the FASTA format, which include the nucleotide sequence of the genome, are required for the target species. From the genome sequence of the target species, protein-coding regions are predicted and translated to protein sequences by Prodigal. Marker proteins are proteins highly conserved within the lineage and encoded by a single-copy gene and are detected by AMPHORA 2. In Genome Identifier, 31 bacterial marker proteins or 104 archaeal marker proteins are used. Using the marker proteins as query sequences, homologs are detected by BLAST+. The database for BLAST+ is constructed in advance and incorporated into Genome Identifier. The threshold of the E-value for the homology search is $10^{-10}$. The default threshold values of the identity and sequence coverage are 25% and 50%, respectively. The average sequence identities to the marker protein for each species are calculated. Based on the average sequence identity, species to be used for phylogenetic analysis are automatically chosen. By default, 25 species having the highest sequence identity (%) are used. Multiple alignments for each protein are performed using MAFFT (the option is “—auto”). The alignment data of each protein are concatenated into a large alignment data set. Then, non-conserved regions are removed by trimAl. A concatenated tree is constructed by FastTree 2. The parameters for FastTree2 were set by default.

The purpose of Genome Identifier is to construct a phylogenetic tree for given species and closely related species. Ideally, the constructed phylogenetic tree should be compared with the genuine species phylogenetic tree to validate the program (or method). However, it is difficult to obtain genuine species phylogenetic trees. Therefore, we checked the taxonomic information of the given strain on the constructed phylogenetic tree and compared it with the taxonomic information in the NCBI taxonomy database. We constructed 36 phylogenetic trees for 36 prokaryotic genomes. The prokaryotes were automatically chosen (Table 1) using “choose one genome for each phylum” in the MBGD taxonomy browser [27], and unclassified species were excluded. Genome sequence data of the 36 species in FASTA format were obtained from the NCBI database. All species were hypothesized to be newly sequenced organisms, and we tested whether the taxon of species could be correctly identified using Genome Identifier. Marker proteins from each genome sequence were detected, and the phylogenetic tree was constructed using Genome Identifier. The parameters for the programs were set by default.

When a given species was located in a taxonomic group (e.g., order, family, genus), the taxon of the species was identified as the taxon of the group. We performed this procedure for identification manually. The genome data of the given species for this test were already registered in the database. Therefore, the registered data were detected as the most closely related species to the given species. The registered data were thus ignored to determine the closest related species. The identified taxa of the
Table 1  List of species and identified taxonomy. The results of identification are shown. Because phylogenetic trees were not constructed, *Nitrosopumilus maritimus SCM1* is indicated as “No tree.”

| Domain | Species | Level | Identified taxonomy |
|--------|---------|-------|---------------------|
| Bacteria | *Acidobacteria bacterium* Ellin345 | Species | Bacteroides thetaiotaomicron |
| | *Aquifex aeolicus* VF5 | Species | Borrelia burgdorferi |
| | *Bacillus subtilis* 168 | Species | Chlamydia trachomatis |
| | *Bacteroides thetaiotaomicron* VPI-5482 | Species | Dehalococcoides mccartyi |
| | *Borrelia burgdorferi* B31 | Species | Escherichia coli |
| | *Caldericium exile* AZM16e01 | Species | Fusobacterium nucleatum |
| | *Chlamydia trachomatis* D/UW-3/CX | Species | Mycobacterium tuberculosis |
| | *Chlorobium tepidum* TLS | Genus | Bacillus |
| | *Cloacamonas acidaminovorans* | Genus | Chlorobaculum |
| | *Dehalococcoides ethenogenes* 195 | Genus | Deinococcus |
| | *Deinococcus radiodurans* R1 | Genus | Dictyoglomus |
| | *Denitrovibrio acetiphilus* DSM 12809 | Genus | Gemmatimonas |
| | *Desulfitospirillum indicum* S5 | Genus | Mycoplasma |
| | *Dictyoglomus thermophilum* H-6-12 | Genus | Rhodopirellula |
| | *Escherichia coli* K-12 MG1655 | Genus | Thermaeoeventibrio |
| | *Fibrobacter succinogenes* S85 | Genus | Thermodesulfobacterium |
| | *Fimbriimonas ginsengisoli* Gsoil 348 | Genus | Thermodesulfovibrio |
| | *Fusobacterium nucleatum* ATCC 25586 | Genus | Thermotoga |
| | *Gemmatimonas aurantiaca* T-27 | Family | Aquificaeae |
| | *Ignivibacterium album* JCM 16511 | Family | Deferribacteraceae |
| | *Myocobacterium tuberculosis* H37Rv | Family | Chrysiogenaeae |
| | *Mycoplasma genitalium* G37 | Order | Acidobacteriales |
| | *Rhodopirellula baltica* SH1 | Order | Ignavibacteriales |
| | *Synechocystis* sp. PCC 6803 | Order | Verrucomicrobiales |
| | *Thermodesulfobacterium* sp. OPB45 | N/A | Unknown |
| | *Thermodesulfobacterium* yellowstonii DSM 11347 | N/A | Unknown |
| | *Thermotoga maritima* MSB8 | N/A | Unknown |
| | *Verrucomicrobium spinosum* DSM 4136 | N/A | Unknown |
| Archaea | *Aeropyrum pernix* K1 DNA | Genus | Methanocaldococcus |
| | *Halophilic archaean* DL31 | Genus | Nitrosopumilus |
| | *Korarchaeum cryptofilum* | Class | Thermoprotei |
| | *Methanocaldococcus jannaschii* DSM 2661 | Class | Halobacteria |
| | *Nanoarchaeum equitans* Kin4-M | N/A | Unknown |
| | *Nitrosopumilus maritimus* SCM1 | N/A | No tree |

