Master's Dissertation

ASSESSMENT PROGRAM FOR SYSTEMATIC ERROR CAUSING PHYLOGENETIC INCONGRUENCE OF GENE MARKERS

바이오인포매틱스 프로그램을 이용한
유전자 마커 선별 및 계통수 오류 평가 연구

August 2021

Graduate School of Natural Sciences
Seoul National University
Interdisciplinary Program in Bioinformatics

Junghwan Lee

ASSESSMENT PROGRAM FOR SYSTEMATIC ERROR CAUSING PHYLOGENETIC INCONGRUENCE OF GENE MARKERS

지도교수 손 현 석

이 논문을 이학석사 학위논문으로 제출함

2021년 7월

서울대학교 대학원

자연과학대학 생물정보학협동과정

이 정 환

이정환의 석사 학위논문을 인준함

2021년 7월

위 원 장 _____

부 위 원 장 _____

위 원 _____

ABSTRACT

Assessment Program for Systematic Error Causing Phylogenetic Incongruence of Gene Markers

Junghwan Lee

Laboratory of Computational Biology and Bioinformatics
Interdisciplinary Program in Bioinformatics
Seoul National University

The steadily increasing volume of biological data with decisive phylogenetic relationship provides unparalleled opportunities in bioinformatics. Phylogenetics based on a large amount of datasets handling an evolutionary history and assigning the placement of taxa in a phylogeny establishes the tree of life. Constructing a phylogeny involving a phylogenetic analysis is implemented in most branches of biology and emphasizing the evolutionary history elucidates the phylogenetical background as a prerequisite interpreting a specific biological system, which is a biologically indispensable process. Due to the advent of computing and sequencing techniques as the phylogenetic approach, phyloinformatics has rapidly advanced at the technical and methodological levels along with phylogenetic reconstruction algorithm and evolutionary models. Unlike the classic approach using morphological data, modern phylogenetic analysis reconstructs a phylogeny using genetic information following the inference of phylogenetic tree from molecular data. Therefore, phylogeneticists have naturally dealt with questions concerning the

accuracy of phylogenetic estimation and carried out studies on the reliability of phylogenies. In terms of molecular systematics, the concerns regarding the assessment of phylogenetic accuracy considering specific evolutionary conditions and the amount of molecular data implemented can now be divided into two types: how phylogenetic method works and how reliable it is under certain circumstances. Moreover, in terms of data quality, assessment for suitability of nuclear marker is required before the phylogenetic inference is performed for confident phylogeny. Recently, the probability of stochastic errors in phylogenetic estimation dealing with a large-scale datasets has decreased, while the probability of systematic errors has increased. Thus, before the implementation of phylogenetic reconstruction, the assessment of sources of systematic errors is indispensable for the improvement and estimation of phylogenetic accuracy. Assessment Program for Systematic Error (APSE) developed by this study will plays a key role in assessment between user datasets and phylogenies for improving the results of phylogenetic reconstruction in systematics and will be able to implement an analysis of the effect on data bearing systematic errors in a phylogeny after the misleading phylogenetic results are produced. This study with APSE will serve as the inference of phylogenetic accuracy and the assessment of systematic errors using an unresolved example showing the contradicting topologies between different gene markers in the same diversity group. Furthermore, by selectively grouping the properties of the existing systematic biases provided by the APSE, it proceeds in the direction of proposing a new protocol that can provide the best gene marker among candidate markers for a specific taxon.

……

Keywords: systematic error, bioinformatics, standalone, phylogenetic reliability, multiple sequence alignment, phylogenetic analysis, data quality

Student ID: 2017-23460

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

A2AB	alpha-2B adrenergic receptor
BMCMC-PP	bayesian posterior probabilities calculated via markov chain monte carlo sampling
C factor	convergence factor
CNC	convergence of nucleotide composition
FASTA	fast-all
IRBP	interphotoreceptor retinoid binding protein
Iss	index of substitution saturation
Iss.c	critical index of substitution saturation
Iss.cSym	critical Iss with symmetric topology
Iss.cAsym	critical Iss with asymmetric topology
LBA	long branch attraction
LRTs	likelihood ratio tests
MAFFT	multiple alignment using fast Fourier transform
MCMC	markov chain monte carlo
MEGA	molecular evolutionary genetics analysis
ML	maximum likelihood
ML-BP	maximum likelihood bootstrap proportion
MP	maximum parsimony
MSA	multiple sequence alignment
MUSCLE	multiple sequence comparison by log-expectation
NCBI	national center for biotechnology information
NDE	node-density effect
NGS	next-generation sequencing
PP	posterior probability
RCFV	relative composition frequency variability
RCV	relative composition variability
RELL	resampling estimated log-likelihood
SBS	standard bootstrap
STL	standard template library
vWF	von willebrand factor

CHAPTER I.

INTRODUCTION

1.1 Background of research

1.1.1 Overview of phylogenetics in bioinformatics

Phylogenetics is the study of acquiring the information of evolutionary relationships within species or groups of organisms and discovering these relationships using a variety of computational phylogenetic methods. The advent of the high-throughput sequencing technique was desirable for biologists and bioinformaticians. Therefore, the utilization of phylogenetic approaches has increased with the progress of computing techniques. As the amount of usable molecular data has steadily increased, phylogenetic study has extends from the single-gene analysis to the phylogenomics using multigene datasets, and this analysis cannot be performed without an automatic computational pipeline. Therefore, evolutionary biologists and phylogeneticists must consider the selection of the bioinformatics program and the sampling of biological data for reliable phylogenetic inferences leading to true phylogenies. Bioinformatics is a science that answers biological questions by using methods of mathematics, statistics, and computer science and interpreting the biological data (Thampi, 2009). Bioinformatics started as a way to build a database to store biological data and to develop algorithms for analyzing molecular data (Dayhoff, Doolittle, Fitch, McLachlan, 1960). The current meaning of bioinformatics, has been extended to the study of all biological information in various biological systems, and

thus, it processes the raw data to derive meaningful results. Furthermore, the results derived from bioinformatics research are being applied in various fields, including phylogenetics, and the bioinformatics programs for these applications are software programs that helps to understand life phenomena by analyzing biological data and the corresponding mechanisms with computer algorithms. Therefore, computational phylogenetics in bioinformatics is a method of performing phylogenetic analysis (Fitch, Margoliash, 1967) by approaching computing or statistical problems such as using computational algorithms and programs. In particular, with the advancement of computing technology and algorithms, the questions regarding biological lineages can be answered with computational phylogenetics, and specifically, it encompasses all problems such as multilocus or large-scale phylogenetic estimation, supertree methods, and multiple sequence alignment (MSA) (Corpet, 1988) techniques. Early phylogenetics was a method of analysis based on the morphology of a species, and evolutionary relationships were inferred by phenotypic traits, such as the body size of an organism, the length of a specific bone, or a specific behavioral sign. Moreover, phylogenetics based on biochemical properties identified the relationships between organisms with clues such as the components of cells, the sequence of synthesis of end products from the metabolic process, and attributes of enzymes. Current phylogenetics selects an appropriate mathematical model to define the evolutionary relationship and at this time, phylogenetic estimation is performed with a computational phylogenetics program in order to efficiently analyze a large amount of sequence data. Thus, the problem that researchers are facing now in phylogenetic analysis is finding a method that can accurately and efficiently handle large amounts of sequence data and define the evolutionary relationship of structures and functions between organisms. On the other hand, the result of the analysis mentioned above is the output of a phylogenetic tree, and the evolutionary tree drawn in this way

becomes a prerequisite for interpreting biological systems and grasping evolutionary patterns by presenting how molecular sequences have evolved from common ancestors. Therefore, in the phylogenetic analysis results, a reliable evolutionary tree must be drawn by the researcher to avoid the fundamental problem of variation in the historical path of the taxonomic species or group of species, and the problem of evaluating the accuracy of the drawn tree has evolved into an important issue.

1.1.2 Assessment of the credibility and accuracy of phylogenies

Since phylogenetics focuses on constructing an accurate phylogenetic tree for estimating evolutionary relationships, it is natural for relevant researchers to question the accuracy of phylogenetic trees and correctly infer that phylogenies are essential to the study of the evolutionary history of living things. Historically, the process of building a phylogenetic tree has valued knowing the evolutionary relationship or the pattern of tree topology itself, and the phenomenon of drawing a phylogenetic tree has emerged in order to know information about the process leading to the observed evolutionary pattern. In this regard, phylogenetic reconstruction facilitated the analysis of gene duplication, evolution rate, polymorphism, recombination, lineage divergence, and population demographics, and the accurate estimation of evolutionary parameters supports the validity of phylogenetic reconstruction. The process of phylogenetic inference is generally divided into two parts. The first part is identifying homologous characters within datasets to be studied (i.e., sampling of sequences from species). The second part of the process is inferring the evolutionary history of organisms through comparison of these characters using tree reconstruction methods. Phylogenetic trees helps biologists to make evolutionary inferences by providing information on expected features (i.e., when or where

various structures, molecules, and behaviors of living organisms evolve in a taxonomic group). Accordingly, it is possible to predict the entire biological system, and when a phylogenetic analysis across various levels of biological organization (i.e., gene, genome, individual, population, species, or clade) is performed, it facilitates the interpretation of comparative observations by explaining the historical non-independence of organisms. In particular, the first reason that the evolutionary history provided by phylogenetic inference is important is that it enables the study of evolutionary relationships for the levels of classification of organisms (i.e., kingdom, phylum, class, family, genus, species, intraspecific population). The second reason is that it not only allows the evolutionary pattern of multigene families to be clearly understood, but also explains the evolution of adaptive traits at the molecular level, ancestral character states, the timing of species divergences and variation in evolutionary rates. Currently used major methods for phylogenetic inference including distance-based methods such as neighbor-joining (NJ), character-based methods such as maximum likelihood (ML), maximum parsimony (MP), Bayesian inference, and minimum evolution, can draw valid phylogenies when performing phylogenetic estimation of enough nucleotides or amino acids but may not provide a true topology when certain situations with different evolutionary rates between branches occur. Inaccurate estimation of phylogenetic trees leads to biased results, that is, erroneous estimation of evolutionary mechanisms, and the complexity of DNA sequence evolution and the molecular force acting on the sequence make phylogenetic inference a complex problem. As a huge number of biological sequences have recently been produced, phylogenetic analysis has become a sequence analysis technique that is meaningfully used in evolutionary studies. Moreover, the number of phylogenetic inference techniques has led to the

complex problem of judging the accuracy of phylogenetic reconstruction and to relevant studies being conducted.

1.1.2.1 Approaches to assessing phylogenetic accuracy

There are actually several properties for phylogenetic estimation, which are used as criteria for statistically evaluating the performance of both character-based and distance-based phylogenetic reconstruction methods. The first property is consistency; that is, a statistically consistent estimator approaches a true value as the amount of data increases. In other words, as the biological sequence becomes longer, a larger amount of data is analyzed, and the likelihood of drawing a true phylogenetic tree increases. According to the property of efficiency, the phylogenetic method outputs accurate results when the amount of data is limited. If the phylogenetic result converges quickly on the true tree, the estimator can be considered efficient. Although computing techniques have accelerated the speed of computers, it remains a practical criterion because several phylogenetic methods are computationally burdensome. The following criterion is robustness, which evaluates whether the phylogenetic methods can derive a true tree under conditions facing violations such as branch length variation and substitution frequencies. Finally, the property of computational speed refers to the speed at which the reconstruction methods reach the best tree. For example, the NJ method draws an optimal phylogenetic tree using the cluster algorithm; thus its computational speed is faster compared to those of other methods, and this speed using the Bayesian method that depends on the length of the chain generated by the Markov Chain Monte Carlo (MCMC) algorithm varies with the sample data. In order to judge phylogenetic accuracy and evaluate the reliability according to the results, several approaches such as computer simulation, experimental phylogenetics, statistical analysis, and congruence analysis, have been

studied, and widely used so far as a method to evaluate phylogenetic performance. First, the simulation study is one of the most important approaches dealing with general questions about phylogenetic accuracy. For example, it is a technique to judge phylogenetic accuracy by comparing the performance of phylogenetic inference methods such as ML, MP, and Bayesian inference. It is also evaluated by analyzing the effects of parameters affecting the performance of the phylogenetic method, tree topologies, and relative or absolute rates of evolution. Numerous simulation studies have already been conducted on the phylogenetic inference method (Felsenstein, 1988), and with the development of computers and algorithms, it is now possible to provide a performance evaluation of more diverse phylogenetic methods analyzing or simulating thousands of datasets. However, many systematists do not accept the results of simulation studies as phylogenetic assessment. Existing studies show that many reconstruction methods that implement phylogenetic inference perform well or do not perform well under specific conditions, and it is easy to check the optimal condition for the phylogenetic inference method that the researcher prefers. Thus, the various factors that affect phylogenetics can limit the evaluation, which can lead to a reliability problem. In addition, the simulation study is affected by several biases such as branching order, branch lengths, and number of terminal taxa, which make the interpretation of simulation results difficult or contradictory. In particular, the evolutionary model in phylogenetic analysis can be regarded as a representative bias of simulation studies, and most simulations studies use this as a method of estimating an evolutionary model consistent with the result of the phylogenetic inference. Nevertheless, many researchers have relied on simulation methods to evaluate their results because evolutionary history in phylogenetic estimation is generally difficult to observe directly, and this approach has provided a high level of insight into various phylogenetic algorithms. Despite

the usefulness of simulation studies, since it is difficult to vaguely manipulate and evaluate phylogenetic estimation and molecular evolution, an experimental approach considering the actual case of evolutionary events was devised to avoid estimating the approximate value of biological evolution. This approach creates the evolutionary history of biological entities (i.e., clades of organisms) in a laboratory based on known history and evaluates the performance of the phylogenetic method for phylogenetic reconstruction. In other words, researchers can determine whether phylogenetic results are accurate, because experimental phylogenetics can reveal evolutionary history a priori in the laboratory. In addition, the accuracy evaluation using the statistical approach mainly deals with phylogenetic accuracy within a particular case rather than general conditions in which the phylogenetic method performs well or poorly. Additionally, the most well-known method of evaluating the performance of the phylogenetic tree using statistical analysis is the bootstrap test. The statistical approach usually maximizes the statistical reliability of the bootstrap tree when the sequence group in the dataset is monophyletic, and evaluating a group that spans multiple regions or the entire phylogenetic tree is less accurate because bootstrap values for unclear relationships are assigned to various groups. Therefore, the bootstrap test can be regarded as suitable for evaluating a small part of the phylogenetic tree, and researchers should use a statistical analysis that fits the condition of the datasets in determining phylogenetic accuracy. The final evaluation method is the congruence study, which approaches the problem by finding a common pattern of the tree topologies within phylogenetic trees created from multiple independent datasets. If multiple phylogenetic trees estimated from independent datasets represent the same pattern of relationship, this is strong evidence of the accuracy of the phylogenetic method. When a historical taxonomic sample sequence is provided, it is the best evidence of evolutionary accuracy if the

phylogeny based on morphological characteristics and the sequence-based phylogeny including orthologous genes show congruence (Jessica W. Leigh, 2011). Congruence justifies multigene phylogeny or phylogenomics and is a concept mainly applied to evolutionary biology as a basis for research on coevolution, lateral gene transfer, and evidence for common descent. However, two contradictory results arose from entering the current phylogenomic era, which considers multigene datasets (Kuck, Struck, 2013). First, congruence can be achieved as a result of phylogenetic analysis due to sufficient data (Gee, 2003). The other result began to occur due to the accumulation of systematic biases such as increased substitution rates, compositional saturation, and heterogeneity (Jeffroy et al., 2006). In other words, it has become indispensable to study the effect of systematic biases in the dataset to evaluate the congruence of phylogenetic analysis.

1.1.2.2 Influences of phylogenetic tree accuracy

The preparation and study of the factors affecting the accuracy of reconstructed trees through the above-mentioned approaches is a significant process. Tree uncertainty is expected inaccuracy, such as systematic error and random error, which is not expressed as a specific element but can be judged by using various parameters, such as the amount of input data (i.e., molecular sequence length for inference of evolutionary time and amount of change), divergence between sequences, model of evolution, and tree searching algorithm. First, in terms of the evaluation of phylogenetic accuracy, taxon sampling that considers the amount of input data includes the taxon informativeness problem for missing data and is the most significant factor affecting the accuracy. Phylogeneticists recognized that datasets containing a large number of taxa have created a more complex computational problem for phylogenetic analyses, and since then, numerous taxon sampling studies

have demonstrated that dense taxon sampling, as a way of introducing additional taxa improves the phylogenetic accuracy as the size of the dataset to be analyzed increases. In particular, dense taxon sampling affects the branch-estimation of the phylogeny, and this branch length not only provides information on the amount of genetic variation occurring across the phylogenetic tree but also plays an important role in direct inferences about evolution. Therefore, if the amount of available information in terms of dataset size is small, it is difficult to infer unobserved substitution, and a node-density effect (NDE) that misleads the relationship between rates of molecular evolution and biodiversity occurs (Webster et al., 2003; Venditti et al., 2006; Hugall, Lee, 2007). Additionally, taxon sampling in terms of the accuracy evaluation of phylogenetic performance in modern bioinformatics includes searching genetic databases to obtain sequences that are of interest or beneficial to the researcher, which means that the reliability of sampling can be biased by the availability of related sequences. The divergence between sequences is one of the main research fields of taxon sampling and evolutionary biology in terms of the relationship between sequence similarity and accurate phylogenetic reconstruction. Many phylogenomic studies rely on the large-scale alignment of nucleotide and amino acid sequences identified by a sequence similarity search as a previous step of phylogenetic tree reconstruction. A number of studies have been conducted on this, and as a group of species with statistically significant similarity between sequences represents lower sequence divergence, the accuracy of the multiple alignment which can be defined as the SOP score, increases (Brandi L. Cantarel et al., 2006). As a result, true alignment is the basis for increasing phylogenetic accuracy, and sequence similarity has a significant impact on phylogenetic accuracy. Therefore, evolutionary biologists and phylogeneticists should not only select an informative sequence that considers the size of the datasets including missing data

and the variation problem that may occur during the sampling process, but also perform a proper MSA to draw the expected performance of phylogenetic inference. Meanwhile, most phylogenetic analysis methods including distance-based and character-based methods in the model-based approach rely on explicit statistical evolution models to build phylogenetic tree, and the appropriate phylogenetic analysis depends on the suitability of the dataset. The specific statistical model selected for the dataset depends on the size of the dataset, the level of divergence between sequences, the pattern of variation resulting from the evolution of the sequence and the nucleotide frequency pattern and is distinguished according to the number of parameters used to represent the evolutionary change in terms of the complexity. Realistic model selection considering these parameters can avoid model overfitting and phylogenetic bias problems, thus yielding a more reliable and accurate phylogenetic tree. For example, one of the model parameters that has a strong influence on phylogenetic estimation is among-site variation, which is a problem when substitution rates between tree branches are differ, and when such variation exists, the use of the best-fit model is essential to obtain an accurate phylogenetic tree. Finally, the phylogenetic reconstruction methods selected for drawing the phylogeny can be reflected in the accuracy. The four methods of the parsimony, distance, likelihood, and Bayesian, divided based on their completely different schema, result in fundamentally distinct phylogenies, and each has both advantages and disadvantages.

1.1.3 Application of phylogenetic evaluation

Phylogeneticists and bioinformaticians enable phylogenetic reconstruction by rapidly incorporating these molecular parameters into phylogenetic analysis through

advanced sequencing and computing techniques to access the molecular characteristics of living organisms. Unlike traditional morphology-based phylogenies, where the number of phylogenetically reliable characters was not sufficient, the number of genes with their own complex history, such as speciation, is gradually increasing in molecular data, which is important to phylogenetic accuracy. Clearly, it is very difficult to obtain and estimate the accuracy of unique events that occurred in the distant past, which play a role in genome evolution, organism diversification, and new cell function expression. As the accuracy evaluation of the printed phylogeny is an indispensable process to objectively understand a biology, various evaluation methods are being studied. However, despite the recent significant advances in the methods for evaluating phylogenetic accuracy, a wide range of defined approaches has not been properly applied to systematics. The evaluation of phylogenetic accuracy from the perspective of systematics can be largely divided into two parts based on the method used for phylogenetic analysis. In general, researchers in phylogenetics, including evolutionary biologists, have used model selection software to find the best-fitting evolutionary model for datasets before the phylogenetic tree building process. They have also estimated the reliability and accuracy of the tree as a result of the assessment method using the built-in statistical component for obtaining a true phylogeny in phylogenetic reconstruction software. Most of the software currently used in computational phylogenetics is limited to the statistical approach and focuses on the process of building the most accurate phylogenetic tree, because the currently available methods can reconstruct a general consensus tree. Nevertheless, the major phylogenetic evaluation methods may not accurately reflect the distribution of underlying characters in the molecular data, and this may be difficult to overcome when systematic biases are introduced in the early stages of analysis. Moreover, if

the appropriateness of the datasets is not considered, an uninformative tree can be produced.

