Ortholog-Finder: A tool for constructing an ortholog dataset

Genome Biology and Evolution Advance Access published January 18, 2016 doi:10.1093/gbe/evw005

Genome Biology and Evolution

SMBE

Article

Ortholog-Finder: A tool for constructing an ortholog dataset

Tokumasa Horiike[1*], Ryoichi Minai[2], Daisuke Miyata[3], Yoji Nakamura[4], and Yoshio Tateno[5]

[1]Department of Biological and Environmental Science, Shizuoka University, Shizuoka 422-8529, Japan; [2]Japanese Foundation of Cancer Research, Tokyo 135-8550, Japan; [3]Department of Economics, Chiba University of Commerce, Ichikawa 272-8512, Japan; [4]Research Center for Aquatic Genomics, National Research Institute of Fisheries Science, Fisheries Research Agency, Kanagawa 236-8648, Japan; [5]School of New Sciences, Daegu Gyoungbook Institute of Science and Technology, Daegu 711-873, Republic of Korea

*Author for Correspondence: Tokumasa Horiike, Graduate School of Agriculture, Shizuoka University, Shizuoka 422-8529, Japan; +81-54-238-3083; horiike.tokumasa@shizuoka.ac.jp

Abstract

Orthologs are widely used for phylogenetic analysis of species; however, identifying genuine orthologs among distantly related species is challenging, because genes obtained through horizontal gene transfer (HGT) and out-paralogs derived from gene duplication before speciation are often present among the predicted orthologs. We developed a program, “Ortholog-Finder,” to obtain ortholog datasets for performing phylogenetic analysis by using all ORF data of species. The program includes 5 processes for minimizing the effects of HGT and out-paralogs in phylogeny construction: (1) HGT filtering: Genes derived from HGT could be detected and deleted from the initial sequence dataset by examining their base compositions. (2) Out-paralog filtering: Out-paralogs are detected and deleted from the dataset based on sequence similarity. (3) Classification of phylogenetic trees: Phylogenetic trees generated for ortholog candidates are classified as monophyletic or polyphyletic trees. (4) Tree splitting: Polyphyletic trees are bisected to obtain monophyletic trees and remove HGT genes and out-paralogs. (5) Threshold changing: Out-paralogs are further excluded from the dataset based on the difference in the similarity scores of genuine orthologs and out-paralogs. We examined how out-paralogs and HGTs affected phylogenetic trees constructed for species based on ortholog datasets obtained by Ortholog-Finder with the use of simulation data, and we determined the effects of confounding factors. We then used Ortholog-Finder in phylogeny construction for 12 gram-positive bacteria from 2 phyla and validated each node of the constructed tree by comparison with individually constructed ortholog trees.

Key words: phylogenetic analysis, horizontal gene transfer, out-paralog, ortholog, prokaryote, eukaryote

Introduction

The construction of phylogenetic trees for species plays a key role in understanding the evolution of life. In evolutionary studies, molecular phylogenetic trees are routinely constructed because phylogenetic trees of genes or genomes are expected to reflect the phylogeny of the species in question. The sequence data of homologous genes used in constructing phylogenetic trees can be readily collected by means of BLAST searches. Although the sequence data used for phylogeny construction should comprise orthologs, homologs obtained from BLAST searches are not always orthologs but rather out-paralogs of one another: they are homologs derived from gene duplication before speciation. This phenomenon causes out-paralogs to be included in the ortholog dataset, and the generated phylogenetic trees are inconsistent with other trees (Koonin 2005; Snel et al. 2005). To avoid this inconsistency, phylogenetic trees have been constructed by concatenating all the orthologs used for phylogenetic analysis of species (Brown et al. 2001; Brochier et al. 2002; Wolf et al. 2002; Gadagkar et al. 2005; Ciccarelli et al. 2006). An alignment-concatenated tree allows phylogeny to be readily interpreted, because it is expected to reflect the true phylogenetic relationships of the species in question and avoid inconsistencies in the topology of the phylogenetic trees constructed for each gene. However, developing concatenated trees for orthologs including hidden out-paralogs is not a fundamental solution for obtaining accurate phylogenetic relationships. To construct a concatenated tree, ortholog datasets without out-paralogs must be prepared very carefully.

Currently, ortholog-prediction methods are classified as pairwise-species, multi-species graph-based, and multi-species tree-based methods (Trachana et al. 2011). For phylogenetic analysis of species, pairwise-species methods (Huynen and Bork 1998; Deluca et al. 2006; Ostlund et al. 2010) are unsuitable because >3 operational taxonomic units (OTUs) are required to construct a phylogenetic tree. The multi-species tree-based method (Li et al. 2006; Huerta-Cepas et al. 2008, Schneider et al. 2007) is also unsuitable for inferring the phylogeny of species because the method requires information from the species tree. Several ortholog databases have been developed using graph-based methods such as COG (Clusters of Orthologous Groups; Tatusov et al. 2000), MBGD (Uchiyama 2003), OrthoMCL-DB (Chen et al. 2006), orthoDB (Kriventseva et al. 2008), and eggNOG (Jensen et al. 2008). Although these graph-based methods are useful for predicting protein functions, they are unsuitable for phylogenetic analysis because they include out-paralogs (Trachana et al. 2011). By contrast, several stand-alone programs can be used for constructing an ortholog dataset. OrthoMCL (Li et al. 2003) is one of the most widely used programs of this type, and it is based on a Markov clustering algorithm (Enright et al. 2002). DomClust—a tool for generating ortholog datasets based on the UPGMA (Unweighted Pair Group Method with Arithmetic mean), which is a hierarchical clustering algorithm, and the domain structures of the protein in use (Uchiyama 2006)—is available in the MBGD database. MultiParanoid, an extended tool of InParanoid, can identify orthology relationships between proteins in >3 species (Alexeyenko et al. 2006). Gclust, another tool used for generating an ortholog dataset by means of single-linkage clustering (Sato 2009), considers the overlap score (which represents the proportion of homologous regions shared by 2 sequences) when clustering potential ortholog pairs into an ortholog group by using the specific E-value of a BLAST search. The aforementioned 4 programs are useful for constructing ortholog datasets that are primarily used for predicting protein function. Therefore, these programs do not aggressively exclude out-paralogs from the dataset, because out-paralogs do not influence the predictions of protein function.

Out-paralogs and genes obtained through HGT tend to hamper the phylogenetic analysis of species (Fitch 2000; Delsuc et al. 2005; Koonin 2005; Snel et al. 2005). To infer the relationships of 17 bacterial phyla, we previously excluded out-paralogs and HGTs from our dataset through manual curation (Horiike et al. 2009). The dataset enabled us to construct the phylogenetic tree with confidence, but the task was extremely complex and could not be applied to a larger dataset. To overcome this problem, we developed—based on considering the aforementioned 5 methods—a new program, Ortholog-Finder, for constructing ortholog datasets that contain no HGT genes or out-paralogs. We examined the effects of out-paralogs and HGTs on phylogenetic trees constructed for species based on the ortholog dataset obtained by Ortholog-Finder with the use of simulation data and determined the effects of the confounding factors (out-paralogs and HGT genes). We then used Ortholog-Finder to generate an ortholog dataset for Actinobacteria and Firmicutes that excluded confounding genes, and thus constructed a highly reliable phylogenetic tree.

Materials and Methods

Implementation

Input files in the GenBank format, which includes nucleotide and amino acid sequences, or multi-FASTA format, which includes amino acid sequences, are required for each species selected. If the input files are in the GenBank format, HGT filtering is available for inferring HGT genes and removing them from the ortholog dataset. If the input files are in multi-FASTA format, HGT filtering is not available because of the lack of nucleotide sequence data. To detect out-paralogs and HGT events, all chosen species must belong to one of 2 taxonomic groups. Two output file formats are available for ortholog sequence data: multi-FASTA format, which includes sequence data for each ortholog; and multi-FASTA-like format, which includes all ortholog sequence data in one file. Moreover, an alignment-concatenated tree can be constructed using the neighbor-joining (NJ) method. If users seek to construct a sequence-concatenated tree by using other programs that support the maximum likelihood method or Bayes method, concatenated sequence data of orthologs in the phylip format can be generated by Ortholog-Finder. After HGT filtering, the inferred HGT sequences and non-HGT sequences are saved separately, and the data can be used for other analyses. Ortholog-Finder requires BLAST (Altschul et al. 1990), ClustalW (Thompson et al. 1994), MAFFT (Katoh et al. 2002), Gblocks (Castresana 2000), Bioperl (Stajich et al. 2002), EMBOSS (Rice et al. 2000), MCL (Enright et al. 2002), OrthoMCL (Li et al. 2003), and Java Runtime Environment. Ortholog-Finder operates on Linux operating systems, as tested on Ubuntu 12.04 LTS and CentOS 6.5. The program and a user guide are available at http://www.grl.shizuoka.ac.jp/~thoriike/Ortholog-Finder.html.

Algorithm

To remove the effects of HGT genes and out-paralogs from the ortholog dataset, we added 5 processes (HGT filtering, out-paralog filtering, classification of tree data, tree splitting, and E-value changing) to the conventional graph-based method of ortholog prediction. All species selected must belong to one of the 2 taxonomic groups because the information is required for estimating out-paralogs. The 2 taxonomic groups selected must be as closely related as possible, but must not be a part of the in-group. A flowchart of the construction of ortholog datasets from sequence data is presented in fig. 1.

