Alignment- and Alignment-refining Algorithms: Effects on Branch-length Estimation and Selection Pattern Analyses

---

A Dissertation Presented to

the Faculty of the Department of Biology

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Yichen Zheng

May 2015

Alignment- and Alignment-refining Algorithms: Effects on Branch-length Estimation and Selection Pattern Analyses

_____

Yichen Zheng

APPROVED:

_____

Dr. Dan Graur, Chair

_____

Dr. Tim F. Cooper

_____

Dr. Rebecca A. Zufall

_____

Dr. Krešimir Josić

_____

Dr. Luay Nakhleh, Rice University

_____

Dr. Dan E. Wells, Dean, College of Natural Sciences and Mathematics

Acknowledgements

I would like to thank the present and former members of my Ph.D. committee: Drs. Tim Cooper, Rebecca Zufall, Ken Whitney, Kresimir Josic, and Luay Nakhleh. They have provided me with many helpful suggestions during committee meetings, and improved the structure of my research projects.

I would also like to thank Drs. Ricardo Azevedo, William Ott and Chinmaya Gupta for their help in my research projects. They helped me doing the analyses and preparing the manuscripts.

I would also like to thank my fellow graduate students during the last five years: Dr. Lara Appleby, Eric Bakota, Kristen Dimond, Ata Kalirad, Dr. Hongan Long, Rebecca Satterwhite, Dr. Xiaopeng Shen, Dr. Jun Wang, Yinghua Wang, Joe West, Bingjun Zhang … As friends and fellow scientists, we have talked during meetings, seminars, lunches and in the SR2 corridors; they gave me valuable academic and life advice. In particular, Dr. Hongan Long has been a mentor to me, from the first day he drove me to school from the airport when I first arrived in Houston. I wish every one of them to be successful in their academic lives and careers.

I would also like to thank my labmates: Dr. Nicholas Price, Wenfu Li, Betsy Salazar, Fei Yuan, and Hoa Nguyen. They are my most valuable friends and peers. From them, I learned how to do group projects and how to communicate scientific progress. They

improved my interpersonal skills and eased my social difficulties. I will forever cherish their friendship.

I would not be able to come to the United States from the other side of Earth and pursue my dreams without the support of my family in China. None of my family members has a Ph.D.; however, they fully understand what science means to me, and backed my choice of becoming a scientist instead of leading an "easier" life.

Finally, I owe my biggest gratitude to my advisor and committee chair, Dr. Dan Graur. He is straightforward, thoughtful, and is not afraid to criticize other people's ideas. He looks formidable from afar, but he is one of the friendliest and nicest people I have ever met. Academically, he taught me how to think like a scientist, how to write a scientific paper, and gave me many opportunities to meet other researchers and share ideas with them. Personally, he chatted with me in topics from travel to food. He also taught me how to appreciate modern art. I will be forever indebted to this awesome mentor.

Alignment- and Alignment-refining Algorithms: Effects on Branch-length Estimation and Selection Pattern Analyses

---

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Biology

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Yichen Zheng

May 2015

Abstract

This dissertation consists of a study of the effects of multiple-alignment method on phylogenetic analyses.

First, I investigated the effects of multiple-sequence alignment quality on branch-length estimation, which can influence downstream bioinformatic analyses such as estimating rates of evolution and divergence times. To quantify the accuracy of branch-length estimates, I devised a scale-free measure of branch length proportionality between two phylogenetic trees that contain the same taxa and topology. This measure was named “normalized tree distance” (NTD). NTD is an ideal measure for detecting coevolutionary processes, in addition to measuring the accuracy of branch-length estimates.

Using NTD as an error measure, I investigated the effects of multiple-sequence alignment quality on branch-length estimation. I simulated coding sequences and estimated the effects of multiple evolution parameters and choice of alignment- and alignment-filtering algorithms on the accuracy of branch-length estimation. I demonstrated that branch-length accuracy is indeed dependent on the method of alignment. Alignments with high-accuracy algorithms combined with methods for filtering out unreliable sites produce significantly better branch-length estimates. The optimal method combination depends on the evolutionary scenario. Thus, different alignment algorithms and different combinations of algorithms yield better branch-length estimates under different evolutionary conditions. A judicious choice of

alignment- and alignment filtering algorithms is recommended for phylogenetic studies.

Second, I studied the correlation between two types of purifying selection: against nonsynonymous mutations and against deletions using mammalian genomic protein-coding sequences. Intuitively, a codon that is intolerant of amino-acid altering substitutions is expected to be also intolerant of deletion. However, there has not been any comprehensive study on this purported correlation. In addition to the nine-species alignments of 8,595 genes, I simulated coding sequences along the same phylogenetic trees. The real data showed a much stronger correlation than the simulated sequences. I demonstrated that the correlation between amino-acid replacement and deletion rates exists and cannot be explained solely by alignment errors. Further investigations on nonsynonymous and synonymous mutations showed that this is most likely due to selection rather than mutation rates. Understanding selection on different types of mutations would help strengthen the link between population genetics and sequence evolution.

Contents

Chapter One: Introduction

A typical molecular macroevolution study starts with the collection of homologous DNA or protein sequences from multiple taxa. In order to directly compare these sequences, the homology must be extended to a positional homology – determining which sites (nucleotides or amino acids) in the sequences are descended from the same ancestral site. This process, called sequence alignment, can be difficult if multiple insertion and deletion events ("indels") have occurred in their evolutionary history (Figure 1.1A, 1.1B). Researchers usually use automated algorithms to produce alignments. As a mathematical tool, alignment algorithms are not able to accurately reconstruct all mutation events, causing alignment errors (Figure 1.1C). Because alignments are produced by optimizing (attempting to maximize) an arbitrary mathematical score, it is possible that the true alignment has an equal (cooptimal) or lower (suboptimal) score compared to the reconstructed one (Figure 1.2A, 1.2B; Landan and Graur 2008). As seen in the figure, cooptimal or suboptimal situations can occur even with a very small number of indels and substitutions. In addition, alignments involving three or more sequences have to resort to heuristic algorithms to avoid prohibitive computation time. These algorithms usually produce a neighbor-joining tree (Saitou and Nei 1987) based on pairwise alignment results, and add sequences to the alignments according to the estimated relatedness. This process cannot fix errors occurring in earlier stages based on data from later stages.

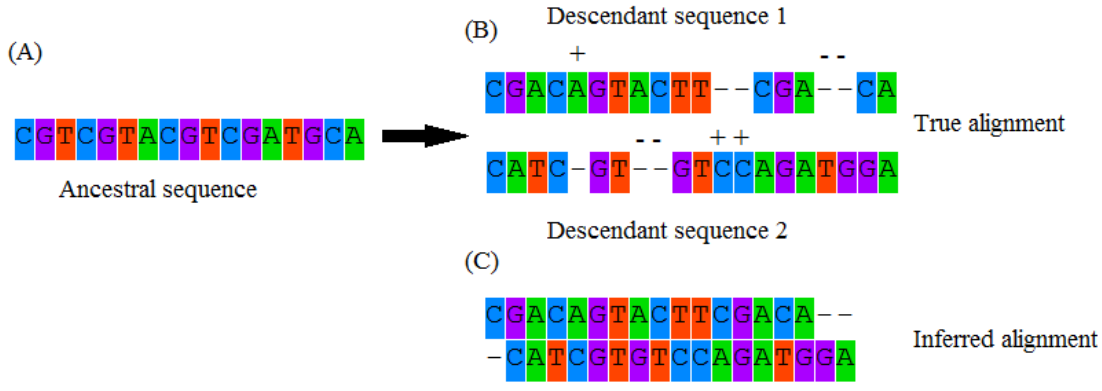

Figure 1.1 Examples of alignments and alignment errors. (A) A hypothetical ancestral DNA sequences. (B) Two descendants of this ancestor. Note that multiple substitutions, insertions, and deletions have occurred in the process; "+" denote inserted nucleotides and "-"denotes gaps left by deleted nucleotides. (C) The same two descendant sequences as aligned by CLUSTALW. Note the difference between this alignment and the true alignment; especially the fact that there are fewer gaps compared to the true alignment.

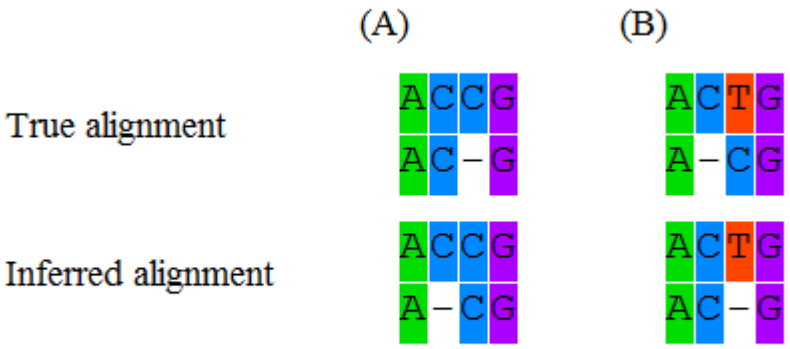

Figure 1.2 Cooptimal and suboptimal alignments. (A) Situation where the reconstructed and true alignments are cooptimal, caused by a deletion of one in two consecutive and identical nucleotides. A mathematical model cannot distinguish which alignment fits the model better due to ambiguity. (B) Situation where the true alignment is suboptimal to the reconstructed alignment, caused by a deletion and a substitution in its neighboring site. A mathematical model will favor the wrong alignment.

In the past two decades, as biologists gather more and more DNA and protein data from thousands of species, programmers have devised dozens of algorithms for multiple sequence alignment. The majority of researchers use CLUSTALW (Thompson et al. 1994) to align their sequences. However, CLUSTALW, a 1994 program, has been repeatedly shown to be outperformed by more recent ones (Nuin et al. 2006, Wang et al. 2011). Alignment errors are likely to cascade into even larger errors in later parts of the study (Markova-Raina and Petrov 2011, Schneider et al. 2009); for example, an erroneous mismatch can produce a positive selection signal where none has occurred. Therefore, the under-usage of more accurate alignment algorithms may indicate a large amount of errors (especially false positives) in the literature.

In addition to choosing a better-performing alignment algorithm, researchers can also employ alignment-filtering algorithms for alignment quality control. These algorithms attempt to identify segments of the alignments as reliable or unreliable based on statistical probabilities (e.g., Misof and Misof 2009), parallel comparisons (e.g., Penn et al. 2010) or conservation levels (Castresana 2000). The removal of sites marked as unreliable can improve the quality of the alignment and reduce error rates in subsequent analysis, as demonstrated by Privman et al. (2012) and Jordan and Goldman (2012).

Table 1.1 shows the number of citations for some alignment and alignment filtering algorithms in the year 2013. I see from this table that alignment-filtering algorithms

are much less used than alignment algorithms, and in each category, older algorithms are much more frequently used than newer ones. This represents a mindset of many evolutionary and phylogenetic biologists: they consider alignment as a trivial technicality that does not warrant scrutiny, and choose alignment algorithms by imitating earlier studies. Unfortunately, as I will demonstrate in my dissertation, this treatment of the alignment step causes a hidden, but real source of error in evolutionary analysis.

Table 1.1 Year of publication and number of citations in 2013 for a few alignment- and alignment filtering algorithms. Some important algorithms have not been included because they were presented in more than one paper and citations may overlap. All data from scholar.google.com.

Type	Alignment algorithms				Alignment filtering algorithms		
Method	CLUSTALW	MUSCLE	PROBCONS	T-COFFEE	GBLOCKS	ALISCORE	ZORRO
Year published	1994	2004	2005	2000	2000	2009	2013
Citations in 2013	2890	1960	68	407	454	20	9
Reference	Thompson et al. (1994)	Edgar (2004)	Do et al. (2005)	Notredame et al. (2000)	Castresana (2000)	Misof and Misof (2009)	Wu et al. (2009)

My dissertation focuses on the effects of alignment errors and choice of alignment- and alignment filtering algorithms on molecular evolution studies. Particularly, I want to demonstrate quantitatively that alignment errors can cause inaccuracies in estimates of branch lengths, and attempt to find methods that can mitigate them.

In Chapter Two, I develop a method to compare the branch lengths of two phylogenetic trees with the same set of taxa and topology. This method, Normalized Tree Distance, is not affected by the scale of either tree and can reflect the overall difference in proportionality.