As mentioned above, the use of a specific evolutionary model can change the results of phylogenetic analysis, and as these model-based approaches stand out in systematic biology, including phylogenetic estimation, the studies of selecting an evolutionary model for the distance-based, likelihood, and Bayesian methods have been the focus. Evolutionary models calculate the probability of change between the characters in the biological sequence underlying phylogenetic tree branches, and they affect phylogeny estimation, molecular clock tests, bootstrap values, posterior probabilities, and substitution rates. In fact, if the model is incorrectly estimated, branch lengths, the transition/transversion ratio, and sequence divergence are underestimated, while the rate variation strength is overestimated. The optimal evolutionary model for a specific dataset is rigorously selected through statistical testing, and it performs estimation to transform a complex problem into a computationally tractable problem. The degree of fit between various models and dataset (i.e., model selection between available evolutionary models) is determined by comparison through likelihood ratio tests (LRTs) or information criteria. In particular, care is required when selecting the best-fit model of heterogeneous data and combining other genes in the coding and noncoding regions. This is because different genomic regions have different selective pressures and evolutionary constraints, and as a result, one substitution model cannot be perfectly suited to all datasets. Given that any evolutionary model cannot be asserted as a true model for the data extracted by the researcher, model selection does not identify reality, but estimates it as closely as possible. Model suitability within phylogenetics is judged by comparative analysis using likelihood function (Felsenstein, 1981; Goldman, 1990) and information criteria approaches, and is implemented by using a

JModelTest (Posada, 2008) as the most representative software among various programs that perform the statistical selection of models. This program provides access to model selection by adding more statistical theories than the ModelTest (Posada, Crandall, 1998), which was most widely used for model selection. Most of the programs, including jModelTest, evaluate model suitability through the likelihood function, which uses the existing ML estimation to find the maximized parameter of the likelihood for a given dataset and select a statistically significant model through hypothesis testing with multiple models. In addition, jModelTest provides a series of LRTs: hierarchical LRTs (Frati et al., 1997; Huelsenbeck, Crandall, 1977; Sullivan et al., 1997) that calculate the difference between models comparing the hypothesis test hierarchically based on the presence of base frequencies, transition and transversion bias, invariable sites, and rate homogeneity among sites, and a dynamic LRT (Posada, Crandall, 2001) that compares the current model and hypothetical model with LRTs. Several model selection methods such as the Akaike information criterion (Akaike, 1973), Bayesian information criterion (Schwarz, 1978), and decision-theoretic performance-based approach (Minin et al., 2003) are also provided.

On the other hand, phylogenetic reconstruction software, which many researchers mainly use for phylogenetic inference, provides accuracy assessment embedded in the program as a component with one function so that statistically significant phylogenetic results can be printed (Table 1.1). The method for evaluating the reliability of phylogenetic trees most commonly used so far is the bootstrap test (Efron, 1982; Felsenstein, 1985), which judges the accuracy of statistical estimation. Standard bootstrap (SBS) is useful in complex nonparametric estimations, and it operates by randomly restoring and extracting bootstrap replicates of the same size as the original datasets. As the bootstrap test is applied to the estimation of the

phylogenetic tree, it is possible to create bootstrap replicates by sampling the restored and extracted sites from observed molecular data; thus, a psudo sequence of the same size as the original molecular data can be produced. Each bootstrap sample is inferred using a specific phylogenetic reconstruction method, and the sampling distribution of the phylogeny estimated above can be known by the empirical distribution of bootstrap estimates. SBS is performed in most phylogenetic tree-making software programs including PAUP* (Swofford, 1993) and Mesquite (Wayne P. Maddison, David R. Maddison, 2001), and these programs provide a reliability evaluation for the observed clades of the tree based on the ratio of bootstrap trees extracting the same clade. Specifically, programs such as RAxML8 (Stamatakis A, 2014) and MOLPHY (Adachi, Hasegawa, 1992), which perform phylogenetic estimation under the ML method, implement the resampling estimated log-likelihood (RELL) method (Kishino et al., 1990) or rapid bootstrap search algorithm (Stamatakis A, 2008), which estimate the bootstrap probability of a tree without repetition of bootstrap resampling, because a computational burden is required for ML estimation for bootstrap samples. Meanwhile, as Bayesian inference was applied to computational phylogenetics, an uncertain phylogeny was inferred as the probability distribution. The Bayesian approach based on Bayes's theorem calculates the posterior probability (PP) P(A|B) for the tree by combining the prior probability P(A) of the tree and the likelihood P(B|A) of the data, and it operates by estimating the best phylogeny with the maximum PP representing the probability of the true tree. As various phylogenetics software use bayesian inference, the number of molecular phylogenetic parameters that need to be estimated has rapidly increased, and thus, the problem of computational complexity has been solved. The newly designed MCMC approach (Yang, Rannala, 1997) is a method of extracting the sample of the desired stationary distribution reaching the known target distribution

from the probability distribution based on the composition of the Markov chain. At this time, the Metropolis-Hasting algorithm is used to perform the step of the MCMC method. After n generations, which randomly propose a new tree topology or a new value for a model parameter, if the Markov chain converges to the stationary distribution, the state generated by the Markov chain can be considered as a sample of the target density and can be repeatedly sampled. For example, MrBayes, which performs Bayesian inference for phylogenetic estimation uses the standard MCMC method to calculate the PP of a tree, and many programs, including a BATWING (Wilson et al., 2003) and BEAST (Drummond et al., 2012), estimate the reliability of a phylogeny using the MCMC approach connected to the Metropolis algorithm.

1.1.4 Source of systematic biases affecting a phylogenetic accuracy and phylogenetic example

The use of the large multigene datasets has been successfully applied to solve the evolutionary question that defines the evolutionary relationship between specific species, leading to the quite accurate estimation of phylogenetic inferences. However, as the size of the datasets increases and the growth of phylogenetically informative positions intensifies as a result of large-scale bioinformatics studies, the potential of systematic errors also rises. Therefore, the phylogenetic inference through expansion to such a large dataset does not necessarily lead to more accurate results. The following factors mainly depending on the accuracy of phylogenetic reconstruction can be classified into five terms: the quality of the sequence, the identification of exact homologous sites by sequence alignment, the regularity of the substitution process, biases such as consistency and efficiency of phylogenetic estimation methods, and sequence divergence. If the existing phylogenetic assessment methods

relied on the accuracy of a series of continuous processes that secure and align sequences and perform phylogenetic inference, the recent evaluation methods focus on understanding the characteristics of genetic markers to analyze the sequence quality caused by systematic biases affecting phylogenetic inference. This means that large-scale datasets provide benefits to phylogeneticists, but when an unexpected evolutionary relationship occurs, it reflects a violation of phylogenetic assumptions due to systematic errors. If the estimated phylogenetic result is violated by a specific evolutionary phenomenon, this result begins to represent a systematic error and leads to an inaccurate phylogenetic relationship (Figure 1.1). Systematic biases refer to inaccuracies that occur as a result of inappropriate modeling of biological phenomena (Romiguier et al., 2016) or methodological issues (Hosner et al., 2016), and these are represented in datasets as non-phylogenetic issues, such as saturation, missing data, compositional heterogeneity among species, and rate variation across lineages.

1.1.4.1 Compositional heterogeneity

Compositional heterogeneity is said to occur when the equilibrium state of nucleotide or amino acid frequencies varies across the phylogenetic tree. It is not defined in the stationarity state where the transition probability does not change, and it occurs as a result of non-stationarity evolution where the substitution pattern of the evolutionary tree is not uniform over time. Therefore, compositional heterogeneity has bias in branch length and topology, and the characteristic of non-stationarity means that compositional attraction occurs, which means that taxa with similar nucleotide compositions are grouped together even though the evolutionary distance between them is great. In fact, compositional biases are more prominent in nucleotide sequences than amino acid sequences, because the third codon positions,

which are rapidly evolving sites, accumulate mutational biases as a result of degeneracy of the genetic code. In terms of systematics, the evaluation of compositional heterogeneity is implemented by analyzing simulated data with different nucleotide compositions, AT-biased or GC-biased compositions, using the P4 phylogenetics library (Foster, 2004). Such quantitative characteristics can be expressed as one of the following skew values representing compositional heterogeneity: AT skew, GC skew, purine skew (G and A), pyrimidine skew (C and T), and keto skew (GT). For example, GC skew indicates that guanine and cytosine are overabundant or underabundant in specific regions of DNA or RNA, because nucleotides randomly distributed within the gene are in a state of nonequilibrium with different mutational and selective pressures. The combination of nucleotide frequencies (i.e., the base composition itself) is a standard as a marker that indicates an unambiguous phylogenetic signal, and GC content, which is the average GC% of one sequence of an alignment, and GC heterogeneity, which is the variance of GC% among sequences of an alignment, function as biases for phylogenetic reconstruction (Romiguier et al., 2013).

1.1.4.2 Saturation

When multiple substitutions occur, where the actual genetic distance of the sequences is underestimated within the multiple alignment, it is said to be saturated. In other words, a single nucleotide undergoing multiple changes before reaching the final nucleotide identity is called multiple substitution, and saturation characteristics are represented by mutations acting on nucleotide changes. Assuming there are no model violations, saturation results in a decrease in phylogenetic accuracy and random sequences leading to poorly resolved trees. Conversely, if there is a model violation, a systematic error emerges, and due to the fast evolution rate or long time span, long branches accumulate multiple substitutions, which is affected by long branch attraction (LBA). When phylogenetic reconstruction is performed without considering saturation in alignment, the probability of multiple substitutions makes the distance between taxa smaller than the true distance, and thus, genetic saturation hinders MSA as an essential process for phylogenetic analyses that relies on the comparison of homologous sequences. The number of substitutions of one nucleotide in each sequence is not indicated in the loci, and substitutions consequently reduce the amount of phylogenetic information contained in the sequence. Saturation can be evaluated by measuring the substitution rate of the biological sequence and the time that has passed since the divergence, and the divergence rate can be estimated from sources including ancestral DNA, fossil records, and biographical events.

1.1.4.3 Missing data

The issue of missing data, which reduce phylogenetic accuracy, has been an important one from the early studies of taxon sampling (1990-1999) to modern phylogenetics, and this problem is considered a significant obstacle to reconstructing

the phylogenetic relationship between fossil taxa and extant taxa. In general, missing data are used for phylogenetic analysis in the form of incomplete taxa or characters, and thus, shared missing data are not distributed randomly among taxa. Specifically, the missing data are expressed as empty cells and the special character "?" of the character-by-taxon data matrix in the nexus file format used in phylogenetic analysis (Figure 1.2). In molecular phylogenetics, the issue of missing data is a complex one that determines the phylogenetic study design, and it is one of the important problems in many molecular data matrices or it is deliberately excluded to avoid missing data. Research on the effect of missing data on phylogenetic analysis has been conducted, yielding two main results. The first is the result that the missing data appearing in the sequence have a deleterious impact on the phylogenetic accuracy (Huelsenbeck, 1991; Lemmon et al., 2009). Although the missing data does not directly affect phylogenetic estimation, using a small number of characters may lead to phylogenetic inaccuracy due to the lack of phylogenetic signal (Wiens, 2003; Philippe et al., 2004). In addition, although many previous studies have demonstrated that an increase in the amount of missing data decreases the phylogenetic accuracy (Wiens, Reeder, 1995), recent results have proven that a more accurate phylogeny can be drawn depending on the number of characters regardless of whether or not the missing data are included (Wei Jiang et al., 2014). However, the result may be significantly different depending on various parameters used to determine the accuracy of the phylogenetic inference, such as incomplete genes, number of characters, and phylogenetic analysis method. Since the amount of missing data has increased considerably compared to the previous one, the comparison of missing data between taxa can provide a more significant phylogenetic estimation than the existing evaluation.

1.2 Necessity of research

The molecular phylogenetics studied today basically involve the comparison of macromolecular sequences to estimate genealogical and evolutionary relationships, and this concept was first proposed by Francis Crick in 1958 (Table 1.2). As the protein sequence is determined with protocols developed by Fred Sanger and access to the sequences increases, many protein biochemists have constructed phylogenetic relatedness maps and phylogenetic trees based on amino acid sequences derived from various organisms. With the expansion of molecular phylogenetics, the next-generation sequencing (NGS) technologies and advanced computing techniques have led to the accumulation of a large number of biological sequences from various living organisms, but the activation of phylogenetics studies estimating the evolutionary relationships has also resulted in a variety of phylogenetic conflicts. In spite of the increased amount of available biological information, dependence on tree-making methods without consideration of data quality is resulting in inaccurate, contradictory results. In addition, the phylogenetic reliability issue is considered one of the important problems to be addressed in phylogenetic reconstruction studies. Phylogenetic inference is an inductive science that relies on sampling empirical data such as nucleotide sequences. In the natural sciences, it is essential to assess the quality of empirical data to detect differences in the quality of sampling data before and after generating results. When various factors that may cause phylogenetic conflicts such as homologous characters, sensitivity of tree-making methods to unequal evolutionary rates, biases of species sampling, unrecognized paralogs, and functional differentiation, are considered, more accurate and reliable phylogenetic reconstruction can be achieved. Evaluating phylogenetic reliability ultimately means inferring a more accurate evolutionary history by generating better phylogenetic

results through assessing the quality of the data to be employed in a study. On the other hand, bootstrap methods, which evaluate the credibility of existing phylogenies from the perspective of computational phylogenetics, are embedded in most phylogenetic reconstruction software programs and directly contribute to the shape and topology of the tree. There are also various model selection programs aimed at improving the accuracy of phylogenies by using different methods of estimating evolutionary models. Furthermore, studies on factors affecting the accuracy of the tree-making process have led to the development of phylogenetic software that evaluates and improves specific elements. However, the programs that developed so far do not perform analysis on molecular data including genetic markers at the beginning of the phylogenetic reconstruction stage. Since there is no program that evaluates the effect of phylogenetic signals that cause various conflicts other than model selection, evaluating the reliability of phylogenetic reconstruction with only existing pipelines of phylogenetic reliability can result in ambiguity problems. Therefore, analyzing nuclear markers used for phylogenetic analysis as well as sequence quality and accuracy of sequence alignment, is an important process to improve phylogenetic reliability. Basically, since the initial datasets of all researchers imply the possibility of containing noise, it is essential to derive true phylogenies by considering potential problems known as systematic biases or phylogenetic biases that cause phylogenetic inaccuracy. Phylogenetic incongruence, in which branch orders collide with different evolutionary mechanisms between phylogenetic trees by several candidate nuclear markers can occur. In addition, an apomorphy that exists in all descendants within a specific clade is estimated by the erroneous grouping by the characteristics shared by two or more taxa with the initial common ancestor (i.e., the symplesiomorphy, which is shared by descendants who diverged from the common ancestor). When this symplesiomorphy is introduced, a

data quality-based program is useful (Figure 1.3). Until now, sequence-based approaches, such as variation in taxon sampling, removal of fast-evolving species, genes, or sites, and efficient detection of multiple substitution, have been applied to continuous research to resolve these systematic errors. Thus, SeqVis (Ho, 2006), which visualizes the composition of biological sequences and detects compositional heterogeneity, and BaCoCa (Kuck, Struck, 2013), which evaluates phylogenetic reconstruction by visualizing and calculating potential biases causing various systematic errors have been developed as a programs that consider phylogenetic biases. However, this program requires the PERL interpreter and R package to be installed and it is complicated to use by entering the instructions, resulting in low usability and efficiency from the perspective of experimental biologists. At this point, phylogeneticists and associated researchers need not only the ability to evaluate and analyze reliability at the sequence-level, but also user-friendly phylogenetic evaluation software. Additionally, since the potential problem of systematic errors still remains, it is vital to develop a reliable program that improves the credibility of phylogenetic reconstruction through multigene datasets.

1.3 Research objectives

The ultimate goal of this study is to develop a software that estimates the reliability and accuracy of the phylogenetic tree as specific values, which is a statistical parameters for molecular characters in the nuclear marker files. Thus, it also aims to estimate the reliability of a specific clade by using phylogenies for evolutionary relationships of living organisms and parameters before the phylogenetic reconstruction through a congruence approach and a statistical approach. Using this program, the phylogenetic accuracy is inferred by finding the optimal combination

within the parameters of systematic biases for gene markers. The first objective of developing software is to design a component that contains the function of converting an multi-aligned dataset of gene markers into a file with tractable structure. In addition, it converts a multiple alignment file format (.fasta) into a simple structured file format and helps facilitate the next analysis using the processed data. Researchers use multigene datasets for phylogenetic estimation, and since the MSA file format is a structure that makes the mathematical approach of molecular characteristics difficult, it is necessary to convert the data for easy handling. The second objective is to calculate and estimate statistical properties, which indicate potential biases that infer phylogenetic accuracy using characteristics extracted from the raw data and provide the corresponding results in the form of an independent file. The last objective is to integrate the data processing logic, which implements parsing, and data analysis logic, which calculates specific properties into a single module program. The researcher estimates the reliability of phylogeny through the values provided as a result and can reconstruct phylogeny with improved accuracy using a program such as Phyutility that manipulates molecular data and alignment data (Smith, Dunn, 2008).

Consequently, the program created by this study will serve as a part of the pipeline that can estimate and improve phylogenetic reliability within the phyloinformatics workflow and will be developed in a user-friendly standalone structure so that it does not use any other installer packages and libraries. In addition, this program will be a standalone software that is intuitively easy to understand from the perspective of experimental biologists who are not familiar with analytical methods using software packages and will be able to estimate phylogenetic accuracy for evolutionary problems as a single integrated program. Furthermore, this study will propose a combination of systematic biases that can distinguish the best gene markers within a

series of taxa and provide contradictory topologies when using this software. If related datasets are accumulated through this study in the future, it will serve as a phylogenetic reliability program that can estimate more accurate phylogeny using systematic biases.

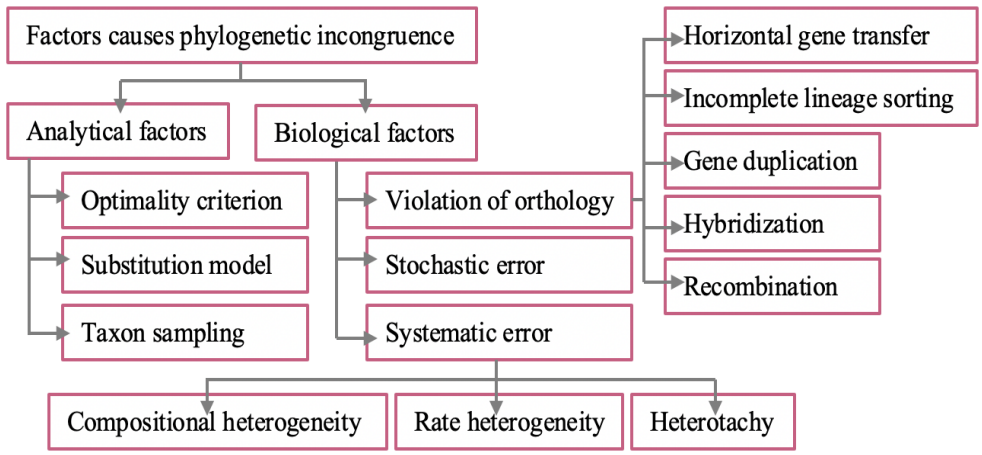


Figure 1.1 Summary of negative factors causing systematic errors. Implication of careful taxon sampling for hypotheses supported by phylogenetic inference.

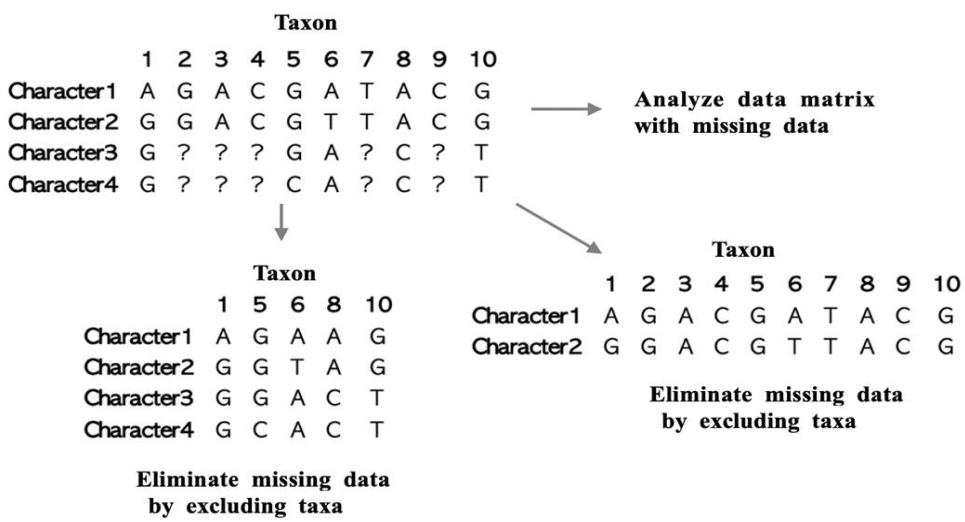


Figure 1.2 Hypothetical example illustrating missing data in phylogenetic analysis. Taxa 2, 3, 4, 7, and 9 lack data for Characters 3 and 4. If the researcher includes all of these data in a single analysis, there will be missing data cells ("?"). A researcher might choose to deal with this situation by deleting these taxa, deleting Characters 3 and 4, or simply including all the characters and taxa.

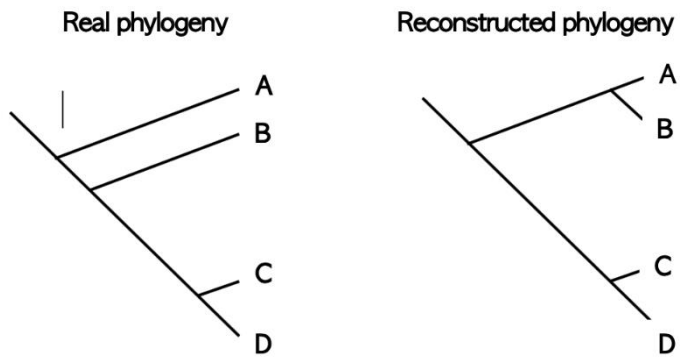


Figure 1.3 Dendrograms of real phylogeny and reconstructed phylogeny by symplesiomorphic trap. A-D are recent species; branch lengths symbolize the divergence time. The wrong grouping {A, B} in the reconstructed phylogeny is produced by analogies or symplesiomorphies shared by {B, C, D} eroded on the stem lineage of {C, D}.