1. Preparation of the sequence dataset

All open reading frame (ORF) data, including amino acid and nucleotide data in GenBank format for a given species, must be prepared in a specific directory in the system. If users do not require HGT filtering (next paragraph), they can use amino acid sequence data in multi-FASTA format files instead of in a GenBank format file.

2. HGT filtering (optional; the default is disabled)

Genes derived from HGT are inferred by the program developed by Nakamura et al. (2004). The process of HGT filtering is based on Bayesian inference performed using training models for nucleotide composition. The threshold of the chi-square test is adjustable. The default threshold is $P = 0.005$. The inferred HGT sequence data are removed from the sequence dataset.

3. Finding ortholog candidates

3-1. Identifying ortholog candidates by means of single linkage (optional; alternative step: 3-2, Markov clustering algorithm implemented using OrthoMCL)

3-1-1. BLAST search

All sequences are compared with each other by performing BLAST searches to detect homologs. The queries are each sequence for all organisms, and the database contains all sequence data for each organism. The default threshold is E = 10^{-10}.

3-1-2. Out-paralog filtering

Out-paralog filtering is a simple process used for detecting and removing out-paralogs from the dataset. It is based on the E-value of the BLAST search result. If a hit sequence belongs to the same group as the query sequence and exhibits a lower similarity than do sequences belonging to another group, the hit sequence is considered an out-paralog and removed from the dataset (fig. 2).

3-1-3. Best-hit search

The sequence in a species that is most similar to a specific sequence is obtained as the best-hit sequence. All best-hit sequences between every pair of species are detected.

3-1-4. Reciprocal best-hit search

Reciprocal best-hit pairs are detected as ortholog pairs based on an E-value between every species pair.

3-1-5. Generating ortholog candidates

All ortholog pairs are clustered using a single-linkage method. The clustered data are the “ortholog candidates.”

3-1-6. Separation of ortholog candidates based on the number of paralogs

If the number of paralogs originating from one species is ≥3, Steps 3-2, 4-1, 4-2, 4-3, 4-4, and 4-5 are skipped, and the ortholog candidates are advanced to Step 5. The parameter used for the number of paralogs is adjustable.

3-2. Identifying ortholog candidates by using OrthoMCL (optional; alternative step: 3-1, single chain clustering)

This step of identifying ortholog candidates is an alternative to Step 3-1. OrthoMCL is based on the Markov clustering algorithm.

4. From ortholog candidate to ortholog

4-1. Phylogenetic tree construction for each ortholog candidate

Multiple alignments are performed for all ortholog candidates by using MAFFT (Katoh et al. 2002), and the non-conserved regions are eliminated using Gblocks (Talavera and Castresana 2007). Using the alignment data, phylogenetic trees are constructed based on the NJ method (Saitou and Nei 1987) by using ClustalW (Thompson et al. 1994).

4-2. Classification of phylogenetic trees

Ortholog candidate trees are constructed as unrooted trees, but the root is placed at the branch between 2 groups. Therefore, the trees are treated as rooted trees. Phylogenetic trees of ortholog candidates are classified as “monophyletic trees for groups” or “polyphyletic trees for groups.” Monophyletic trees are advanced to Step 4-4. Polyphyletic trees for groups contain at least 2 sequences that belong to the same group and share a polyphyletic relationship (an example is shown in fig. 1). In this case, at least one out-paralog or HGT is included in the ortholog candidate data, and the data are advanced to Step 4-3.

4-3. Tree splitting of polyphyletic trees for groups (optional; the default is enabled)

Tree splitting is a novel method used for removing out-paralogs from the ortholog candidates. The concept of tree splitting is described in fig. 3. In the tree-splitting procedure, branch-cutting is attempted in single steps—starting from the longest branch and then proceeding to the next longest uncut branch if the first attempt is unsuccessful, and so on—until a monophyletic tree is obtained. If the monophyletic tree for groups contains at least one species for each group, the tree data are advanced to Step 4-4. If a monophyletic tree cannot be obtained, the data are restored and advanced to Step 5.

4-4. Classification of polyphyletic trees for groups

These trees are classified as “monophyletic tree for species” or “polyphyletic tree for species.” The OTUs in the monophyletic tree are advanced to Step 4-6. Polyphyletic trees for species contain at least 2 sequences that are from the same species and share a polyphyletic relationship (an example is shown in fig. 1). In this case, at least one out-paralog or HGT is included in the ortholog candidate dataset, and the data are advanced to Step 4-5.

4-5. Tree splitting of polyphyletic trees for species (optional; the default is enabled)

This step is almost identical to Step 4-3. Here, branch-cutting is attempted in single steps starting from the longest branch until the tree contains at most one gene per species. If the monophyletic tree for species contains at least one species for each group, the OTUs in the monophyletic tree are advanced to Step 4-6. If a monophyletic tree cannot be obtained, the data are restored and advanced to Step 5.

4-6. Checking for species number

Users can set the minimum number of species for each group in an ortholog dataset. The default value is 2. If the number of species for a group is less than the minimum number, the data are not added to the ortholog dataset and are advanced to Step 5. If the number of species is ≥2, the OTUs in the monophyletic tree are judged to be orthologs and are added to the ortholog dataset.

5. Threshold changing

All data are returned to Step 3-1-3 (best-hit search) and the threshold is changed. This process is cycled with the threshold (BLAST E-value) being changed from 10^{-10} to 10^{-100} at a 10^{-10} decrement. In theory, the threshold for homology must be set appropriately to detect orthologs, because the sequence similarities of each protein family depend on evolutionary rates. Therefore, at the first cycle, Ortholog-Finder uses a loose threshold (E-value = 10^{-10}) for proteins that evolve rapidly. At the next cycle, a stricter threshold than the first one is used and orthologs are detected in protein families that evolve slowly. The concept of threshold changing is depicted in fig. 4. Users can set the thresholds.

6. Phylogenetic tree construction

Ortholog-Finder can construct a concatenated tree by using the ortholog dataset and the NJ method. Multiple alignments are performed for each ortholog by using MAFFT, and the non-conserved regions in the alignment are eliminated using Gblocks. The alignment data are concatenated into one alignment dataset. The orthologs do not always include all species. If certain multiple-alignment datasets lack several species, gaps are inserted at the site of the missing data in the concatenated multiple alignment. Using a large amount of alignment data, a concatenated tree is constructed using ClustalW, based on the NJ method. Phylogenetic trees for each ortholog are constructed using the NJ method in order to calculate the branch-support percentage, which reflects the percentage of ortholog trees that exhibits the same branching pattern as the concatenated tree, and this is calculated at each internal branch. If a branch of a concatenated tree is shared with an internal branch of several orthologs, the branch-support percentage will be high (i.e., the reliability of the branch will be high). Furthermore, the bootstrap test is performed to illustrate the reliability of each internal branch of the concatenated tree. Ortholog-Finder does not support the maximum likelihood method or Bayes method because the calculation times required for choosing the optimal substitution model and constructing phylogenetic trees are extremely long. For users who wish to use these methods, Ortholog-Finder provides concatenated sequence data of orthologs in the phylip format and the data can be used in another tree-building program.

Simulations for gene loss after gene duplication

To evaluate the quality of the ortholog data obtained using Ortholog-Finder and the phylogenetic tree constructed based on the ortholog data, we tested them by using simulated data under several conditions. We began by testing the effect of out-paralogs on the generation of ortholog data and the phylogenetic analysis, because out-paralogs hamper the estimation of species phylogeny (Fitch 2000; Delsuc et al. 2005; Koonin 2005; Snel et al. 2005). We performed the simulation test by using “Ortholog-Finder with single linkage” (OF-S; *Algorithm* 3-1), “Ortholog-Finder with OrthoMCL” (OF-M; *Algorithm* 3-2), and OrthoMCL alone (OM), as discussed in the following paragraph.

First, we generated a tree featuring 32 OTUs as the model tree. The lengths of branches between neighboring OTUs or nodes were 0.04. To generate simulation data containing out-paralogs, the model tree was duplicated to simulate the gene-duplication event of the common ancestor, and this yielded 64 OTUs. Branch lengths between the roots of 2 duplicates were set to 0.02, 0.04, or 0.08, and each condition was tested. The simulated amino acid sequences, which contained 200 amino acids corresponding to each OTU in the duplicated tree, were generated by using indel-Seq-Gen Ver. 2.01.03 (Strope et al. 2007) with default parameters. To simulate gene losses, certain percentages of genes among the ortholog candidates were randomly deleted (30%–85% deleted, at 5% increments). The sequence data of 100 homologs were generated by the simulations, after which we removed the information on individual homologs by sorting sequences according to species when testing the performance of Ortholog-Finder in the construction of ortholog datasets. Using Ortholog-Finder, ortholog groups were identified and the ortholog-sequence-concatenated trees were constructed using the NJ method. All parameters used with Ortholog-Finder were the defaults (default parameters in *Algorithm*). One hundred concatenated trees were constructed by running 100 trials, and, using the NJ method, phylogenetic trees were also constructed for each of the orthologs that were generated by the 100 trials. To evaluate the quality of the ortholog dataset generated by Ortholog-Finder, we calculated the accuracy rates of the concatenated trees (percentage of ortholog-sequence-concatenated trees exhibiting the same topology as the model tree) and each ortholog tree (percentage of ortholog trees exhibiting the same topology as the model tree) at each tested branch length and gene-loss percentage.