In Chapter Three, I study the effects of alignment algorithm and alignment filtering on the accuracy of branch-length estimates in maximum likelihood phylogenetic trees, using the vector-cosine score. I simulated coding sequence under a variety of evolutionary scenarios, and used a number of different alignment and alignment filtering algorithms before producing maximum likelihood trees with defined branch lengths. I demonstrated the effects of algorithm choice, as well as its interaction with evolutionary scenarios, on the accuracy of branch-length estimates.

In Chapter Four, I analyze the correlation between the patterns of substitutions and deletions in mammalian protein-coding sequences. To measure the extent of alignment-induced artifacts, I included a parallel simulation dataset to serve as controls.

Chapter Two: A scale-free method for testing the proportionality of branch lengths between two phylogenetic trees

Introduction

In species groups with established phylogenetic relationships, such as the great apes, branch length is used to compare the rates and patterns of evolution among different lineages and among different genes. Let us consider a case of two orthologous groups of genes from the same taxa. On each branch, gene *a* will evolve at a rate determined by the rate of mutation on that branch and the selective constraints that are dictated by its function. The same applies to gene *b* that performs a different function. If the lineages under study experience mutation rates that do not change with time (but vary between genes) and if the two genes maintain their respective functions in the lineages under study, then the branch lengths of the phylogenetic tree for gene *a* (tree ***A***) will most probably be different from the branch lengths of the phylogenetic tree for gene *b* (tree ***B***); however, the corresponding branches on the two different trees will be proportional to each other. That is, dividing the length of a branch in tree ***A*** by the length of the corresponding branch in tree ***B*** will yield the same result regardless of which branch pair is chosen (Figure 2.1A, B). If, on the other hand, the selective constraints or the mutation patterns change in one or more branches, proportionality will be violated (Figure 2.1A, C).

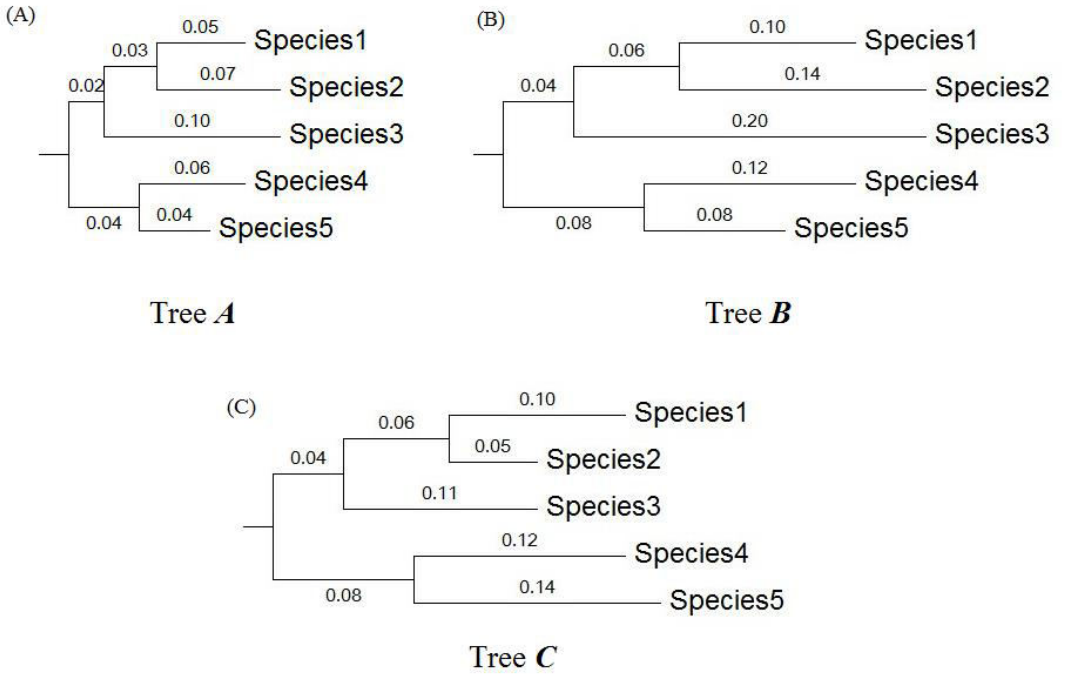

Figure 2.1 Phylogenetic trees *A*, *B*, and *C* for three hypothetical proteins, *a*, *b,* and *c*. Between *a* and *b*, there are no lineage-specific changes in the selection scheme, so the branch lengths are completely proportional; however in protein *c*, the selection has strengthened or relaxed in some branches, causing a deviation from proportionality.

Methods for comparing phylogenetic trees, especially branch lengths, can be used in studying the patterns of molecular evolution. For example, Pazos et al. (2008) compared phylogenetic trees of bacterial proteins and found that tree similarity can be predictive for protein interaction. Lovell and Robertson (2010) also suggested that the similarity of branch length ratios, or "evolution rate correlation," is an indicator of protein-protein interactions. Rosa et al. (2013) used branch length comparison as one of the methods to characterize the evolution of "barcode sequences" (stretches of mitochondrial DNA used to identify species.) When determining the accuracy of

phylogenetic tree reconstruction, the accuracy of branch lengths is an important aspect to consider (e.g., García-Pereira et al. 2011, Knowles et al. 2012).

Measures of proportionality between two trees should be free of bias caused by the scale of the two trees. Mathematically, there are two issues that must be dealt with. First, the distance between two trees should be independent of scale: resizing one (or both) of the trees by multiplying all branch lengths by a fixed number should leave the distance between the trees unchanged. For example, since trees ***A*** and ***B*** in Figure 2.1A and 2.1B are different in scale but perfectly proportional to each other, the distance between them should be zero (or their "similarity" should be 100%). Second, the distance function should be a metric in the mathematical sense, meaning that it should be symmetric and satisfy the triangle inequality. The triangle inequality implies that the distance between trees ***A*** and ***C*** in Figure 2.1 should be equal to that between trees ***B*** and ***C***. Such a scale-independent, mathematically rigorous notion of distance would be useful in a variety of contexts. In particular, scale independence prevents bias due to longer trees appearing to have larger distances from one another.

There are a few methods in the literature that compare phylogenetic trees; however, most of them only take differences in topology into account (e.g., Robinson and Foulds 1981; Nye et al. 2006). One of the very few distance measures that take both topology and branch length into consideration is the Branch Length Score (BLS) by Kuhner and Felsenstein (1994):

$$BLS = \sqrt{\sum(a_i - b_i)^2} \qquad (2.1)$$

Here a_i and b_i are the branch lengths corresponding to the *i*-th possible bipartition of all the taxa in trees ***A*** and ***B***, respectively. This measure is implemented in the popular phylogenetic package PHYLIP (Felsenstein 2005). The BLS measure, however, depends on scale: trees with longer branch lengths will produce larger BLS values. In addition, a large BLS value will be produced if the trees are proportional to each other but the rates of evolution are different. For example, the BLS between trees ***A*** and ***B*** in Figure 2.1 is 0.0255, while the BLS between trees ***B*** and ***C*** is 0.0197. Therefore, BLS can be affected by non-lineage-specific variation of evolution rate, i.e., tree scale, which makes it an inappropriate measurement of proportionality.

To counter this problem, Soria-Carrasco et al. (2007) made an ingenious modification to BLS by scaling one of the trees with a parameter K that minimizes BLS. This modified distance measure is called the "K tree Score" (KTS).

$$KTS = \sqrt{\sum_{i=1}^{N}(a_i - Kb_i)^2}\ , where\ K = \frac{\sum_{i=1}^{N}(a_i b_i)}{\sum_{i=1}^{N} {b_i}^2} \qquad (2.2)$$

This value of K is chosen to minimize the score. Because only one of the trees is scaled, this measure is not symmetrical. Here the capital "*N*" is used to denote the total number of branches, as opposed to the lower-case "*n*" which will be used to represent number of species.

Another tree-comparing algorithm that uses branch length data is Hall's CompareTrees program (as used in Hall 2005). Unlike BLS, one of the two trees has to be designated the "true tree." This branch length score (CompareTrees Score, CTS)

is calculated by averaging the relative differences between the lengths of the same branches in the two trees:

$$CTS = (\sum_{i=1}^{N} 1 - \frac{|a_i - b_i|}{a_i})/N \quad (2.3)$$

Here N is the number of branches shared by the two trees and a_i and b_i are the lengths of the shared branches. In this method, ***A*** is designated the "true tree" or reference tree, and ***B*** is the tree compared to it. This method is intended to be used when one of the trees is known to be true. Similar to KTS, this method is asymmetrical. Furthermore, if there is a very short branch in ***A***, it may produce an extremely large value due to being a denominator; this may obscure the comparisons of other branches.

Here, I propose a method for comparing phylogenetic trees that solves the mathematical challenges outlined above. The method uses what I call the normalized tree distance (NTD) and is suitable for comparing trees with the same topology and set of taxa.

Materials and Methods

The NTD method

Imagine two unrooted phylogenetic trees, ***A*** and ***B***, with the same topology and the same set of n taxa. Since the topology is identical, each tree can be described by $N = 2n–3$ branch lengths. They are denoted by $a_1, a_2, a_3, \ldots, a_N$ and $b_1, b_2, b_3, \ldots, b_N$. As a

consequence, each phylogenetic tree is represented by a vector in an *N*-dimensional space: $\boldsymbol{A} = (a_1, a_2, \ldots, a_N)$, and $\boldsymbol{B} = (b_1, b_2, \ldots, b_N)$. Comparing the trees can be done by comparing the two vectors. The measure I choose, the NTD, is derived by adding up numerical differences between each pair of branches after both trees are scaled to a total branch length of 1, then dividing the sum by 2:

$$NTD = \left(\sum\left(\left|\frac{a_i}{\sum_{j=1}^{N} a_j} - \frac{b_i}{\sum_{j=1}^{N} b_j}\right|\right)\right)/2 \tag{2.4}$$

As the added differences are all absolute values, the NTD will always be greater than or equal to 0. At the same time, the theoretical maximal value is 1; this happens when all branches with non-zero length in tree ***A*** have zero length in tree ***B*** and vice versa. In this situation, the differences will add to 2; after dividing by 2, the NTD will be 1. The range [0,1] of the NTD does not change with either number of taxa or the total length of the trees; therefore, this dimensionless measure is fully normalized.

In mathematical terms, the calculation of NTD after scaling is the L1 metric on the set of N-vectors whose nonnegative entries sum to 1. Like all mathematical metrics, NTD is therefore symmetric and satisfies the triangle inequality.

Simulated Example

Let us first compare my NTD with scores obtained by the three other methods: BLS, KTS ,and CTS. Two eight-taxon phylogenetic trees (Figure 2.2) with the same topology but different branch lengths were randomly generated and named ***X1*** and ***Y1***.

All branch lengths in tree ***Y1*** were doubled to produce tree ***Y2***, and tripled to produce tree ***Y3***. Here I will examine the properties of different tree-comparing methods using these simulated phylogenetic trees.

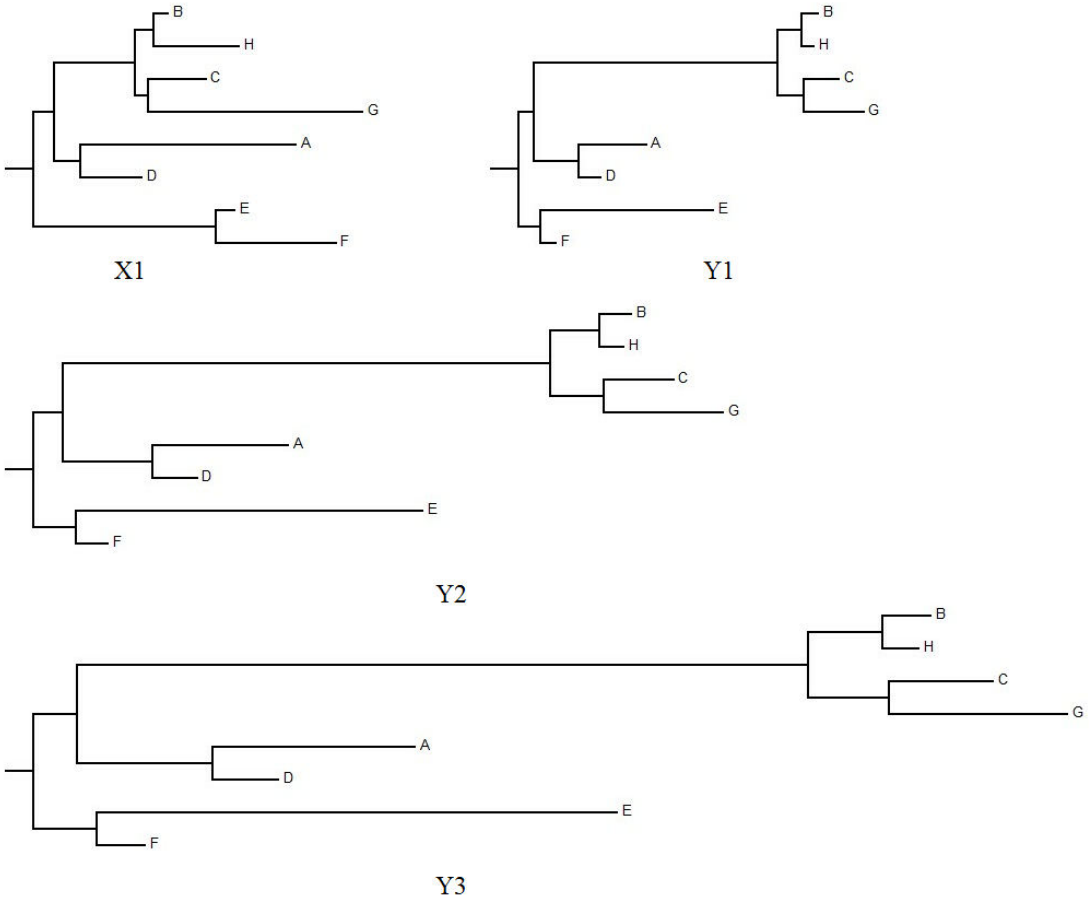

Figure 2.2 Simulated 8-taxa trees used to compare the four measures. *X1* and *Y1* are produced independently, while *Y2* and *Y3* are produced respectively by doubling and tripling each branch length of *Y1*.

