AN INTRODUCTION TO MOLECULAR PHYLOGENETIC ANALYSIS

Article · October 2016

DOI: 10.7831/ras.4.36

Tokumasa Horiike
Shizuoka University

Please cite this article as:
Horiike, Reviews in Agricultural Science, 4: 36- 45, 2016.
Doi: 10.7831/ras.4.36

INVITED MINI-REVIEW

OPEN ACCESS

AN INTRODUCTION TO MOLECULAR PHYLOGENETIC ANALYSIS

Tokumasa Horiike

Department of Biological and Environmental Science, Shizuoka University, Shizuoka 422-8529, Japan

ABSTRACT

Phylogenetic analysis using molecular data such as DNA sequence for genes and amino acid sequence for proteins is very common not only in the field of evolutionary biology but also in the wide fields of molecular biology. The reason is that DNA sequencing became very popular and a huge amount of sequence data of genes and proteins are available in the public online database. Since many molecules (genes or proteins) which have various evolutionary rates are available, it is important to choose the suitable molecule for the phylogenetic analysis of a given lineage. For example, when the evolutionary rate of the gene (or protein) is too much higher for a given lineage, the substitution of nucleotide (or amino acid) is saturated. In this case, the accuracy of the phylogenetic analysis decreases. The methods for phylogenetic analysis are improving along with the evolution of computer science. Thus, there are many methods to infer phylogenetic tree, and many programs for each method are available. This mini review shows that general pattern of phylogenetic analysis, and explains some representative methods (Unweighted pair group method with arithmetic, Neighbor-joining method, Maximum parsimony method, maximum likelihood method, and Bayesian method). In the phylogenetic analysis, the most important feature is the interpretation of the phylogenetic tree. Therefore, several distinct points to evaluate a phylogenetic tree are also explained. These include, "validity of the tree shape", "evolutionary distance", and "validation of each internal branch". Towards the end, the procedure of evaluating a phylogenetic tree with an example using MEGA 7 is presented.

Keywords: molecular evolution, phylogenetic analysis, phylogenetic tree.

Introduction

Phylogenetic analysis is a method to elucidate the evolutionary history and relationship among a group of organisms. Previously, phylogenetic analysis was based on morphological comparison among the fossils, but the information from fossils was limited. Now, molecular phylogenetic analysis using molecular data such as DNA or proteins become popular. There are several reasons (Nei and Kumar, 2000). These include, (1) popularity of DNA sequencing method, (2) establishment of methods for phylogenetic tree construction using gene or protein sequences, (3) The results of a phylogenetic analysis being treated in a quantitative pattern, (4) Availability of many programs for constructing phylogenetic tree. The knowledge from phylogenetic analysis contributes to basic biology (e.g. evolutionary history of species, the evolution of genes, and identification of sampled species) as well as applied biology (e.g. investigation of the route of the infection of pathogenic microorganisms). Phylogenetic trees are commonly constructed to figure out the evolutionary relationship among species.

Selection of the molecules (genes or proteins)

DNA sequences of genes, RNA sequences of functional RNA, or amino acid sequences of proteins are used for phylogenetic analysis. To choose the molecule for phylogenetic analysis, there are two focal points. First, the genes must be shared by all of the given species. Secondly, the genes have the proper evolutionary rates, because proteins have varied evolutionary rates (Miyata *et al*., 1980). If a species has a distant relationship, the molecule which has low evolutionary rate should be chosen. This is because nucleotide or amino acid substitution of gene or proteins reaches to saturation between distant species when the evolutionary rate is high. Note that nucleotide sequence of a gene is easy to reach to saturation than an amino acid sequence of the coded protein. In this case, housekeeping genes which have low evolutionary rate are suitable. For example, 16SrRNA and gyrB for prokaryotes, 18SrRNA and Histone 3 for eukaryotes are available. If the species has a close relationship, the molecules which have high evolutionary rate should be chosen. This is because the accuracy reduces few substitution events when the evolutionary rate is low.

Received: March 16, 2016. Revised: August 18, 2016. Accepted: August 24, 2016.
Published on line: October 24, 2016. Correspondence to T.H.: horiike.tokumasa@shizuoka.ac.jp

In this case, tissue-specific genes having high evolutionary rate are suitable (Hastings, 1996; Duret, 2000).