We calculated the accuracy rates for the simulated data generated not only by OF-S (as noted in the previous paragraph), but also by OF-M, which uses an external OrthoMCL program to cluster amino acid sequences as candidates of orthologs based on a Markov clustering algorithm. To compare the quality of the ortholog data generated by Ortholog-Finder with that of the data generated by OrthoMCL, we calculated the accuracy rates of the ortholog-sequence-concatenated trees and those of the ortholog trees constructed using the ortholog data generated by OrthoMCL alone (i.e., by OM). The BLAST threshold used in OrthoMCL was E = 10^{-10}.

Simulations for horizontal gene transfer

We tested the influence of HGT on ortholog data and the phylogenetic analysis by using simulated sequence data. The model tree was a simple bifurcated tree featuring 32 OTUs. The branch lengths between neighboring OTUs or nodes were 0.04. We simulated HGT events on the model tree at random lineages to generate “horizontally gene-transferred trees” or “HGT trees” (Horiike et al., 2011). We assumed that HGT occurrence follows Poisson distribution, and HGT trees were constructed using the parameter mean HGT frequency (0.0–5.0; Results and Discussion). Amino acid sequence data based on the HGT trees were generated using indel-Seq-Gen. The original sequence contained 200 amino acids and all other parameters were defaults. After the simulated ortholog sequence data were generated, we removed the information on individual orthologs by sorting the sequences according to species when testing the performance of Ortholog-Finder in identifying orthologs. Using Ortholog-Finder, ortholog groups were identified once again and the ortholog-sequence-concatenated trees were constructed using the

NJ method. All parameters of Ortholog-Finder were the defaults. Phylogenetic trees for each ortholog were also constructed using the NJ method. To evaluate the quality of the ortholog datasets obtained by Ortholog-Finder, the accuracy rates of concatenated trees and each ortholog tree were calculated at each HGT frequency.

We calculated the accuracy rates for the simulated data generated by not only OF-S but also OF-M and OM; the calculation was exactly as described in the last paragraph of the previous subsection.

Simulations for long evolutionary distance

We tested Ortholog-Finder by using sequences that were increasingly more divergent (branch lengths were increased until the sequences were in saturation). The model tree was a simple bifurcated tree featuring 32 OTUs. The branch lengths between neighboring OTUs or nodes were set to 0.02, 0.04, 0.06, 0.08, 0.10, 0.12, 0.14, 0.16, 0.18, 0.20, 0.22, or 0.24. Amino acid sequence data based on the trees were generated using indel-Seq-Gen. The processes of the simulation test after sequence generation were same as those of “Simulations for horizontal gene transfer.”

Phylogenetic analysis for 12 gram-positive bacteria

We used Ortholog-Finder to analyze the phylogeny of 12 gram-positive bacteria for which genome data were available (table 1). The 12 bacteria belong to 2 phyla, Actinobacteria (5 species from distinct families) and Firmicutes (7 species from distinct families), and therefore these could be used as out-groups for each other. The protein sequences, translated from predicted ORFs, were obtained from the NCBI site ftp://ftp.ncbi.nlm.nih.gov. Ortholog-Finder parameters were set to the defaults, with the exception being that HGT filtering was enabled. The ortholog-sequence-concatenated tree was constructed according to the NJ method by using Ortholog-Finder. Bootstrap values and branch-support percentages were calculated to examine the reliability of the internal branches of the concatenated tree.

The ortholog-sequence-concatenated tree created based on the maximum likelihood method was constructed using RAxML Version 8.0.20 (Stamatakis et al. 2005) in order to compare the phylogenetic tree shape with the shape of the tree generated using the NJ method. The substitution model and the parameters were estimated automatically by RAxML.

Results and Discussion

Simulations for gene loss after gene duplication

To test the performance of Ortholog-Finder against out-paralogs, we calculated the accuracy rate of ortholog-sequence-concatenated trees, which revealed the percentage of concatenated trees that exhibited the accurate topology. Because we assumed that out-paralogs in ortholog data are generated by gene loss after gene duplication, the model trees were duplicated in the common ancestor, and certain genes were randomly deleted from the ortholog candidates to simulate gene-loss events. Simulation tests were performed using OF-S (Ortholog-Finder with single linkage), OF-M (Ortholog-Finder with OrthoMCL), and OM (OrthoMCL alone). OrthoMCL is a program used for constructing ortholog datasets (Introduction). In OF-M, Ortholog-Finder uses OrthoMCL to cluster sequences by employing a Markov clustering algorithm (Materials and Methods). The accuracy rate allows investigators to examine the influence of out-paralogs generated as a result of gene loss after gene duplication (fig. 5). When the branch length between the 2 duplicated trees was 0.02 (fig. 5*A*), most trees were accurately reproduced when gene-loss percentage was <40%. However, the accuracy of trees generated by OM gradually decreased when the gene-loss percentage was >40%. The accuracies of the remaining trees (OF-S, OF-M) decreased sharply starting from a gene-loss percentage of 70%. When the branch length between the 2 duplicated trees was 0.04 (fig. 5*B*), most of the trees were accurately reproduced when the gene-loss percentage was <45%. The accuracy of trees generated by OM decreased starting from 45% gene loss, whereas the accuracies of the other trees (OF-S, OF-M) decreased sharply starting from 75% gene loss. At 80% gene loss, the accuracy rate of the tree generated based on the Markov clustering algorithm (OF-M) was greater than that of the tree generated by means of single linkage (OF-S). When the branch length between the roots of 2 duplicated trees was 0.08

(fig. 5*C*), most of the trees were accurately reproduced when the gene-loss percentage was <45%. The accuracy of the trees generated by OM and OF-S declined rapidly starting from 45% and 80% gene loss, respectively, but that of the tree generated by OF-M did not decrease markedly. When the gene-loss percentage was 85%, the accuracy rates of all trees were 0%. We also calculated the accuracy rate of the phylogenetic tree of each ortholog, which showed the percentage of ortholog trees that exhibited the accurate topology for each tested branch length between the roots: 0.02, 0.04, and 0.08 (fig. 5*D*, *E*, and *F*). The accuracy rates measured for OM decreased sharply at 60% or 65% gene loss, but those measured for OF-S or OF-M gradually decreased when the branch lengths were 0.02, 0.04, and 0.08. These results reveal 5 key points: (1) When the gene-loss percentage was <45%, the model tree was accurately reconstructed under all conditions. (2) When the gene-loss percentage was >45%, the accuracy rates of the tree generated by Ortholog-Finder were greater than those of the tree generated by OrthoMCL alone. (3) When the gene-loss percentage was 75% or 80%, accuracy decreased drastically. (4) When the branch length between the duplicated trees was increased, the accuracy rates also increased at the same gene-loss percentage. (5) When gene-loss percentage was <60%, most of the ortholog trees generated by OF-S, OF-M, and OM were consistent with the model tree; however, when the gene-loss percentage was >60%, the accuracy rate of only the ortholog tree generated by OM decreased sharply.

Simulations for horizontal gene transfer

To test the performance of Ortholog-Finder against HGT, we calculated the accuracy rate (%) by conducting computer simulations. A bifurcated tree featuring 32 OTUs was used as a model tree. Because we assumed that HGT occurrence follows a Poisson distribution, HGT trees were constructed using the parameter of mean HGT frequency (0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, and 5.0). Amino acid sequence data corresponding to the HGT tree were generated using indel-Seq-Gen. When the mean HGT frequency was ≤1.5, the model tree was perfectly reproduced by OF-S, OF-M, and OM (fig. 6). When the mean HGT frequency was ≤3, the model tree was mostly reproduced, with the exception being the case of the ortholog data generated by OM (fig. 6). When the mean HGT frequency was >3, the accuracy rates of the trees generated by OF-S and OF-M decreased with an increase in HGT frequency. When the mean HGT frequency was 5, an accurate tree could not be readily constructed by OF-S, OF-M, or OM. These results suggest that Ortholog-Finder (OF-M and OF-S) can tolerate HGT events, because most of the phylogenetic trees were accurately reproduced even when the mean HGT frequency was 3. Thus, even if HGT events occur >3 times, HGT-filter (Materials and Methods) removes some of the HGT genes and reduces the effect of HGT.

We calculated the accuracy rates of the phylogenetic trees constructed for each ortholog. When mean HGT frequency was ≤0.5, the accuracy rates of ortholog trees generated by OF-S, OF-M, and OM were all approximately 80%. The accuracy rate of ortholog trees generated by OF-S stayed around 80%, but that of the trees generated by OF-M decreased to around 65%; the accuracy rate of the ortholog tree generated by OM decreased to <10%.

Simulations for long evolutionary distance

To test the performance of Ortholog-Finder in a near-saturation scenario, we calculated the accuracy rate (%) by conducting computer simulations. The accuracy rates for the tree featuring distinct branch lengths (0.02–0.24, 0.02 increment) are shown in fig. 7. When the branch length ranged from 0.02 to 0.18, the phylogenetic trees were accurately generated by OF-S, OF-M, and

OM. The accuracy rate of the tree generated by OF-M decreased starting from a branch length of 0.18, and from a branch length of 0.20, the accuracies of all trees (OF-S, OF-M, OM) decreased sharply. When the branch length was 0.24, the accuracy rates of all trees were 0%. The effects of saturation on OF-S, OF-M, and OM were almost identical. Thus, the genuine tree cannot be readily reproduced by constructing concatenated trees by using orthologs.