Table 1.1 Program for phylogenetic reconstruction and accuracy

Software	Discription	Assessment
PhyloBayes3	A software package can be used for conducting Bayesian phylogenetic reconstruction and molecular dating analyses, using a large variety of amino acid replacement and nucleotide substitution models	Markov Chain Monte Carlo (MCMC)
RAxML	A program for sequential and parallel Maximum Likelihood based inference of large phylogenetic trees	Bootstrap (RELL method)
BEAST	A software architecture for Bayesian analysis of molecular sequences related by an evolutionary tree	Modified MCMC (BF value)
PAUP*	A computational phylogenetics program for inferring evolutionary trees by implementing parsimony and other methods	Bootstrap support
MrBayes	A program for Bayesian inference and model choice across a wide range of phylogenetic and evolutionary models	Markov Chain Monte Carlo (MCMC)
PHYLIP	A computational phylogenetics package of programs for inferring evolutionary trees by using parsimony, distance matrix, and likelihood methods	Bootstrap support
BATWING	A phylogenetic inference program about population histories, evolutionary processes, and forensic match probabilities from DNA sequence	Markov Chain Monte Carlo (MCMC)
Mesquite	A modular, extendible software for phylogenetic analysis, designed to help biologists organize and analyze comparative data about organisms	Bootstrap support
MOLPHY	A package of programs for molecular phylogenetics based on the ML method	Bootstrap (RELL method)
MEGA	A program for estimating evolutionary distances, reconstructing phylogenetic trees and computing basic statistical quantities from molecular data	Bootstrap, standard error (least-squares method)
TREE-PUZZLE	A program package for quartet-based maximum likelihood phylogenetic analysis that provides methods for reconstruction, comparison, and testing of trees and models on DNA as well as protein sequences	Bootstrap (quartet puzzling values)

Table 1.2 Basic phylogenetic glossary

Molecular phylogenetics: the study of evolutionary relationships among organisms or genes by a combination of molecular biology and statistical techniques.

Phylogenomics: reconstruction of phylogenies using a large number of genes or genomics regions. Two fundamentally different approaches are used for reconstructing phylogenies from multiple datasets. In one, the phylogenetic reconstruction is done after the gene sequences are concatenated head-to-tail to form a super-gene alignment – called 'supermatrix' approach. In the other, phylogenies are inferred separately for each gene and the resulting gene trees are used to generate a majority rule consensus phylogeny – called 'supertree' approach. The size of homologous to several thousand species or genes.

Multiple sequence alignment (MSA): the process of aligning three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a linkage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequence's shared evolutionary origins.

phylogenetic tree: A graph that represents the branching patterns of evolution and relationships among organisms.

Tree topology: the particular branching pattern for a tree. A labeled topology represents the relationships among the taxa at the tips. An unlabeled topology has no taxa at the tips and thus consists only of the abstract tree shape.

Polymorphism: the different form arise from the same genotype. According to the theory of evolution, polymorphism results from evolutionary processes, as does any aspect of a species. It is heritable and is modified by natural selection. In genetic polymorphism, the genetic makeup determines the morph.

Recombination: the exchange of a segment of DNA between two homologous chromosomes during meiosis leading to a novel combination of genetic material in the offspring.

Bootstrap: A procedure that involves resampling with replacement of the characters of the phylogenetic matrix to reproduce a number of matrices. Phylogenetic trees are then inferred from each resampled phylogenetic matrix. The number of times that a node appears in each of the resampled matrices is the 'bootstrap value' of the node.

Phylogenetic incongruence: two or more phylogenetic trees are said to be incongruent when they exhibit conflicting branching orders (i.e. topologies) and cannot be superimposed. This implies that at least one node (bipartition) present in one tree is not found in the other(s), where it is replaced by alternative groupings of taxa.

Model of sequence evolution: a statistical description of the process of substitution in nucleotide or amino acid sequences. Complex models better approximate the evolutionary process but at the expense of more parameters and computational time.

Table 1.2 Basic phylogenetic glossary (continued)

Systematic error: the error in phylogenetic estimation that is due to the failure of the reconstruction method to account fully for the properties of the data.

Stochastic error: the error in phylogenetic estimation caused by the finite length of the sequences used in the inference. As the size of the sequences increases, the magnitude of the error decreases (stochastic error $\propto \frac{1}{\sqrt{n}}$).

Phylogenetic networks: a graph used to visualize evolutionary relationships between nucleotide sequences, genes, chromosomes, genomes or species. They are used when reticulate events such as hybridization, horizontal gene transfer, recombination or gene duplication and loss are believed to be involved.

Monophyletic: monophyletic taxa include all the species that are derived from a single common ancestor.

Polyphyletic: taxon is composed of unrelated organisms descended from more than one ancestor.

Non-phylogenetic signal: The combination of different kinds of structured noise (e.g., undetected homoplasies) that compete with the genuine phylogenetic signal during tree reconstruction. Even if the non-phylogenetic content is partly a property of a multiple sequence alignment, the non-phylogenetic signal actually inferred heavily depends on the method and the model of evolution selected.

CHAPTER II.

MATERIALS AND METHODS

2.1 Dataset definition and data collection

Prior to designing the data collection, the topology of taxa in nuclear markers affected by systematic errors may be misplaced, and these biases also affect the placement of unbiased taxa. The datasets used in this study belong to the somewhat well-known Terebelliformia (Annelida), Daphniid (Arthoropoda) and Mammalia, and taxa with misplacement problem within each clade were included in the analysis. The first dataset used is a suborder, consisting of species belonging to Ampharetidae, Pectinariidae, Terebellidae, Trichobranchidae and Alvinellidae, which are the five family of terebelliformia (Zhong et al., 2011). Depending on the position of Trichobranchidae within this clade, it is divided into TriAA hypothesis forming a sister relationship with Alvinellidae and Ampharetidae, and TriTer hypothesis forming a sister group with Terebellidae. In this study, the two hypotheses caused by gene markers representing incongruent topologies as one dataset were assessed using systematic biases to select more accurate gene marker and to evaluate phylogenetic reliability. Specifically, Sipuncula was added as an outgroup taxon in addition to the 5 families mentioned above, and four gene markers including EF1α (1163 positions, 6 species), 18S rDNA (1897 positions, 7 species), mtDNA (16580 positions, 7 species), and 28S rDNA (4387 positions, 7 species) were collected (Zhong et al., 2011). First, elongation factor 1α (EF1α) is a highly conserved protein

involved in translation and has been significantly used as a phylogenetic marker (Roger et al., 1999). 18S ribosomal DNA (18S rDNA) is one of the most frequently used marker in phylogenetic studies and is particularly widely used to reconstruct the relationships between vertbrates due to slow evolutionary rates. Mitochondrial DNA (mtDNA) is also one of the well-known markers, and its molecular content is highly conserved among organisms, and has no introns, very few duplications, and very short intergenic regions (Gissi et al., 2008). 28S ribosomal DNA (28S rDNA), together with 18S rDNA, constitutes a subunit of the eukaryotic cytoplasm ribosomes and functions as a marker responsible for phylogenetic reconstruction of various organisms such as protists, fungi, and vertbrates. The second dataset consists of species belonging to Daphniid (Crustacea), and divided into two controversial topologies with generally expected clade closely related to Daphnia laevis and Daphnia dentifera and a clade containing systematic error closely related to Daphnia laevis and Daphnia occidentalis. 16S rDNA (502 positions, 9 species), 28S rDNA (4702 positions, 10 species) were used to estimate more accurate phylogeny between gene markers that cause controversial clades. The last dataset used in this study focused on taxa to identify Glire hypothesis, which belongs to the Mammalia class but has been controversial in misplacement between the Lagomorpha (rabbits) and the Rodentia (mice, rats and guinea pigs). Lagomorphs and Rodents forming a clade of Glires in Mammals have been strongly supported as sister group by specific gene markers, whereas other markers do not support their monophyletic group. First, Lagomorpha consists of two recognized families: Ochotonidae (pikas) and Leporidae (hares and rabbits), which consists of 29 species in 1 genus and 58 species in 11 genera. On the other hand, Rodentia contains 2055 living species in 27 families that comprise 454 genera and accounts for about 40% of the total mammal species. The lack of diversity in Lagomorpha as described above represents a remarkable

contrast with evolutionary radiation occurring in Rodentia, and as a member of Glires, Lagomorpha increased the need for research on the evolutionary mechanism that influences the diversification of Lagomorphs (Halanych, 1998). Reflecting this controversial clades, the dataset was based on the utilization of candidate genetic markers including nuclear gene and mitochondrial gene of the Lagomorphs, Rodents, Primates and tried to reveal whether the phylogenetic position is effective in Glires through a molecular systematics approach. Since a limited amount of data in taxon sampling prior to phylogenetic analysis can lead to stochastic errors, 18 phyla of Mammalia including 52 lineages were collected (Madsen et al., 2001). The marker alpha-2B adrenergic receptor (A2AB) gene (2391 positions, 30 species), marker interphotoreceptor retinoid binding protein (IRBP) gene (9719 positions, 27 species), marker von Willebrand factor (vWF) gene (8663 positions, 30 species) and a concatenation of three marker mitochondrial genes (2789 positions, 26 species), which are 12s ribosomal RNA, tRNA valine (MT-TV), and 16s ribosomal RNA, were collected to infer the Glires hypothesis (Table 2.1). A2AB is a G-protein coupled receptor and a sub-type of the adrenergic receptor family that regulates neurotransmitter release from sympathetic nerves and adrenergic neurons. IRBP is a gene identified in most eutherians and is involved in the visual process that transports retinoids between the retinal pigment epithelium and photoreceptors of organisms. vWF is a blood glycoprotein involved in hemostasis and plays an important role in platelet adhesion of wound sites by binding to specific proteins such as factor VIII. The primary reason for collecting the A2AB, IRBP, and vWF gene markers in this study is that all three genes have successfully performed phylogenetic reconstruction of Eutherians at various taxonomic levels. Second, the sizes of the three genes and the number of variable sites are similar, and this feature is useful for comparison of phylogenetic performance. Finally, each gene is not genetically linked and does not

exhibit any biological interactions. In addition to the three nuclear genes, a mitochondrial gene marker that could be compared and analyzed was prepared to carry out a congruence approach. For mitochondrial genes to be used for this study, concatenated regions of 12S rRNA, tRNA valine and 16S rRNA were sampled, and this dataset used in the clade study of the Placental mammals, which is included all living mammals except marsupials and monotremes, was referenced (Madsen et al., 2001). First, 16S ribosomal RNA is the most useful macromolecule that performs phylogenetic analysis. Due to its high information content, conservative nature, and universal distribution, it is possible to explore distant relationships between organisms, and it has been used to analyze the Mammalia class and Glires clade. The 12S ribosomal RNA can be used to trace the history of more recent evolutionary events and can perform phylogenetic analysis at different levels of taxa such as families, genera, and species.

The data used in this study were collected from the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov). Corresponding fasta-formatted data was collected using a ncbi-acc-download version 0.2.6 Python tool. It has a mechanism to download sequences from GenBank, which is accessible through the NCBI entrez retrieval system, and execute them with Mac operation system terminal-based instruction such as ncbi-acc-download --format fasta “accession number” to collect all raw data to be used in the study.

2.2 Data processing and bioinformatics software used

Before defining raw data collected from the GenBank database as a dataset, an MSA that analyzes the homology of sequences is required. Many tree reconstruction methods and the current state-of-the-art phylogenetic approaches use the two-step

process of MSA to perform phylogenetic inference. The goal of the first step is to identify homologous characters between sequences to make a heuristic estimate of homologies for MSA, and the second step is to calculate the best-fit tree of observed sequences using the fixed MSA and probabilistic substitution model chosen by the researcher. Therefore, MSA is the fundamental approach of this study to evaluate systematic biases of multi-aligned sequences, and the true alignment of multiple sequences is very important because it directly affects the output of accurate phylogenetic trees. Multiple sequence comparision by log-expectation (MUSCLE) (Edgar, 2004), ClustalW (Thompson et al., 2002), multiple alignment using fast Fourier transform (MAFFT) (Katoh et al., 2002), Kalign (Lassmann et al., 2005), T-Coffee (Notredame et al., 2000) and so on are widely used as the software for sequence alignment, identifies similarities between query sequences, and MSA is performed with a bioinformatics algorithm that is distinct from each other. In this study, the ClustalW was used in the MSA software list, and MSA was performed by a setting gap open penalty of 15, a gap extension penalty of 6.66, and an IUB scoring matrix including a match score of 1.9 and a mismatch score of 0 as the alignment parameters. Before performing phylogenetic analysis, which estimates the evolutionary relationships among groups of organisms based on the dataset created through MSA, the MSA-associated format files such as the fasta format file (.fasta) and the ClustalW format file (.aln) were converted to a nexus format file (.nex) and a meg format file (.meg). A seqmagick version 0.6.2 (Matsen Group, 2016) was used in the conversion of file format by command-line instruction such as seqmagick convert fasta format file nex format file –alphabet DNA. Currently, as computer programs for phylogenetic reconstruction such as PAUP*4.0 (Wilgenbusch, Swofford, 2003), MrBayes (Huelsenbeck et al., 2001), Mesquite (Wayne P. Maddison, David R. Maddison, 2001), PHYLIP (Felsenstein J, 1989), and molecular

evolutionary genetics analysis (MEGA) (Kumar et al., 1994; Kumar et al., 2001) are widely used, PAUP*4.0, MEGA X, and MrBayes were used in this study. On the other hand, heterogeneous sequence divergence causes taxa misplacement along with strong biases in tree reconstruction. To improve this part with the Assessment Program for Systematic Error (APSE) developed through this study, the relationship between the results of the parameters for each gene and each phylogeny was analyzed and evaluated. An AliGROOVE version 1.07 (Meid et al., 2013) was used to calculate a random similarity score for heterogeneity in dataset. At this time, a pairwise sequence comparison analysis of nucleotide data was performed using Monte Carlo resampling with a simple match/mismatch score provided by AliGROOVE. Thus, the level of taxonomically heterogeneous alignment ambiguity was evaluated through the resulting scoring values. Additionally, the saturation level was evaluated using the substitution saturation test of DAMBE version 7.2.43 (Xia, 2018) to represent the validity of the convergence factor (C factor) as saturation potential biases by APSE in four datasets including three nuclear genes and a concatenated mitochondrial gene. This level was interpreted using the entropy-based index approach of substitution saturation (Xia et al, 2003). The substitution saturation approach is a method to test whether observed entropy has a significantly smaller value than full substitution saturation in a biological sequence. The ratio of entropy of observed substitution saturation to entropy of full substitution saturation was calculated by index of substitution saturation (Iss) value, and Iss.c value was defined as a critical index of substitution saturation. If Iss is greater than Iss.c, the dataset is in severe substitution saturation, and the opposite is evaluated as it experiencing little substitution saturation.

2.3 Phylogenetic reconstruction and accuracy assessment

For the character of the clade that can be confirmed in this study, three datasets such as Terebelliformia, Daphniid and Mammalia were constructed through taxon sampling, multiple sequence alignment, and data format conversion, and the optimal sequence evolutionary model for datasets was explored prior to phylogenetic analysis. These substitutional models represent a process in which one nucleotide of the DNA sequence is replaced with another character during evolution, and the use of a specific model changes the outcome of phylogenetic analyses. The statistical approaches that determine the appropriate model for true phylogeny are well-established through various computer programs, and in this study, statistical model selection was performed using the jModelTest version 2.1.7. For the phylogenetic analysis of all three aligned datasets, the maximum likelihood (ML) (Felsenstein, 1981; Kishino et al., 1990) was used by MEGA X. As the ML analysis in this study, the results of model selection for each modified dataset were selected differently according to the dataset. In addition, bootstrap values can be used to evaluate the accuracy of the tree generated, and the ML tree was performed with 100 bootstrap replicates. The Bayesian approach, based on Bayes theorem, is a method of calculating the PP distribution P(A|B) for the tree by combining the prior probability P(A) of the phylogenetic tree and the likelihood P(B|A) of the data. The PP of the tree indicates the probability that the tree is correct, and the tree with the highest PP is selected as the best phylogeny. Bayesian analysis in this study was performed using MrBayes version 3.2.7 and a Metropolis-coupled MCMC (MC^3) sampling approach to calculate the PP of phylogenetic trees from the posterior distribution. The prior probabilities of all

trees were given the default parameters including the flat Dirichlet probability density, and tree sampling was assigned every 20 generations with starting trees being randomly assigned. Finally, to check the consistency of phylogenetic results, Markov chains were performed for 200000 and 500000 generations, and the phylogenetic trees were printed by giving a burn-in value of 25% of the sampling.

2.4 Software development environment and allowable data

The APSE computer program constructed in this study was written in a high-level, general-purpose C++ programming language and was developed using the C++ standard template library (STL). Additionally, it was compiled using the g++ compiler based on Windows, Mac, and Linux operation systems. The program was built with specifications of a 2.3 GHz octa-core Intel i9 processor (CPU) and used 32 GB 2667 MHZ DDR4 RAM. The input logic (console class) with included parsing logic (parse class), the output logic (spread class), and the calculation logic (calvalue class) were separated into each component. A console class was placed between the I/O processing logic and the logic in charge of the central calculation of the program, and this program was designed to efficiently access the desired function through the menu linking system (Figure 2.1). Since the APSE has a stand-alone architecture, it does not require an additional external module or library program, and it is written to facilitate reusability and maintenance.

The phylogenetic reconstruction is generally analyzed in a collective unit such as collection of genes, and accordingly, the MSA file is used to prepare the initial dataset for phylogenetic analysis and has a significant effect on the result of

phylogeny. Multi-aligned files consist of the alignment of three or more biological sequences, and homology and relatedness between sequences can be inferred from the characters including nucleotides, amino acids, indel events, and gaps of the file. The fast-all (FASTA) file contains various alignment characters, and the multi-aligned file is a simple structure in which each sequence is separated by a description line through concatenates of multiple single FASTA files. The APSE reads a FASTA file and provides the calculated parameters as an output file through text format file (.txt), as it is in a one-dimensional or two-dimensional vector structure. In the case of filtering gap-rich taxa with high gap frequencies by user-defined threshold, a modified multi-aligned file is provided as a new dataset and a reasonable tree can be reconstructed by entering this newly designed file.

2.5 Assessment of systematic errors

The data quality was evaluated by calculating systematic errors of the entered dataset before the phylogenetic reconstruction process. First, the taxon-specific base composition, GC/AT content, and gap frequency were calculated and provided to confirm the nonstationarity of the base composition, which may cause erroneous phylogenetic inference. In addition to studying the effect on the phylogenetic accuracy represented by GC-rich or AT-rich genes, it has been found that the GC-rich region has a higher recombination rate than the AT-rich region, which can result in phylogenetic error. In the relationships between the number of gapped sites and phylogeny, a large number of gap was considered to have a strong influence on the phylogenetic accuracy, and the gaps resulting from deletions could contribute to the phylogenetic inaccuracy (Dwivedi et al, 2009).

The divergence and change of the base frequency in the DNA sequence represent unrelated clades as similar groups with strong statistical support due to evolutionarily unrelated similarities. Therefore, the convergence of nucleotide composition (CNC) among unrelated lineages can play a role as a factor affecting the performance of accurate phylogenetic reconstruction. Second, the compositional heterogeneity or compositional biases were calculated using the formula for relative composition frequency variability (RCFV) value (Zhong et al., 2011). RCFV is derived from relative composition variability (RCV), which defines the average composition variability between dataset separating taxa and is a method of calculating the result using base frequencies in addition to the RCV.

$$\text{RCFV} = \sum_{i=1}^{n} \frac{|\mu_{Ai} - \tilde{\mu}_A| + |\mu_{Ci} - \tilde{\mu}_C| + |\mu_{Gi} - \tilde{\mu}_G| + |\mu_{Ti} - \tilde{\mu}_T|}{n} \quad (1)$$

μ_{Ai} means the base frequency of adenosine for the ith taxon, and $\tilde{\mu}_A$ defines the average base frequency of A in the entire n taxa. Additionally, from the perspective of statistics, since the heterogeneity is related to the validity of the difference in statistical properties between any one part and any other part of the entire dataset, compositional biases can be evaluated by calculating the skewness change and skew value of the base composition. Various skew values can be defined to determine whether base composition biases exist between two nucleotides frequencies, and GC skew and AT skew are mainly used to describe the overall pattern of nucleotides. In particular, since major mutational biases in the mitochondrial genome exist at purine and pyrimidine frequencies respectively, AG skew and CT skew have been newly defined (Zhong et al., 2011).

$$\text{AT skew} = \frac{\mu_A - \mu_T}{\mu_A + \mu_T}; \quad \text{GC skew} = \frac{\mu_G - \mu_C}{\mu_G + \mu_C}; \quad \text{AG skew} = \frac{\mu_A - \mu_G}{\mu_A + \mu_G};$$

$$\text{CT skew} = \frac{\mu_C - \mu_T}{\mu_C + \mu_T} \quad (2)$$

The compositional heterogeneity in the sequences between taxa causes a systematic error of phylogenetic inference, and researchers use strategies to reduce the potential impact of this compositional heterogeneity in datasets or tree-building methods to calculate this bias. As a result, researchers can calculate the taxon-specific RCFV or specific skew to understand the compositional heterogeneity of datasets and evaluate the accuracy of the phylogenetic tree. Whether some or all of the sequences in the datasets have lost phylogenetic information due to substitution saturation is an important factor. Saturation as a statistical property can be defined as the result of multiple substitutions occurring at the same position in a sequence or identical substitutions in different sequences. In this case, the assessment of saturation is defined as the C factor, and the C factor can evaluate the degree of convergence of transition-transversion ratios as the genetic p distance increases.

$$C = \frac{\sigma(\frac{Ti}{Tv})}{\sigma(p)} \quad (3)$$

By calculating this formula, the degree of saturation of the DNA sequence can be determined. Saturation occurs when the transition no longer increases despite an increase in the genetic distance, because it indicates that multiple substitutions have occurred in the nucleotide position. Finally, the proportion of shared missing data was defined as the indel and gap event "-", with ambiguity state "N" and missing data "?" representing the uninformative state at the same position between two taxa as a pair. The uninformative state is a statistical property that

evaluates data quality by calculating the ratio of the region of an ambiguous state where the character state is not defined. It is important to understand the influence of the statistical parameters as deciding factors of biological sequences representing specific evolutionary phenomena, which shows phylogenetic conflicts in order to determine the reliable nuclear marker for the accuracy of tree reconstruction by overcoming systematic errors.