Table 2.1 shows the scores given by all four methods. From the comparison between ***X1*** and ***Y1*** and between ***Y1*** and ***X1***, it is clear that both KTS and CTS produce asymmetrical results. From the comparison between ***Y1*** and ***Y2***, I see that only NTD and KTS recognize perfectly proportional trees. Also, when one of the compared trees

has very long branches (e.g., when ***X1*** and ***Y3*** are compared), the BLS and the CTS will have large absolute values, while NTD is always between 0 and 1. No matter in which order are they compared or if the branches are proportionally changed, as long as the comparison is done between an "***X***" and a "***Y***" tree, NTD will remain the same.

Table 2.1 Comparison of four tree-comparing measures used simulated eight-taxon trees (Figure 2.2). These scores are, the Normalized Tree Distance (NTD), the branch length score (BLS), the K tree score (KTS), and the CompareTrees score (CTS).

Tree Pairs	NTD	BLS	KTS	CTS
X1 / X1	0	0	0	1
Y1 / X1	0.08144	2.3054	1.71401	-1.09598
X1 / Y1	0.08144	2.3054	2.18746	-0.34057
X1 / Y2	0.08144	3.47463	2.18746	-1.63611
X1 / Y3	0.08144	5.15815	2.18746	-3.00397
Y3 / X1	0.08144	5.15815	5.14204	0.27316
Y1 / Y2	0	1.97171	0	0
Y2 / Y1	0	1.97171	0	0.5

Deriving distribution of NTD

Here I will provide a profile of how the distribution of NTD looks from real DNA and protein sequence data. I downloaded CDS and corresponding protein alignments that contain 12 well-sequenced mammal species from the online database ORTHOMAM

(Douzery et al. 2014). These species are: human (*Homo*), chimpanzee (*Pan*), macaque (*Macaca*), marmoset (*Callithrix*), rat (*Rattus*), mouse (*Mus*), guinea pig (*Cavia*), dog (*Canis*), horse (*Equus*), cow (*Bos*), pig (*Sus*), and elephant (*Loxodonta*). The website also provides the topological phylogenetic relationship among these species (Figure 2.3). Although there is controversy on the placement of horse (e.g., Zhou et al. 2012), I decided to use this external tree as the user tree for simplicity. All alignments containing unknown nucleotides or amino acids were removed. 5,140 pairs of DNA/protein alignments remained in the dataset.

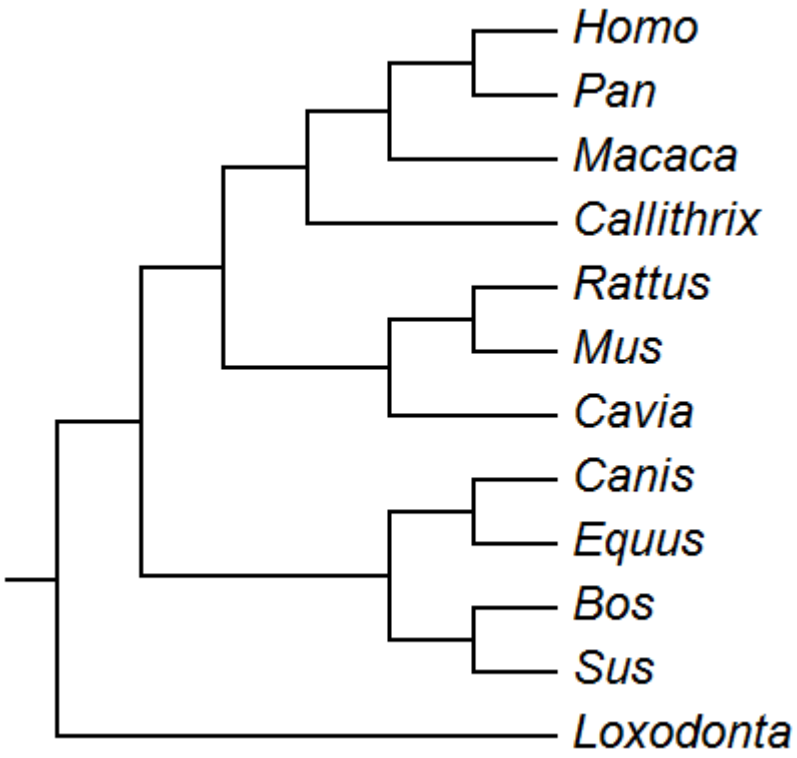

Figure 2.3 The phylogenetic tree topology of 12 mammalian species. This is a commonly accepted topology of their phylogenetic relationship, used as a guide tree for the maximum likelihood tree reconstruction.

All CDS and protein alignments were used to produce maximum likelihood trees with RAxML (Stamatakis 2006), using the GTRGAMMA model for DNA sequences and PROTGAMMAJTT model for protein sequences. A user tree (Figure 2.3) was used to guide the tree topology. Branch lengths were collected from the result for calculating NTD. Three empirical distributions were derived: among all DNA trees, among all protein trees, and between DNA and protein trees for each gene. The distributions were fit to beta, gamma, and lognormal distributions, using log likelihood as a measure of goodness-of-fit.

Distributions were fit to the NTD data by using a maximum likelihood estimation of the distributions parameters, together with resampling to ensure that the fitted distributions are robust to outliers in the given data.

For a given probability distribution with probability density function *f*, the log likelihood (LL) of data $X_1, \ldots, X_M$, given a parameter θ, is computed as $\mathrm{LL}(\theta) = \sum_{i=1}^{M} \log f(X_i|\theta)$. I normalized this log likelihood by the sample size to better compare the LL for data sets of different sizes. Therefore, instead of using LL(θ), I used $\mathrm{L}(\theta) := \mathrm{LL}(\theta)/M$.

To fit distributions to the NTD data, I used maximum likelihood estimation, together with a resampling technique to ensure that the fitted distributions are robust against outliers in the datasets. Given the original dataset $X_1, \ldots, X_P$, a family of subsamples $X_1^j, \ldots, X_{P_j}^j$ ($j = 1,\ldots, 1000$, $P_j = 1000$ for all j) was taken. For each subsample, the maximum likelihood estimate for the unknown parameter θ was computed. I call this

estimate θ_j. I then used a resampling technique to compute a log likelihood value for each θ_j. Specifically, from the original data I then sampled k = 1000 subsamples, each of size K = 1000, and used these subsamples to obtain repeated estimates $\{L(\theta_j)_i\}_{i=1}^{k}$. The log likelihood for the parameter θ_j was obtained by taking the mean $\hat{L}(\theta_j) = \frac{1}{k}\sum_{i=1}^{k} L(\theta_j)_i$. Finally, I chose the best fit estimator $\hat{\theta}$ as $\hat{\theta} = \arg\max_{\theta_j} \hat{L}(\theta_j)$.

The error on the estimate $\hat{L}(\hat{\theta})$ was obtained by computing $\sigma/\sqrt{k}$ where σ is the sample standard deviation of $\{L(\hat{\theta})_i\}_{i=1}^{k}$.

Results

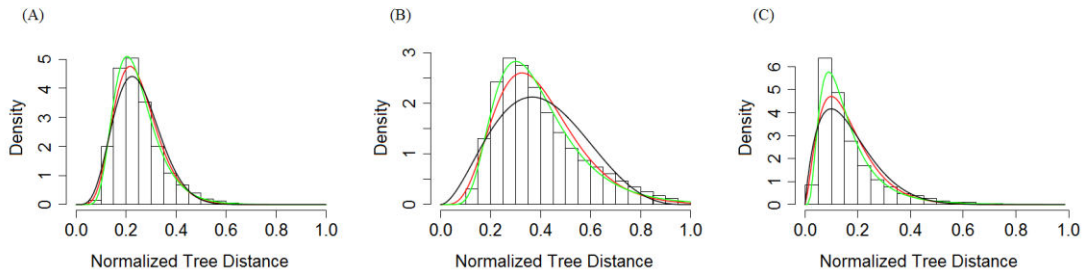

Figure 2.4 The distribution of NTD scores from a sample of mammalian gene trees. (A) Trees produced from DNA sequences are compared with one another; (B) trees produced from protein sequences are compared with one another; (C) trees produced from corresponding DNA and protein sequences are compared with each other for each gene. The curves are theoretical distributions fit to the data with a maximum likelihood method. The black curves represent the best-fit beta distribution, green curves the best-fit lognormal distribution, and red curves the best-fit gamma distribution. In all three cases, the lognormal distribution fits most appropriately.

Figure 2.4 shows empirical distributions of NTD scores in three different datasets and approximation of them to established distribution families. Although they are all

unimodally distributed, the mean score for comparisons between corresponding DNA and protein trees (Figure 2.4C) is smaller than the mean score for comparisons among DNA trees (Figure 2.4A), which is in turn smaller than the mean score for comparisons among protein trees (Figure 2.4B). This can be explained biologically, as DNA sequences have more neutral-evolving characters than proteins and the lineage-specific selection effects are weaker. Trees produced from the DNA and protein alignments of the same gene are more similar because these two sources are dependent on each other.

Table 2.2 The log-likelihood of three distribution families fit to the NTD data from protein-protein, DNA-DNA, and protein-DNA tree comparisons. In all cases, lognormal distribution appears to have the highest log-likelihood values, indicating a good fit. Error estimates are in parentheses.

Dataset	Beta distribution	Lognormal distribution	Gamma distribution
Protein-protein	0.3612 (<0.0001)	0.4482 (0.0004)	0.4372 (0.0004)
DNA-DNA	1.0107 (0.0011)	1.0644 (0.0009)	1.0447 (0.0008)
Protein-DNA pairs	0.9369 (0.0010)	1.0515 (0.0009)	0.9918 (0.0009)

Table 2.2 shows the log likelihood scores for distribution fitting. For all three datasets, lognormal distribution fits better than both beta and gamma distributions, though the scores are not highly different. Similarly, in Figure 2.4, one can see that the green curves (lognormal) fit the histograms better than the black (beta) and red (gamma) curves. Thus, the NTD scores of a sample of phylogenetic trees produced from real sequences are distributed most closely to lognormal.

Discussion

Differences between phylogenetic trees can be classified into three categories: differences in topology, scale, or proportionality. While most comparison methods (Robinson and Foulds 1981; Nye et al. 2006) focus on topology, the ones that compare branch lengths do not distinguish between scale (Kuhner and Felsenstein 1994) and proportionality, or are not symmetrical (Soria-Carrasco et al. 2007, Hall 2005). My NTD measure is useful when only proportionality information (but not scale) is needed.