Phylogenetic analysis of gram-positive bacteria

We used the sequence data of 12 gram-positive bacteria (table 1) and created the ortholog dataset by using OF-M and constructed the sequence-concatenated tree by using the NJ method. We first obtained 843 sets of ortholog candidates after implementation of the Markov clustering algorithm by OrthoMCL (see *Algorithm* and fig. 1). When the threshold was $E = 10^{-10}$, 425 trees were identified as monophyletic trees (table 2). After tree spitting, 188 polyphyletic trees became monophyletic trees and they were joined as orthologs. In total, 613 trees were obtained as orthologs when the threshold was reached at the first cycle ($E = 10^{-10}$). Furthermore, 65 orthologs were obtained through threshold changing, and 56 orthologs were generated by applying the threshold-changing and tree-splitting operations. In the end, 734 orthologs were generated by Ortholog-Finder.

The 734 orthologs obtained were used for constructing the concatenated tree for Actinobacteria and Firmicutes (fig. 8*A*). The bootstrap values on each internal branch were 100%. This result indicates that the bootstrap test may not be adequate for testing the reliability of each internal branch of the concatenated tree because the various reliabilities of each internal branch were not sufficiently estimated (Salichos and Rokas 2013). The “branch-support percentage” independently reflects the number of orthologs that support the branching pattern for each internal branch on the concatenated tree. The branch-support percentages indicated that the support percentage of short branches in the Firmicutes clade tended to be low. Therefore, we caution users that short branches in a tree indicate low reliability, and thus biological inferences made from such a tree might be suspect.

Similarly, 146 orthologs were obtained by OF-S, and the concatenated phylogenetic tree was constructed using the NJ method (fig. 8*B*); the shape of this tree is identical to that of the tree generated by OF-M. The number of orthologs obtained using OF-M (734) and OF-S (146) differed substantially because of the difference in the algorithms used for finding orthologs. In OF-S, we selected reciprocal best hit for single linkage in order to avoid false-positive orthologs in our phylogenetic analysis. However, the generated phylogenetic trees show that the orthologs obtained using OrthoMCL (OF-M) are also sufficient for phylogenetic analysis.

We constructed sequence-concatenated trees by using the maximum likelihood method; we used RAxML Version 8.0.20 and the ortholog datasets generated by OF-M and OF-S (fig. 8*C* and *D*). The maximum likelihood tree constructed using the ortholog data generated by OF-M was identical in topology to the 2 NJ concatenated trees generated by OF-M and OF-S. The topologies of the 3 concatenated trees are more reliable than the topology of the maximum likelihood tree generated by OF-S. A part of the maximum likelihood tree generated by OF-S differed slightly from the other trees, but most of the branching patterns in the tree were the same as those in the other concatenated trees.

Removing out-paralog and HGT genes

The existence of in-paralogs in an ortholog dataset does not generally affect the estimation of the species phylogenetic tree. Therefore, if >1 in-paralog is included among the ortholog candidates, one of the in-paralogs is randomly selected to obtain the ortholog dataset. Out-paralogs generated before the divergence of 2 groups are removed by means of out-paralog filtering, tree splitting, and threshold changing. Conversely, out-paralogs generated after the divergence of 2 groups are not removed by out-paralog filtering and threshold changing, but tree splitting is useful for removing the out-paralogs. By using HGT filtering, HGTs within and between groups are removed from the dataset, and some of the remaining HGTs are removed through tree splitting. If the timespan between gene duplication and speciation is extremely short, identifying in/out-paralogs is challenging. However, the bootstrap value or the branch-support percentage of internal branches of the concatenated tree indirectly shows the quality of an ortholog dataset.

Limitations of Ortholog-Finder

Certain limitations of Ortholog-Finder are the following. (1) To obtain orthologs, Ortholog-Finder requires 2 known groups of species, because the program identifies and removes out-paralogs based on the phylogenetic tree constructed with a root. Therefore, if a user plans to use only one group of species, another group of species must be added as an out-group. (2) Most of the run time of Ortholog-Finder is related to the HGT filter, which requires a long calculation time. If a user selects an extremely large amount of species data to use with Ortholog-Finder, a sequence dataset without HGT must be prepared for each species. Users can remove HGT genes from the sequence dataset generated by Ortholog-Finder by selecting the option “HGT filtering only” in advance. (3) NJ trees are recognized to feature long branches and suffer from long-branch attraction. Therefore, certain orthologs affected by such trees might be included in the ortholog database. (4) We were able to use Ortholog-Finder and obtain ortholog data for eukaryotes, but an extremely long time was required for completing the process (e.g., 2 weeks for 12 species); the results obtained for mammals are presented in Supplementary Data. We believe that our program can be used for eukaryotes, but the number of species that can be included is limited.

Conclusion

In this study, we developed Ortholog-Finder, a program for constructing ortholog datasets for phylogenetic analysis. To eliminate the effects of HGT genes and out-paralogs from the ortholog dataset, this program includes 5 processes: HGT filtering, out-paralog filtering, classification of tree data, tree splitting, and E-value changing. We examined the effects of out-paralogs and HGTs on phylogenetic trees constructed for species based on the ortholog dataset obtained by Ortholog-Finder with the use of simulation data, and we determined the effects of confounding factors. Our results suggest that Ortholog-Finder can tolerate gene loss after gene duplication and HGT events, because most of the phylogenetic trees were accurately reproduced even when these events occurred. Lastly, we used Ortholog-Finder to generate an ortholog dataset for 12 gram-positive bacteria (Actinobacteria and Firmicutes) and validated each node of the constructed tree by comparison with individually constructed ortholog trees.

Acknowledgments

This work was supported by the Japan Society for the Promotion of Science, Grant Number 22710184.

References

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL. 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22:e9–e15.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Brochier C, Bapteste E, Moreira D, Philippe H. 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends Genet. 18:1–5.

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. 2001. Universal trees based on large combined protein sequence data sets. Nat Genet. 28:281–285.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 34:D363–D368.

Ciccarelli FD, et al. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361–375.

Deluca TF, et al. 2006. Roundup: a multi-genome repository of orthologs and evolutionary distances. Bioinformatics 22:2044–2046.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.

Fitch WM. 2000. Homology a personal view on some of the problems. Trends Genet. 16:227–231.

Gadagkar SR, Rosenberg MS, Kumar S. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J Exp Zool B Mol Dev Evol. 304:64–74.

Horiike T, et al. 2009. Phylogenetic construction of 17 bacterial phyla by new method and carefully selected orthologs. Gene 429:59–64.

Horiike T, Miyata D, Tateno Y, Minai R. 2011. HGT-Gen: a tool for generating a phylogenetic tree with horizontal gene transfer. Bioinformation 7:211–213.

Huerta-Cepas J, Bueno A, Dopazo J, Gabaldón T. 2008. PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Res. 36:D491–D496.

Huynen MA, Bork P. 1998. Measuring genome evolution. Proc Natl Acad Sci U S A. 95:5849–5856.

Jensen LJ, et al. 2008. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 36:D250–D254.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 39:309–338.

Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM. 2008. OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res. 36:D271–D275.

Li H, et al. 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. 34:D572–D580.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet. 36:760–766.

Ostlund G, et al. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38:D196–D203.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276–277.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Sato N. 2009. Gclust: trans-kingdom classification of proteins using automatic individual threshold setting. Bioinformatics 25:599–605.

Schneider A, Dessimoz C, Gonnet GH. 2007. OMA Browser--exploring orthologous relations across 352 complete genomes. Bioinformatics 23:2180–2182.

Snel B, Huynen M, Dutilh BE. 2005. Genome trees and the nature of genome evolution. Annu Rev Microbiol. 59:191–209.

Stajich JE, et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12:1611–1618.

Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 15:456–463.

Strope CL, Scott SD, Moriyama EN. 2007. indel-Seq-Gen: a new protein family simulator incorporating domains, motifs, and indels. Mol Biol Evol. 24:640–649.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 56:564–577.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28:33–36.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Trachana K, et al. 2011. Orthology prediction methods: a quality assessment using curated protein families. Bioessays 33:769–780.

Uchiyama I. 2003. MBGD: microbial genome database for comparative analysis. Nucleic Acids Res. 31:58–62.

Uchiyama I. 2006. Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. Nucleic Acids Res. 34:647–658.

Wolf YI, Rogozin IB, Grishin NV, Koonin EV. 2002. Genome trees and the tree of life. Trends Genet. 18:472–429.

Figure Legends

FIG. 1. From ortholog candidate to ortholog: a flowchart.

FIG. 2. Out-paralog filtering.

An example of the BLAST search is shown. The top 3 hit sequences are from Actinobacteria, which is the group that also contains the query sequence. The next 2 sequences are from Firmicutes. The 6th hit sequence is from Actinobacteria again, but the similarity is lower than that of sequences from the other group. In this case, the 6th sequence was considered an out-paralog and removed from the dataset.

FIG. 3. Example of tree splitting for gene loss after gene duplication.