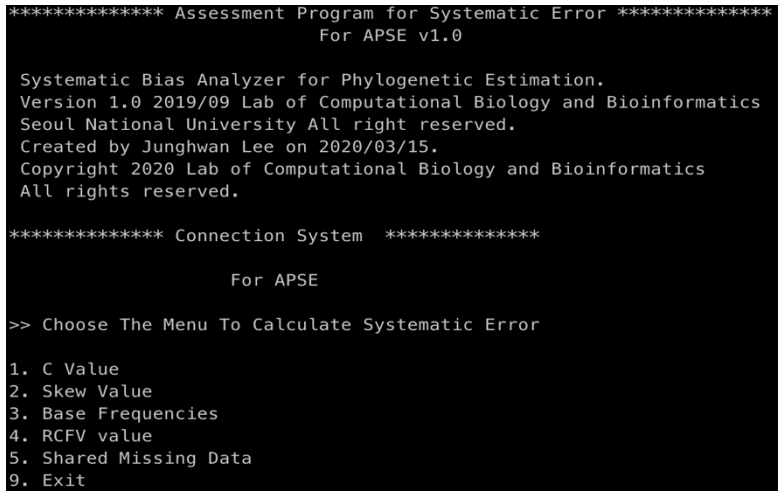


Figure 2.1 Console class for linking system between I/O logic and main function of program. The console class links the main function and component class. The program implements the calculation of potential bias for systematic errors depending on the user input.

Table 2.1 Mammalia taxa used in phylogenetic analysis

Order	Species	A2AB	IRBP	vWF
Carnivora	Cat, Felis catus	AJ251174	Z11811	U31613
	Dog, Canis familiaris			L76227
	Fox, Vulpes velox		AF179293	
	Harbor seal, Phoca vitulina	AJ251176		
Cetartiodactyla	Cow, Bos taurus	Y15944		AF004285
	Fin whale, Balaenoptera physalus	AJ251175		
	Hippo, Hippopotamus amphibius	AJ251178	AF108837	AF108832
	Humpback whale, Megaptera novaeangliae			AF226849
	Minke whale, Balaenoptera acutorostrata		U50820	
	Pig, Sus scrofa	AJ251177	U48588	S78431
Chiroptera	Big eared bat, Macrotus californicus	AJ251180		
	Flying fox, Pteropus hypomelanus		Z11809	
	Fruit bat, Dobsonia moluccensis			U31609
	Round eared bat, Tonatia bidens		Z11810	U31622
Dermoptera	Flying lemur, Cynocephalus variegatus	AJ251182	Z11807	U31606
Didelphimorphia	Large American Opossum, Didelphis virginiana		Z11814	AF226848
	Large American Opossum, Didelphis marsupialis	Y15943		
Diprotodontia	Kangaroo, Macropus rufus	AJ251183		
	Kangaroo, Macropus gaganteus			AJ224670
Hyracoidea	Rock Hyrax, Procavia capensis	Y12523	U48586	U31619
Insectivora	Eastern mole, Scalopus aquaticus			AF076479
	European mole, Talpa europaea	Y12520		
	Golden mole, Amblysomus hottentotus	Y12526		U97534
	Hedgehog, Erinaceus europaeus	Y12521	AF025390	U97536
	Madagascar hedgehog, Echinops telfairi	Y17692		AF076478
	Shrew, Sorex palustris		U48587	
Lagomorpha	Rabbit, Oryctolagus cuniculus	Y15946	Z11812	U31618
Macroscelidea	Long-eared Elephant shrew, Macroscelides proboscideus	Y12524		
Perissodactyla	Black Rhino, Diceros bicornis	AJ251184		
	Donkey, Equus asinus			U31604
	Horse, Equus caballus	Y15945	U48710	
	Tapir, Tapirus pinchaque		AF179294	
	White Rhino, Ceratotherium simum			U31604
Pholidota	Pangolin, Manis sp	AJ251185	AF025389	U97535

Table 2.1 (continued) Mammalia taxa used in phylogenetic analysis

Order	Species	A2AB	IRBP	vWF
Primates	Galago, Otolemur crassicaudatus		Z11805	AF061064
	Human, Homo sapiens	M34041	J05253	M25851
	Slow loris, Nycticebus coucang	AJ251186		
Proboscidea	African elephant, Loxodonta africana		U48711	U31615
	Asian elephant, Elephas maximus	Y12525		
Rodentia	Agouti, Dasyprocta agouti			U31607
	Guinea pig, Cavia porcellus	AJ271336		
	Mouse, Mus musculus		Z11813	
	North American Porcupine, Erethizon dorsatum		AF179292	
	Rat, Rattus norvegicus	M32061		U50044
Scandentia	Tree shrew, Tupaia glis		Z11808	AF061063
	Tree shrew, Tupaia tana	AJ251187		
Sirenia	Dugong, Dugong dugon	Y15947	U48583	U31608
Tubulidentata	Aardvark, Orycteropus afer	Y12522	U48712	U31617
Xenarthra	Three toed sloth, Bradypus tridactylus	AJ251179	U48708	U31603

CHAPTER III.

RESULTS

3.1 Phylogenetic analysis for incongruence between gene markers

Phylogenetic tree reconstruction is preceded by data quality assessment using posteriori criteria, which represents the suitability between the phylogeny and the nuclear marker. In this study, analyses of three datasets including Terebelliformia, Daphniid, and Glires clade, which has been a phylogenetic issue due to inconsistency between different genetic markers, was performed. First, through the ML method, Terebelliformia phylogenies were reconstructed to represent the controversial question of this clade, and a significant difference was shown between the gene marker EF1 α and 28S rDNA, indicating each specific hypothesis. In addition, the interior bootstrap proportion provided by ML analysis was basically judged by ML-BP (Maximum Likelihood Bootstrap Proportion, BP_{ML}) resampling 100 replicates. The results of the phylogenetic analysis for each gene marker are as follows. The tree of ML analysis for EF1α supported the TriAA hypothesis by indicating that Trichobranchidae (Terebellides sp.) had a closely related relationship with the group of Ampharetidae (Auchenoplax crinite) and Alvinellidae (Paralvinella hessleri). On the other hand, the phylogeny of 28S rDNA supported the TriTer hypothesis as the position of Trichobranchidae (Terebellides stroemi) forms a sister relationship with Terebellidae (pista cristata).

Unlike the phylogenies for the above two gene markers, 18S rDNA and mtDNA showed topologies that did not support a specific hypothesis generally accepted by researchers, and it was estimated that phylogenetic reliabilities were lowered compared to other markers due to the position of Alvinellidae (Paralvinella sulfincola) and Pectinariidae (Pectinaria gouldi) (figure 3.1). Second, The Daphniid phylogenies were reconstructed by ML method for two gene markers including 16S rDNA and 28S rDNA, and it was distinguished by specific association between the following three species such as Daphnia laevis, Daphnia dentifera, Daphnia occidentalis. As one of the two simulated marker data, 16S rDNA phylogeny represented an expected clade that has been recognized by existing studies as Daphnia laevis and Daphnia dentifera forming a monophyly (Angela R. Omilian et al., 2001). As for the phylogeny of 28S rDNA, Daphnia laevis builds a closer relationship with Daphnia occidentalis than Daphnia dentifera, and this topology is a result of the misleading clade caused by long branch attraction due to saturation as one of the systematic errors (figure 3.2).

Finally, Glires phylogenies showed a significant difference, which is a distinct incongruence, between nuclear gene IRBP and mitochondrial gene 12S rRNA-tRNA val-16S rRNA, and it is particularly important to understand the information biases and quality parameters of the sample when poorly supported relationships appear in clades including controversial phylogenies. The phylogenies that posed questions about the Glires hypothesis were reconstructed through the ML analysis and Bayesian inference to confirm the difference in support of clade for the hypothesis. Each tree building method provided the same results for each gene, while different topologies were resulted depending on the gene. For the comparative purpose of statistical confidence of interior branches provided as a result of Bayesian inference, the results were basically measured by

BMCMC-PP (Bayesian Posterior Probabilities calculated via MCMC sampling, PP_{Bay}). In general, the PP of Bayesian analysis prints a higher value than the ML bootstrap, and the PP, which determines the node reliability of mammal phylogenies constructed in this study, was also higher than the ML bootstrap frequency, which means that the PP strongly support these phylogenies. Therefore, the conflict between ML and Bayesian analyses can be considered to be caused by the overconfidence of statistical support shown by PP, whereas the result of the ML bootstrap was underestimated compared to PP. In other words, since the ML bootstrap support is relatively more conservative than the Bayesian inference support, the PPs are greater than the bootstrap proportion, and accordingly, since the two values indicate an inequivalent tendency in the measurement of confidence, the result is that type I error rates, which frequently reject true phylogeny, are higher in ML-BP than in BMCMC-PP. In addition, in the case of short internodes shown in the mammal clade of the ML method, the underestimate of the bootstrap can be estimated, and when analyzing the same number of characters, the result was more sensitive than that of the Bayesian method.

The results of phylogenetic tree analysis for each gene are as follows. The phylogeny of the ML method in which the first, second, and third codon positions for the A2AB were selected is the Lagomorphs (rabbits), which have been supported as a Glires member, forms more closely related relationships with Primates (humans, slow lorises, galagos), Scandentia (tree shrews), and Dermoptera (flying lemurs) than Rodents (guinea pigs, rats, mice). Lagomorphs showed a sister relationship with Scandentia (BP_{ML} = 29), and supported a monophyly with Primates, Dermoptera, Scandentia, and Lagomorphs. The Bayesian phylogeny also represented monophyly in Primates, Dermoptera,

Scandentia, and Lagomorphs (PP_{Bay}= 0.78), and the same result as those of the ML phylogeny were printed (Figure 3.3). The A2AB for mammals analyzed by the two tree-making methods showed that the position of the Lagomorphs was more closely related to Euarchonta, which is mammals containing the Scandentia, Dermoptera, and Primates, than Glires. It has been shown that this result does not support the EuarchontoGlires, which contains the Euarchonta and Glires and has been accepted in molecular studies based on branch lengths. Unlike the A2AB, which does not support the Glires hypothesis, the ML tree of IRBP has a closer distance to Rodents than Primates and Dermoptera (BP_{ML}= 25), and the Bayesian tree is also more closely related to Rodents than Euarchonta (PP_{Bay}= 1.0). The IRBP-represented Glires tree was supported by morphological synapomorphy, and monophyly of the Boreoeutheria clade including Euarchontoglires and Laurasiatheria was supported (PP_{Bay}= 1.0) (Figure 3.4). In terms of the two gene markers, the ML method did not guarantee sufficient reliability compared to the Bayesian method (BP_{ML}< 50, PP_{Bay}< 0.9), and IRBP was found to be the basis for supporting the Glires hypothesis and supports Euarchontoglires together. The vWF tree using two tree-making methods also showed a clade that did not support the monophyly of Lagomorphs and Rodents like the A2AB and formed a paraphyly with both Primates and Rodents from the viewpoint of Lagomorphs (Figure 3.5). However, the vWF phylogeny shown by ML analysis was not guaranteed to be accurate as it indicated a weakly supported node. In particular, the branch where the Lagomorphs is located acts as a cause of collapse of the cluster of the Lagomorphs, Rodents, Primates, and Scandentia (PP_{Bay}= 0.56). As can be seen, all phylogenies generated by the ML method other than the IRBP strongly or weakly support that the evolutionary distance shown by the topology

of Lagomorphs is closer to Scandentia, Dermoptera, and Primates than Rodents, and since the Glires tree was not constructed, this result does not support the Glires hypothesis. In the phylogenies analyzed by Bayesian inference, A2AB and vWF excluding IRBP also represented a topology similar to the ML method in terms of the position of Lagomorph. Next, to conduct a congruence study on the Glires hypothesis with the three gene markers used previously, the phylogenies analyzed by ML and Bayesian method were constructed as mitochondrial gene markers including the genomic regions of 12S rRNA-tRNA valine-16S rRNA. First, the The position of Lagomorphs formed a sister relationship with Primates (BP_{ML} = 35), and formed monophyly with several Primates and Scandentia (BP_{ML}= 13) in the phylogeny by ML method. This is similar to the phylogeny of concatenated three nuclear genes, which indicates that Lagomorphs are closer in evolutionary distance to Primates than to Rodents. The concatenated mitochondrial gene tree analyzed by Bayesian inference also represented similar results to the ML method. A sister relationship was found between Lagomorphs and Primates (PP_{Bay}= 0.96), and the position of Lagomorphs was more closely related to Primates, Dermoptera, and Bradypodidae (three-toed sloths) than Rodents (Figure 3.6). As a result, it can be concluded that the phylogenies of mitochondrial genes represented a well-supported clade compared to the nuclear genes, and do not support the Glires hypothesis because the Glires tree was not generated.

3.2 Data-quality analysis using systematic errors

In general, phylogenies constructed from large datasets tend to detect nonphylogenetic signals proving false results of phylogenetic analysis due to the

potential for weakly supported branches. Therefore, the systematic biases as potential errors of phylogenies were analyzed using APSE. First, by calculating the Convergence factor (C factor), which is the statistical property that compresses the information for the saturation level, for the taxa of Terebelliformia, the gene marker estimated to be the most accurate among the four markers was analyzed (Table 3.1). Within the Terebelliformia group, the min/max values of the C factor in EF1α supporting the TriAA hypothesis were 3.1534/4.2051 and 2.3844/16.4249 in 28S rDNA supporting the TriTer hypothesis were calculated. Therefore, it can be estimated that EF1α is a more accurate marker because the biases for the C factor of EF1α that satisfies the TriAA hypothesis, which is currently accepted among researchers, were significantly lower than that of 28S rDNA. However, it was confirmed that the biases for C factor of mtDNA were the lowest among the four gene markers, and C factor as a parameter reflecting systematic error did not select EF1α in Terebelliformia. Next, the validity of Terebelliformia phylogeny was confirmed through the difference in RCFV potential biases between gene markers reflecting base compositional heterogeneity. The average RCFV of EF1α for each taxon was 0.1639, which was significantly higher than 0.1131 of 28S rDNA and was the most heterogeneous among all markers. Similarly for RCFV, 0.0975 calculated by mtDNA is the lowest among markers, and as a result, it is estimated that the average of RCFV cannot accurately select gene marker within Terebelliformia. As a remarkable result from the viewpoint of taxon-specific RCFV, when the RCFV of taxa for all markers was compared, the more closely related relationship was formed between the taxa of three gene markers except for 28S rDNA, the more similar the RCFV was. In addition, nucleotide frequencies for the purpose of base composition analysis represented significant fluctuations across organisms, and such

divergence of nucleotide frequencies results in an erroneous phylogenetic inference. In particular, the analysis was performed focusing on the GC content and gap proportion, which could affect the selection of accurate markers by casing phylogenetic errors as biases for base frequencies. In terms of GC content, the average of EF1α was 51.10%, which was lower than 57.10% of 28S rDNA. However, as the average of mtDNA was 35.15%, the GC content could not select EF1α as a parameter reflecting systematic error. Interpreting the results for GC content, it can be concluded that 28S rDNA with a relatively high GC content will mutate faster than mtDNA or EF1α with a high AT content ratio due to the methylation tendancy of cytosine nucleotides, which will result in stronger saturation. On the other hand, when comparing the gap proportion, the average of EF1α was 3.39%, which was significantly lower than 24.54% of 28S rDNA. In addition, the average gap of EF1α showed the lowest value among all markers, and the result of selecting EF1α, which is estimated to have the most accurate, was represented by gap proportion. As a result of calculating shared missing data as a parameter for comparative analysis, there was a significant difference between gene markers. The average of shared missing data among the taxa of the Terebelliformia was 1.99% in EF1α, which was the lowest value among all markers, and the average of 28S rDNA was 16.57%. Consequentially, shared missing data discriminated EF1α, which has been accepted as accurate among other markers that build relationships in the taxa of the Terebelliformia. As the second dataset, systematic biases were analyzed for two gene markers including 16S rDNA and 28S rDNA that build Daphniid phylogenies (Table 3.2). First of all, the C factor of 16S rDNA was significantly higher than the value indicated by the 28S rDNA in all species except two species, and the average min/max was

also 6.1322/18.3071, indicating a relatively higher value than that of 28S rDNA. Therefore, it was estimated that C factor is a parameter that cannot select the correct gene marker 16S rDNA that has been accepted as Daphniid phylognies. In addition, to confirm the potential of compositional heterogeneity, RCFV between the two markers was comparatively analyzed. The average RCFV of 16S rDNA was 0.0876, which was significantly lower than that of 28S rDNA, and furthermore, it was confirmed that RCFV acts as a parameter for selecting accurate marker of the taxa of Daphniid. However, the more closely related within the Daphniid divided into the excepted clade or the misleading clade according to the position of Daphnia laevis, Daphnia dentifera, and Daphnia occidentalis, the taxon-specific RCFV was not similar. Next, the average GC content of 16S rDNA and 28S rDNA was 35.25% and 55.96%, respectively, and the difference in these ratios was significantly lower in 16S rDNA. The gap proportion also represented a significant difference when compared with 1.89% of 16S rDNA and 9.90% of 28S rDNA. As biases belonging to the base frequencies, the GC content and gap proportion were analyzed as parameters for selecting the gene marker 16S rDNA, which has been recognized to reconstruct the correct Daphniid phylogeny. In addition, when the proportion of the shared missing data of the taxa of the Daphniid was analyzed, 16S rDNA showed a ratio of less than 1.0% in the remaining species except Daphnia dentifera, and the average of 28S rDNA was 6.41%. Through the previous results, it was estimated that the proportion of shared missing data can select 16S rDNA with low biases, and it was analyzed that it funtions as a parameter of systematic error responsible for phylogenetic reliability.