Correction of homologs by homology search

Homologs are sequences having a common ancestor. Generally, homologs maintained the sequence similarity each other, then the phylogenetic analysis is available using the homolog sequences. Homologs were detected from DNA or protein database by homology search programs such as BLAST (Altschul *et al*., 1997). In a homology search, a query sequence is compared to all sequences in a DNA or protein database, and the sequences in the database with high similarity to the query sequence are reported. BLAST, which is based on pairwise sequence alignment, uses approximate calculation for speeding up, because the query sequence is compared to huge amount of sequences in the database. First, BLAST searches for "words" shared by the query sequence and a database sequence. The default word length for nucleotide is 11, and for amino acid is 3. For proteins, BLAST searches not only identical oligopeptides but also similar oligopeptides, and add them to the words list. Subsequently BLAST tries to find the identical region of the sequence in the database to the "words" in the words list. The identical region is called a "hit". When a "hit" has been found, BLAST attempts to extend the "hit" in both directions. The extending alignment is scored, using for nucleic acids a match/mismatch scoring scheme and for proteins a scoring matrix. When the score decreases a certain amount below the already found highest score, BLAST will stop extending and obtain a local alignment without gaps, called "High Scoring segment Pair" (HSP). Finally, BLAST makes a local sequence alignment with gaps for the HSP using Smith-Waterman algorithm. The HSP is retained if its E value is below a chosen threshold.

E-value is the expect number of sequences yielding a similarity score of at least S that one would expect to find by chance from the artificial database which contains random sequences and has the same number of residues as the current database. E-value is defined as follows.

$\text{E-value} = m * n * K * e^{-\lambda * S}$

Where m is the length of the query sequence, n is the total length of the database. K and λ are the so-called "Karlin and Altschul" parameters, which depend on the scoring scheme and the base/amino acid composition of the database. A lower E-value (near to 0) indicates higher significance of the similarity. The E-value takes into account the length of the query sequence, because shorter sequences have a high probability of occurring in the database by chance. Generally, the threshold of E-value is 10^{-5} for phylogenetic analysis.

Homologs are classified in to orthologs, paralogs, and xenologs. These are homologs derived from speciation, gene duplication, and lateral transfer, respectively. Thus, if a researcher focuses on the phylogenetic relationship of a species, paralogs and xenologs should be removed or ignored from the analysis.

To determine the position of a root, outgroup should be added to the analysis data. The chosen outgroup are species which are closely related to group under study but less closely related than the relationship among the species under study (Fig. 1). The outgroup is essential for understanding the evolution of traits along a phylogeny.

Alignment of genes or proteins

A Multiple Sequence Alignment (MSA) is a sequence alignment of three or more biological sequences including protein, DNA, or RNA. It is used to calculate the best match for the selected sequences, and lines them up so that the identities, similarities, and differences can be seen (Fig. 2). The input set of query sequences is assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor. If partial sequences are included in the dataset, the correlated region (matched region) should be used for the phylogenetic analysis, because the lacking part of the partial sequence may be treated as a "deletion" event. There are several popular programs for MSA such as ClustalW (Thompson *et al*., 1994), ClustalX (Thompson *et al*., 1997), Clustal Omega (Sievers *et al*., 2011), MAFFT (Katoh *et al*., 2002), and MUSCLE (Edgar, 2004). MEGA 7 (Kumar S *et al*., 2016) which includes MUSCLE is also popular software because it has a user-friendly interface. The software list is presented in Table 1.

Methods for inferring of phylogenetic tree

Methods for inferring the phylogenetic tree are classified into two groups. First is the distance-based method which uses evolutionary distance matrix. UPGMA (Unweighted Pair Group Method with Arithmetic) and NJ (Neighbor-joining) method are the representative methods. UPGMA was first developed, but it was hardly used due to a demerit that it assumes the molecular clock is constant. However, NJ method has been widely used, because the prevailing demerit of UPGMA was overcome by this method. The advantage of the distance-based method is its short calculation time. Therefore, large amount of data can be handled. Second is the character-based method which uses the aligned sequences directly during tree inference. Maximum parsimony, Maximum likelihood, and Bayes method are the representative methods. Akin to distance based method, Maximum parsimony was developed first, but it came to hardly be used due to a disadvantage mentioned below. Maximum likelihood and Bayes method are the most widely used methods, because these methods