The tree on the left is a duplicated phylogenetic tree from the common ancestor. This tree contains 2 lineages (a group containing Species 1–8 and one containing Species 9 and 10). After gene duplication, the alpha and beta genes appeared; these genes diverged because of species divergence. Subsequently, gene losses occurred in the lineages that are displayed using a shadow. The tree on the right was constructed using the existing genes. Information on the paralogs alpha and beta was lost (shown in the black bar), and thus the paralogs were not recognized. Here, the topology of the tree on the right is not consistent with the species tree indicated with a black frame on the tree on the left. Therefore, the tree on the right should not be used as an ortholog tree. However, in the case where information was available on the position of out-groups (Species 9 and 10 form an out-group of the Species 1–8 group), the gene-duplication point shown in the closed square can be inferred. When the tree on the right is bisected at the duplication point, 2 trees are obtained, and because one of these trees now contains few species, this segment is eliminated. The topology of the other tree, displayed in a gray frame, is consistent with a species tree. Consequently, the OTUs in the tree shown in the gray frame are available as orthologs.

FIG. 4. Example of threshold changing.

These graphs show the linkage of proteins based on similarity. The *sp* number represents each species, and the *seq* number represents each protein. Broad and narrow arrows indicate high and low similarities, respectively. In cases of E = 10^{-10}, an ortholog cannot be determined, because out-paralogs link to the genuine ortholog as a part of the ortholog candidate (*A*). When the threshold becomes more strict (E = 10^{-20}), the linkage between genuine orthologs and out-paralogs breaks, and the genuine orthologs are detected (*B*). In this case, the proteins of sp1-seq2, sp2-seq2, and sp3-seq2 cannot be detected as orthologs because the number of species is insufficient (see Materials and Methods).

FIG. 5. Accuracy rates of the phylogenetic tree constructed for out-paralog simulation data.

Accuracy rates of the phylogenetic trees were calculated at various gene-loss percentages. OF-S (black circles): Ortholog-Finder with single linkage; OF-M (gray circles): Ortholog-Finder with OrthoMCL; and OM (gray squares): OrthoMCL. Accuracy rates of the sequence-concatenated tree when the branch lengths between the roots of 2 duplicates were 0.02 (*A*), 0.04 (*B*), and 0.08 (*C*). Accuracy rates of the tree of individual orthologs when the branch lengths between the roots of 2 duplicates were 0.02 (*D*), 0.04 (*E*), and 0.08 (*F*).

FIG. 6. Accuracy rates of the phylogenetic tree constructed for the HGT simulation data.

Accuracy rates of the phylogenetic trees were calculated at various HGT frequencies. OF-S (black circles): Ortholog-Finder with single linkage; OF-M (gray circles): Ortholog-Finder with OrthoMCL; OM (gray squares): OrthoMCL.

FIG. 7. Accuracy rates of the phylogenetic tree constructed for the evolutionary-distance simulation data.

Accuracy rates of the phylogenetic trees were calculated at various evolutionary distances. OF-S (black circles): Ortholog-Finder with single linkage; OF-M (gray circles): Ortholog-Finder with OrthoMCL; and OM (gray squares): OrthoMCL.

FIG. 8. Phylogenetic tree constructed for 12 gram-positive bacteria

A phylogenetic tree for 12 gram-positive bacteria was constructed based on the NJ method and using the ortholog dataset generated by OF-M (*A*) and OF-S (*B*). The bootstrap values and the branch-support percentage for each internal branch are displayed on left and right, respectively. A phylogenetic tree for the 12 gram-positive bacteria was constructed based on the maximum likelihood method and using the ortholog dataset generated by OF-M (*C*) and OF-S (*D*).

Table 1. List of organisms.

Phylum	Species
Actinobacteria	*Acidimicrobium ferrooxidans* (DSM 10331)
	Bifidobacterium longum (NCC2705)
	Cryptobacterium curtum (DSM 15641)
	Mycobacterium tuberculosis (H37Rv)
	Rubrobacter xylanophilus (DSM 9941)
Firmicutes	*Bacillus subtilis* (168)
	Clostridium acetobutylicum (ATCC 824)
	Halothermothrix orenii (H 168)
	Natranaerobius thermophilus (JW/NM-WN-LF)
	Staphylococcus aureus (N315)
	Streptococcus pyogenes (SF370)
	Thermoanaerobacter tengcongensis (MB4T)

Table 2. Number of phylogenetic trees and orthologs generated by OF-M under each tested threshold.

E-value	Monophyletic trees	Monophyletic trees after tree splitting	orthologs
10^{-10}	425	188	613
10^{-20}	25	22	47
10^{-30}	13	10	23
10^{-40}	9	9	18
10^{-50}	8	8	16
10^{-60}	4	3	7
10^{-70}	3	1	4
10^{-80}	1	2	3
10^{-90}	1	0	1
10^{-100}	1	1	2
Total	490	244	734

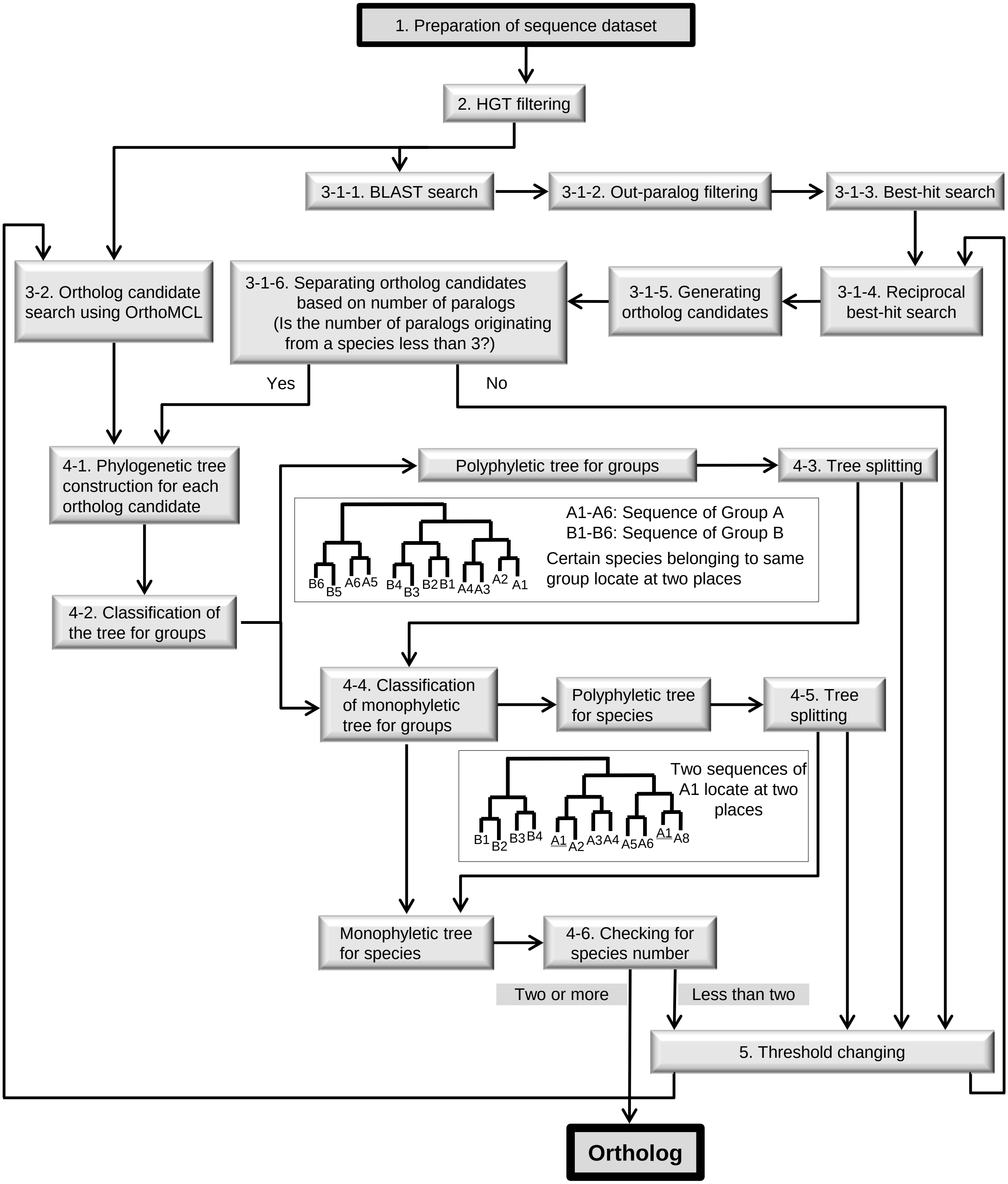
1. Preparation of sequence dataset
2. HGT filtering
3-1-1. BLAST search
3-1-2. Out-paralog filtering
3-1-3. Best-hit search
3-2. Ortholog candidate search using OrthoMCL
3-1-6. Separating ortholog candidates based on number of paralogs (Is the number of paralogs originating from a species less than 3?)
3-1-5. Generating ortholog candidates
3-1-4. Reciprocal best-hit search
Yes
No
4-1. Phylogenetic tree construction for each ortholog candidate
Polyphyletic tree for groups
4-3. Tree splitting
B6 B5 A6 A5 B4 B3 B2 B1 A4 A3 A2 A1
A1-A6: Sequence of Group A
B1-B6: Sequence of Group B
Certain species belonging to same group locate at two places
4-2. Classification of the tree for groups
4-4. Classification of monophyletic tree for groups
Polyphyletic tree for species
4-5. Tree splitting
B1 B2 B3 B4 A1 A2 A3 A4 A5 A6 A1 A8
Two sequences of A1 locate at two places
Monophyletic tree for species
4-6. Checking for species number
Two or more
Less than two
5. Threshold changing
Ortholog

FIG. 1.