Finally, for the purpose of identifying the best phylogeny in the controversial Glires clade, systematic biases were analyzed for the four gene markers in the

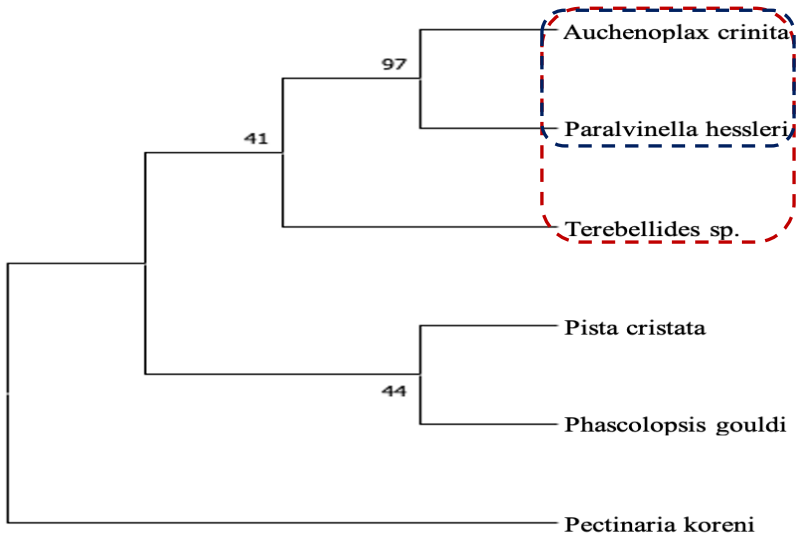
mammal dataset including the taxa of the Glires (Table 3.3). Since the Glires phylogeny has not been clearly identified for the position of the Lagomorphs so far, the corresponding phylogenetic accuracy was estimated through the parameters reflecting the systematic biases. In particular, the gene marker IRBP, which reconstructed the phylogeny supporting the Glires hypothesis of Lagomorphs having a morphologically closer evolutionary relationship with Rodents, and the concatenated mitochondrial gene marker 12S rDNA-tRNA valine-16S rDNA, which supported the phylogeny that Lagomorphs were closer to Primates than Rodents, were comparatively analyzed. First, from the perspective of the C factor, IRBP shows a large deviation such as 2.4641/82.9405 as a min/max value, and also represents high biases compared to 2.1348/6.0840 of the mitochondrial marker. In addition, the average C factor of IRBP was 50.6555, which is high compared to 3.2704 of the mitochondrial marker, and the saturation caused in this way reduces the accuracy of the phylogenetic signal. Therefore, through this analysis, it is estimated that the C factor functions as a parameter for selecting the mitochondrial markers that reconstruct the phylogeny, which is recognized as more appropriate in modern research. Next, the taxon-specific RCFV calculated as a property to evaluate the compositional biases of gene markers also shows a significant difference between the two markers, thereby reflecting each hypothesis. The average RCFV of mitochondrial markers was 0.0347, which had IRBP biases higher than 0.0061, thereby indicating rather large compositional heterogeneity. In the case of IRBP, the RCFV of 0.0041 of rabbits (Lagomorpha), 0.0053 of mice and 0.0052 of porcupines (Rodentia), and 0.0051 of tree shrews (Scandentia), were similar to those of 0.0390 of humans (Primates). Therefore, similar RCFV appears as each node is closely related within the clusters of Lagomorphs, Rodents, and Scandentia, which supports the

Glires clade in IRBP, and the topology for phylogenies is significantly affected by RCFV. On the other hand, in the mitochondrial markers that do not support the Glires tree, it can be estimated that correlations between the topology of the mitochondrial gene tree and the RCFV were insignificant through RCFV of the sister relationship between Lagomorphs and Primates and RCFV of neighbor taxa, including Rodentia and Scandentia. Therefore, mitochondrial markers through RCFV showed that the substitution pattern was not uniform across lineages of phylogenies, and that RCFV-selected IRBP as an accurate marker in the Glires clade. In addition, when comparing the biases of nucleotide frequencies between gene markers, the GC content and gap proportion showed significant differences and were thus important biases that could produce contradicting topologies. The average GC content and gap proportion of IRBP were 61.36% and 85.11%, respectively, which were significantly higher than the values of 39.69% and 8.37% of the mitochondrial markers, respectively, and they are thus the two biases of the selected mitochondrial markers that do not support the Glires hypothesis as the best marker. As the last biases, when comparing the character proportion with ambiguity state among taxa, which refers to shared missing data, the average values of 81.34% in IRBP and 4.81% in mitochondrial markers were seen. As a result of this analysis, the biases of the shared missing data of the mitochondrial marker were smaller than those of the IRBP, and the mitochondrial marker was distinguished as the best marker by the shared missing data. Additionally, although the criterion is different for each phylogenetic inference method, all methods treat gapped positions as missing data, and most phylogenetic software including MrBayes also includes gaps as ambiguity data for analysis. Since the result of the missing data proportion by APSE also treated the gap in alignment as missing data, this may estimate less accurate phylogenies when compared to

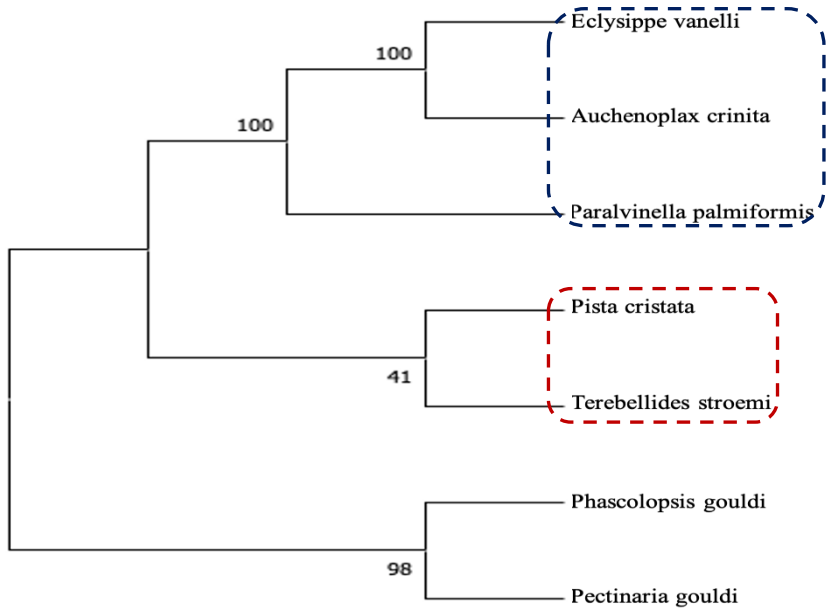
the probabilistic methods for phylogenetic analysis that deal with the gap as a phylogenetic signal (Dwivedi et al., 2009). As a result, it is estimated that the phylogenetic reliability of mitochondrial markers was higher when analyzing the relationship between parameters for the remaining genes that do not support the Glires hypothesis and IRBP that supports the Glires hypothesis.

When all systematic biases of the datasets for the three clades were summarized, the same single gene marker was not selected through all parameters provided by APSE (Table 3.4). First, combining all systematic biases on the EF1α and 28S rDNA, which draw a controversial topologies for Terebelliformia phylogenies, the max value of C factor, gap proportion, and shared missing data were a combination of parameters that are reflected in selecting the correct gene marker. Second, in the case of Daphniid, the gene marker 16S rDNA, which has been accepted as the best marker was selected through a combination of RCFV, GC content, gap proportion, and shared missing data. Finally, within the unresolved controversial problem among the Glires phylogenies, the reliability of phylogeny using the concatenated 12S rDNA-tRNA valine-16S rDNA marker was highly estimated by the combination of C factor, GC content, gap proportion, and shared missing data. Through this result, it was confirmed that phylogenetic reliability based on data quality can be estimated with the parameters provided by APSE, which evaluates systematic biases.

(a)


Auchenoplax crinita
Paralvinella hessleri
Terebellides sp.
Pista cristata
Phascolopsis gouldi
Pectinaria koreni
97
41
44

(b)


Eclysippe vanelli
Auchenoplax crinita
Paralvinella palmiformis
Pista cristata
Terebellides stroemi
Phascolopsis gouldi
Pectinaria gouldi
100
100
41
98

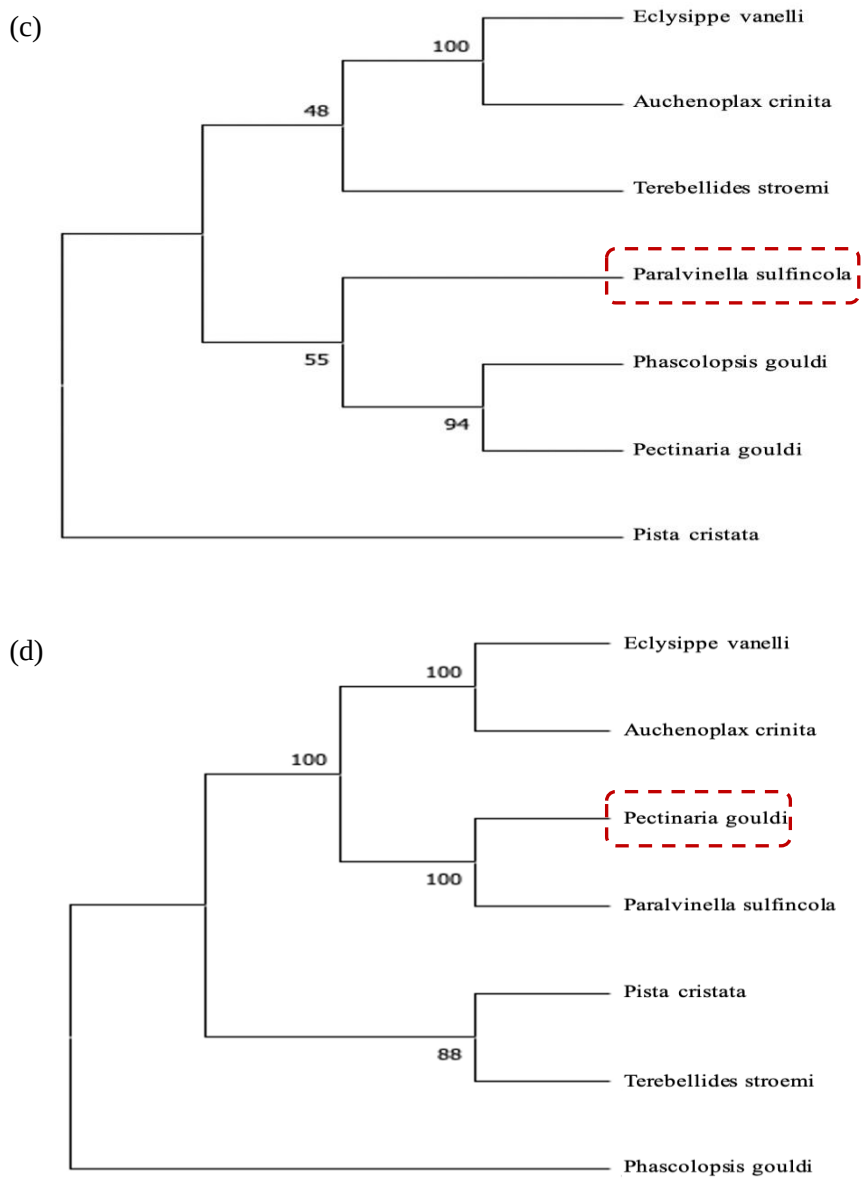


Figure 3.1 Phylogenetic trees based on four gene markers of Terebelliformia taxa. (a) The result of the ML method is visualized for interpreting the TriAA hypothesis based on EF1α. (b) The phylogeny of 28S rDNA is visualized for interpreting the TriTer hypothesis. (c) (d) The results of the phylogenetic tree of 18S rDNA and mtDNA are visualized.

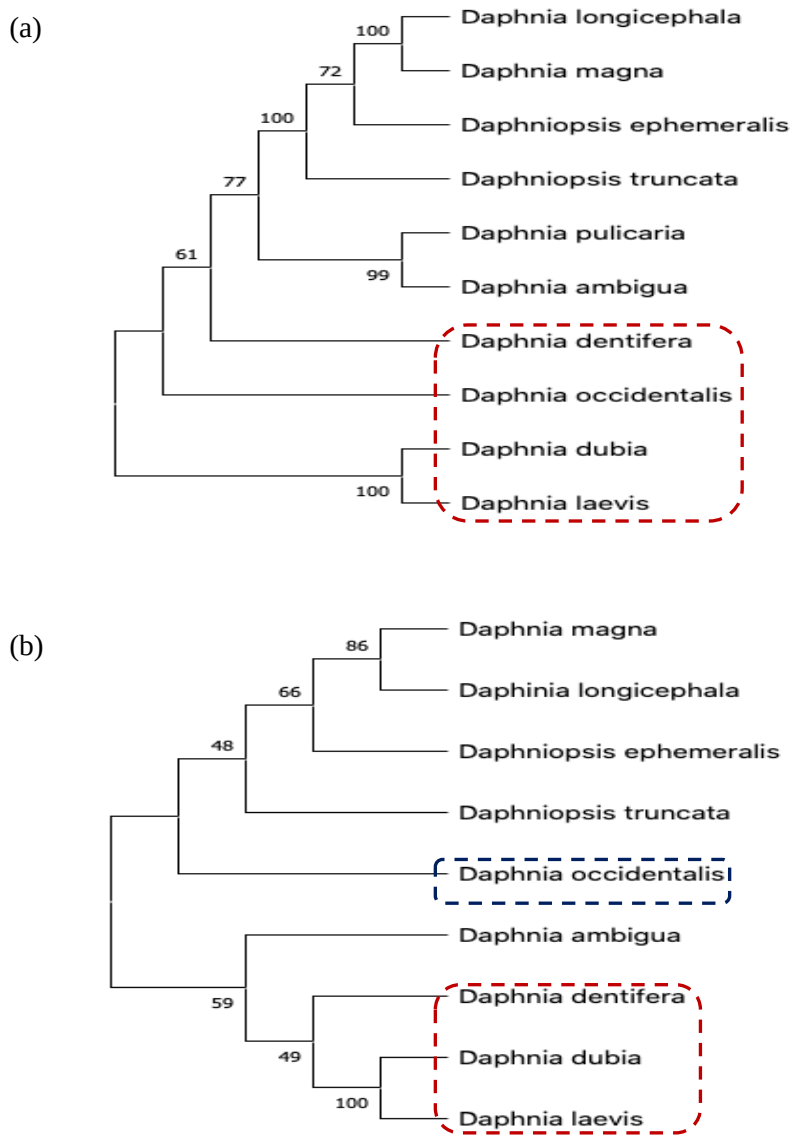


Figure 3.2 Phylogenetic trees based on two gene markers of Daphniid taxa. (a) The result of the phylogenetic tree based on 28S rDNA is visualized for interpreting the misleading clade that is grouped with Daphnia laevis and Daphnia occidentalis. (b) The result of the phylogenetic tree based on 16S rDNA is visualized for interpreting the expected clade that is grouped with Daphnia laevis and Daphnia dentifera.

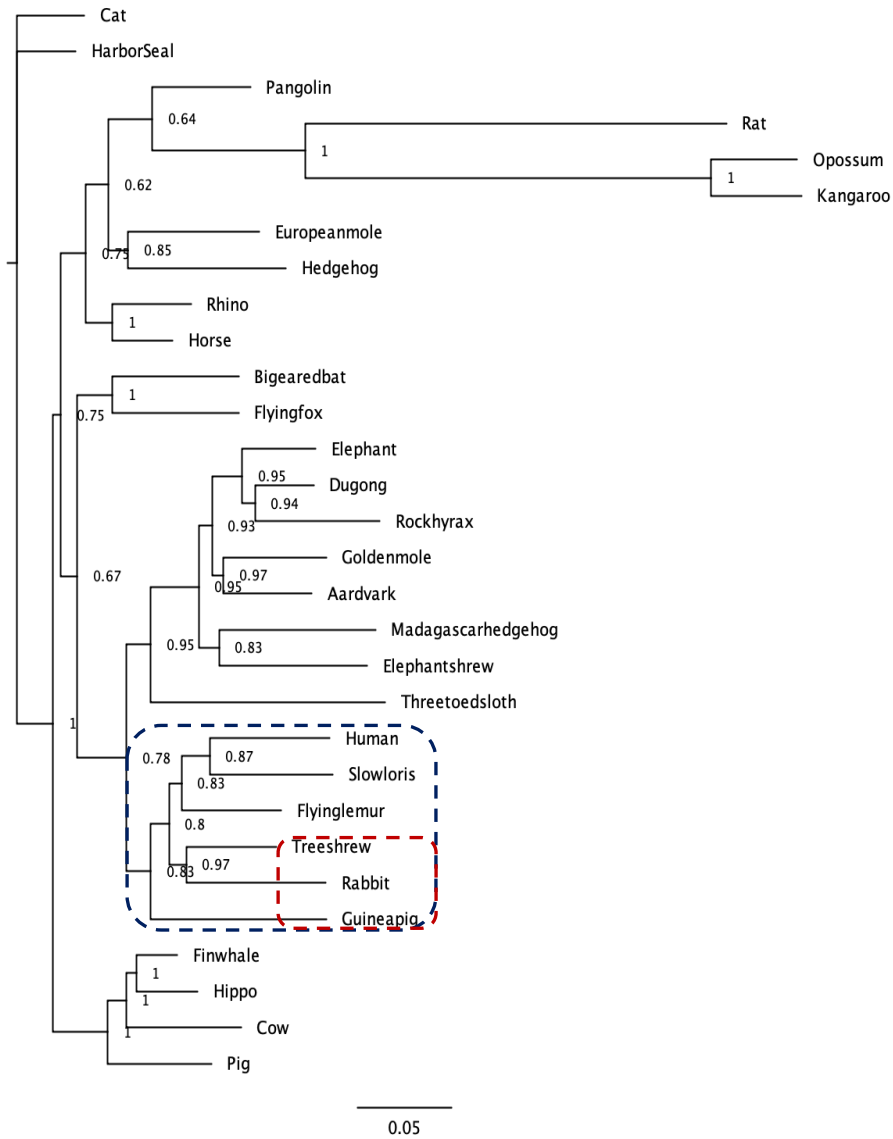


Figure 3.3 Phylogenetic trees based on A2AB of 30 mammal taxa. The result of the Bayesian inference method is visualized for interpreting the Glires hypothesis based on A2AB genes. It demonstrates that the evolutionary relationship between Lagomorph (rabbit) and Scandentia (tree shrew) is closer than that between Primates (human, slow loris).

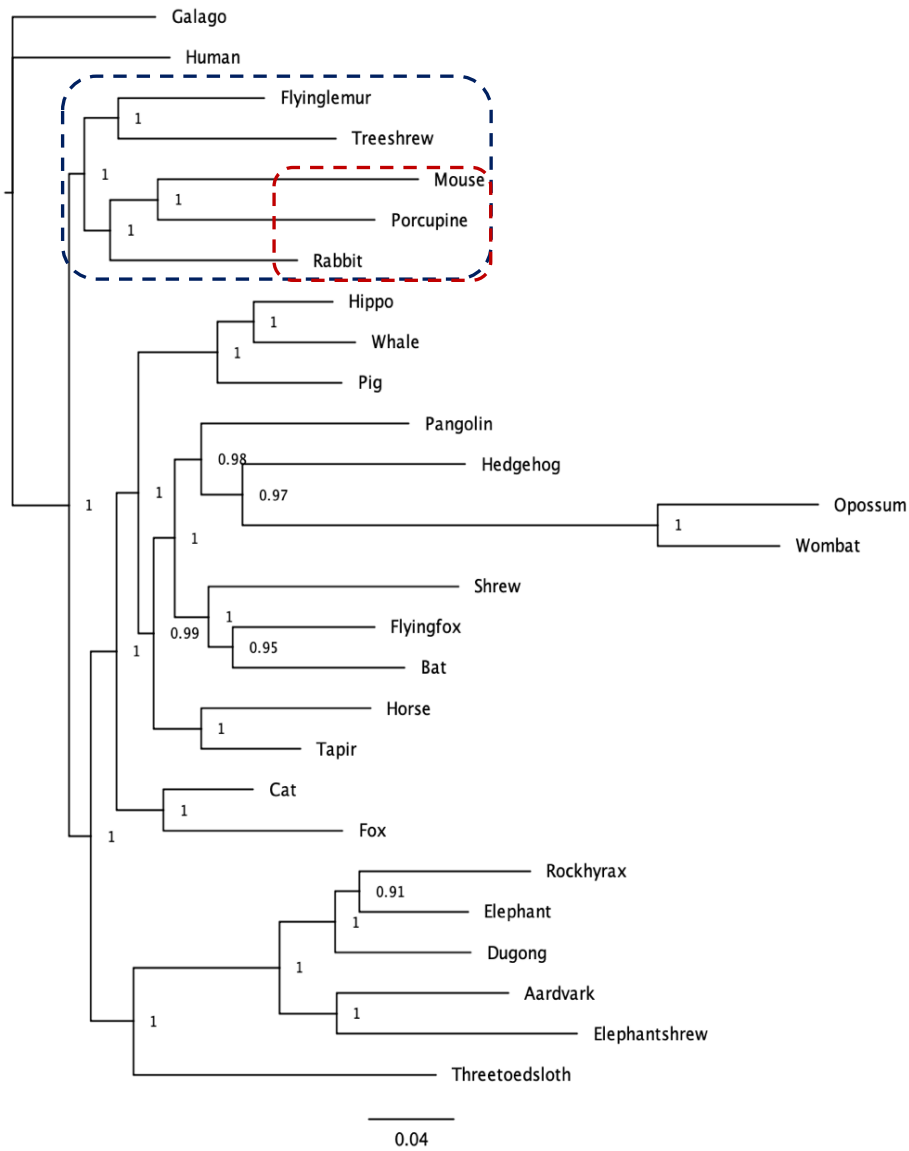


Figure 3.4 Phylogenetic trees based on IRBP of 27 mammal taxa. The result of the Bayesian inference method is constructed based on IRBP genes. It demonstrates that the evolutionary relationship between Lagomorph (rabbit) and Rodents (mouse, porcupine) is closer than that between Primates (human, galago). Although the phylogram visualizes a Glires tree, it is not a reliable tree because of its low values of PP.