given species were compared with the genuine taxon based on
the NCBI taxonomy database. The total calculation time of the
test for the 36 species was 35 min using a PC (CPU: Intel Core
i5-4670 3.40 GHz, memory: 16 GB, OS: Ubuntu 16.04 LTS).
In the case of viral genomes, signature genes cannot be de-
tected because there is no gene that is widely shared in viral
genomes. Genome Identifier estimates ORFs from the genome
using Prodigal and detects the homologous genes of each ORF by
homology search using BLAST+ (the threshold E-value is 10^{-10}).
Using the homologs, phylogenetic trees of each homolog are con-
structed. From the results of Genome Identifier, the phylogenetic
relationship of a given genome and its closely related species or
strains can be estimated.

3. Results and Discussion

Phylogenetic trees of 35 out of 36 prokaryotes were con-
structured (an example is shown in Supplemental Fig. 1); the tree of *Nanoarchaeum equitans* Kin4-M was not constructed (Table 1). *Nanoarchaeum equitans* Kin4-M is the smallest known archaeon. It lacks not only 11 genes for marker proteins but also almost all the genes required for the synthesis of amino acids, nucleotides, cofactors, and lipids. Using Genome Identifier, the phylogenetic tree of organisms that do not have all marker proteins cannot be constructed by using the remaining marker proteins, because most organisms have all the marker proteins. The taxonomy of 29 out of 35 prokaryotes was identified from the phylogenetic tree. Out of 29 prokaryotes, seven could be identified at the species level; 13 prokaryotes, at the genus level; three prokaryotes, at the family level; another three, at the order level; two, at the class level; and one, at the phylum level. However, the taxonomy of the other six prokaryotes could not be identified because they did not have closely related species. All taxa inferred from the phylogenetic tree constructed by Genome Identifier were consistent with the information in the NCBI taxonomy database, implying that the phylogenetic trees with each query prokaryote were correctly constructed using Genome Identifier. Therefore, Genome Identifier can be widely used for inferring the phylogenetic relationship of query species and their closely related species because the 36 prokaryotes were collected from different phyla.

Acknowledgments We thank Hisae Hirata and Misako Kashihara for their advice.

References

[1] Chaudhuri, R.R., Loman, N.J., Snyder, L.A., Bailey, C.M., Stelk, D.J. and Pallen, M.J.: xBASE2: A comprehensive resource for comparative bacterial genomics, *Nucleic Acids Res.*, Vol.36, pp.D535–D546 (2008).

[2] Aziz, R.K., Bartels, D., Best, A.A., Dejongh, M., Disz, T., Edwards, R.A., Glass, E.M., Kubal, M., et al.: The RAST server: Rapid annotations using subsystems technology, *BMC Genomics*, Vol.9, No.75 (2008).

[3] Seemann, T.: Prokka: Rapid prokaryotic genome annotation, *Bioinformatics*, Vol.30, pp.2068–2069 (2014).

[4] Wu, M. and Scott, A.J.: Phylogenomic analysis of bacterial and archeal sequences with AMPPHOR2, *Bioinformatics*, Vol.28, pp.1033–1034 (2012).

[5] Segata, N., Börsnigen, D., Morgan, X.C. and Huttenhower, C.: PhyloPhAn is a new method for improved phylogenetic and taxonomic placement of microbes, *Nat. Commun.*, Vol.4, 2304 (2014).

[6] Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V. and Polz, M.F.: Dissimilarity and redundancy of 16S rRNA sequences in genomes with multiple rrn operons, *J. Bacteriol.*, Vol.186, pp.2629–2635 (2004).

[7] Kitahara, K. and Miyazaki, K.: Revisiting bacterial phylogeny: Natural and experimental evidence for horizontal gene transfer of 16S rRNA, *Mol. Gen. Elements*, Vol.3, e24210 (2013).