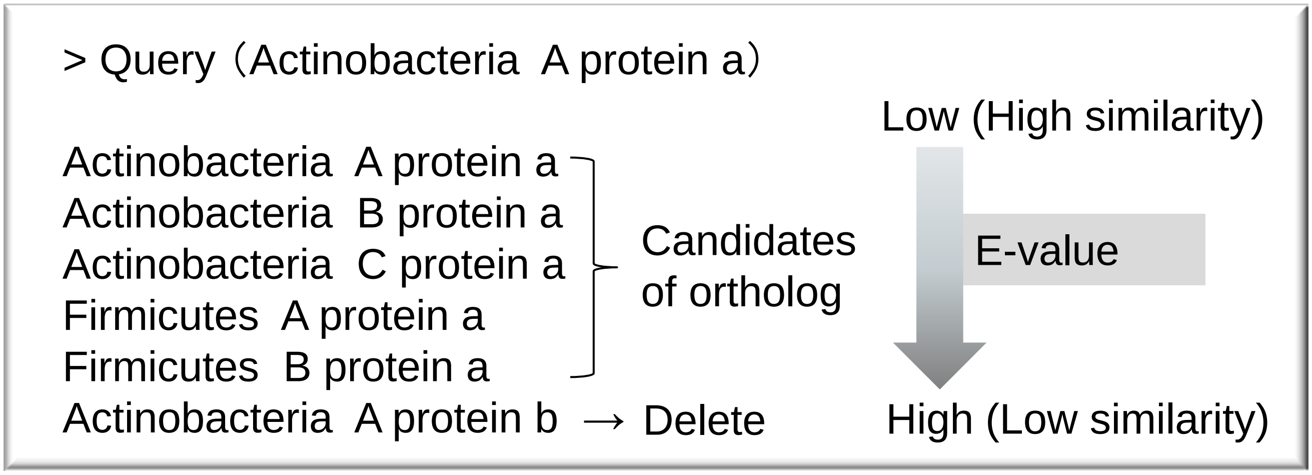
> Query（Actinobacteria A protein a）
Low (High similarity)
Actinobacteria A protein a
Actinobacteria B protein a
Actinobacteria C protein a
Firmicutes A protein a
Firmicutes B protein a
Candidates of ortholog
E-value
Actinobacteria A protein b → Delete
High (Low similarity)

FIG. 2.

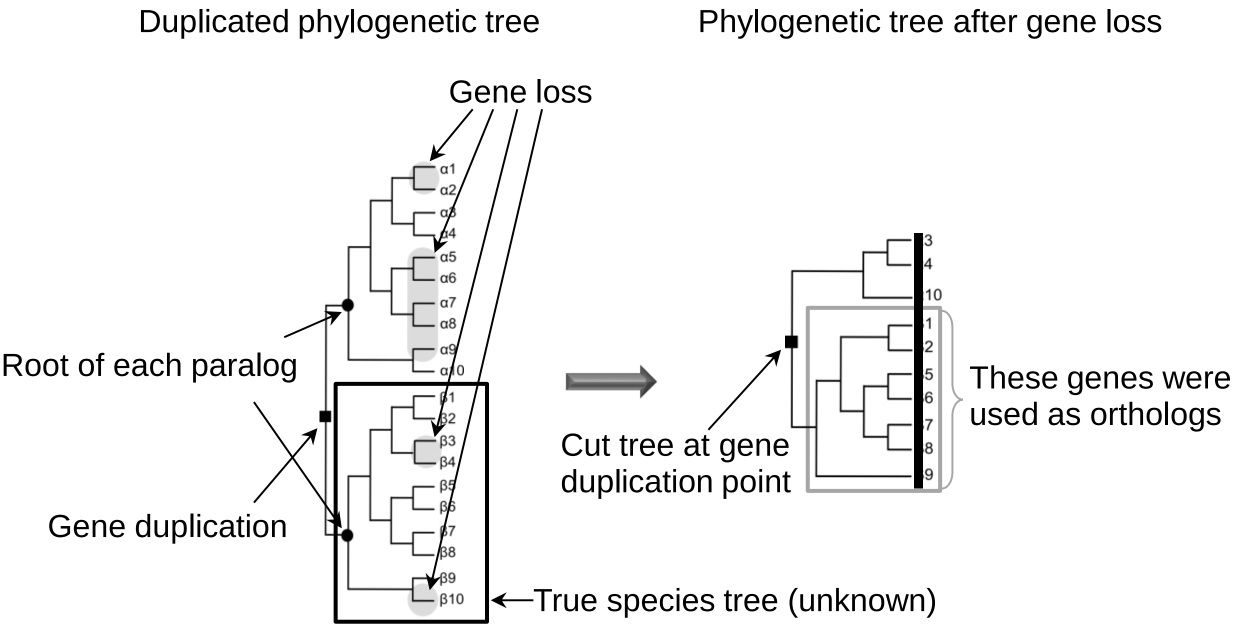
Duplicated phylogenetic tree
Phylogenetic tree after gene loss
Gene loss
α1
α2
α3
α4
α5
α6
α7
α8
α9
α10
β1
β2
β3
β4
β5
β6
β7
β8
β9
β10
Root of each paralog
Gene duplication
True species tree (unknown)
Cut tree at gene duplication point
3
4
10
1
2
5
6
7
8
9
These genes were used as orthologs

FIG. 3.

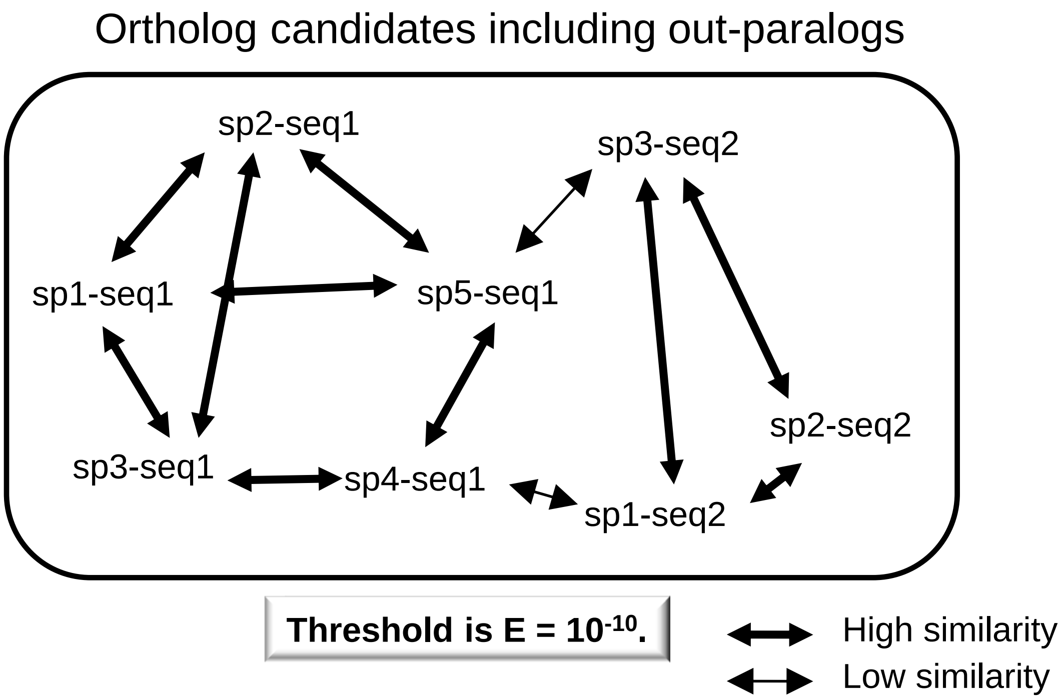
Ortholog candidates including out-paralogs
sp2-seq1
sp3-seq2
sp1-seq1
sp5-seq1
sp2-seq2
sp3-seq1
sp4-seq1
sp1-seq2
Threshold is E = 10-10.
High similarity
Low similarity

FIG. 4*A*.