For the NTD method, I chose to use the total tree length to scale each tree to total length one before comparison. If the tree is considered a vector, the scaling factor is known as the L1-norm ($a'_i = a_i / \sum_{j=1}^{N} a_j$). In mathematical algorithms that deal with vectors, however, a popular scaling alternative is the L2-norm ($a'_i = a_i / \sqrt{\sum_{j=1}^{N} a^2}$), which is the square root of the sum of the squared values. I compared the L1-scaled trees branch by branch and used the sum of the differences, known as the L1-distance ($\sum_{i=1}^{N}(|a'_i - b'_i|)$). I certainly could have used the L2-distance (Euclidean distance, $\sqrt{\sum_{i=1}^{N}(a'_i - b'_i)^2}$), since Euclidean distance has a natural geometric meaning. The main reason I chose L1 measures over L2 is the consideration of the biological meaning of branch lengths. Since branch length signifies the amount of evolutionary change from one point in the tree to another, the total length (L1-norm) is the total

amount of evolutionary change that occurred along the entire tree, while the L2-norm does not have a biological meaning.

Unlike previous methods, the NTD is itself normalized, because both trees are scaled before the comparison. All possible NTD scores are between 0 and 1. This enables the NTD to become a standardized measure of tree proportionality, which can be compared across different taxa. Another advantage of this method is that it does not need too much computation time. To calculate the NTD between two n-taxa trees, only $8n - 15$ additions/deductions and $4n - 5$ divisions are needed. No iterations are required, and the computation time increases linearly with the number of taxa.

The potential applications for such a measure include comparing the evolutionary histories of two proteins where the phylogeny among the species studied is uncontroversial or known (e.g., experimentally evolved organisms). For example, if the NTD between gene *a* and gene *b* is much lower than those between gene *a* and other genes, it is possible that genes *a* and *b* coevolve. In addition, because the absolute values of branch lengths are not important, NTD can be used to compare trees computed from different kinds of data, e.g., from DNA sequences and protein sequences. Even if the absolute branch-length values in a DNA tree and a protein tree are not comparable directly, the trees can be compared using NTD. If comparing phylogenetic trees produced by a coding gene and its protein product gives a high NTD score, it is likely that the selection pressure on this gene differs among lineages.

As the NTD is a global measure that compares entire trees, it can only be used to compare the branch length patterns over the entire trees. To pinpoint in which lineages the selective changes occur, lineage-specific measures such as dN/dS analysis are required. However, since NTD is easy and fast to compute, it is computationally efficient to first identify "interesting" tree pairs with NTD before analyzing them in depth with more sophisticated methods like maximum likelihood or Bayesian analysis. In the future, a statistical test can be devised that uses NTD scores to identify genes that evolve under different situations in a single set of species. For example, researchers can establish a lognormal distribution for one-to-one comparisons in a collection of gene trees, and look for trees that produce significantly more scores in the tail of the distribution compared to the mean.

Finally, I want to mention the possibility of using NTD for trees that are not topologically identical. In the BLS method (Kuhner and Felsenstein 1994), branches that are present in one tree but not in the other are treated as zero length in the latter tree. This can also work in my NTD measure; however, I decided not to include this aspect here, since I am skeptical as to how well a zero-length branch can represent a non-existent one in a phylogenetic tree.

Chapter Three: Multiple sequence alignment quality control improves estimates of branch length

Introduction

In phylogenetic analyses, branch lengths indicate evolutionary distance, i.e., how many changes had occurred between a hypothetical ancestral state and either another ancestral state or an extant state. Under the molecular clock hypothesis, branch lengths are expected to be proportional to the length of time between the two states and can, therefore, be used to estimate divergence time between species (Hasegawa et al. 1985). Even when the molecular clock is relaxed or absent, researchers still use branch lengths for dating speciation events, evolutionary rates, and coalescent times (Sanderson 2002; Lepage et al. 2007; Smith and Donoghue 2008; Edwards 2009). Notwithstanding the importance of branch lengths, phylogenetic studies usually emphasize the reliability of the tree topology, while under-emphasizing the reliability of branch length estimations.

Various means to improve the accuracy of branch length estimations are used. The most conspicuous such trend in the literature is the use of likelihood and Bayesian methods instead of simpler methods of phylogenetic reconstruction, such as parsimony. However, simply using a more complicated algorithm does not ensure increased accuracy. Bayesian methods are usually computationally intensive and dependent on prior settings. Indeed, priors can have an inordinate effect on branch length estimation (Marshall et al. 2006; Leaché and Mulcahy 2007; Gamble et al. 2008; Brown et al. 2010). Moreover, because of inherent characteristics of Markov-Chain Monte-Carlo

(MCMC) algorithms, the computation can easily become "trapped" in a parameter space that does not include the true value of the parameter (Marshall 2010).

Alignment quality may be a factor of branch-length accuracy. Although a theoretical optimal solution to a multiple sequence alignment problem exists, the time needed to achieve such solution increases exponentially with both sequence length and number of sequences. As a NP-complete problem (Wang and Jiang 1994), such an algorithm can exceed the computation capacity of the best computers even when no more than three or four short sequences are aligned. Therefore, for all algorithms attempting to produce multiple sequence alignments, a heuristic method must be used. Typically, the program produces a guide tree from a pairwise distance matrix, and adds sequences one by one to that guide tree. Because of the heuristic nature as well as the difficulty of exactly duplicating a biological process with a mathematical model, multiple sequence alignments usually contain a large number of errors. Landan and Graur (2009) characterized common alignment error types. One of the most common errors is to favor mismatches over gaps, thereby yielding an alignment that is shorter than it should be. Mathematical over-fitting of models may give optimal placements of gaps, while the true natural process is co-optimal or sub-optimal (Landan and Graur 2008). Alignment error rate increases with the level divergence, but the accuracy of guide trees is largely independent from it.

It is widely known that different alignment algorithms produce multiple sequence alignments of different qualities (e.g., Wang et al. 2011), and this difference can affect

downstream analysis such as phylogenetic reconstruction (Ogden and Rosenberg 2006) and estimation of positive selection (Markova-Raina and Petrov 2011, Schneider et al. 2009). Unfortunately, in studies focusing on phylogeny and divergence time, the accuracy of alignment is usually ignored (Morrison 2009).

Alignment filtering algorithms (also known as refining or masking algorithms) aim to identify regions of a multiple sequence alignment likely to contain errors and remove them from the dataset. These algorithms identify sequence conservation (e.g., Castresana 2000), use statistical hypothesis testing (e.g., Misof and Misof 2009), or compare parallel alignment attempts (e.g., Penn et al. 2010). Filtering either unreliable columns or characters can significantly decrease the false positive rate in detecting positive selection (Privman et al. 2012). However, simply removing uncertain columns does not increase the accuracy of topology by much (Landan 2005).

Unfortunately, researchers use these algorithms much less than they should when estimating phylogenies. Most phylogenetic studies take automatic alignments for granted. They may check their alignments by eye (which is not reproducible), and even if they use filtering, they are likely to use GBLOCKS (Castresana 2000), which is less accurate than newer algorithms (Privman et al. 2012).

In this study, I tested how difference in choice of alignment algorithm and refinement algorithm affect the accuracy of branch length estimates.

Materials and Methods

Simulation of coding sequences

INDELible (Fletcher and Yang 2009) was used to simulate coding sequences, because it contains a large number of controllable variables, and because its model setting is highly flexible. Below I list the constants and variables chosen for this study (Table 3.1).

Table 3.1 The list of all six variables in producing the simulated alignments.

Category	Variable	Value 1	Value 2	Value 3
Phylogenetic tree	Tree topology	Balanced	Pectinate	
	Divergence level	1 substitution per codon root-to-tip	2 substitution per codon root-to-tip	
	Internal and terminal branch lengths (branching time)	Longer internal branches ("ancient" branching)	Equal internal and terminal branches	Longer terminal branches ("recent" branching)
	Variation in branch lengths	Small variation (longest is 2.25× shortest)	Large variation (longest is 9× shortest)	
Indel model	Indel frequency relative to substitutions	0.0429	0.0857	0.1286
	Insertion vs deletion rate	Equal	Unequal	

The site-wise distribution of dN/dS values was drawn from the data of Lindblad-Toh et al. (2011) which contain more than 12,000 genes in 29 mammal species. The part of

dN/dS distribution above 1 was truncated. Due to INDELible's requirement, the distribution was further binned into 50 categories, each represented by the middle value (0.01, 0.03… 0.99).

The ratio of transition to transversion mutations was set to 3.6. The rates of insertion and deletion (frequency compared to substitution) were produced from an earlier study on human pseudogenes (Zhang and Gerstein 2003), which gave the insertion rate to be 0.0291 and deletion rate to be 0.0566. To examine the effects of total indel rates and relationship between insertion and deletion rates, these numbers were varied in simulation. First, using × 0.5, × 1, and × 1.5 of the pseudogene indel rate, three levels of indel rates were produced; second, both "equal" (using a mean value for both insertion and deletion) and "unequal" insertion and deletion rates were used. For the distribution of indel size, I used Jordan and Goldman's (2012) value, a truncated power-law distribution with a parameter of 1.8 and a maximum size of 40 codons.

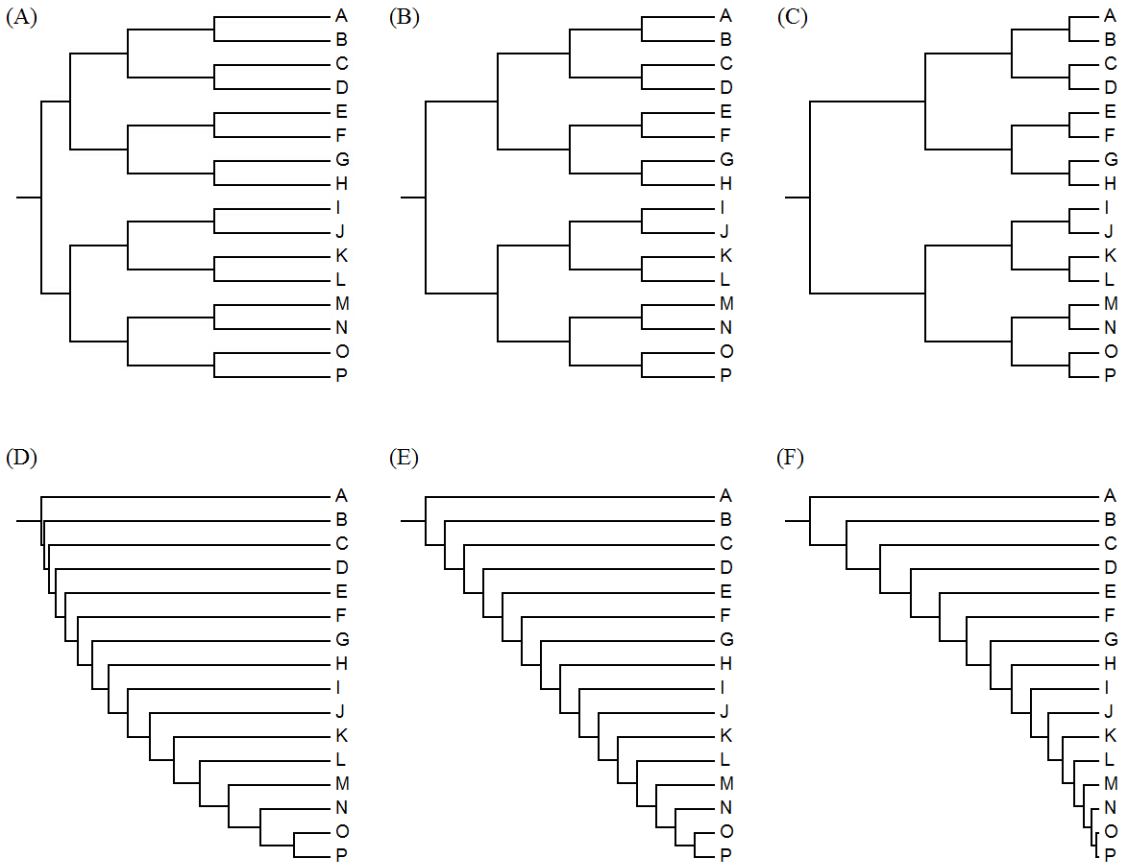

Figure 3.1 The phylogenetic trees before introducing of random variation in branch lengths. The topologies are balanced (A,B,C) and pectinate (D,E,F). The "branching times" are "ancient" (A,D), "intermediate" (B,E), and "recent" (C,F).