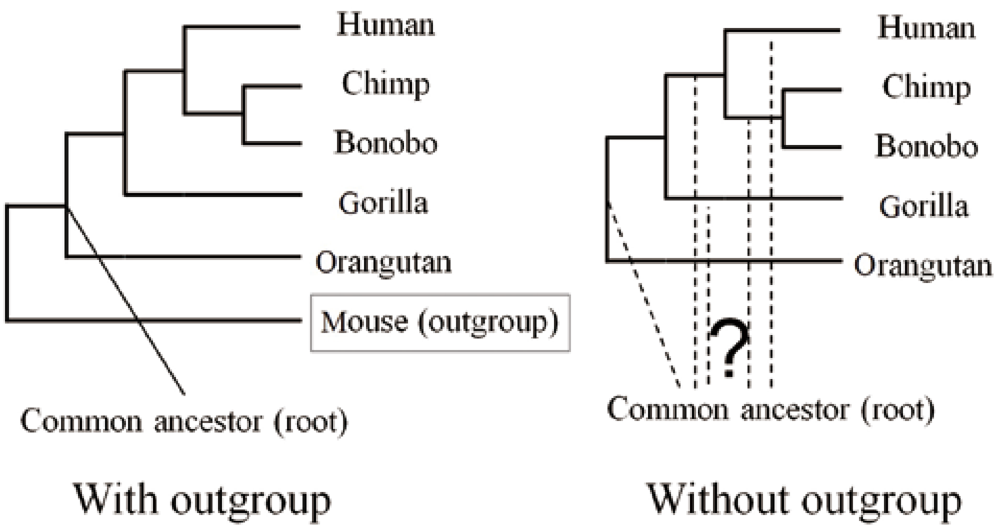
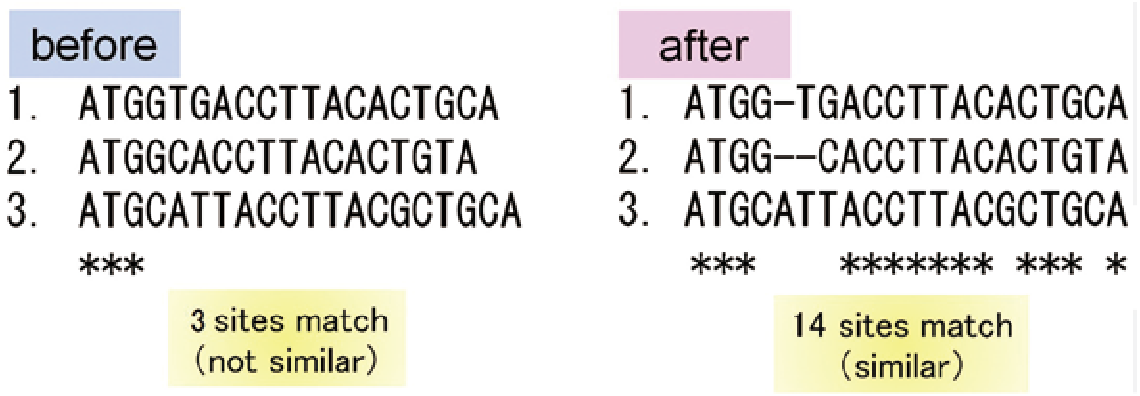


Fig.1. Phylogenetic trees with outgroup and without outgroup



```
before
1.  ATGGTGACCTTACACTGCA
2.  ATGGCACCTTACACTGTA
3.  ATGCATTACCTTACGCTGCA
    ***

after
1.  ATGG-TGACCTTACACTGCA
2.  ATGG--CACCTTACACTGTA
3.  ATGCATTACCTTACGCTGCA
    ***   ******* *** *
```

Fig.2. An example of multiple alignment. Gaps are added to adjust sequences. In this example, there are 3 site matches before alignment. After multiple alignment, there are 14 site match.

introduce the elaborated evolutionary-models based on statistical method. Unfortunately, it is difficult to gain clear consensus which method is the best for inferring a phylogenetic tree. Therefore, researchers choose several methods to infer the phylogenetic tree, and compare the results. If the results are consistent, the phylogenetic tree is thought to be reliable. The software and representative methods are listed in Table 1 and Table 2, respectively.

UPGMA (Sokal and Michener, 1958) is the original method used for reconstructing phylogenetic trees using evolutionary distance matrix. Clustering is done by searching for the shortest evolutionary distance in the distance matrix. The newly formed cluster replaces the OTUs (Operational Taxonomic Unit, the group of organisms currently being analyzed) it represents in the distance matrix, and distances between the newly formed cluster and each of the remaining OTUs are calculated. This process is repeated until all OTUs are clustered. The distance of the newly formed cluster is the average of the distances of the original OTUs. This process assumes that the evolutionary rate from the node of the two clustered OTUs to each of the two OTUs is identical. The whole process of clustering thus assumes that the evolutionary rate is the same in all branches. However, it is

Table 1: List of programs for MSA, homology search, and phylogenetic analysis

Software	Category	Interface	OS	Web service	Remarks
BLAST	homology search	command-line	Win, Mac, Linux	available	standard program for homology search
ClustalW	MSA, phylogenetic tree	command-line	Win, Mac, Linux	available	integrated in MEGA, no maintenance
ClustalX	MSA, phylogenetic tree	graphical	Win, Mac, Linux	not available	no maintenance
Clustal Omega	MSA	command-line	Win, Mac, Linux	available	successor of ClustalW
MAFFT	MSA	command-line	Win, Mac, Linux	available	available for >2,000 sequences
MUSCLE	MSA	command-line	Win, Mac, Linux	available	integrated in MEGA, more accurate than ClustalW
PhyML	phylogenetic tree	command-line	Win, Mac, Linux	available	widely used
RAxML	phylogenetic tree	command-line	Win, Mac, Linux	available	possible to estimate the best substitution model
FastTree	phylogenetic tree	command-line	Win, Mac, Linux	not available	100–1,000 times faster than PhyML 3.0 or RAxML 7
MrBayes	phylogenetic tree	command-line	Win, Mac, Linux	not available	standard program for Bayesian method
PHYLIP	multipurpose	command-line	Win, Mac, Linux	not available	multifunctional
MEGA 7	multipurpose	graphical	Win, Mac, Linux	not available	friendly user interface, frequently updated
TOPALi v2	multipurpose	graphical	Win	not available	friendly user interface