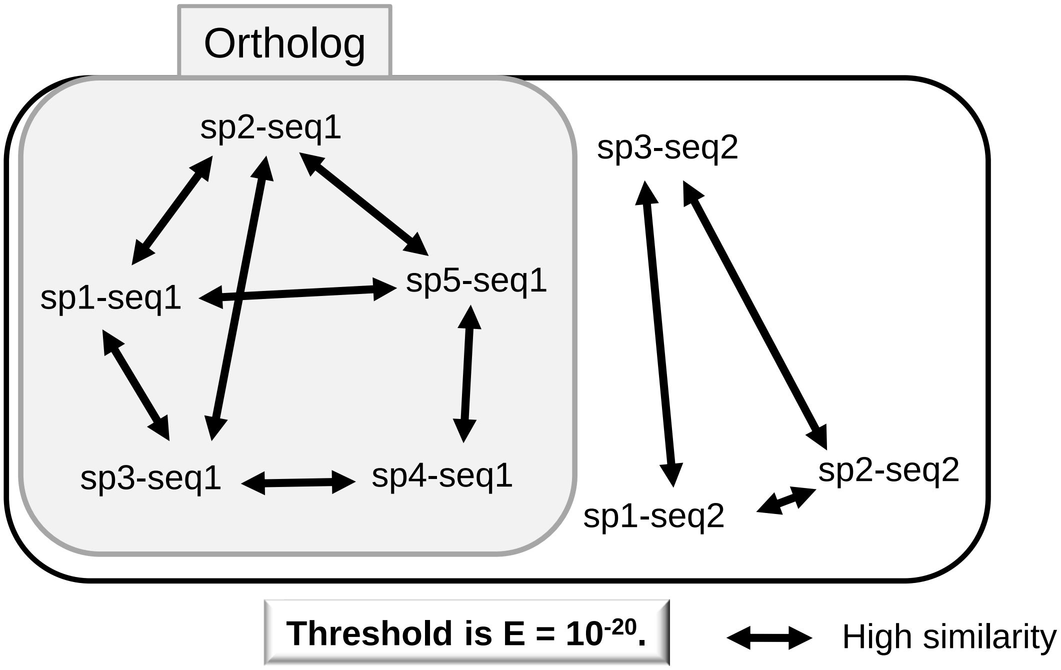
Ortholog
sp2-seq1
sp3-seq2
sp1-seq1
sp5-seq1
sp3-seq1
sp4-seq1
sp2-seq2
sp1-seq2
Threshold is E = 10-20.
High similarity

FIG. 4*B*.

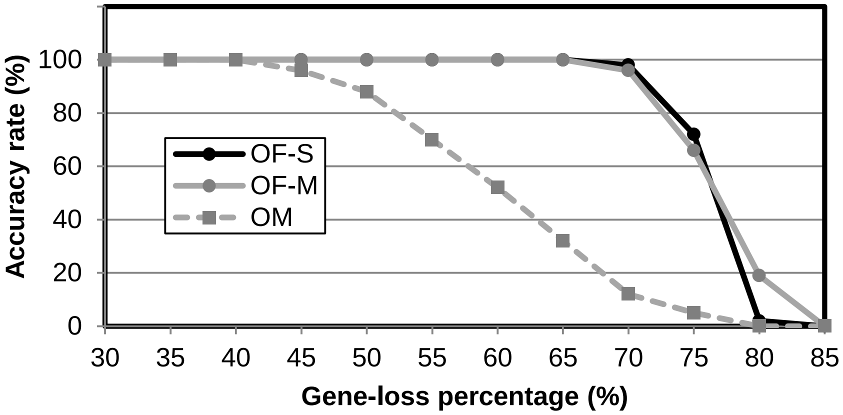

FIG. 5*A*.

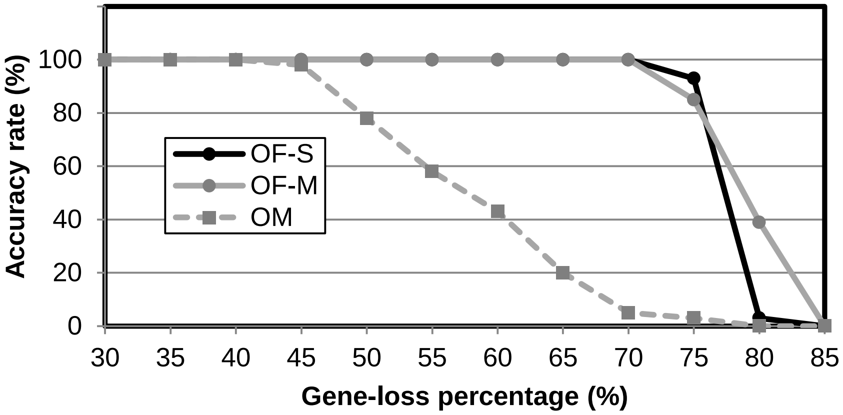

FIG. 5*B*.

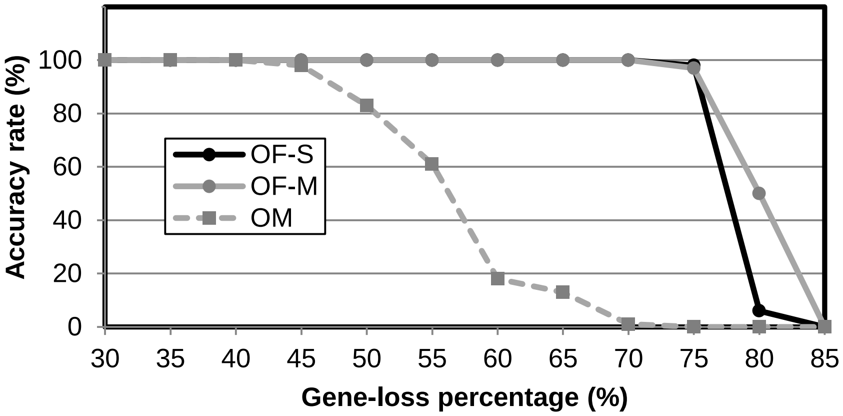

FIG. 5*C*.

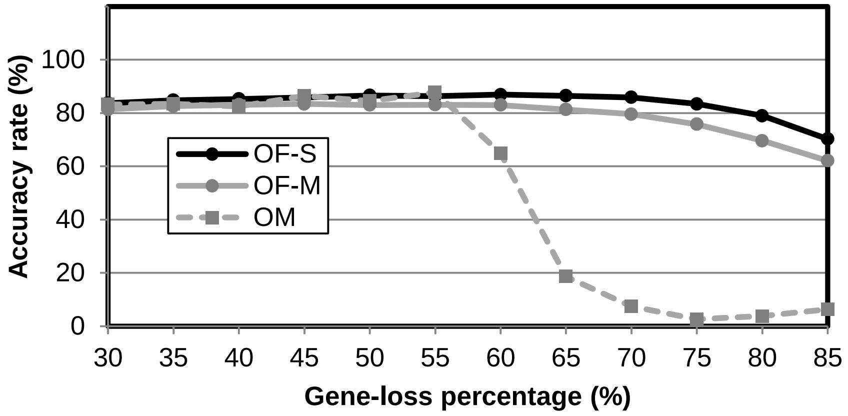

FIG. 5*D*.

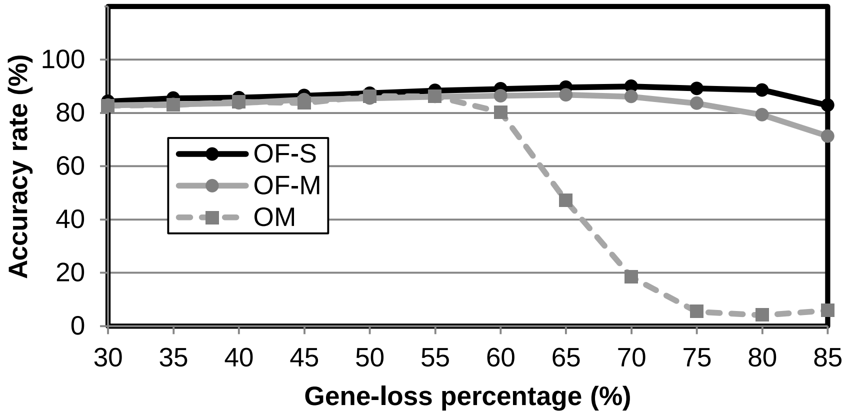

FIG. 5*E*.

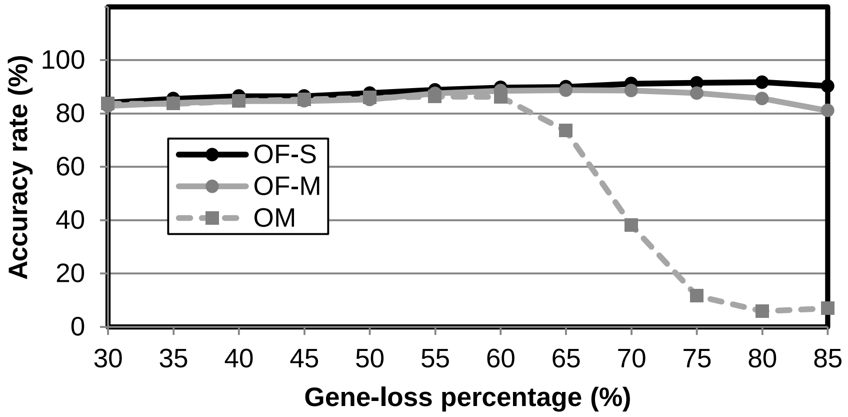

FIG. 5*F*.