Next, 16-species phylogenetic trees were used as the "correct trees" for the simulations. First, in half of the trees, the topology was perfectly balanced (Figure 3.1A,B,C); in the other half it was thoroughly pectinate (Figure 3.1D,E,F). Second, two different divergence levels were used, measured by mean root-to-tip distance, being 1 and 2 substitutions per codon. Third, "ancient," (Figure 3.1A,D) "intermediate," (Figure 3.1B,E) and "recent" (Figure 3.1C,F) branching was achieved by varying the ratio of internal and terminal branch lengths. In "ancient" branching, I

set the ratio of branch lengths from the most internal to the terminal is 1:2:3:4 for balanced trees, and 1:2:3:…:15 for pectinate trees (for early branched taxa, like A~N in Figure 3.1D, the terminal branch is set to make root-to-tip distance same in each taxon; same for "recent" branching.) In "intermediate" branching, the branch lengths of every level were the same. In "recent" branching, the ratio was 4:3:2:1 for balanced trees and 15:14:13:…:1 for pectinate trees. I did this because it was previously demonstrated that trees that have "ancient" branching (longer terminal than internal branches) are more difficult to reconstruct (Cantarel et al. 2006). At this step, all trees were ultrametric. Finally, every branch length was multiplied with an independent random number log-uniformly distributed between (1/K, K) (density function: $f(x) = \frac{1}{2x \times \ln(K)}$), then with a constant ($\frac{K^2-1}{2K \times \ln(K)}$). K is 1.5 for "small branch length variation" and 3 for "large branch length variation." In this way, the randomized branch length has an expectation of the original branch length in the ultrametric tree. Therefore, 24 trees are produced, each with a different combination of these four variables.

With 6 different indel models and 24 different trees, there are a total of 6 variables and 144 sets of simulation data (see Table 3.1 for the list of variables and their values); for each set, 50 replicates were produced, and for each replicate, the initial length is 1,000 codons.

Alignment, filtering, and phylogenetic reconstruction

ClustalW (Thompson et al. 1994) is one of earliest multiple sequence alignment programs. It is the most commonly used alignment program in phylogenetic studies (Yamamoto et al. 2000; Regier et al. 2008; Morrison 2009). However, comparisons showed that ClustalW produces relatively less accurate alignments (Edgar and Batzoglou 2006, Thompson et al. 2011) than more recently developed algorithms such as MAFFT (Katoh et al. 2005), MUSCLE (Edgar 2004), and ProbCons (Do et al. 2005). In addition, MAFFT has an option called L-ins-I (abbreviated as MALINSI in this chapter) where a few more iterative steps are added after the default MAFFT alignment. This option has been shown to be superior to the default option of MAFFT (Katoh and Toh 2008, Nuin et al. 2006).

Seven different alignment algorithms were used to align the simulated sequences. In addition to the ones introduced in the last paragraph, I used T-COFFEE (Notredame et al. 2000) and Clustal-Omega (abbreviated as ClustalΩ) in this chapter, Sievers et al. 2011). Because of the mechanism of one refinement program (GUIDANCE), CLUSTALW, MAFFT and MUSCLE were used in joint with GUIDANCE in codon align mode, which produce both pre-refining and post-refining alignments. For the other algorithms, which are PROBCONS, MALINSI, ClustalΩ, and T-COFFEE, the unaligned sequences were first translated into proteins and aligned, and DNA alignments are back-constructed from protein alignments. In addition, the true alignment was included as a control to see how much variation in the final results is

due to alignment errors. The accuracy of each reconstructed alignment was calculated as the proportion of nucleotide pairs that are correctly aligned, and "alignment error" is one minus the accuracy.

All alignments, including the true alignment, were processed through a number of different refinement algorithms.

One of the first alignment-refining algorithms was GBLOCKS (Castresana 2000). It was written specifically to find conserved regions (blocks) in an alignment (e.g., Kück et al. 2010; Göker et al. 2011; Privman et al. 2012). It is the most frequently used alignment filtering program in phylogenetic studies (e.g., Rodríguez-Ezpeleta et al. 2005; Fitzpatrick et al. 2006; Philippe et al. 2009). The option for gap permission were set to either "all" (more lenient) or "half" (more stringent), producing two parallel sets of results.

GUIDANCE (Penn et al. 2010) was developed using an approach of comparing alignments of same sequences under different guide trees, based on an earlier program named HoT (Landan and Graur 2007), which simply compared the alignment produced from the 5' to 3' sequence to that produced from the 3' to 5' sequence. The percentage of "unreliably aligned" columns according to GUIDANCE is usually similar to that identified by HoT. Because GUIDANCE must be used in conjunction with compatible alignment algorithms, I could only filter CLUSTALW, MUSCLE, and MAFFT alignment with GUIDANCE. Because the default cut-off value removes virtually all sites in many alignments, alternative criteria were used: 30%, 50% or 70%

of lowest-scoring codons are masked (GUIDANCE-30%, GUIDANCE-50%, and GUIDANCE-70% respectively), changed to "X," and the corresponding nucleotides are changed to "NNN" indicating missing data.

ALISCORE (Misof and Misof 2009) is a probability-based alignment filtering algorithm. It tests the statistical significance of aligned columns against a null model of random association. ALISCORE uses a random permutation of small sliding windows, and tends to remove a large proportion of sequences among conservative but distant sequences (von Reumont et al. 2009). Only the default option was used in my study.

ZORRO (Wu et al. 2012) calculates the posterior probability of a pair of aligned characters. A higher probability indicates that they are reliably aligned. Currently, ZORRO can only be used on protein sequences. The cut-off score was set to 4, 5 or 6 (henceforth referred to as ZORRO-4, ZORRO-5, and ZORRO-6, respectively).

The multiple sequence alignment package T-COFFEE (Notredame et al. 2000) includes an "evaluation mode" that can assign scores to each character of an externally provided alignment. In the evaluation mode, a score ("CORE," Consistency of Overall Residue Evaluation) is calculated based on the appearance of character pairs in a library of pairwise alignments (Notredame and Abergel 2003). This function can be used as an independent alignment filtering algorithm. I masked low-scored amino acid residues, and the cut-off scores were set to 3, 5, or 7 (T-COFFEE-3, T-COFFEE-5,

and T-COFFEE-7 respectively). In both ZORRO and T-COFFEE, a lower cut-off value indicates a more lenient filtering. Both algorithms produced three sets of results.

In highly diverged sequence alignments, refining may cause too few sites to remain for phylogenetic purposes. Therefore, I removed all filtered alignments with less than 100 columns of 2 or more non-gap codons.

Maximum likelihood trees were produced from all refined and unrefined alignments, with the program RAxML (Stamatakis 2006). This program was used because it can generate maximum likelihood trees rapidly. The user tree option was used to produce trees that are topologically identical as the true trees.

Evaluation and analysis of branch length estimates

A fully resolved phylogenetic tree of 16 taxa has 29 branches, so the data would be very difficult to analyze if the accuracies of branch lengths are estimated one by one. Instead, a measure that will produce one value from the comparison of two trees is needed. I chose the NTD (Normalized Tree Distance) score (see Chapter 2), a symmetric score that measures the proportionality of trees without interference from scale. To calculate NTD, both trees are scaled to a total length of 1, and the absolute differences between corresponding branches are added up and halved. NTD will always be in the range of [0, 1]. For two n-taxa trees ***A*** and ***B*** whose branches are a_1 to a_{2n-3} and b_1 to b_{2n-3} respectively, the calculation is:

$NTD = (\sum(\left|\frac{a_i}{\sum a} - \frac{b_i}{\sum b}\right|))/2$, where $\sum a = a_1 + a_2 + \cdots + a_{2n-3}$, and $\sum b = b_1 + b_2 + \cdots + b_{2n-3}$

The scores were analyzed with ANOVA to determine which factors (6 simulation variables and 2 method variables) affect the accuracy of branch length estimation. During the preliminary test, 2 of these variables (variance of branch lengths within a tree and ratio of insertion to deletion rates) did not express any large individual or two-way interaction effects (all mean squares <15). These variables were subsequently removed from the analysis. Another ANOVA was done with the six remaining variables: tree topology, branching time (relative internal branch length), divergence level, indel rate, alignment algorithm, and alignment refining algorithm. Single-variable and two-way interaction effects are studied. Paired t-tests for NTD score among alignment algorithms were performed.

A similar ANOVA was also performed on alignment accuracy (in which only unfiltered alignments are used), to see if the effects of alignment choice is consistent in alignment accuracy and branch-length score. Linear regression was used to analyze the correlation between alignment accuracy and branch length accuracy. Paired t-tests for alignment accuracy among algorithms were performed.

To find which combination of alignment and refinement algorithms could produce the most accurate estimation of branch lengths, a ranking was done for each evolutionary scenario. I use the phrase “evolutionary scenario” to mean a combination of tree topology, relative internal branch length, divergence time, and indel rate. Paired t-tests

(with Bonferroni correction) were performed to see if the difference between the best method combination and its unfiltered/T-COFFEE-3 filtered counterpart is significant.

To see how alignment quality correlates with the estimation of branch length, regression and correlation analysis was performed between alignment score and branch length score. Because alignment score was calculated only for unrefined alignments, only the branch length score from unrefined alignments were used here. For each evolutionary scenario the coefficient of determination (r^2) was calculated.

All branch length scores were recalculated using the tree reconstructed from the true alignment instead of the true tree as the reference. ANOVA and ranking analysis were similarly conducted, to disentangle the effect of errors from phylogenetic reconstruction algorithm from that of alignment errors.

Changes in tree scale after filtering

To evaluate to what extent alignment filtering causes the tree to shrink in scale, all trees produced from filtered alignments were compared to their unfiltered counterpart. The score used to represent scale change is mean of the logs of the ratios between filtered and unfiltered tree branches is $Scale - score = \frac{\sum_{i=1}^{2n-3} \ln(a_i/b_i)}{2n-3}$, where a_i and b_i are corresponding branches from filtered and unfiltered alignments. Mean scores for each filtering methods were computed, and paired t-tests were conducted.

Results and Discussion

Alignment and filtering

I produced by simulation 144 datasets differing in six variables, each with 50 replicates. There are 7,200 true alignments in total. Including the true alignments, 57,600 alignments are generated. Although one set of sequences can produce 97 filtered and unfiltered alignments, some of them have too many columns removed to be used in phylogenetic studies. I removed all filtered alignments with less than 100 codons and classify them as missing data.

Factors affecting branch length accuracy

Table 3.2 ANOVA showing effects of topology, degree of divergence, indel rate, alignment method, and filtering, as well as their two-way interactions on branch length accuracy. All variables and interactions have p-value below 3.3×10^{-15}after Bonferroni correction.

Variable	Sum of squares	Percentage of Variation Explained
Topology	23.58	1.18
Branching Time	30.54	1.53
Divergence	160.04	8.02
Indel Rate	415.09	20.79
Alignment Algorithm	487.75	24.43
Alignment Filtering Algorithm	34.52	1.73
Topology × Branching Time	1.73	0.09
Topology × Divergence	1.99	0.10
Branching Time × Divergence	2.86	0.14
Topology × Indel Rate	3.14	0.16
Branching Time × Indel Rate	1.53	0.08
Divergence × Indel Rate	47.31	2.37
Topology × Alignment Algorithm	14.05	0.70
Branching Time × Alignment Algorithm	4.2	0.21
Divergence × Alignment Algorithm	48.48	2.43
Indel Rate × Alignment Algorithm	104.92	5.25
Topology × Alignment Filtering Algorithm	14.36	0.72
Branching Time × Alignment Filtering Algorithm	6.27	0.31
Divergence × Alignment Filtering Algorithm	20.61	1.03
Indel Rate × Alignment Filtering Algorithm	13.54	0.68
Alignment Algorithm × Alignment Filtering Algorithm	51.82	2.60
Higher Interactions	151.12	7.57
Residuals	357.16	17.89
Total	1996.61	100

After a series of preliminary ANOVA tests, I removed two variables from the analyses (variance of individual branch lengths in the tree and ratio of insertion to deletion rates) due to weak direct and interaction effects. I then did another ANOVA with the remaining six variables. The combinations of tree topology, branching time (relative internal branch length), divergence, and indel rate are called evolutionary scenarios (total 36 scenarios). All six variables and their 15 two-way interactions are highly significant ($p < 0.0001$), but the sums of squares, which tell us how much variance is explained, are vastly different (Table 3.2). The choice of alignment method alone explains 24.4% of the total variance. The choice of alignment filtering explains 1.7%, and the interaction between alignment and filtering methods explains 2.6%. Therefore, the total effect of method is 28.7%. On the other hand, the largest evolutionary-scenario-rated source of variation is indel rate, which is 20.8%; the total effect of evolutionary scenarios is 34.8%. Since the residuals are 17.9%, the interaction between method and evolutionary scenario explains 18.6% of the variance. Because my evolutionary scenarios cover a very large range of indel rates and divergence levels, these results shows that choices of alignment and filtering methods have a very strong effect on the branch-length score.