Table 2: List of methods for inferring phylogenetic trees

Method	Group	Algorithm	Software
UPGMA	distance-matrix	Clustering for the shortest evolutionary distance	MEGA 7
Neighbor-joining	distance-matrix	Clustering for minimum total branch length	PHYLIP, Clustal X, MEGA 7
Maximum parsimony	character-based	Searching tree with minimum total number of character-state changes	PHYLIP, MEGA 7
Maximum likelihood	character-based	Searching tree with maximum likelihood	PHYLIP, PhyML, RAxML, FastTree, MEGA 7, TOPALi v2
Bayesian	character-based	Searching tree with maximum posterior probability	MrBayes, TOPALi v2

known that most evolutionary rate of molecules are not constant, thus UPGMA is not suitable in many cases. This is the demerit of UPGMA. MEGA 7 is available to infer phylogenetic tree by this method.

NJ method (Saitou and Nei, 1987) is a commonly used distance-based method for inferring phylogenetic trees, which overcomes the disadvantage of UPGMA. It requires an evolutionary distance matrix calculated from multiple sequence alignment but it does not require a molecular clock. The principle of NJ method is to find pairs of operational taxonomic units (OTUs) that minimize the total branch length at each stage of clustering of OTUs starting with a star-like tree. NJ method is, therefore, a special case of the star decomposition method. The merit of NJ method is fast. Therefore, it is practical to analyze a large dataset (hundreds or thousands of taxa, Tamura 2004; Niimura and Nei 2005). In case, if the taxa number is not large (<100), other methods which are not based on evolutionary distance supersede NJ method. This is because they offer superior accuracy and NJ method often assigns negative lengths to some of the branches. It is also known that the accuracy of the NJ method is lower in case of short DNA or amino acid length (Page and Holmes 1998). PHYLIP (Felsenstein 1989), Clustalw, Clustal X and MEGA 7 are widely used for inferring phylogenetic tree by this method.

Maximum parsimony is the origin of character-based methods. It was developed by Henning (1966) using morphological data. Later, Maximum parsimony with an amino acid or nucleotide data was developed, respectively by Eck and Dayhoff (1966) as well as Fitch (1971). It finds the tree topology for a set of aligned sequences that is subject to the constraint of invoking the fewest possible evolutionary changes. The Maximum parsimony algorithm deduces for each site the minimum number of character changes required along its branches to explain the observed states at the OTUs. When some reasonable topologies have been obtained, the tree that requires the minimum number of changes is selected as the maximum parsimony tree. Maximum parsimony assumes that a common character is derived from a common ancestor and thus underestimates the real divergence between distantly related taxa. This is the disadvantage of the Maximum parsimony. PHYLIP and MEGA 7 is available to infer phylogenetic tree by this method.

The maximum likelihood method is a statistical method (character based) for inferring phylogenetic trees (Felsenstein, 1981). It uses statistical techniques for inferring probability distributions to assign probabilities to particular possible phylogenetic trees. Therefore, the calculation time is longer than those of other methods. This method requires a substitution model to assess the probability of particular mutations. Thus, finding best substitution model have to be carried out in advance. PHYLIP and PhyML (Guindon *et al*., 2010) are the most famous programs for maximum likelihood method. Nowadays, RAxML (Stamatakis *et al*., 2005), MEGA 7, TOPALi V2 (Milne *et al*., 2009) are widely used because it can find the best substitution model and use the model to infer the phylogenetic tree by maximum likelihood method. Long calculation time is the demerit of the maximum likelihood method. If the data is very large, FastTree, which infers approximately-maximum-likelihood phylogenetic trees very quickly, is useful (Price *et al*., 2009).

Bayesian method is a character-state method for inferring phylogenetic tree which is based on posterior probabilities under the estimated best model. Bayesian method employs the concept of likelihood and searches for a set of plausible trees. Bayesian method requires the information of a prior distribution on the model parameters, such as substitution model parameters, branch lengths, and tree topology. Posterior probabilities are obtained by exploring tree space using a sampling technique, called Markov chain Monte Carlo (MCMC) algorithms. The Bayesian approach has become popular due to advances in computing speeds and is one of the most widely used methods now. MrBayes (Huelsenbeck et al., 2001) is the most widely used program for inferring phylogenetic tree by Bayes method. Due to its graphical user interface, TOPALi v2 has been popular as well.