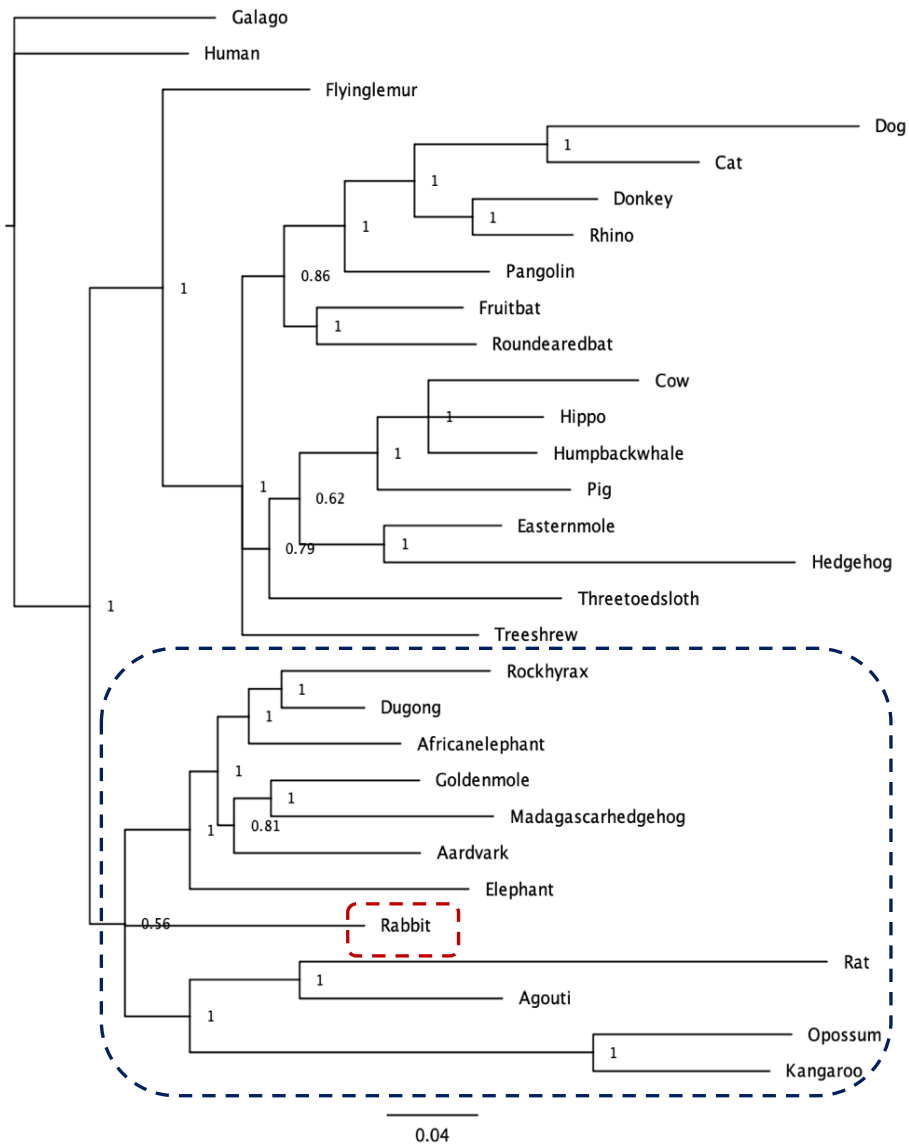


Figure 3.5 Phylogenetic trees based on vWF of 30 mammal taxa. The result of the Bayesian inference method demonstrates that the evolutionary relationship between Lagomorph (rabbit) and Rodents (rat) is closer than that between Primates (human, galago). Unlike the A2AB and IRBP genes, the branch of the red box (position of Lagomorph) leading to the Lagomorph, Primates, Rodents, and Scandentia is collapsed.

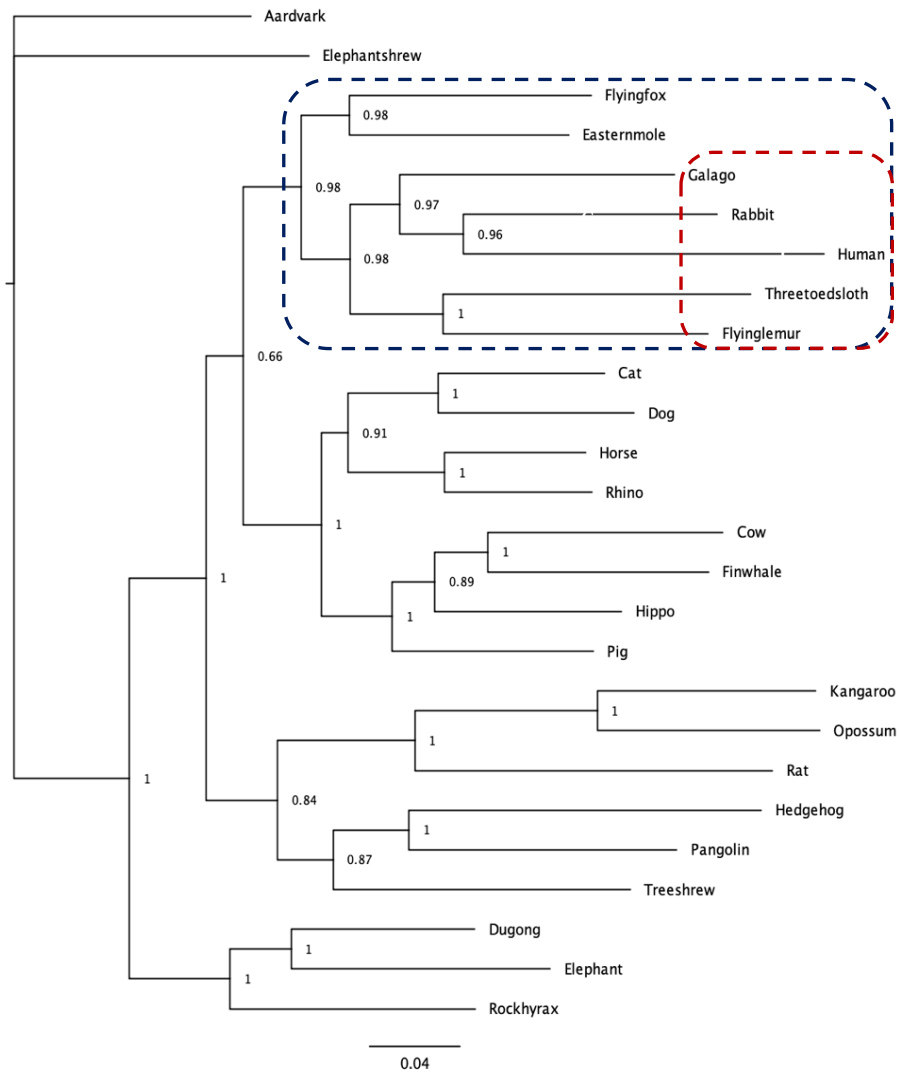


Figure 3.6 Phylogenetic tree based on mitochondrial data from 26 mammal taxa. Result of phylogenetic tree using Bayesian inference based on 12s rRNA-tRNA valine-16s rRNA concatenated gene. Unlike phylogenies from nuclear data, it cannot support the Glires hypothesis because the position of Lagomorph (rabbit) is more closely related with Primates (human, galago) than with Rodents (rat) and Scandentia (tree shrew).

Table 3.1 Four systematic biases for all taxa within Terebelliformia.

(a) C factor

Taxon	Species	EF1α (accepted)	18S rDNA	mtDNA	28S rDNA
Terebellidae	Pista cristata	4.2051	9.3189	0.9761	4.0405
Trichobranchidae	Terebellides stroemi		6.7234	0.9948	3.9428
	Terebellides sp.	4.1171			
Ampharetidae	Auchenoplax crinita	4.2242	7.8458	1.0801	7.9911
	Eclysippe vanelli		6.7568	1.0854	16.4249
Alvinellidae	Paralvinella sulfinacola		6.9538	1.1768	
	Paralvinella palmiformis				5.1764
	Paralvinella hessleri	3.6998			
Pectinariidae	Pectinaria gouldi		5.3702	1.1801	3.5787
	Pectinaria koreni	3.9005			
Sipuncula	Phascolopsis gouldi	3.1534	5.3119	2.2059	2.3844

(b) RCFV

Taxon	Species	EF1α (accepted)	18S rDNA	mtDNA	28S rDNA
Terebellidae	Pista cristata	0.1656	0.1257	0.1130	0.1164
Trichobranchidae	Terebellides stroemi		0.1349	0.1159	0.1199
	Terebellides sp.	0.1611			
Ampharetidae	Auchenoplax crinita	0.1578	0.1338	0.0867	0.1196
	Eclysippe vanelli		0.1348	0.0840	0.0763
Alvinellidae	Paralvinella sulfinacola		0.1390	0.1177	
	Paralvinella palmiformis				0.1024
	Paralvinella hessleri	0.1614			
Pectinariidae	Pectinaria gouldi		0.1318	0.1073	0.1080
	Pectinaria koreni	0.1734			
Sipuncula	Phascolopsis gouldi	0.1639	0.1336	0.0581	0.1493

(c) Nucleotide frequencies (GC content / gap proportions)

Taxon	Species	EF1α (accepted)	18S rDNA	mtDNA	28S rDNA
Terebellidae	Pista cristata	53.77 / 6.03	50.30 / 10.72	31.88 / 2.55	55.01 / 20.64
Trichobranchidae	Terebellides stroemi		50.33 / 3.00	32.85 / 3.40	58.79 / 21.29
	Terebellides sp.	51.19 / 4.90			
Ampharetidae	Auchenoplax crinita	49.95 / 4.28	51.56 / 5.52	31.54 / 15.64	57.80 / 21.27
	Eclysippe vanelli		49.13 / 1.71	30.92 / 15.70	57.10 / 49.24
Alvinellidae	Paralvinella sulfinacola		51.89 / 2.30	42.22 / 16.37	
	Paralvinella palmiformis				56.90 / 30.65
	Paralvinella hessleri	48.14 / 1.40			
Pectinariidae	Pectinaria gouldi		49.44 / 4.72	39.74 / 17.62	53.33 / 24.93
	Pectinaria koreni	56.16 / 3.50			
Sipuncula	Phascolopsis gouldi	47.41 / 0.26	51.13 / 5.47	36.89 / 54.21	60.75 / 3.78

(d) Shared missing data

Taxon	Species	EF1α (accepted)	18S rDNA	mtDNA	28S rDNA
Terebellidae	Pista cristata	3.04	4.01	1.43	16.82
Trichobranchidae	Terebellides stroemi		2.10	1.83	17.29
	Terebellides sp.	2.94			
Ampharetidae	Auchenoplax crinita	2.65	3.12	9.76	17.43
	Eclysippe vanelli		1.24	9.80	23.75
Alvinellidae	Paralvinella sulfinacola		1.65	9.72	
	Paralvinella palmiformis				20.49
	Paralvinella hessleri	0.92			
Pectinariidae	Pectinaria gouldi		2.83	9.93	17.69
	Pectinaria koreni	2.14			
Sipuncula	Phascolopsis gouldi	0.22	3.33	15.18	2.56

Table 3.2 Four systematic biases for all taxa in Daphniid.

(a) C factor

Taxon	Species	16S rDNA (accepted)	28S rDNA
Daphnia (Ctenodaphnia)	Daphnia longicephala	18.3071	7.3941
	Daphnia magna	15.4636	7.1191
Daphnia (Daphnia)	Daphnia ambigua	6.7592	4.6620
	Daphnia dubia	6.1322	7.2332
	Daphnia laevis	6.7254	6.8424
	Daphnia occidentalis	6.4375	5.2300
	Daphnia pulicaria		4.3785
	Daphnia dentifera	15.4189	5.2714
Daphniopsis	Daphniopsis ephemeralis	10.1942	7.2658
	Daphniopsis truncata	7.4402	5.0012

(b) RCFV

Taxon	Species	16S rDNA (accepted)	28S rDNA
Daphnia (Ctenodaphnia)	Daphnia longicephala	0.0864	0.0913
	Daphnia magna	0.0891	0.0930
Daphnia (Daphnia)	Daphnia ambigua	0.0904	0.0942
	Daphnia dubia	0.0882	0.0818
	Daphnia laevis	0.0877	0.0932
	Daphnia occidentalis	0.0888	0.0948
	Daphnia pulicaria		0.0958
	Daphnia dentifera	0.0794	0.0930
Daphniopsis	Daphniopsis ephemeralis	0.0868	0.0905
	Daphniopsis truncata	0.0920	0.0958

(c) Nucleotide frequencies (GC content / gap proportions)

Taxon	Species	16S rDNA (accepted)	28S rDNA
Daphnia (Ctenodaphnia)	Daphnia longicephala	32.58 / 1.82	56.89 / 12.13
	Daphnia magna	35.03 / 0.61	56.41 / 9.77
Daphnia (Daphnia)	Daphnia ambigua	36.20 / 1.01	55.28 / 7.42
	Daphnia dubia	35.17 / 1.01	55.33 / 20.04
	Daphnia laevis	34.36 / 1.01	55.70 / 8.63
	Daphnia occidentalis	33.95 / 1.01	56.48 / 7.55
	Daphnia pulicaria		54.94 / 5.38
	Daphnia dentifera	37.33 / 8.91	55.81 / 8.76
Daphniopsis	Daphniopsis ephemeralis	35.17 / 1.01	56.64 / 12.69
	Daphniopsis truncata	37.47 / 0.61	56.11 / 6.59

(d) Shared missing data

Taxon	Species	16S rDNA (accepted)	28S rDNA
Daphnia (Ctenodaphnia)	Daphnia longicephala	0.85	8.27
	Daphnia magna	0.43	7.13
Daphnia (Daphnia)	Daphnia ambigua	0.58	5.79
	Daphnia dubia	0.79	7.90
	Daphnia laevis	0.79	6.20
	Daphnia occidentalis	0.67	4.98
	Daphnia pulicaria		4.14
	Daphnia dentifera	1.66	6.44
Daphniopsis	Daphniopsis ephemeralis	0.76	8.39
	Daphniopsis truncata	0.52	4.85

Table 3.3 Four systematic biases for all taxa in Mammals.
(a) C factor

Taxon	Species	A2AB	IRBP	vWF	12S rRNA-tRNA valine-16S rRNA
Carnivora	Cat	14.6902	52.6945	28.0348	3.2503
	Harbor seal	16.8912			
	Fox		35.3954		
	Dog			3.2423	3.2906
Pholidota	Pangoline	17.2971	55.7803	24.2531	3.2021
Chiroptera	Big eared bat	15.8131			
	Flying fox	16.4732	82.9405		3.6334
	Round eared bat		70.5909	20.9106	
	Fruit bat			29.5074	
Insectivora	European mole	16.3491			
	Hedgehog	16.0289	77.4798	13.5697	2.6240
	Golden mole	21.7271		27.1668	
	Madagascar hedgehog	11.8678		24.1904	
	Shrew		37.3290		
	Eastern mole			20.1551	2.7316
Cetartiodactyla	Fin whale	14.4141			3.7529
	Hippo	16.0234	46.7625	23.0741	
	Cow	13.1500		62.3413	3.7287
	Pig	17.3147	40.8732	18.6526	3.3611
	Minke whale		53.8989		
	Humpback whale			25.9979	
Perissodactyla	Black Rhino	17.3501			
	Horse	14.9789	30.5147		4.7453
	Tapir		57.3463		
	Donkey			25.6294	
	White Rhino			30.32	5.0860

(a) (continued) C factor

Taxon	Species	A2AB	IRBP	vWF	12S rRNA-tRNA valine-16S rRNA
Proboscidea	Asian elephant	20.1080			
	African elephant		63.6684	30.8007 / 25.8594	4.1349
Sirenia	Dugong	18.0068	48.7893	26.7521	3.7836
Hyracoidea	Rock Hyrax	18.1559	67.4240	27.5626	3.5100
Tubuildentata	Aardvark	20.2226	47.6981	27.8578	2.5217
Macroscelidea	Rond-eared Elephant shrew	18.7290			
	Long-eared Elephant shrew		63.7776		2.4849
Primates	Human	4.1977	2.4641	20.3764	2.7564
	Slow loris	18.1236			
	Galago		50.6555	22.5251	2.5550
Dermoptera	Flying lemur	19.3250	54.0463	26.1241	6.0840
Scandentia	Tree shrew	15.7635	47.0846	26.5827	3.3137
Lagomorpha	Rabbit	12.9820	73.4257	18.2164	3.0197
Xenarthra	Three toed sloth	12.5853	56.1495	17.9196	3.0042
Rodentia	Guinea pig	10.4813			
	Rat	3.4258		10.3902	2.1352
	Mouse		25.8104		
	North American Porcupine		30.4650		
	Agouti			23.1679	
Didelphimorphia	Large American Opossum	12.7522	22.0719	19.8941	2.1348
Diprotodontia	Kangaroo	12.7789		16.7634	2.1392
	Wombat		21.1917		

(b) RCFV

Taxon	Species	A2AB	IRBP	vWF	12S rRNA-tRNA valine-16S rRNA
Carnivora	Cat	0.0184	0.0050	0.0049	0.0356
	Harbor seal	0.0186			
	Fox		0.0056		
	Dog			0.0357	0.0347
Pholidota	Pangoline	0.0182	0.0046	0.0054	0.0336
Chiroptera	Big eared bat	0.0185			
	Flying fox	0.0182	0.0046		0.0361
	Round eared bat		0.0047	0.0049	
	Fruit bat			0.0053	
Insectivora	European mole	0.0184			
	Hedgehog	0.0184	0.0038	0.0049	0.0328
	Golden mole	0.0178		0.0052	
	Madagascar hedgehog	0.0183		0.0055	
	Shrew		0.0050		
	Eastern mole			0.0056	0.0348
Cetartiodactyla	Fin whale	0.0190			0.0358
	Hippo	0.0189	0.0047	0.0055	
	Cow	0.0187		0.0019	0.0351
	Pig	0.0180	0.0052	0.0055	0.0353
	Minke whale		0.0046		
	Humpback whale			0.0055	
Perissodactyla	Black Rhino	0.0185			
	Horse	0.0188	0.0051		0.0357
	Tapir		0.0054		
	Donkey			0.0047	
	White Rhino			0.0047	0.0361

(b) (continued) RCFV

Taxon	Species	A2AB	IRBP	vWF	12S rRNA-tRNA valine-16S rRNA
Proboscidea	Asian elephant	0.0182			
	African elephant		0.0046	0.0047 / 0.0046	0.0323
Sirenia	Dugong	0.0184	0.0046	0.0054	0.0359
Hyracoidea	Rock Hyrax	0.0176	0.0042	0.0053	0.0359
Tubuildentata	Aardvark	0.0181	0.0050	0.0051	0.0341
Macroscelidea	Rond-eared Elephant shrew	0.0184			
	Long-eared Elephant shrew		0.0040		0.0344
Primates	Human	0.0333	0.0390	0.0073	0.0370
	Slow loris	0.0182			
	Galago		0.0051	0.0051	0.0366
Dermoptera	Flying lemur	0.0184	0.0052	0.0056	0.0237
Scandentia	Tree shrew	0.0187	0.0051	0.0047	0.0353
Lagomorpha	Rabbit	0.0190	0.0041	0.0049	0.0345
Xenarthra	Three toed sloth	0.0194	0.0044	0.0056	0.0385
Rodentia	Guinea pig	0.0186			
	Rat	0.0342		0.0078	0.0346
	Mouse		0.0053		
	North American Porcupine		0.0052		
	Agouti			0.0051	
Didelphimorphia	Large American Opossum	0.0172	0.0050	0.0043	0.0332
Diprotodontia	Kangaroo	0.0168		0.0049	0.0348
	Wombat		0.0052		

(c) Nucleotide frequencies (GC content / gap proportions)

Taxon	Species	A2AB	IRBP	vWF	12S rRNA-tRNA valine-16S rRNA
Carnivora	Cat	63.68 / 51.40	63.74 / 88.17	58.43 / 86.59	40.28 / 6.70
	Harbor seal	64.13 / 51.15			
	Fox		67.47 / 87.25		
	Dog			57.60 / 0.100	38.99 / 6.74
Pholidota	Pangoline	62.14 / 50.40	61.17 / 88.90	63.25 / 85.80	36.85 / 7.57
Chiroptera	Big eared bat	62.80 / 50.65			
	Flying fox	63.65 / 51.90	60.35 / 88.76		41.27 / 6.60
	Round eared bat		58.37 / 88.14	63.05 / 87.22	
	Fruit bat			62.10 / 85.93	
Insectivora	European mole	60.40 / 50.15			
	Hedgehog	61.93 / 50.90	59.87 / 90.62	65.68 / 87.52	35.74 / 6.70
	Golden mole	60.31 / 51.53		56.75 / 85.56	
	Madagascar hedgehog	64.12 / 52.32		63.87 / 85.85	
	Shrew		60.39 / 87.87		
	Eastern mole			67.77 / 86.00	39.36 / 6.63

(c) (continued) Nucleotide frequencies (GC content / gap proportions)

Taxon	Species	A2AB	IRBP	vWF	12S rRNA-tRNA valine-16S rRNA
Proboscidea	Asian elephant	63.37 / 51.82			
	African elephant		62.24 / 88.91	59.05 / 87.12 / 55.56 / 87.22	38.86 / 12.62
Sirenia	Dugong	62.74 / 51.07	62.91 / 89.24	58.10 / 85.18	42.24 / 7.14
Hyracoidea	Rock Hyrax	59.93 / 51.15	59.68 / 89.64	56.18 / 85.25	40.29 / 6.92
Tubuildentata	Aardvark	61.17 / 51.32	56.74 / 87.32	56.64 / 85.84	37.37 / 6.74
Macroscelidea	Rond-eared Elephant shrew	62.74 / 51.40			
	Long-eared Elephant shrew		57.70 / 89.91		39.36 / 7.46
Primates	Human	65.88 / 13.34	54.71 / 0.080	58.82 / 79.40	43.78 / 7.46
	Slow loris	61.44 / 50.65			
	Galago		63.92 / 87.77	59.02 / 85.92	41.80 / 6.17
Dermoptera	Flying lemur	64.65 / 51.02	62.88 / 87.78	62.54 / 85.18	43.47 / 41.52
Scandentia	Tree shrew	64.73 / 51.15	66.20 / 88.25	62.29 / 87.60	40.85 / 7.57
Lagomorpha	Rabbit	65.68 / 50.52	65.67 / 90.38	63.31 / 87.29	39.05 / 6.70
Xenarthra	Three toed sloth	69.65 / 51.90	63.70 / 89.60	64.62 / 85.32	43.79 / 6.24

(c) (continued) Nucleotide frequencies (GC content / gap proportions)