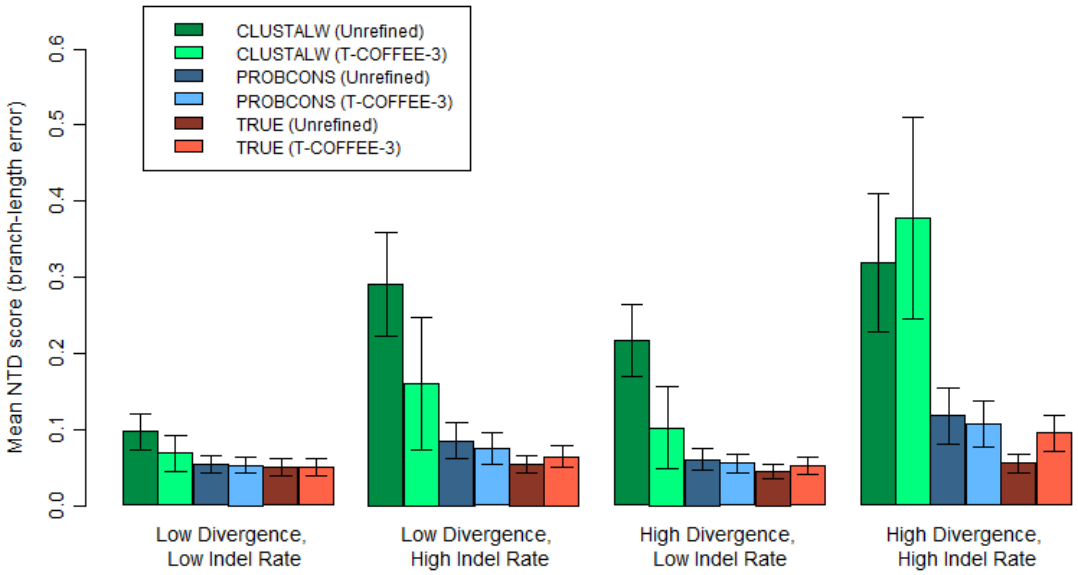

Figure 3.2 Mean Normalized Tree Distance scores under different divergence, indel rate, and a few chosen method combinations. Error bars are standard deviations. A higher score indicates less accurate reconstruction. In all 4 combinations of parameters, PROBCONS provide better results than CLUSTALW, and T-COFFEE refining further improves the accuracy.

Figure 3.2 shows the a few examples of method combinations. No matter what the level of divergence and indel rate are, PROBCONS alignments produce lower NTD than CLUSTALW – closer to true alignment. At the same time, T-COFFEE-3 filtering improves the estimation in PROBCONS alignments by a smaller amount. Judging by the mean for each alignment algorithm (Figure 3.3), PROBCONS and T-COFFEE produce the most accurate branch-length estimates in all reconstructed alignments; MALINSI, CLUSTALΩ, and MUSCLE are closely followed. While MAFFT produces lower branch-length accuracies, CLUSTALW is the worst of all these algorithms by a wide margin.

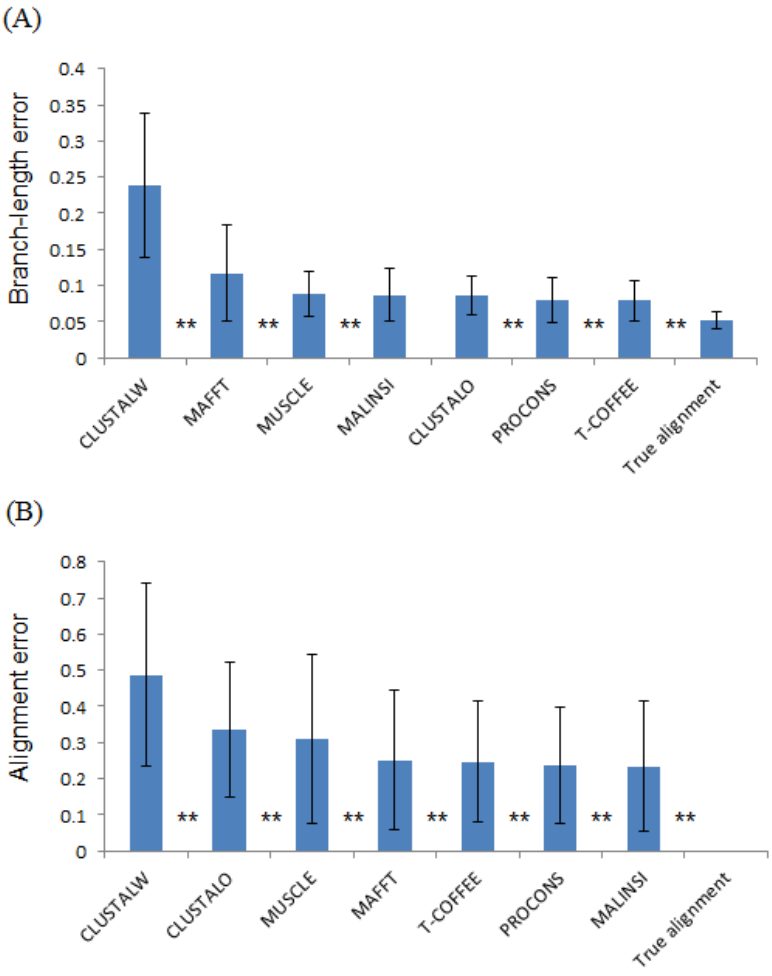

Figure 3.3 Mean (A) alignment error and (B) Normalized Tree Distance scores (from unfiltered alignments) by alignment algorithm, ordered from the most to least amount of error. Error bars are standard deviations. ""between two adjacent columns indicating significant difference ($p < 0.01$) after Bonferroni correction.**

Interactions among evolution scenario and methodology variables

The interaction between evolutionary scenario and method combination is evident. I found that when the tree topology, divergence, and indel rate varies, the optimal method combination also changes. In Table 3.3, the best methods of each evolutionary

scenario are provided, and t-tests between the best method and its filtered/unfiltered counterparts are described.

Table 3.3 The best method combination for each evolutionary scenario (by mean Normalized Tree Distance score). The p-values of filtering are based on t-test between the best method and its T-COFFEE-3/unfiltered counterpart. For example, PROBCONS is compared to PROBCONS_T-COFFEE-3, while MAFFT_T-COFFEE-3 is compared to MAFFT. Significant difference between filtered and unfiltered alignments are marked with "" (α = 0.01) or "*" (α = 0.05), after Bonferroni correction.**

Tree Topology	Relative Terminal Branch	Divergence	Indel Rate	Best Method Combination	P-Value of filtering
Balanced	Long (Ancient Branching)	Low	Low	PROBCONS_T-COFFEE-3	0.006905
			Medium	PROBCONS_T-COFFEE-3	0.00011 **
			High	PROBCONS_T-COFFEE-3	<0.0001 **
		High	Low	PROBCONS	<0.0001 **
			Medium	PROBCONS	0.002196
			High	T-COFFEE	0.00133 *
	Medium (Intermediate Branching)	Low	Low	PROBCONS_T-COFFEE-3	0.475491
			Medium	PROBCONS_T-COFFEE-3	0.000562 *
			High	PROBCONS_T-COFFEE-3	<0.0001 **
		High	Low	PROBCONS	0.094791
			Medium	T-COFFEE	<0.0001 **
			High	T-COFFEE	<0.0001 **
	Short (Recent Branching)	Low	Low	PROBCONS	0.320608
			Medium	PROBCONS_ZORRO-4	0.022583
			High	T-COFFEE_T-COFFEE-3	<0.0001 **
		High	Low	PROBCONS	0.008921
			Medium	T-COFFEE_T-COFFEE-3	0.227178
			High	T-COFFEE	<0.0001 **
Pectinate	Long (Ancient Branching)	Low	Low	MAFFT_T-COFFEE-5	<0.0001 **
			Medium	MAFFT_T-COFFEE-3	<0.0001 **
			High	MAFFT_T-COFFEE-3	<0.0001 **
		High	Low	MALINSI_T-COFFEE-3	<0.0001 **
			Medium	PROBCONS_T-COFFEE-3	<0.0001 **
			High	CLUSTALO	<0.0001 **
	Medium (Intermediate Branching)	Low	Low	MAFFT_T-COFFEE-5	<0.0001 **
			Medium	MAFFT_T-COFFEE-3	<0.0001 **
			High	MAFFT_T-COFFEE-3	<0.0001 **
		High	Low	MAFFT_T-COFFEE-3	<0.0001 **
			Medium	MAFFT_T-COFFEE-3	<0.0001 **
			High	CLUSTALO	<0.0001 **
	Short (Recent Branching)	Low	Low	MAFFT_T-COFFEE-5	<0.0001 **
			Medium	MAFFT_T-COFFEE-5	<0.0001 **
			High	MAFFT_T-COFFEE-5	<0.0001 **
		High	Low	MAFFT_T-COFFEE-5	<0.0001 **
			Medium	MAFFT_T-COFFEE-3	<0.0001 **
			High	CLUSTALO	<0.0001 **

The data suggested that, in different evolutionary scenarios, different method combinations produce the most accurate branch lengths. With balanced trees, PROBCONS appears the most the top list; in a few scenarios T-COFFEE alignments are in the top, and MALINSI and MUSCLE occasionally occupying the second or third place. In balanced trees and low level of divergence T-COFFEE-3 filtering is used in the best combination, although ZORRO-4 and unfiltered alignments top the list once each. In 5 of the 9 scenarios (mostly with higher indel rates), T-COFFEE-3 provides a significant benefit. On the other hand, sequence alignments evolved along balanced trees with high level of divergence are best left unfiltered, being top 8 of 9 cases and significantly better than their T-COFFEE-3 filtered counterpart in 5 of them.

Pectinate trees produce a very different set of outcomes. Instead of PROBCONS or T-COFFEE, alignments by MAFFT (default option) occupy almost all top spots, except when the divergence and indel rate are both high. These best methods are often MAFFT alignments (with two exceptions) filtered by T-COFFEE-3 or T-COFFEE-5, and are highly significantly better than unfiltered MAFFT alignments ($P < 0.0001$ in 13 scenarios). Under high divergence and high indel rate (scenarios highly improbable in real data), the best method is unfiltered CLUSTALΩ.

I have shown that the best alignment algorithm depends on the tree shape. Therefore, at least in my case, an objectively “best” alignment algorithm does not exist; a sensible researcher would choose their algorithm(s) based on their needs. Of course, sometimes the tree shape cannot be known without aligning the sequence and building a

phylogeny, especially in taxa that are not studied extensively before. To break this catch-22, a preliminary phylogeny can be made and the resulting tree topology can be used to guide the main alignment.

When different filtering algorithms are compared, in almost all evolutionary scenarios, the best were filtered with the evaluation mode of T-COFFEE. With one exception, T-COFFEE produces the best filtered alignment in all evolutionary scenarios; most of them are under the stringency level T-COFFEE-3. Lenient filtering performs better than stringent filtering, likely because less informative sites are removed.

A recent study (Chang et al. 2014) also demonstrated that T-COFFEE evaluation have better sensitivity and specificity compared to other alignment filtering algorithms. They used an improvement of the T-COFFEE CORE evaluation which I did not use because it was developed after I conducted the studies.

There are studies arguing that MALINSI provides better results than MAFFT's default mode (Katoh and Toh 2008, Nuin et al. 2006), and some even claim it is consistently superior to PROBCONS (Ahola et al. 2006). However, I showed that, while in term of mean effects MALINSI performs better than the MAFFT default mode in both the alignment accuracy and branch-length score, it appears less frequently in the top three lists (Table 3.3). Specifically, in the scenarios where the tree topology is pectinate, filtered MAFFT alignments often yield higher branch-length score than MALINSI alignments filtered by the same method. However, in all but one case, unfiltered MALINSI yields higher branch-length score than MAFFT default, and even in that

case the difference is very marginal. It is possible that although MAFFT's default mode is less accurate than MALINSI, it has more space for improvement for alignment filtering, especially with pectinate trees.

Correlation between alignment accuracy and branch-length accuracy

Table 3.4 ANOVA showing effects of topology, degree of divergence, indel rate, and alignment method, as well as their two-way interactions, on alignment accuracy. All variables and interactions have p-value below 0.0001.