Evaluation of phylogenetic tree

After the inference of a phylogenetic tree, the tree should be evaluated in some specific points. First is the validity of the tree shape. If the molecules used in the analysis are derived from the common ancestor by speciation, the phylogenetic tree may represent the species phylogeny (Fig. 3A). If two or more sequences of a species are located separately, the tree does not represent the genuine phylogeny of the species (Fig. 3B). Second is evolutionary distance. Branch length of the tree reflects the evolutionary distance of each branch. Evolutionary distance tentatively correlated with the evolutionary time. Therefore, the periods of evolutionary events such as speciation and gene duplication, and change of evolutionary rate can be estimated from the information in branch length (Fig. 4). Third is the validation of each internal branch. The bootstrap test is a commonly-used method for evaluation the reliability of specific clades in the tree. The procedure of the bootstrap test is as follows: First, an artificial alignment is obtained from the original alignment data by random choice of columns. Each column in the original alignment can be selected more than once until the same length as the original one has been constructed. Secondly, for each artificial alignment data, a tree is constructed, and the proportion of each internal branch to all the artificial trees is computed. If an internal branch has high proportion (bootstrap value), the internal branch is thought to be reliable. However, the bootstrap test is not statistical test for evaluation of each branch. Empirically, when the bootstrap value is 90 % or more, the internal branch thought to be reliable (Fig. 5).

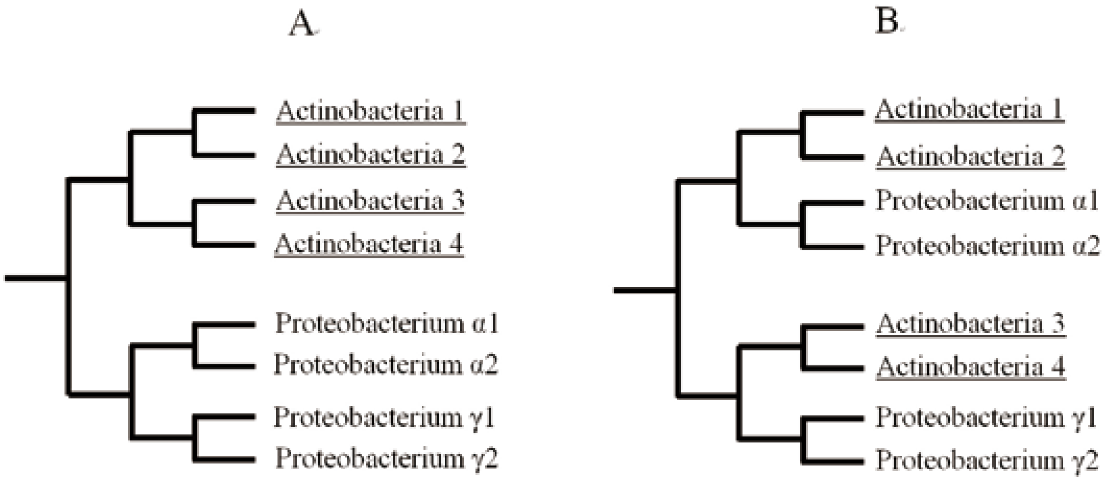


Fig.3. Examples of the monophyletic tree (A) and polyphyletic tree (B).

To infer phylogenetic tree of a gene, many programs for each method are presented in Tables 1 and 2. Researchers need to be cautious with using many programs for phylogenetic analysis which have not been updated. The process of inferring phylogenetic tree by maximum likelihood method using MEGA 7 is presented in Fig. 6, consisting of 9 components. Please see the details for these components, presented in page 45).

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25: 3389-3402.

Duret L and Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol., 17: 68-74.

Eck RV and Dayhoff MO (1966) Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, D.C..

Edgar CR (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res., 32: 1792-1797.

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol., 17: 368-376.

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20: 406-416.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, and Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307-321.

Hastings KE (1996) Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. J. Mol. Evol., 42: 631-640.

Hennig (1966) Phylogenetic Systematics. University of Illinois Press, Urbana.

Huelsenbeck JP and Ronquist F (2001) MrBayes: Bayesian inference of phylogeny. Bioinformatics, 17: 754-755.

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

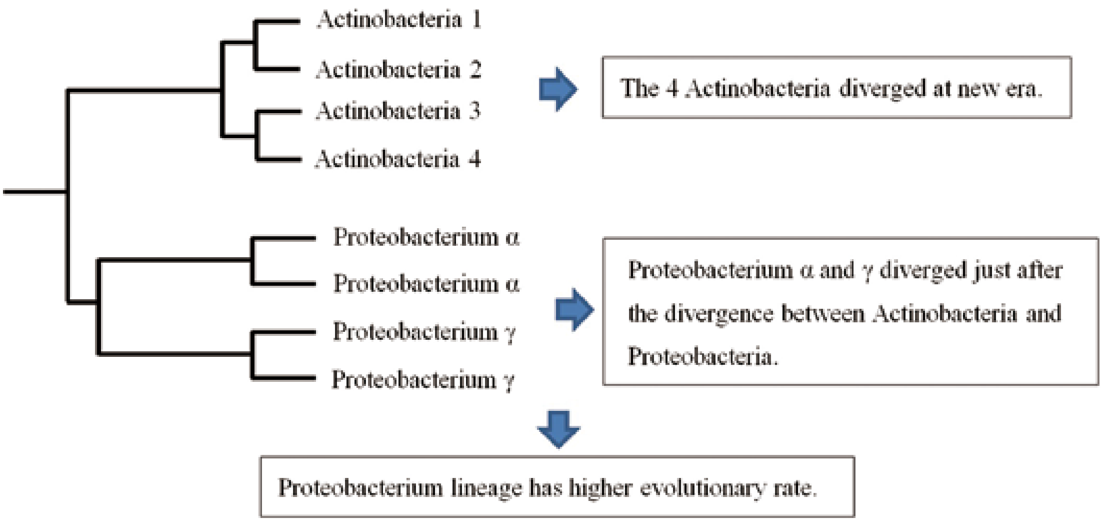
Kumar S, Stecher G, and Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 33:1870-1874.