[8] Saruhashi, S., Hamada, K., Horike, T. and Shinozawa, T.: Determination of whole prokaryotic phylogeny by the development of a random extraction method, *Gene*, Vol.392, pp.157–163 (2007).

[9] Tian, R.M., Cai, L., Zhang, W.P., Cao, H.L. and Qian, P.Y.: Rare events of intragenus and intraspecies horizontal transfer of the 16S rRNA gene, *Genome Biol. Evol.*, Vol.7, pp.2310–2320 (2015).

[10] Finch, W.M.: Homology: A personal view on some of the problems, *Trends Genet.*, Vol.16, pp.227–231 (2000).

[11] Delsuc, F., Brinkmann, H. and Philippe, H.: Phylogenomics and the reconstruction of the tree of life, *Nat. Rev. Genet.*, Vol.6, pp.361–375 (2005).

[12] Koonin, E.V.: Orthologs, paralogs, and evolutionary genomics, *Ann. Rev. Genet.*, Vol.39, pp.309–338 (2005).

[13] Snell, B., Huynen, M. and Dutilh, B.E.: Genome trees and the nature of genome evolution, *Ann. Rev. Microbiol.*, Vol.59, pp.191–209 (2005).

[14] Horike, T., Miyata, D., Hamada, K., Saruhashi, S., Shinozawa, T., Kumar, S., Chakraborty, R., Komiyama, T. and Tateeno, Y.: Phylogenetic construction of 17 bacterial phyla by new method and carefully selected orthologs, *Gene*, Vol.429, pp.59–64 (2009).

[15] Baldauf, S.L., Roger, A.J., Worch-Siebert, I. and Doolittle, W.F.: A kingdom-level phylogeny of eukaryotes based on combined protein data, *Science*, Vol.290, pp.972–977 (2000).

[16] Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E. and Stanhope, M.J.: Universal trees based on large combined protein sequence data sets, *Nat. Genet.*, Vol.28, pp.281–285 (2000).

[17] Horike, T., Minai, R., Miyata, D., Nakamura, Y. and Tateeno, Y.: Ortholog-Finder: A tool for constructing an ortholog data set, *Genome Biol. Evol.*, Vol.8, pp.446–457 (2016).

[18] Gadagkar, S.R., Rosenberg, M.S. and Kumar, S.: Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree, *J. Exp. Zool. B. Mol. Dev. Evol.*, Vol.304, No.1, pp.64–74 (2005).

[19] Thiergart, T., Landan, G. and Martin, W.F.: Concatenated alignments and the case of the disappearing tree, *BMC Evol. Biol.*, Vol.14, p.266 (2014).

[20] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L.: BLAST+: Architecture and applications, *BMC Bioinformatics*, Vol.10, p.421 (2009).

[21] Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigen, C., Fuellen, G., Gilbert, J.G., Korl, I., Lapp, H., et al.: The Bioperl toolkit: Perl modules for the life sciences, *Genome Res.*, Vol.12, pp.1611–1618 (2002).

[22] Finn, R.D., Clements, J. and Eddy, S.R.: HMMER web server: Interactive sequence similarity searching, *Nucleic Acids Res.*, Vol.39, pp.29–37 (2011).

[23] Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J.: Prodigal: A prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics*, Vol.11, p.119 (2010).

[24] Katoh, K. and Standley, D.M.: MAFFT multiple sequence alignment software version 7: Improvements in performance and usability, *Mol. Biol. Evol.*, Vol.30, pp.772–780 (2013).

[25] Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T.: trimAL: A tool for automated alignment trimming in large-scale phylogenetic analyses, *Bioinformatics*, Vol.25, pp.1972–1973 (2009).

[26] Price, M.N., Dehal, P.S. and Arkin, A.P.: FastTree 2–approximately maximum-likelihood trees for large alignments, *PLoS One 5*, e9490 (2010).

[27] Uchiyama, I., Mihara, M., Nishide, H. and Chiba, H.: MBGD update 2015: Microbial genome database for flexible orthology analysis utilizing a diverse set of genomic data, *Nucleic Acids Res.*, Vol.43, pp.D270–D276 (2015).

Yuki Shimoyma received his M.Agr. degree from Shizuoka University in 2017. He currently works at a private company.

Tokumasa Horike received his D.Eng. degree from Gunma University in 2002. He worked as a researcher (Gunma University from 2003 to 2004, National Institute of Genetics from 2004 to 2006, University of Cincinnati from 2006 to 2007, and National Institute of Genetics from 2007 to 2008). He worked at Shizuoka University from 2008 to 2013 as an assistant professor. Since 2008, he has worked at Shizuoka University as an associate professor. His research interests include evolutionary analysis of genes and genomes.

(Communicated by Yoshihiro Taguchi)