Taxon	Species	A2AB	IRBP	vWF	12S rRNA-tRNA valine-16S rRNA
Cetartiodactyla	Fin whale	65.34 / 50.77			40.35 / 6.17
	Hippo	64.75 / 50.65	63.11 / 88.90	63.88 / 85.78	
	Cow	63.38 / 50.77		64.04 / 95.02	39.22 / 7.03
	Pig	61.50 / 50.90	60.73 / 87.24	63.61 / 85.66	38.61 / 6.85
	Minke whale		63.04 / 88.95		
	Humpback whale			66.40 / 85.78	
Perissodactyla	Black Rhino	64.03 / 51.28			
	Horse	63.61 / 49.90	62.36 / 87.89		39.34 / 5.95
	Tapir		63.68 / 87.25		
	Donkey			58.64 / 87.24	
	White Rhino			59.71 / 87.22	39.66 / 6.24
Rodentia	Guinea pig	63.39 / 50.65			
	Rat	56.83 / 3.01		50.86 / 76.62	37.99 / 7.31
	Mouse		60.34 / 87.16		
	North American porcupine		64.21 / 87.84		
	Agouti			56.86 / 85.52	

(c) (continued) Nucleotide frequencies (GC content / gap proportions)

Taxon	Species	A2AB	IRBP	vWF	12S rRNA-tRNA valine-16S rRNA
Didelphimorphia	Large American Opossum	56.06 / 52.03	54.95 / 87.21	49.47 / 87.05	35.31 / 6.99
Diprotodontia	Kangaroo	55.81 / 52.49		50.83 / 85.47	38.10 / 7.03
	Wombat		56.71 / 86.88		

(d) Shared missing data

Taxon	Species	A2AB	IRBP	vWF	12S rRNA-tRNA valine-16S rRNA
Carnivora	Cat	47.88	84.45	82.27	5.02
	Harbor seal	47.68			
	Fox		84.00		
	Dog			0.100	4.93
Pholidota	Pangoline	47.20	84.78	82.22	4.80
Chiroptera	Big eared bat	47.42			
	Flying fox	47.81	84.73		4.83
	Round eared bat		84.50	82.64	
	Fruit bat			82.21	
Insectivora	European mole	47.25			
	Hedgehog	47.61	85.07	82.36	4.29
	Golden mole	47.98		82.08	
	Madagascar hedgehog	48.02		82.22	
	Shrew		84.20		
	Eastern mole			82.27	4.80
Cetartiodactyla	Fin whale	47.64			4.49
	Hippo	47.51	84.78	82.20	
	Cow	47.58		82.63	4.98
	Pig	47.71	83.99	82.11	4.90
	Minke whale		84.79		
	Humpback whale			82.21	
Perissodactyla	Black Rhino	47.76			
	Horse	46.95	84.19		4.44
	Tapir		84.00		
	Donkey			82.63	
	White Rhino			82.64	4.61

(d) (continued) Shared missing data

Taxon	Species	A2AB	IRBP	vWF	12S rRNA-tRNA valine-16S rRNA
Proboscidea	Asian elephant	48.00			
	African elephant		84.78	82.61 / 82.62	5.23
Sirenia	Dugong	47.82	84.89	81.79	4.84
Hyracoidea	Rock Hyrax	47.88	85.00	81.80	4.74
Tubuildentata	Aardvark	47.89	84.05	82.24	4.81
Macroscelidea	Rond-eared Elephant shrew	47.83			
	Long-eared Elephant shrew		85.03		4.87
Primates	Human	12.66	0.080	76.48	5.11
	Slow loris	47.60			
	Galago		84.28	82.24	4.22
Dermoptera	Flying lemur	47.84	84.32	81.79	6.73
Scandentia	Tree shrew	47.88	84.53	82.70	4.77
Lagomorpha	Rabbit	47.49	85.06	82.61	4.92
Xenarthra	Three toed sloth	47.93	84.83	81.86	4.43
Rodentia	Guinea pig	47.63			
	Rat	2.76		72.62	4.20
	Mouse		83.92		
	North American Porcupine		84.35		
	Agouti			82.05	
Didelphimorphia	Large American Opossum	47.88	83.96	82.53	4.72
Diprotodontia	Kangaroo	47.94		81.95	4.85
	Wombat		83.65		

Table 3.4 Summary of systematic biases for gene markers.

(a) Average of systematic biases for each gene marker in Terebelliformia

Marker gene	C factor (min / max)	RCFV	Base Frequencies (GC / GAP) (%)	Shared missing data (%)
EF1α (accepted)	3.1534 / 4.2051	0.1639	51.10 / 3.39	1.99
18S rDNA	5.3119 / 9.3189	0.1334	50.54 / 4.78	2.61
mtDNA	0.9761 / 2.2059	0.0975	35.15 / 17.92	8.23
28S rDNA	2.3844 / 16.4249	0.1131	57.10 / 24.54	16.57

(b) Average of systematic biases for each gene marker in Daphniid

Marker gene	C factor (min / max)	RCFV	Base Frequencies (GC / GAP) (%)	Shared missing data (%)
28S rDNA	4.3785 / 7.3941	0.0923	55.96 / 9.90	6.41
16S rDNA (accepted)	6.1322 / 18.3071	0.0876	35.25 / 1.89	0.78

(c) Average of systematic biases for each gene marker in Glires

Marker gene	C factor (min / max)	RCFV	Base Frequencies (GC / GAP) (%)	Shared missing data (%)
A2AB	3.4258 / 21.7271	0.0194	62.66 / 48.27	45.04
IRBP (past)	2.4641 / 82.9405	0.0061	61.36 / 85.11	81.34
vWF	3.2423 / 62.3413	0.0062	59.96 / 83.11	79.02
12S rRNA-tRNA valine-16S rRNA (modern)	2.1348 / 6.0840	0.0347	39.69 / 8.37	4.82

CHAPTER IV.

DISCUSSION

4.1 Significance and implications of study

In this study, within the three datasets including Terebelliformia, Daphniid, and Mammals where phylogenetic analysis was performed, the same gene marker was not selected for each systematic biases provided by the existing developed program. Accordingly, by analyzing the parameters of all taxa in the dataset, the optimal combination of systematic biases indicating the best marker was found using APSE. The performance of phylogenetic inference made on genetic markers can be misled by systematic errors, and the analytical results derived from the evaluation of phylogenetic accuracy are based on the potential biases such as heterogeneity through RCFV, saturation level through C factor, base frequencies, and shared missing data. It is possible to estimate systematic errors that evaluate the results of the each controversial dataset. All parameters that cause systematic errors were calculated by the APSE developed through this study, and a large size of each factor indicates that it is likely that phylogenies were reconstructed with datasets with the potential to cause systematic errors. The result of the analysis in this paper is that it is not right to indiscriminately use the integreated systematic biases previously presented for phylogenetic reliability, and it is important to analyze the combination of parameters to find the optimal marker. In addition, if the inflow of molecular data for gene markers of a specific

clade is made in the APSE, it will be possible to provide a more sophisticated combination of parameters for reflecting an accurate gene marker.

4.2 Application to bioinformatics research

The direction of phylogenetic studies is changing from that of previous studies, which caused incongruence of phylogenetic reconstruction by using limited loci with single or few genes, to phylogenomic analyses using hundreds to thousands of loci. As described above, this trend solved stochastic errors for dataset in terms of phylogenetic reliability and accelerated the probability of systematic errors at the same time. As the size of the dataset increases and the support of phylogenetic relationship increases, it is important for researchers to understand systematic biases, and the effect they have on phylogenetic analysis.

In addition, it has become meaningful to analyze biological phenomena that are evidence of systematic errors, and accordingly, various software programs that help calculate and estimate biases that cause nonphylogenetic signals are being developed. From a technical point of view, an automatic pipeline of phylogenetic reconstruction is also required to handle the sequence data of multiple genes. After phylogenetic analyses using probabilistic inferences, evaluating the effect of systematic biases on phylogenetic accuracy is part of the analytical pipeline and plays a crucial role in the entire phylogenetic reconstruction protocol for the datasets to be analyzed.

The APSE developed through this study suggests that it can serve as part of the pipeline in phylogenetic reconstruction by providing numeric information on systematic errors that affect phylogenetic accuracy. Until now, when a strong incongruence problem has been caused by the accumulation of systematic biases,

software programs that calculate systematic biases to improve this problem have been released. However, parameters that become indicators of various systematic errors are not provided by the execution of the single process, and some programs have a problem that requires additionally installing the corresponding interpreter, and thus, it was difficult for experimental researchers to use the programs. For example, multi-purpose packages for phylogenetic analysis such as Mesquite, DAMBE, and DRUIDS (Fedrigo et al., 2005) do not provide all known systematic biases, so there is a hassle of installing another program to obtain the necessary parameters. In addition, PERL programs such as BaCoCa have disadvantages of not only focusing researchers to install an interpreter themselves, but also having poor accessibility in terms of instructions and usage methods. In comparison, the APSE proposed as a result of this study can perform parallel assessment on various parameters to estimate systematic errors such as base composition biases, nucleotide frequencies, skew value, saturation, and shared missing data. In this case, the phylogenetic reliability is estimated by suggesting a combination of systematic biases that selects the gene marker that provides the best phylogeny. Ultimately, this program can be used as software that becomes a part of the pipeline to improve the reliability of phylogenetic reconstruction in various studies. In addition to this, time-consuming issues can be avoided by simplifying and providing a comprehensive analysis of various systematic biases in a single process in technical aspects, and extensibility in program application can be achieved by allowing the output file to be used as secondary data for a specific study.

4.3 Improvement and achievement

Through the APSE, which provides useful information on various phylogenetic biases, biological researchers will be able to evaluate phylogenetic accuracy and reliability by focusing on datasets. However, there are some drawbacks to the APSE that require improvement. First, since APSE receives multi-aligned format files as input, it does not provide a quality assessment for raw data before sequence alignment. In general, it is assumed that the phylogenetic relationship is estimated based on orthologous sequences, and if this assumption is violated, a phylogenetic error may occur. Therefore, it is necessary to use various orthology assessment software programs previously developed for identification of orthologs for the datasets to be analyzed. Second, since APSE does not perform cluster analysis of output for systematic biases, a correlation assessment with properties such as node support or branch length of phylogenies must be manually implemented. For example, the correlation between similar values can be identified by analyzing the calculated systematic biases by hierarchical clustering. If a component is added to receive the phylogenetic tree format file as an input, it is thought that the correlations assessment between datasets and phylogenies can be provided more efficiently. Finally, since thresholds for RCFV and C factor are not defined in the APSE, objective assessment for base composition biases and saturation level is not possible. For more elaborate analysis, relative assessment should be performed according to each taxon or dataset, and in this process, researchers should carefully evaluate the results at their own discretion. As a result, if the improvements suggested above are implemented, it is expected that there will be a beneficial outcome for the assessment of systematic errors by the APSE.

CHAPTER V.

CONCLUSION AND SUMMARY

5.1 Conclusion

The APSE developed through this study and the entire process of the study are largely divided into two specific implications. One is the construction of a specialized standalone program for phylogenetic reliability, and the other is on the reliability, which is the effect that systematic biases give to the evolutionary relationship of the controversial group in different genetic markers. First, the APSE structured the information system into a single module that combines data and processes it based on an objective-oriented approach. In addition, phylogenetic information containing parameters for systematic errors of aligned dataset can be provided with only a simple keyword without the use of complex or cumbersome instructions. The APSE was constructed to be able to independently estimate the reliability of phylogenies for the choice of species and phylogenetic reconstruction program and to be flexibly introduced into the entire phylogenetic analyses pipeline. Critical analysis of systematic biases using APSE will accelerate tree reconstruction and provide empirical evaluation for the reliability of phylogenies. Second, extensive evolutionary analyses were performed on Terebelliformia, Daphniid, and Mammals including Lagomorphs, Rodents, Primates and their allies with APSE, and incongruent phylogenies were constructed according to each gene marker. In order to estimate the reliability of

phylogenies representing controversial topologies in three datasets, gene markers for all taxa were analyzed in systematic bias units. Among the parameters corresponding to the systematic biases provided by APSE, EF1α, which has been accepted as reconstructing the best phylogeny in Terebelliformia, was selected by combining the max value of C factor, gap proportion, and shared missing data. In the case of Daphniid, the 16S rDNA marker, which has been recognized to build more accurate phylogeny than 28S rDNA, was selected by combining RCFV, GC content, gap proportion, and shared missing data. Finally, the combination of C factor, GC content, gap proportion, and shared missing data between the incongruent four gene markers representing the Glires hypothesis indicates that the systematic biases of the mitochondrial marker were low, and this result does not support the Glires hypothesis. As a result, it can be judged that the phylogeny of the mitochondrial marker with a low probability of systematic errors among four gene markers was the most reliable, and with this study, since the position of Lagomorphs is shown to be more related to Primates than Rodents, it is estimated that Lagomorphs do not belong to the Glires monophyly, and therefore, the Glires clade is not established.

5.2 Summary

In this study, a program was developed and applied research on the concept of systematic errors that can cause unresolved phylogenies and how these biases can affect phylogenetic reliability was implemented. The steps performed in the overall process were data collection, data preprocessing, phylogenetic analyses, software development, and systematic errors analysis. First, for three datasets of the study, gene markers used in the paper “Detecting the symplesiomorphy trap: a multigene

phylogenetic analysis of terebelliform annelids" (Zhing et al., 2011), "Rate acceleration and long-branch attraction in a conserved gene of cryptic Daphniid (Crustacea) species" (Omilian et al., 2001), and "Parallel adaptive radiations in two major clades of placental mammals" (Madsen et al., 2001) were collected and modified. Second, multiple alignment of each dataset was performed using ClustalW, and file conversion for phylogenetic analyses was implemented with the seqmagick tool. Third, ML analysis was performed with MEGA for phylogenetic reconstruction, and quartet puzzling analysis was carried out for fully resolved assessment with TREE-PUZZLE. Additionally, Bayesian inference using MrBayes and PP was visualized using the Figtree program. Finally, bioinformatics software was developed to evaluate the effect of systematic errors on incongruent and unresolved phylogenies. The phylogenetic reliability was evaluated by performing a heterogeneity test and relationship analysis using AliGROOVE based on the potential sources of systematic biases from the APSE.

As an applied study utilizing the developed program, a bioinformatics analytical pipeline was used to examine three datasets that have created controversial issues for a long time. Phylogenetic reconstruction was implemented using maximum likelihood and Bayesian inference approach, and incongruent result was printed according to the gene markers for each dataset. Therefore, an approach was carried out to select the best gene marker to evaluate the phylogenetic reliability in consideration of the probabilities of systematic error. Systematic errors are a problem that misleads a true evolutionary relationship in a phylogenomic context based on a large dataset. Although identifying or evaluating these potential sources is generally quite a complex and difficult task, it is crucial process for defining phylogenetic reliability and its accuracy. The APSE developed for this task helps estimate potential biases that can lead to systematic errors, such as base compositional biases,

nucleotide frequencies, skew value, substitution saturation level, and shared missing data. Unlike software that provides systematic biases that are currently being developed, the APSE is specialized in systematic errors and has been developed to operate flexibly within the existing phylogenetic analysis pipeline. Furthermore, using this constructed program, combinations of parameters were found for each dataset to select the best gene marker among controversial phylogenies. In terms of academic value, the APSE selects gene markers by considering various combinations of each parameter, rather than simply calculating systematic biases or performing comprehensive analysis. As more molecular data flows in the future, these combinations calculated by the program are automatically provided to verify gene markers.

This study focused on proposing a new protocol that uses a combination of systematic biases in validating accurate gene markers rather than focusing on the development of the software program. When the systematic biases of Terebelliformia and Daphniid were analyzed, the gene marker that have recognized as the best was selected. Since the Glires phylogenies, which contain the controversial Glires hypothesis, were not clearly accepted among researchers, it was estimated that the phylogenetic reliability of mitochondrial marker was high through the combination of parameters with low biases.

BIBLIOGRAPHY

Yuan J, Zhu Q, Liu B. Phylogenetic and biological significance of evolutionary elements from metazoan mitochondrial genomes. PLoS One. 9(1):e84330. Doi:10.1371/journal.pone.0084330 (2014).

Chatzou M, Magis C, Chang JM, Kemena C, Bussotti G, Erb I, Notredame C. Multiple sequence alignment modeling: methods and applications. Briefings in Bioinformatics. 17(6):1009-1023. Doi:10.1093/bib/bbv099 (2015).

Kuck P, Struck TH. BaCoCa – A heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. Molecular phylogenetics and evolution. 70(1):94-98. Doi:10.1016/j.ympev.2013.09.011 (2014).

Ballesteros JA, Sharma PP. A critical appraisal of the placemetn of xiphosura (chelicerata) with account of known sources of phylogenetic error. Systematic biology. 68(6):896-917. Doi:10.1093/sysbio/syz011 (2019).

Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nature Reviews Genetics. 13(5):303-314. Doi:10.1038/nrg3186 (2012).

Struck TH, Nesnidal MP, Purschke G, Halanych KM. Detecting possibly saturated positions in 18s and 28s sequences and their influence on phylogenetic reconstruction of annelida (lophotrochozoa). Molecular phylogenetics and evolution. 48(2):628-645. Doi:10.1016/j.ympev.2008.05.015 (2008).

Dwivedi B, Gadagkar SR. The impact of sequence parameter values on phylogenetic accuracy. Biology and Medicine. 1(3):50-62. (2009).

Rokas A, Carroll SB. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Molecular biology and evolution. 22(5):1137-44. Doi:10.1093/molbev/msi121 (2005).

Hillis DM. Approaches for assessing phylogenetic accuracy. Systematic biology. 44(1):3-16. Doi:10.1093/sysbio/44.1.3 (1995).

Perna NT, Kocher YD. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. Journal of molecular evolution. 41(3):353-358. Doi:10.1007/BF01215182 (1995).

Xia X, Xie Z, Salemi M, Chen L, Wang Y. An index of substitution saturation and its application. Molecular phylogenetics and evolution. 26(1):1-7. Doi:10.1016/S1055-7903(02)00326-3 (2003).

Zhong M, Hansen B, Nesnidal M, Golombek A, Halanych KM, Struck TH. Detecting the symplesiomorphy trap: a multigene phylogenetic analysis of terebelliform annelids. Evolutionary Biology. 11(1):369. Doi:10.1186/1471-2148-11-369 (2011).

Leache AD, Rannala B. The accuracy of species tree estimation under simulation: a comparision of methods. Systematic biology. 60(2):126-137. Doi:10.1093/sysbio/syq073 (2010).

Xia X, Lemey P. Assessing substitution saturation with DAMBE. Cambridge University Press. 615-630. Doi:10.1017/CBO9780511819049.022 (2009).

Sota T, Vogler AP. Incongruence of mitochondrial and nuclear gene trees in the carabid beetles ohomopterus. Systematic biology. 50(1):39-59. Doi:10.1093/sysbio/50.1.39 (2001).

Wagele JW, Mayer C. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. Evolutionary Biology. 7(1):147. Doi:10.1186/1471-2148-7-147 (2007).

Wiens JJ. Does adding characters with missing data increase or decrease phylogenetic accuracy?. Systematic biology. 47(4):625-640. Doi:10.1080/106351598260635 (1998).

Jiang W, Chen SY, Wang H, Li DZ, Wiens JJ. Should genes with missing data be excluded from phylogenetic analyses?. Molecular phylogenetics and evolution. 80:308-318. Doi:10.1016/j.ympev.2014.08.006 (2014).

Heath TA, Hedtke SM, Hillis DM. Taxon sampling and the accuracy of phylogenetic analyses. Journal of systematics and evolution. 46(3):239-257. Doi:10.3724/SP.J.1002.2008.08016 (2008).

Nabhan AR, Sarkar IN. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Briefings in Bioinformatics. 13(1):122-134. Doi:10.1093/bib/bbr014 (2011).

Lartillot N, Lepage T, Blanquart S. PhyloBayes 3: a bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 25(17):2286-2288. Doi:10.1093/bioinformatics/btp368 (2009).

Stamatakis A. RaxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30(9):1312-1313. Doi:10.1093/bioinformatics/btu033 (2014).

Scotland RW, Olmstead RG, Bennett JR. Phylogeny reconstruction: the role of morphology. Systematic biology. 52(4):539-548. Doi:10.1080/10635150390223613 (2003).

Penny D, Hendy M. Estimating the reliability of evolutionary trees. Molecular biology and evolution. 3(5):403-417. Doi:10.1093/oxfordjournals.molbev.a040407 (1986).

Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. Evolutionary Biology. 7(1):214. Doi:10.1186/1471-2148-7-214 (2007).

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18(3):502-504. Doi:10.1093/bioinformatics/18.3.502 (2002).

Adachi J, Hasegawa M. Molphy version 2.3 programs for molecular phylogenetics based on maximum likelihood. The institute of statistical mathematics. (1996).

Kumar S, Tamura K, Nei M. MEGA: Molecular evolutionary genetics analysis software for microcomputers. Computer applications in the biosciences. 10(2):189-191. Doi:10.1093/bioinformatics/10.2.189 (1994).

Lartillot N, Philippe H. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philosophical Transactions of the Royal Society B: Biological Science. 363(1496):1463-1472. Doi:10.1098/rstb.2007.2236 (2008).

Doyon JP, Ranwez V, Daubin V, Berry V. Models, algorithms and programs for phylogeny reconciliation. Brief. Bioinformatics. 12(5):392-400. Doi:10.1093/bib/bbr045 (2011).

Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. Proceedings of the National Academy of Sciences of the United States of America. 93(14):7085-7090. Doi:10.1073/pnas.93.14.7085 (1996).

Baele G, Lemey P, Rambaut A, Suchard MA. Adaptive MCMC in bayesian phylogenetics: an application to analyzing partitioned data in BEAST. Bioinformatics. 33(12):1798-1805. Doi:10.1093/bioinformatics/btx088 (2017).