Milne I, Lindner D, Bayer M, Husmeier D, McGuire G, Marshall DF, and Wright F (2009) TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. Bioinformatics. 25:126-127.

Miyata T, Yasunaga T, and Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci U S A. 77: 7328-7332.

Nei M and Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press, Inc., New York.

Niimura Y and Nei M (2005) Evolutionary dynamics of olfactory



The 4 Actinobacteria diverged at new era.

Proteobacterium α and γ diverged just after the divergence between Actinobacteria and Proteobacteria.

Proteobacterium lineage has higher evolutionary rate.

Fig.4. An example of the phylogenetic tree with various branch lengths.

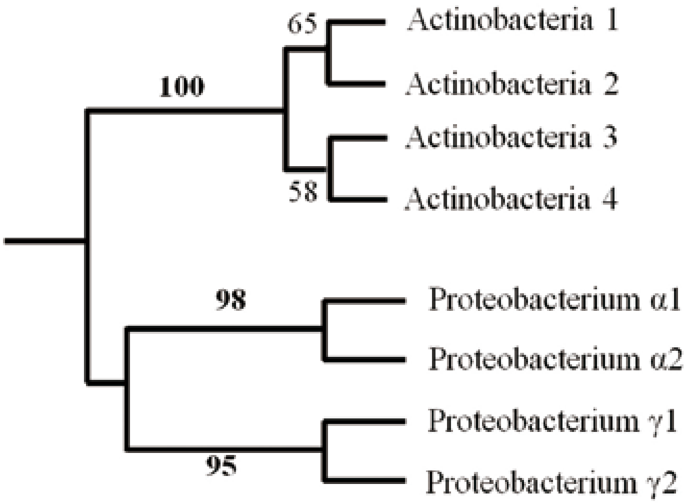


Fig.5. An example of the phylogenetic tree with bootstrap values. The internal branches with values in bold are more than 95%.

receptor genes in fishes and tetrapods. Proc Natl Acad Sci U S A. 102: 6039-6044.

Page RDM and Holmes EC (1998) Molecular Evolution: A Phylogenetic Approach. Blackwell Science, Oxford.

Price MN, Dehal PS, and Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 26:1641-1650.

Saitou N and Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol., 4: 406-25.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, and Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol., 7: 539.

Sokal R and Michener C (1958) A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 38: 1409-1438.