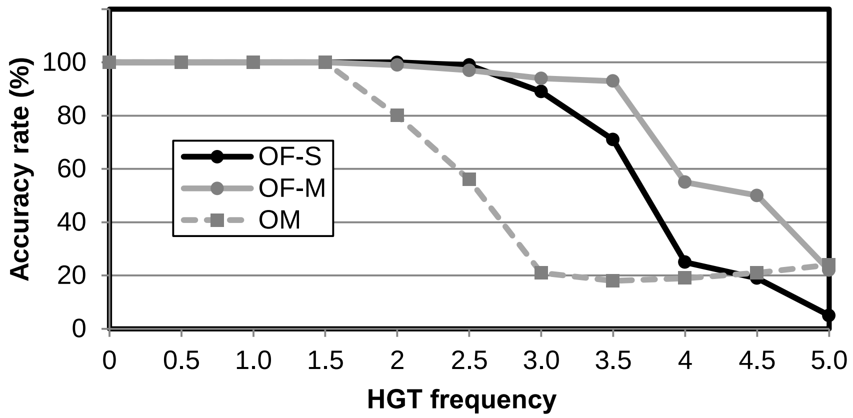
Accuracy rate (%)
0
20
40
60
80
100
0
0.5
1.0
1.5
2
2.5
3.0
3.5
4
4.5
5.0
HGT frequency
OF-S
OF-M
OM

FIG. 6*A*.

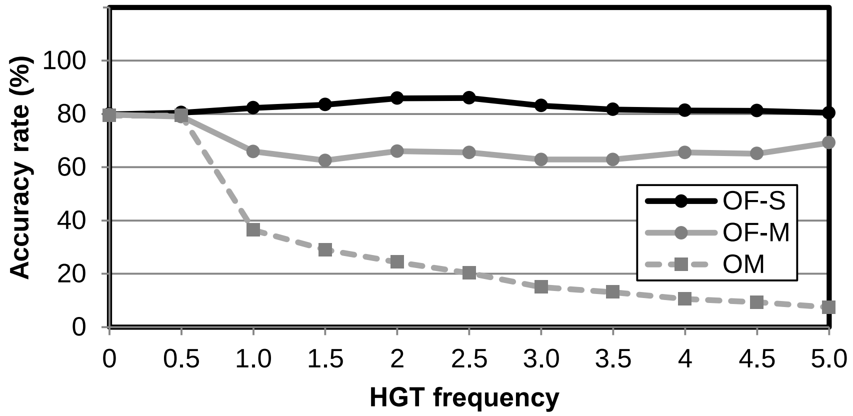
Accuracy rate (%)
0
20
40
60
80
100
0
0.5
1.0
1.5
2
2.5
3.0
3.5
4
4.5
5.0
HGT frequency
OF-S
OF-M
OM

FIG. 6*B*.

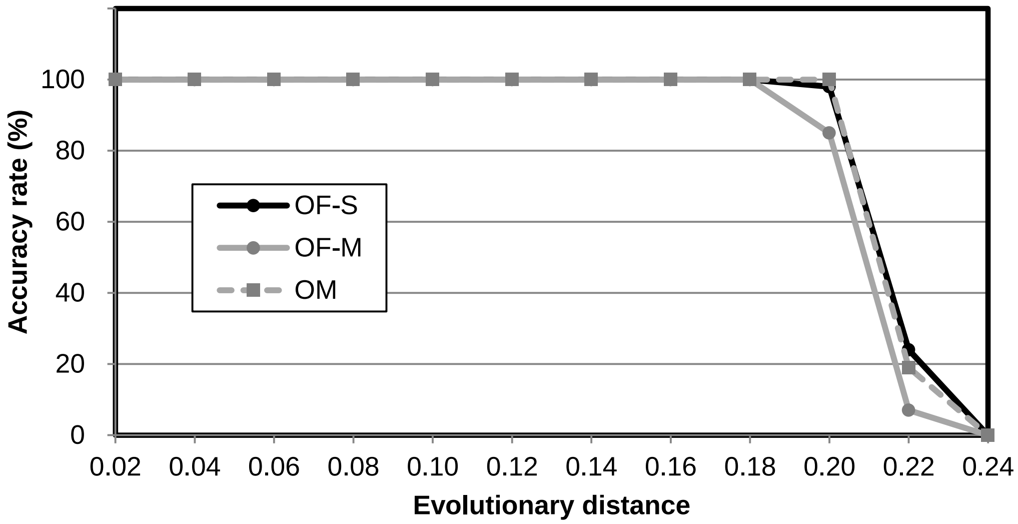
100
80
60
40
20
0
Accuracy rate (%)
OF-S
OF-M
OM
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24
Evolutionary distance

FIG. 7.

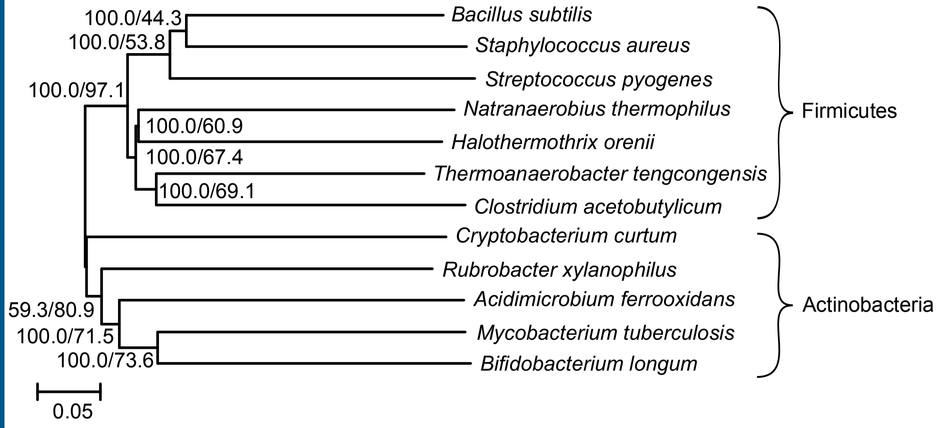

FIG. 8*A*.

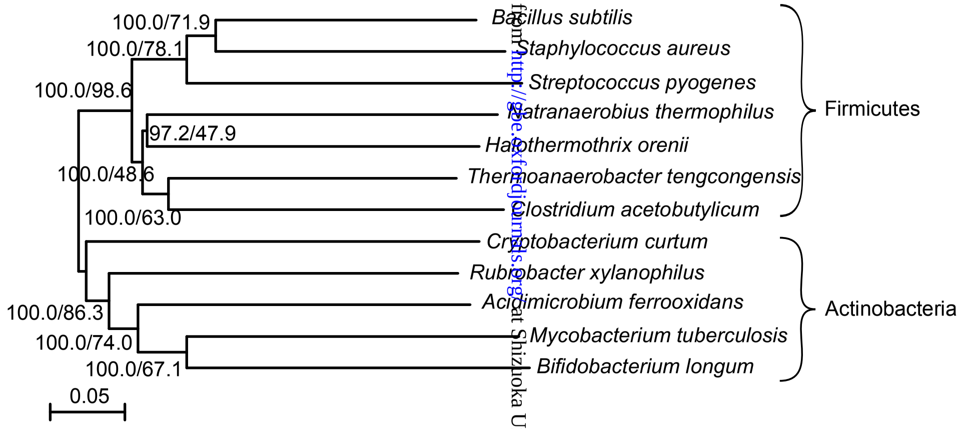

FIG. 8*B*.

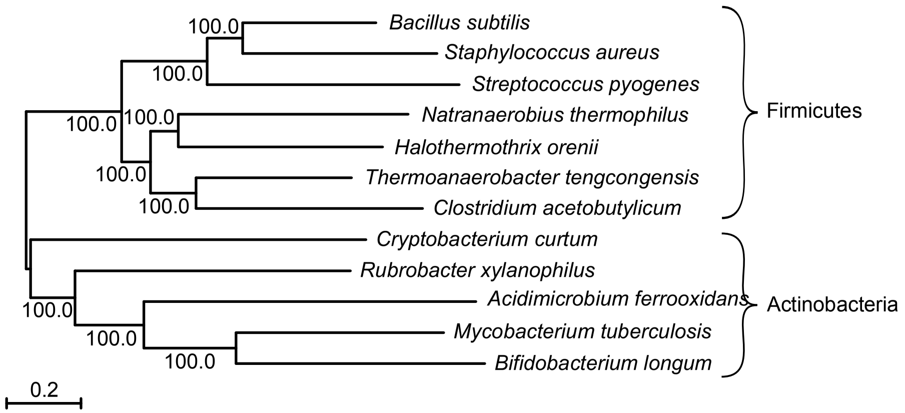

FIG. 8*C*.

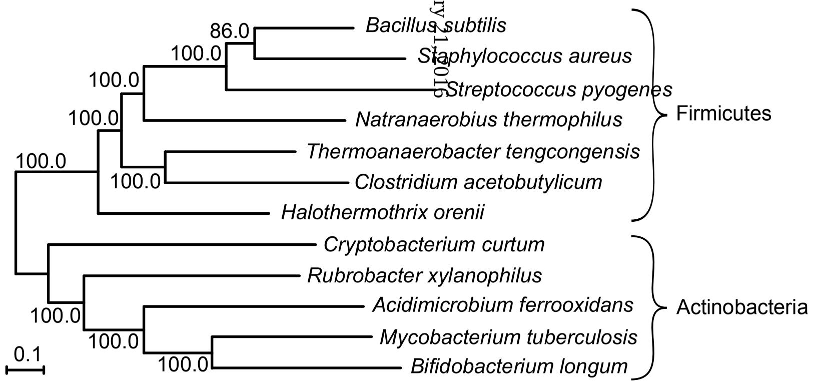

FIG. 8*D*.