Hasegawa M, Kishino H. Accuracies of the simple methods for estimating the bootstrap probability of a maximum-likelihood tree. Molecular biology and evolution. 11(1):142-145. Doi:10.1093/oxfordjournals.molbev.a040097 (1994).

Goloboff PA, Farris JS, Nixon KC. TNT, a free program for phylogenetic analysis. Cladistics. 24(5):774-786. Doi:10.1111/j.1096-0031.2008.00217.x (2008).

Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. Bioinformatics. 14(9):817-818. Doi:10.1093/bioinformatics/14.9.817 (1998).

Posada D. jModelTest: Phylogenetic model averaging. Molevular biology and evolution. 25(7):1253-1256. Doi:10.1093/molbev/msn083 (2008).

Struck TH. TreSpEx-Detection of misleading signal in phylogenetic reconstructions based on tree information. Evolutionary Bioinformatics Online. 10(10):51-67. Doi:10.4137/EBO.S14239 (2014).

Ho JWK, Adams CE, Lew JB, Matthews TJ, Ng CC, Shahabi-Sirjani A, Tan LH, Zhao Y, Easteal S, Wilson SR, Jermiin LS. SeqVis: Visualization of compositional heterogeneity in large alignments of nucleotides. Bioinformatics. 22(17):2162-2163. Doi:10.1093/bioinformatics/btl283 (2006).

Som A. Causes, consequences and solutions of phylogenetic incongruence. Briefings in Bioinformatics. 16(3):536-548. Dio:10.1093/bib/bbu015 (2014).

Roderiguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. Detecting and overcoming systematic errors in genome-scale phylogenies. Systematic biology. 56(3):389-399. Doi: 10.1080/10635150701397643 (2007).

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Worheide G, Baurain D. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biology. 9(3):e1000602. Doi: 10.1371/journal.pbio.1000602 (2011).

Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. Systematic biology. 58(1):130-145. Doi: 10.1093/sysbio/syp017 (2009).

Nesnidal MP, Helmkampf M, Meyer A, Witek A, Bruchhaus I, Ebersberger I, Hankeln T, Lieb B, struck TH, Hausdorf B. New phylogenomic data support the monophyly of lophophorata and an ectoproct-phoronid clade and indicate that polyzoa and kryptrochozoa are caused by systematic bias. Evolutionary Biology. 13(1):253. Doi:10.1186/1471-2148-13-253 (2013).

Rosenberg MS, Kumar S. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. Molecular biology and evolution. 20(4):610-621. Doi:10.1093/molbev/msg067 (2003).

Yang H, Li T, Dang K, Bu W. Compositional and mutational rate heterogeneity in mitochondrial genomes and its effect on the phylogenetic inferences of cimicomorpha (hemiptera: heteroptera). Genomics. 19(1):264. Doi:10.1186/s12864-018-4650-9 (2018).

Liu Y, Cox CJ, Wang W, Goffinet B. Mitochondrial phylogenomics of early land plants: mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. Systematic biology. 63(6):862-878. Doi:10.1093/sysbio/syu049 (2014).

Huchon D, Madsen O, Sibbald MJJB, Ament K, Stanhope MJ, Catzeflis F, de Jong WW, Douzery EJP. Rodent phylogeny and a timescale for the evolution of glires: evidence from an extensive taxon sampling using three nuclear genes. Molecular biology and evolution. 19(7):1053-1065. Doi:10.1093/oxfordjournals.molbev.a004164 (2002).

Blanga-Kanfi S, Miranda H, Penn O, Pupko T, DeBry RW, Huchon D. Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. Evolutionary Biology. 9(1):71. Doi:10.1186/1471-2148-9-71 (2009).

Halanych KM. Lagomorphs misplaced by more characters and fewer taxa. Systematic biology. 47(1):138-146. Doi:10.1080/106351598261085 (1998).

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJP. Less is more in mammalian phylogenomics: at-rich genes minimize tree conflicts and unravel the root of placental mammals. Molecular biology and evolution. 30(9):2134-2144. Doi:10.1093/molbev/mst116 (2013).

Bosset S, Murray EA, Blaimer BB, Danforth BN. The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. Molecular phylogenetics and evolution. 111:149-157. Doi:10.1016/j.ympev.2017.03.022 (2017).

Halanych KM, Robinson TJ. Multiple substitutions affect the phylogenetic utility of cytochrome b and 12s rdna data: examining a rapid radiation in leporid (lagomorpha) evolution. Journal of molecular evolution. 48(3):369-379. Doi:10.1007/pl00006481 (1999).

Wiens JJ. Missing data and the design of phylogenetic analyses. Journal of biomedical informatics. 39(1):34-42. Doi:10.1016/j.jbi.2005.04.001 (2006).

Graur D, Duret L, Gouy M. Phylogenetic position of the order lagomorpha (rabbits, hares and allies). Nature. 379(6563):333-335. Doi:10.1038/379333a0 (1996).

Young AD, Gillung JP. Phylogenomics – principles, opportunities and pitfalls of big-data phylogenetics. Systematic entomology. 45(2):225-247. Doi:10.1111/syen.12406 (2019).

Duchene S, Ho SYW, Holmes EC. Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. Evolutionary Biology. 15(1):36. Doi:10.1186/s12862-015-0312-6 (2015).

DeBry RW, Sagel RM. Phylogeny of rodentia (mammalia) inferred from the nuclear-encoded gene irbp. Molecular phylogenetics and evolution. 19(2):290-301. Doi:10.1006/mpev.2001.0945 (2001).

Douzery EJP, Huchon D. Rabbits, if anything, are likely glires. Molecular phylogenetics and evolution. 33(3):922-935. Doi:10.1016/j.ympev.2004.07.014 (2004).

Kocot KM, Struck TH, Merkel J, Waits DS, Todt C, Brannock PM, Weese DA, Cannon JT, Moroz LL, Lieb B, Halanych KM. Phylogenomics of lophotrochozoa with consideration of systematic error. Systematic biology. 66(2):256-282. Doi:10.1093/sysbio/syw079 (2016).

Wagele JW. Foundations of phylogenetic systematics. Verlag Dr. Friedrich Pfeil. (2005).

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25(15):1972-1973. Doi:10.1093/bioinformatics/btp348 (2009).

Saurabh K, Holland BR, Gibb GC, Penny D. Gaps: An elusive source of phylogenetic information. Systematic biology. 61(5):1075-1082. Doi:10.1093/sysbio/sys043 (2012).

Nagy LG, Kocsube S, Csanadi Z, Kovacs GM, Petkovits T, Vagvolgyi C, Papp T. Re-mind the gap! Insertion – deletion data reveal neglected phylogenetic potential of the nuclear ribosomal internal transcribed spacer (ITS) of fungi. PLoS ONE. 7(11):e49794. Doi:10.1371/journal.pone.0049794 (2012).

Davalos LM, Perkins SL. Saturation and base composition bias explain phylogenomic conflict in plasmodium. Genomics. 91(5):433-442. Doi:10.1016/j.ygeno.2008.01.006 (2008).

Kuck P, Meid SA, GroB C, Wagele JW, Misof B. AliGROOVE – visualization of heterogeneous sequence divergence within multiple sequence alignments and detection of inflated branch support. Bioinformatics. 15(1):294. Doi:10.1186/1471-2105-15-294 (2014).

Mooers AO, Holmes EC. The evolution of base composition and phylogenetic inference. Trends in ecology and evolution. 15(9):365-369. Doi:10.1016/S0169-5347(00)01934-0 (2000).

Foster PG, Hickey DA. Compositional bias may affect both dna-based and protein-based phylogenetic reconstructions. Journal of molecular evolution. 48(3):284-290. Doi:10.1007/pl00006471 (1999).

Sheffield NC. The interaction between base compositional heterogeneity and among-site rate variation in models of molecular evolution. Evolutionary Biology. 2013:8. Doi:10.5402/2013/391561 (2013).

Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution. 17(4):540-552. Doi:10.1093/oxfordjournals.molbev.a026334 (2000).

Tan G, Muffato M, Ledergerber C, Herrero J, Goldman N, Gil M, Dessimoz C. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Systematic biology. 64(5):778-791. Doi:10.1093/sysbio/syv033 (2015).

Phillips MJ, Delsuc F, Penny D. Genome-scale phylogeny and the detection of systematic biases. Molecular biology and evolution. 21(7):1455-1458. Doi:10.1093/molbev/msh137 (2004).

Young ND, Healy J. GapCoder automates the use of indel characters in phylogenetic analysis. Bioinformatics. 4(1):6. Doi:10.1186/1471-2105-4-6 (2003).

Soltis PS, Soltis DE. Applying the bootstrap in phylogeny reconstruction. Statistical Science. 18(2):256-267. Doi:10.1214/ss/1063994980 (2003).

Penzar D, Krivozubov M, Spirin S. PQ, a new program for phylogeny reconstruction. Bioinformatics. 19:374. Doi:10.1186/s12859-018-2399-4 (2018).

Jones MO, Koutsovoulos GD, Blaxter ML. iPhy: an integrated phylogenetic workbench for supermatrix analyses. Bioinformatics. 12(1):30. Doi:10.1186/1471-2105-12-30 (2011).

Harrison CJ, Langdale JA. A step by step guide to phylogeny reconstruction. The Plant Journal. 45(4):561-572. Doi:10.1111/j.1365-313X.2005.02611.x (2006).

Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. Bioinformatics. 5:113. Doi:10.1186/1471-2105-5-113 (2004).

Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research. 22(22):4673-4680. Doi:10.1093/nar/22.22.4673 (1994).

Saitou N, Nei M. The neighbor-joining method: a new method for reconstruction phylogenetic trees. Molecular biology and evolution. 4(4):406-425. Doi:10.1093/OXFORDJOURNALS.MOLBEV.A040454 (1987).

Mansour A. Phylip and phylogenetics. Genes, Genomes and Genomics. 3(1):46-49.

Mehmood MA, Sehar U, Ahmad N. Use of bioinformatics tools in different spheres of life sciences. Journal of data mining in genomics and proteomics. 5(2). Doi:10.4172/2153-0602.1000158 (2014).

Wiens JJ. Missing data, incomplete taxa, and phylogenetic accuracy. Systematic biology. 52(4):528-538. Doi:10.1080/10635150390218330 (2003).

Montgelard C, Forty E, Arnal V, Matthee CA. Suprafamilial relationships among rodentia and the phylogenetic effect of removing fast-evolving nucleotides in mitochondrial, exon and intron fragments. Evolutionary Biology. 8(1):321. Doi:10.1186/1471-2148-8-321 (2008).

Ogden TH, Rosenberg MS. Multiple sequence alignment accuracy and phylogenetic inference. Systematic biology. 55(2):314-328. Doi:10.1080/10635150500541730 (2006).

Kupczok A, Schmidt HA, von Haeseler A. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. Algorithms for Molecular Biology. 5(1):37. Doi:10.1186/1748-7188-5-37 (2010).

Zheng Y, Wiens JJ. Do missing data influence the accuracy of divergence-time estimation with BEAST? Molecular phylogenetics and evolution. 85:41-49. Doi:10.1016/j.ympev.2015.02.002 (2015).

Felsenstein J. Confidence limits on phylogeneies: an approach using the bootstrap. Evolutions. 39(4):783-791. Doi:10.1111/j.1558-5646.1985.tb00420.x. (1985).

Smith SA, Dunn CW. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics. 24(5):715-716. Doi:10.1093/bioinformatics/btm619 (2008).

Hall BG. Building phylogenetic trees from molecular data with MEGA. Molecular biology and evolution. 30(5):1229-1235. Doi:10.1093/molbev/mst012 (2013).

Gadagkar SR, Rosenberg MS, Kumar S. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. Journal of experimental zoology. 304(1):64-74. Doi:10.1002/jez.b.21026 (2005).

Kearney M, Clark JM. Problems due to missing data in phylogenetic analyses including fossils: a critical review. Journal of vertebrate paleontology. 23(2):263-274. Doi:10.1671/0272-4634 (2003).

Timmermans MJTN, Barton C, Haran J, Ahrens D, Culverwell CL, Ollikainen A, Dodsworth S, Foster PG, Bocak L, Vogler AP. Family-level sampling of mitochondrial genomes in coleoptera: compositional heterogeneity and phylogenetics. Genome biology and evolution. 8(1):161-175. Doi:10.1093/gbe/evv241 (2015).

Seplyarskiy VB, Kharchenko P, Kondrashov AS, Bazykin GA. Heterogeneity of the transition/transversion ratio in drosophila and hominidae genomes. Molecular biology and evolution. 29(8):1943-1955. Doi:10.1093/molbev/mss071 (2012).

Xi Z, Liu L, Davis CC. The impact of missing data on species tree estimation. Molecular biology and evolution. 33(3):838-860. Doi:10.1093/molbev/msv266 (2015).

Wiens JJ. Incomplete taxa, incomplete characters, and phylogenetic accuracy: is there a missing data problem? Journal of vertebrate paleontology. 23(2):297-310. Doi:10.1671/0272-4634 (2003).

Zhou HQ, Ning LW, Zhang HX, Guo FB. Analysis of the relationship between genomic gc content and patterns of base usage, codon usage and amino acid usage in prokaryotes: similar gc content adopts similar compositional frequencies regardless of the phylogenetic lineages. PloS ONE. 9(9):e107319. Doi:10.1371/journal.pone.0107319 (2014).

Alfaro ME, Zoller S, Lutzoni F. Bayes or bootstrap? A simulation study comparing the performance of bayesian markov chain monte carlo sampling and bootstrapping in assessing phylogenetic confidence. Molecular biology and evolution. 20(2):255-266. Doi:10.1093/molbev/msg028 (2003).

Song H, Sheffield NC, Cameron SL, Miller KB, Whiting MF. When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. Systematic Entomology. 35(3):429-448. Doi:10.1111/j.1365-3113.2009.00517.x (2010).

Wiens JJ, Morrill MC. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Systematic biology. 60(5):719-731. Doi:10.1093/sysbio/syr025 (2011).

Takezaki N, Gojobori T. Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial dna sequences. Molecular biology and evolution. 16(5):590-601. Doi:10.1093/oxfordjournals.molbev.a026141 (1999).

Wang LS, Mack JL, Wall PK, Beckmann K. The impact of multiple protein sequence alignment on phylogenetic estimation. Transactions on Computational Biology and Bioinformatics. 8(4):1108-1119. Doi:10.1109/TCBB.2009.68 (2011).

Cantarel BL, Morrison HG, Pearson W. Exploring the relationship between sequence similarity and accurate phylogenetic trees. Molecular biology and evolution. 23(11):2090-2100. Doi:10.1093/molbev/msl080 (2006).

Prasad AB, Allard MW, NISC Comparative Sequencing Program, Green ED. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Molecular biology and evolution. 25(9):1795-1808. Doi:10.1093/molbev/msn104 (2008).

Dwivedi B, Gadagkar SR. Phylogenetic inference under varying proportions of indel-induced alignment gaps. Evolutionary Biology. 9(1):211. Doi:10.1186/1471-2148-9-211 (2009).

Wilgenbusch JC, Swofford D. Inferring evolutionary trees with PAUP*. Current protocols in bioinformatics. 6-4. Doi:10.1002/0471250953.bi0604s00 (2003).

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Systematic biology. 61(3):539-542. Doi:10.1093/sysbio/sys029 (2012).

Wilson IJ, Weale ME, Balding DJ. Inferences from DNA data: population histories evolutionary processes and forensic match probabilities. Journal of the Royal Statistical Society Series A. 166(2):155-188 (2003).

Arakawa K, Tomita M. Measures of compositional strand bias related to replication machinery and its applications. Current Genomics. 13(1):4-15. Doi:10.2174/138920212799034749 (2012).

Claudia AMR, Barbara OA, Alexandre PS. Selecting molecular markers for a specific phylogenetic problem. MOJ Proteomics & Bioinformatics. Doi:10.15406/mojpb.2017.06.00196 (2017).

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. Phylogenomics: the beginning of incongruence? Trends in Genetics. 22(4):225-231. Doi:10.1016/j.tig.2006.02.003 (2006).

Cartwright RA. Problems and solutions for estimating indel rates and length distributions. Molecular biology and evolution. 26(2):473-480. Doi:10.1093/molbev/msn275 (2008).

Omilian AR, Taylor DJ. Rate acceleration and long-branch attraction in a conserved gene of cryptic daphniid (Crustacea) species. Molecular biology and evolution. 18(12):2201-2212. Doi:10.1093/oxfordjournals.molbev.a003767 (2001).

Roger AJ, Sandblom O, Doolittle WF, Philippe H. An evaluation of elongation factor 1 alpha as a phylogenetic marker for eukaryotes. Molecular biology and evolution. 16(2):218-233. Doi:10.1093/oxfordjournals.molbev.a026104 (1999).

Galtier N, Nabholz B, Glemin S, Hurst GDD. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Molecular Ecology. 18(22):4541-4550. Doi:10.1111/j.1365-294X.2009.04380.x (2009).

Abadi S, Azouri D, Pupko T, Mayrose I. Model selection may not be a mandatory step for phylogeny reconstruction. Nature Communications. 10(1):934. Doi:10.1038/s41467-019-08822-2 (2019).

ABSTRACT (Korean)

바이오인포매틱스 프로그램을 이용한 유전자 마커 선별 및 계통수 오류 평가 연구

이 정 환
서울대학교 자연과학대학 생물정보협동과정
바이오인포매틱스 전공

지속적으로 산출되는 엄청난 양의 생물학적 서열 데이터는 유기체 사이의 진화적 역사와 계통학적 관계(phylogenetic relationship)를 유추할 수 있는 기회를 제공한다. 이제 계통수 구축은 거의 모든 생물학 연구에서 수행되는 과정의 하나가 되었다. 여기서 계통정보학(phyloinformatics)은 계통수 생성 알고리즘과 진화적 모델 개발과 같은 기술적 또는 방법론적 연구를 중심으로 발전되어 왔다. 현재의 계통수 분석은 서열 데이터, 즉 유전적 마커를 이용하여 계통수를 생성함으로써 실제에 가까운 계통수를 추론하는 것을 목표로 한다. 그러나 유전적 마커를 비롯한 데이터의 크기가 기하급수적으로 증가하고 따라오는 계통수 분석의 정확성에 대한 의문이 점차 중요하게 다루어 지기 시작하면서 계통수의 정확성 및 신뢰성을 평가하기 위한 연구가 다수 이루어지고 있는 상황이다. 분자 시스템학 관점에서 계통수에 대한 정확성 평가는 두 가지 갈래로 나누어 접근할 수 있는데, 하나는 진화 조건, 분자데이터의 양과 같은 특정 환경 아래에서 계통 분석 알고리즘이 얼마나 잘 작동하는지를 다루는 것이고, 또 다른 하나는 특정 계통수를 얼마나

신뢰할 수 있는지에 집중하는 것이다. 그리고 데이터셋의 퀄리티 관점에서 신뢰할 만한 계통수를 획득하기 위해 계통수 분석을 수행한 후, 사용한 데이터셋과의 적절성을 평가하는 것도 중요하다. 대규모 데이터를 기본으로 취급하는 최근 계통수 분석에서 확률론적 오류의 가능성은 낮아졌지만, 시스템 오류의 가능성은 오히려 높아졌으므로, 계통수 정확성을 평가 및 개선하기 위해 계통 분석 결과 후에 데이터셋이 가지는 시스템 오류의 근원을 평가하는 것이 매우 중요한 과정이 되었기 때문이다. 이에 본 연구에서는 데이터 퀄리티 관점에서 계통수의 신뢰도 향상을 가져오기 위해 APSE (Assessment Program for Systematic Error, tentative)라는 프로그램을 개발하였다. APSE를 활용하면 분류군 특이적 상대적 구성 빈도 변이(RCFV)와 대칭적 왜곡값(skew)을 산출하여 염기서열의 구성적 편향성에 대한 정보를 얻고, 이를 통해 연구하고자 하는 데이터의 유전적 이질성(heterogeneity) 및 유전적 변이 편향성(mutational bias)을 추정할 수 있다. 뿐만 아니라 다양한 염기 그룹의 빈도, 변이에 의한 다수 치환을 의미하는 포화(saturation)와 공유 결측 데이터(shared missing data) 변수를 통해 시스템 오류를 유발할 수 있는 편향성 정보들을 계산하는 것이 가능하다. 또한, 시스템 성능을 평가하기 위해 다양한 유전자 마커 사이의 모순되는 계통수를 출력하고 있는, 특이적 예시(Terebelliformia, Daphniid, Glires)를 APSE에 적용하여 마커 데이터셋의 시스템 오류 평가와 그에 따라 선별된 마커 계통수의 정확성 추론에 대한 분석이 제대로 수행될 수 있음을 확인하였다. 따라서 향후 APSE는 시스템학적 관점에서 데이터 퀄리티에 집중하여 생성된 계통수가 보다 정확한 결과를 이끌어낼 수 있도록 사용자의 데이터와 계통수 사이의 정확성을 평가하는 역할을 할 것이고, 유전적

마커에 따라 오해의 소지가 있는 계통수가 출력되었을 때, 시스템 오류의 근원에 대한 철저한 분석과 해당 오류의 영향을 받은 데이터가 계통수에 주는 효과를 파악하는 일을 수행할 수 있을 것이라 기대한다.

.....

주요어: 시스템 오류, 바이오인포매틱스, 독립 프로그램, 계통학적 신뢰성, 다중서열정렬, 계통분류학적 분석, 데이터 퀄리티

학 번: 2017-23460