Variable	Sum of squares	Percentage of Variation Explained
Topology	27.92	1.19
Branching Time	117.19	4.99
Divergence	808.24	34.41
Indel Rate	748.81	31.88
Alignment Algorithm	361.42	15.39
Topology × Branching Time	36.42	1.55
Topology × Divergence	3.65	0.16
Branching Time × Divergence	19.56	0.83
Topology × Indel Rate	3.89	0.17
Branching Time × Indel Rate	16.22	0.69
Divergence × Indel Rate	58.74	2.50
Topology × Alignment Algorithm	2.71	0.12
Branching Time × Alignment Algorithm	9.97	0.42
Divergence × Alignment Algorithm	21.22	0.90
Indel Rate × Alignment Algorithm	19.02	0.81
Higher Interactions	28.58	1.22
Residuals	65.48	2.79
Total	2349.04	100

In the ANOVA test for alignment accuracy (Table 3.4), all four variables of evolutionary scenario, choice of alignment algorithm, as well as all second-order interactions have $p < 0.0001$. However, evolutionary scenarios place a larger role in alignment accuracy variance compared to the NTD data; both divergence and indel rate explain over 30%, while alignment algorithm explains only 15.3%. Judged by the mean (Figure 3.3), PROBCONS and MALINSI are most accurate, followed by T-COFFEE and MAFFT (default), then MUSCLE and Clustal-Omega, and CLUSTALW is the worst. This is consistent with previous studies (e.g., Wang et al. 2011) in which PROBCONS and MAFFT were described as superior while CLUSTALW are less accurate. I can observe that the order of alignment accuracy and branch-length accuracy are not necessarily the same; for example, MAFFT alignments are substantially more accurate than MUSCLE, but the NTD score shows a higher level of error.

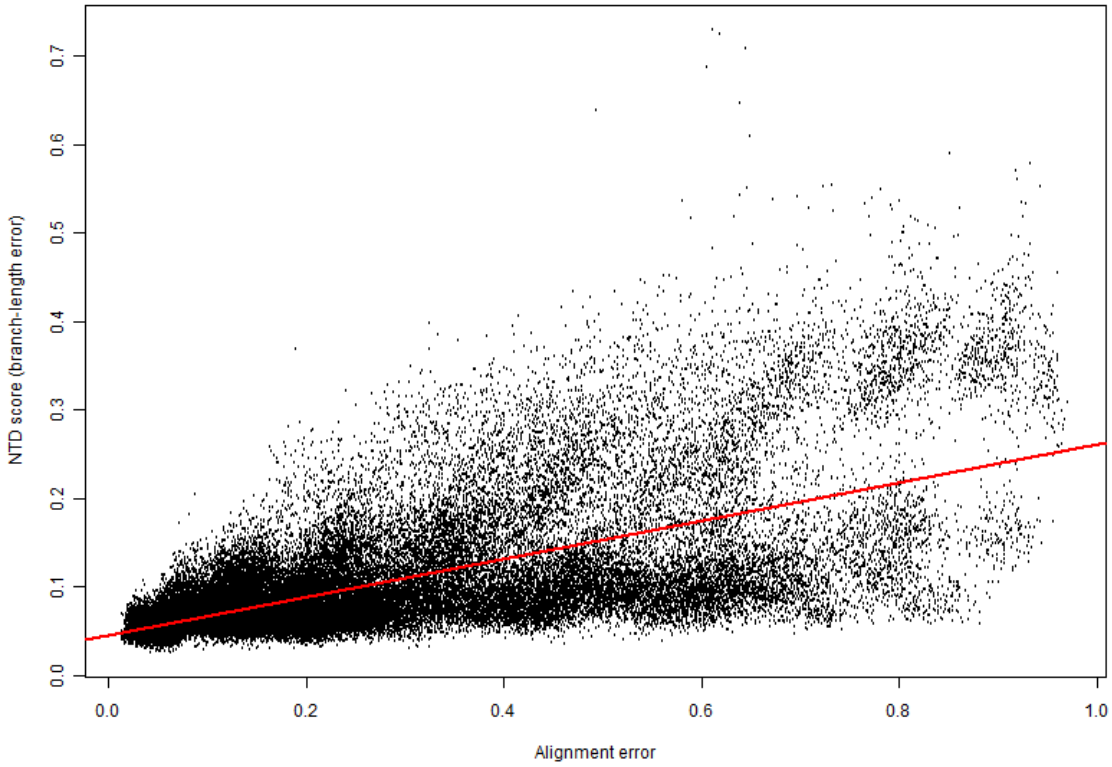

Figure 3.4 The relationship between alignment error and Normalized Tree Distance score (indicating error of branch-length estimates). The red line indicates the linear regression. There is a clear trend that with higher alignment error, the error score is also higher.

The linear regression between the alignment error and NTD score is described in Figure 3.4. I use the error instead of accuracy because NTD is an error measure here, and using the accuracy will give a negative correlation. There is a significant ($p < 2\times10^{-16}$) correlation, and the coefficient of correlation is $r = 0.62$, with the regression formula of NTD = 0.26119 – 0.21500 × (alignment accuracy). The linear regressions by each evolutionary scenario rate are all significant (all $p < 2\times10^{-16}$). Table 3.5 and 3.6 show the correlation coefficients associated with each alignment scenario. The r

(Table 3.5) ranges from 0.51 to 0.91; however, a visible trend is that the easier the alignment is (with low indel rate and level of divergence, and preferably balanced tree), the higher the coefficient of determination is.

Table 3.5 The correlation coefficient between alignment accuracy and Normalized Tree Distance (measure of branch-length error) in each evolutionary scenario.

		Balanced			Pectinate		
Divergence	Branching \Indel	Ancient	Inter-mediate	Recent	Ancient	Inter-mediate	Recent
	Low	0.82	0.78	0.80	0.74	0.71	0.61
	Medium	0.88	0.88	0.91	0.86	0.79	0.72
Low	High	0.86	0.84	0.86	0.87	0.80	0.74
	Low	0.91	0.87	0.91	0.91	0.88	0.83
	Medium	0.78	0.78	0.80	0.69	0.65	0.70
High	High	0.61	0.64	0.60	0.57	0.51	0.53

Table 3.6 The correlation coefficient between alignment accuracy and Normalized Tree Distance (measure of branch-length error) in each evolutionary scenario, after controlling for alignment algorithm choice.

		Balanced			Pectinate		
Divergence	Branching \Indel	Ancient	Inter-mediate	Recent	Ancient	Inter-mediate	Recent
	Low	0.13	0.30	0.23	0.12	0.26	0.17
	Medium	0.11	0.24	0.14	0.16	0.20	0.22
Low	High	0.04	0.13	0.11	0.13	0.12	0.09
	Low	0.06	0.16	0.15	0.12	0.10	0.12
	Medium	0.06	0.08	0.11	0.10	0.07	0.06
High	High	0.08	0.11	0.17	0.05	0.04	0.08

To test if the choice of alignment algorithm plays a great role in this correlation, I re-calculated r controlling for alignment algorithm. In the complete dataset, r dropped to 0.41, while in individual evolutionary scenarios (Table 3.6), r dropped below 0.3, with the lowest being 0.04. Therefore, it is most likely that this correlation largely result from the choice of alignment algorithm; a good algorithm (like PROBCONS, T-COFFEE) produces good alignments and accurate branch-lengths, while a bad one (like CLUSTALW) produces low-quality alignments and inaccurate branch-lengths.

Branch length accuracy measured in reference to the reconstructed tree from true alignment

I produced branch-length scores using the trees generated from true alignments instead of true input trees as references, to separate the effect of alignment errors from sampling error during simulation and systematic bias of the tree-producing algorithm. I performed an ANOVA test, but the effects of all variables are very close to those calculated from the main dataset. The only major difference being tree topology only explains 0.54% instead of 1.18% of variance. I reason that, because the sequences are relatively long (1,000 codons) and RAxML is a high-accuracy phylogenetic algorithm, the effects of sampling error and tree bias when using the true alignment are minimal. However, the effects may be a bit different between balanced and pectinate trees; since pectinate trees are a bit difficult to reconstruct, this effect may also exist in the trees built from true alignments. This may partially offset the errors in the trees by inferred alignments, reducing the difference of NTD scores between balanced and pectinate trees.

Change in scale of phylogenetic trees

The NTD method I adopted to evaluate branch length accuracy does not incorporate increases or decreases in the tree scale. For example, if the three branches in one tree are 0.1, 0.2, and 0.3, while their counterparts are 0.2, 0.4, 0.6, the score will be 0. To observe how alignment filtering causes the tree scale to change, I used the log

geometric mean of all branch length ratios to score a tree produced by a filtered alignment. A negative score indicates decreased tree branch lengths.

Table 3.7 describes the effects of each filtering method on the branch lengths in general. All 12 filtering methods and leniency levels significantly reduce the scale of the tree ($p < 0.0001$ for all). The filtering method that was shown to be most helpful to increase accuracy, T-COFFEE-3, can cause a reconstructed branch to be ~26% shorter on the geometric average.

Table 3.7 Mean proportional branch length reduction for each filtering method, ordered by paired t-test result. Note some methods have lower mean scale-score but are significantly higher in t-test comparison; this is mainly because of the interaction among evolutionary scenario, alignment, and filtering methods. Different letters in "t-test significance" column indicates significant difference.

Filtering Method	Mean Scale-score	Expected Branch length if the unfiltered alignment gives 1.0	t-test significance
GUIDANCE-30%	-0.30	0.74	A
T-COFFEE-3	-0.30	0.74	B
GBLOCKSA	-0.29	0.75	C
GBLOCKSH	-0.33	0.72	D
ALISCORE	-0.36	0.70	E
ZORRO-4	-0.32	0.72	F
GUIDANCE-50%	-0.48	0.62	F
ZORRO-5	-0.36	0.70	G
ZORRO-6	-0.47	0.62	H
GUIDANCE-70%	-0.70	0.49	H
T-COFFEE-5	-0.61	0.55	I
T-COFFEE-7	-0.86	0.43	J

This tendency of alignment filtering to decrease branch lengths in reconstructed phylogenies, due to the removal of fast-evolving regions, is seldom mentioned. This study shows that alignment filtering does cause a global bias towards shorter branches. This may cause a number of problems in phylogenetic studies that require absolute values of branch lengths, for example estimation of mutation rates and divergence time.

Caveats and limitations

While the only phylogeny method I used is maximum likelihood, there are other commonly used ones such as neighbor-joining (Saitou and Nei 1987) and Bayesian (e.g., Ronquist and Huelsenbeck 2003). However, neighbor-joining does not allow the input of a user tree, and the Bayesian methods take so much time that it is infeasible to use it on all the alignments even if a large computational cluster is used.

Because an incorrect topology will complicate the scoring of branch lengths, user trees are used to ensure that there is no topology incongruence. Admitted, this is unrealistic, because I do not know the branching order of real species for sure. This problem can be resolved if one score that take both topology and branch length into consideration (e.g., Kuhner and Felsenstein 1994) is used; however, I have chosen to explorer the accuracy of branch lengths alone (independent from topology) in this study.

Conclusion

This study has shown that alignment quality has a profound effect on the accuracy of branch length estimation in maximum-likelihood phylogenetic trees. It is also shown this accuracy can be significantly improved by choosing a suitable combination of alignment and filtering algorithms. The choice of algorithms depends on the nature of the sequence data, particularly the tree topology (if known), and the degree of divergence.

The application of alignment-filtering algorithms can improve the accuracy in estimation of branch lengths. Therefore, researchers should choose their methods of obtaining multiple-sequence alignments carefully, and based on their needs, rather than use the most popular method (i.e., ClustalW without filtering) for all purposes.

Chapter Four: Correlated selection on amino-acid deletion and replacement in mammalian protein sequences

Introduction

Functional constraint is defined as the limitation of nucleotides that can appear in a site while keeping the gene's function undisrupted (Miyata et al. 1980, Jukes and Kimura 1984). Sites with stronger functional constraints have lower rate of evolution, and usually perform functions more important to the organism's fitness. Because mutations consist of point mutations and indels (insertions and deletions), functional constraint can be defined separately with respect to each type of mutation. A naïve expectation would be that the functional constraints against nucleotide substitutions and against indels will be correlated; because if the function of a genomic site can be disrupted by a substitution, it is likely that it can also be disrupted by an indel. However, this is not always the case. For example, if a nucleotide substitution occurs in a fourfold degenerate site in a protein-coding gene, it will be selectively neutral because the protein product is not affected; but if that fourfold-degenerate site is deleted, it will cause a frameshift which may completely disrupt the function of that protein. Graur et al. (2015) have called such regions, "indifferent DNA," to be evolutionarily between functional and junk DNA; these sites are not under selection as long as they are not deleted.