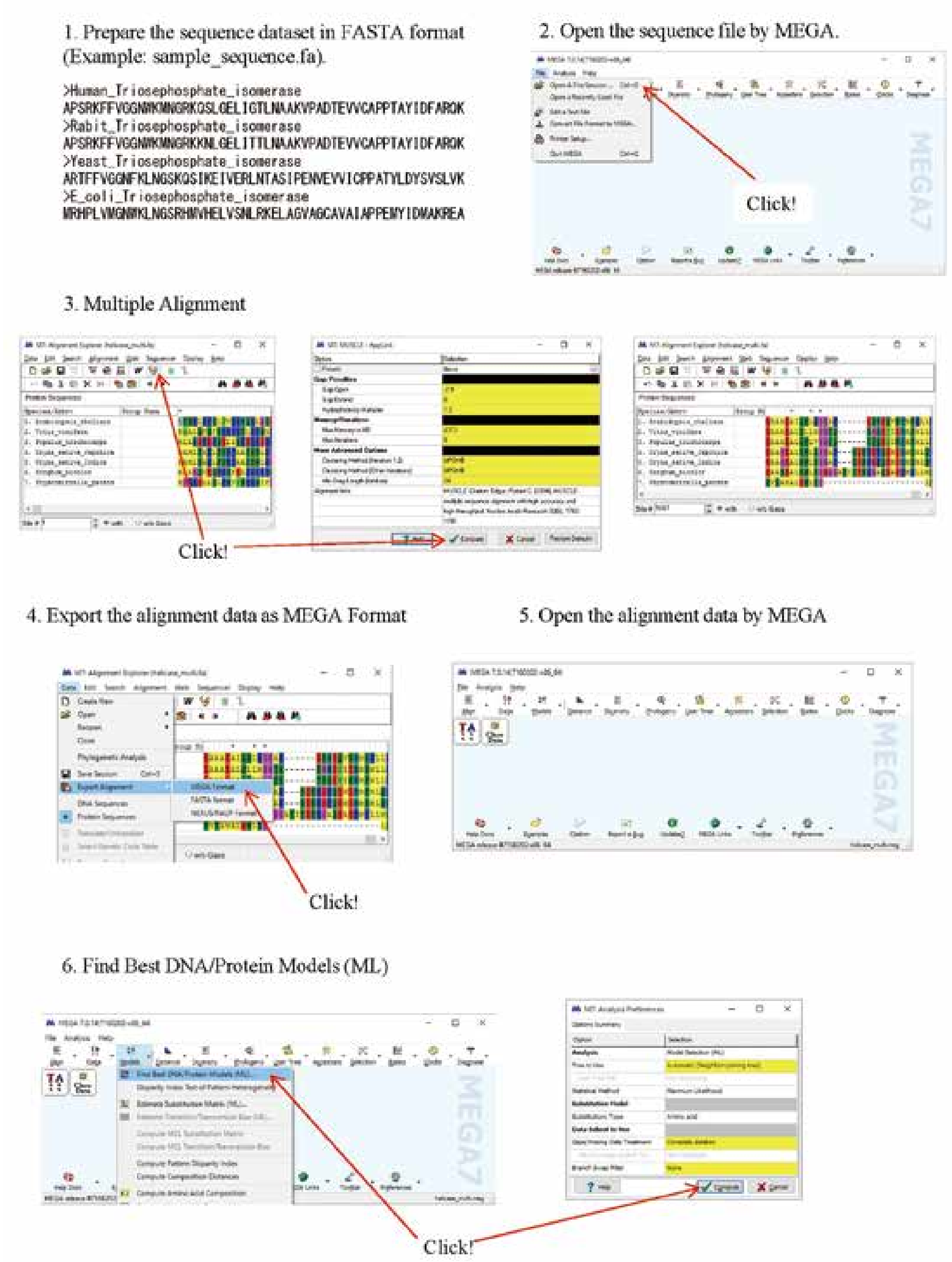
Stamatakis A, Ludwig T, and Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics, 21: 456-463.

Tamura K, Nei M, and Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A. 101: 11030-11035.

Thompson JD, Higgins DG, and 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.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, and Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res., 25: 4876-4882.

Please cite this article as:
Horiike, Reviews in Agricultural Science, 4: 36- 45, 2016.
Doi: 10.7831/ras.4.36



1. Prepare the sequence dataset in FASTA format (Example: sample_sequence.fa).

```
>Human_Triosephosphate_isomerase
APSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPTAYIDFARQK
>Rabit_Triosephosphate_isomerase
APSRKFFVGGNWKMNGRKKNLGELITTLNAAKVPADTEVVCAPPTAYIDFARQK
>Yeast_Triosephosphate_isomerase
ARTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVVICPPATYLDYSVSLVK
>E_coli_Triosephosphate_isomerase
MRHPLVMGNWKLNGSRHMVHELVSNLRKELAGVAGCAVAIAPPEMYIDMAKREA
```

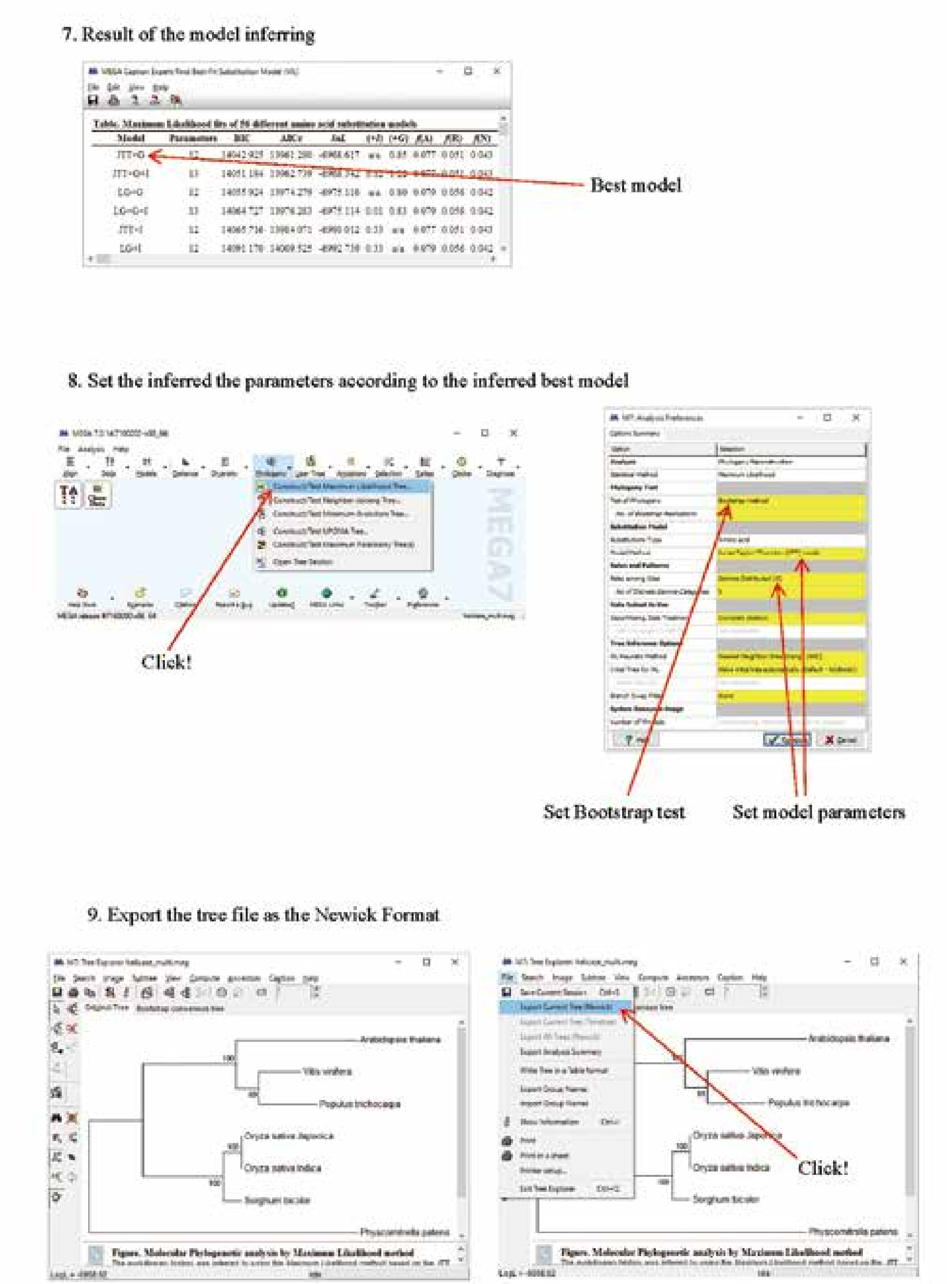
2. Open the sequence file by MEGA.

3. Multiple Alignment

4. Export the alignment data as MEGA Format

5. Open the alignment data by MEGA

6. Find Best DNA/Protein Models (ML)



7. Result of the model inferring

Table. Maximum Likelihood fits of 56 different amino acid substitution models

Model	Parameters	BIC	AICc	lnL	(+I)	(+G)	f(A)	f(R)	f(N)
JTT+G	12	14042.925	13961.280	-6968.617	n/a	0.85	0.077	0.051	0.043
JTT+G+I	13	14051.184	13962.739	-6968.342	0.02	1.22	0.077	0.051	0.043
LG+G	12	14055.924	13974.279	-6975.118	n/a	0.80	0.079	0.056	0.042
LG+G+I	13	14064.727	13976.283	-6975.114	0.01	0.83	0.079	0.056	0.042
JTT+I	12	14065.716	13984.071	-6980.012	0.33	n/a	0.077	0.051	0.043
LG+I	12	14091.170	14009.525	-6992.738	0.33	n/a	0.079	0.056	0.042

Best model

8. Set the inferred the parameters according to the inferred best model

Click!

Set Bootstrap test

Set model parameters

9. Export the tree file as the Newick Format

Click!

Fig.6. The process of an inferring phylogenetic tree by maximum likelihood method using MEGA 7. The steps 1 to 9 of this process are, presenled in page 45.

1. Prepare the sequence dataset in FASTA format.
   The sequence data of orthologs were prepared in FASTA format. The extension of the file should be "fas".
2. Open the sequence file by MEGA.
   Click "Open A File/Session" and choose the sequence file, then click "Align" to open.
3. Multiple Alignment
   Click "Align selected with MUSCLE" button, "Align Protein", and "Compute". The sequences are aligned by MUSCLE.
4. Export the alignment data as MEGA format.
   Click "Data", "Export", and "MEGA format". Then, name the file, and save. The extension of the file is "meg". Close the MEGA window.
5. Open the alignment data by MEGA.
   The multiple alignment file in MEGA format is opened.
6. Find Best DNA/Protein Models (ML).
   Click "Model" and "Compute". The best substitution model for the multiple alignment data is inferred by maximum likelihood method.
7. Result of the model inferring
   The model on the top is the best substitution model for the alignment data. In this example, JTT model with gamma distribution is the best model.
8. Set the inferred the parameters according to the inferred best model
   Click "Construct/Test Maximum Likelihood Tree. Set "Test of Phylogeny" as "Bootstrap method", and "No of Bootstrap Replications" as 1000. Set "Model/Method" and "Rates among Sites" according to the result of the model inferring. In this case, they set as "Jones-Taylar-Thornton (JTT) model" and "Gamma Distributed (G)", respectively. After setting, click "Compute".
9. Export the tree file as the Newick Format
   Click "Export Current Tree (Newick) to save the tree file in Newick format. If researchers want to save the edited tree, click "Save Current Session" to save the all information of the tree.