Although a rich literature exists on the selection pattern on substitutions, and to a lesser extent deletion, the two have only rarely been systematically compared. Substitutions are studied more extensively than deletions, partly because current multiple-sequence alignment methods are better in modeling them. Multiple sequence

alignment algorithms are often unreliable in locating insertions and deletions (Landan and Graur 2009).

Taylor et al. (2004) identified 1,743 indel events in 1,282 genes (out of a dataset of 8,148 genes) from human-mouse-rat triple alignments. They compared indel rates in different gene functions (using Gene Ontology), and found that intracellular proteins and enzymes are less likely to have indels. When they compared the indel rate differences with an earlier study of substitution rates (Waterston et al. 2002), it was found that their distribution among categories were highly similar. Unfortunately, their study used gene categories instead of directly separate the data by gene. Another study (Miller et al. 2007) used a 28-vertebrate alignment to study coding-sequence conservation over these species. In one section, the authors tested the hypothesis that amino acids that are conserved in long-term evolution are more likely to cause diseases when a deletion mutation happens on them. They used the gene PAH (phenylalanine hydroxylase), whose deficiency causes a well-studied genetic disease PKU (phenylketonuria). However, the conservation levels of the codons involved in disease-causing deletions (that are multiples of three nucleotides) are not significantly higher than the gene overall, failing to show a correlation. This study only used a single gene, presumably because disease data on amino acid deletions are difficult to obtain. This makes the sample size rather small, but the usage of disease data is an advantage in that the deleterious effects are confirmed and not inferred. Finally, Chen et al. (2009) studied a different measure, the ratio of nucleotide substitution to indel rates, across mammalian and bacterial genomes. A higher ratio indicates a lower

relative indel rate. First, they found that coding sequences have less indels relative to point mutations; this is expected given frameshifts will destroy a protein's function. Within coding regions, more conservative genes have a higher substitution/indel ratios, i.e., less tolerant to indels. This suggests indels (even those that are times of three nucleotides) are subject to stronger selection than substitutions in conservative genes; however, since the "substitution" they used includes both synonymous and nonsynonymous ones, the ratio does not necessary compare two kinds of selection because of neutral sites mixing in the data.

The null hypothesis to be tested here is that purifying selection against nonsynonymous substitutions and purifying selection against indel events are not correlated. If there is a correlation, one can expect that the dN/dS value of a codon would be proportional to the probability of that being deleted in one (or more) of the studied lineages; otherwise, the two would be independent. dN will also behave similarly because it is also under selection; dS would not because it measures neutral substitutions. It is also possible that the correlation occurs only on a gene level: genes with higher dN/dS would have higher mutation rate, but the correlation is absent within genes. However, selection is not the only evolutionary force that may cause a correlation between substitutions and indels; it is possible that regions with high point mutation rates would also have high indel mutation rates. In this case, I would see that the deletion rate would be correlated to both dN and dS, but the dN/dS ratio would not have such an effect.

In this study, I used mammalian protein-coding sequences and simulated sequences to study the correlation between deletion rates and dN/dS, to understand how similar or different the patterns of the two types of selection are. In addition, I used dN and dS separately to see their correlation with deletion rates, to test my hypothesis on whether or not mutation plays a role. I have found that there is indeed a correlation, and it is more caused by selection than mutation; the correlation is mostly due to the difference of selection patterns between genes.

Materials and Methods

Data collection and analysis of dN, dS, and dN/dS

A list of aligned mammalian protein sequences was taken from Lindblad-Toh et al. (2011). To make sure that only good-quality genome sequences are used, I only used data from 9 mammalian species (Figure 4.1): human (*Homo sapiens*), chimpanzee (*Pan tryglodytes*), macaque (*Macaca mulatta*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), guinea pig (*Cavia porcellus*), dog (*Canis lupus familiaris*), cow (*Bos taurus*), and horse (*Equus caballus*). I retained 8,605 alignments. Coding DNA sequences that correspond to these sequences were retrieved from ENSEMBL 2011 archive (Flicek et al. 2011).

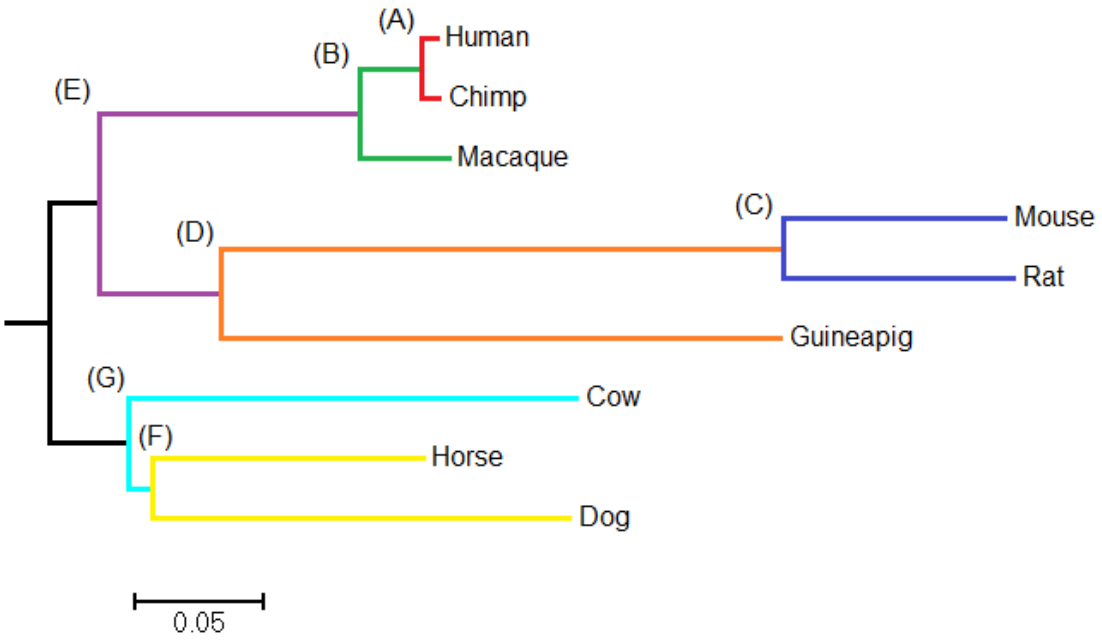

Figure 4.1 The commonly accepted phylogenetic relationship among the 9 species used in this study. This tree will be called the external reference tree throughout the paper. Seven different colors denote seven pairs of branches/lineages (A ~ G) on which deletions were estimated. The black-colored branches are the root of the tree. The branch lengths of the 9-species tree are derived from UCSC Human/hg19/GRCh37 46-way multiple alignment (Kent et al. 2002). These branch lengths are used as guidance for simulation and estimation of deletion rates.

All protein sequences were realigned with CLUSTALW (Thompson et al. 1994), PROBCONS (Do et al. 2005), and T-COFFEE (Notredame et al. 2000), and DNA alignments were threaded through the protein alignments. Maximum likelihood trees were produced with RAxML (Stamatakis 2006) from PROBCONS alignments and 10 genes that produced a total tree length above 5 were removed. Throughout the study, I used the remaining 8,595 genes. The DNA alignments were processed through the program HyPhy using the FUBAR script (Murrell et al. 2013), which estimated the dN and dS of each site using an approximate Bayesian algorithm. Their ratio dN/dS

(Omega) is calculated from the output of FUBAR. Hereafter dN, dS, and dN/dS are collectively called "substitution measures."

Deletion Identification and statistical analysis

I searched seven pairs of branches for deletions of one to eight amino acids (Figure 4.1). These pairs of branches are: (A) human and chimpanzee lineages (red branches, macaque as outgroup); (B) ape and macaque lineages (green branches, cow as outgroup); (C) rat and mouse lineages (indigo branches, guinea pig as outgroup); (D) murid and guinea pig lineages (orange branches, human as outgroup); (E) primates and rodents lineages (purple branches, cow as outgroup); (F) dog and horse lineages (yellow branches, cow as outgroup); (G) (dog+horse) and cow lineages (cyan branches, human as outgroup). The outgroup was used to determine if an indel is an insertion or a deletion. In branch pair (B), the closest outgroup is a rodent, but cow was chosen because rodents have long branch lengths. For a lineage containing multiple species (e.g., apes), only the branch before the divergence (e.g., divergence between human and chimpanzee) was analyzed. This was done by combining multiple sequences into an "ancestral" sequence. If the site is not a gap in at least one of these sequences, it was treated as a non-gap in the whole ingroup. In this way, every branch in the nine-species tree, excluding the root branch, was searched for deletions without repetition. The root branch (the branch separating Euarchontoglires and Laurasiatheria) was not searched for deletions because the directions of its indels could not be determined.

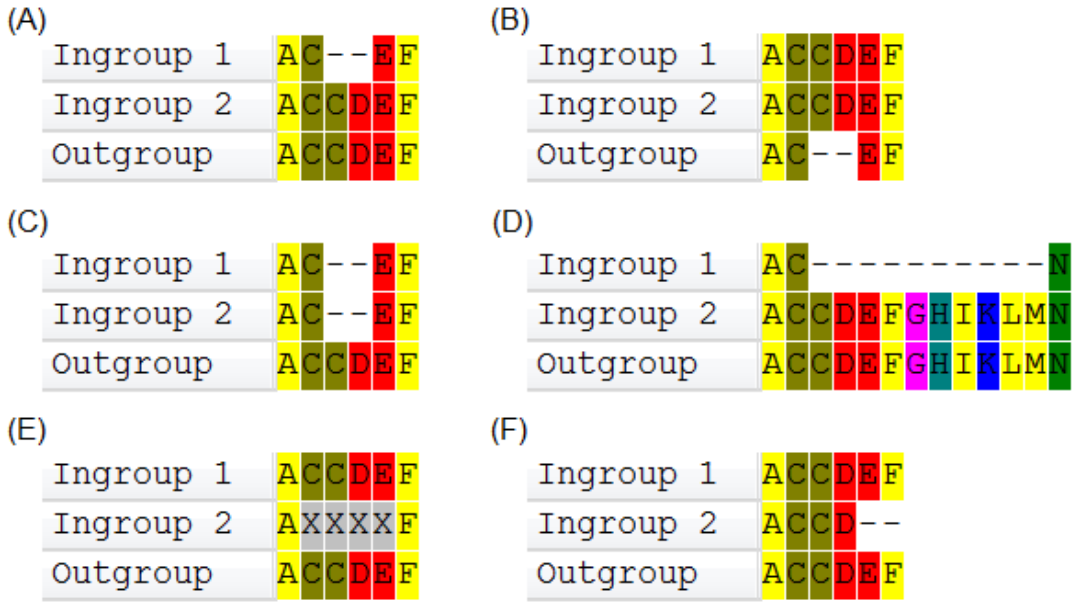

Figure 4.2 Illustration of how I identified deletion events and non-used sites in protein sequences for each pair of lineages. (A) Identified short deletion in ingroup 1. (B) Non-used sites because of gaps in the outgroup. (C) Non-used sites because both ingroups are gaps, thus direction and time of indel unknown. (D) Non-used sites because of long (> 8aa) deletion. (E) Non-used sites because of unknown amino acids. (F) Non-used sites because of terminal gaps. Figure made with MEGA version 6 (Tamura et al. 2013).

Each amino acid site in each pair of branches was also determined to be a "used site" or not; a used site should be a non-gap in the outgroup (Figure 4.2B), and cannot be gaps in both ingroup lineages (Figure 4.2C), part of a long deletion (>8 amino acids, Figure 4.2D), unidentified amino acid (Figure 4.2E) or part of terminal gaps (Figure 4.2F). A detected deletion (Figure 4.2A) always happens in used sites.

Then the weighted "deletion rate" of each amino acid site in all alignments are calculated. For each branch pair, its weight is the sum of the corresponding branch lengths in the reference tree (Figure 4.1). The branch lengths are derived from human/hg19/GRCh37 46 species multiple alignment, using the placental tree without

chromosome X (http://genomewiki.ucsc.edu/index.php/Human/hg19/GRCh37_46-way_multiple_alignment, Kent et al. 2002). The "deletion rate" of that site is number of deletion(s) that occurs in this site, divided by the sum of weights of branch pairs in which the site is a "used site." Similarly, the "deletion rate" of the whole gene is the number of deleted sites in the whole gene divided by the total sum of weights of corresponding branch pairs from each used site.

For each amino acid site in each alignment, four values were obtained: a weighted "deletion rate" and three substitution measures (dN, dS, and dN/dS). Spearman correlation coefficients (Spearman coefficients hereafter) were calculated between "deletion rate" and all three substitution measures, for each alignment method. These datasets use all sites, so I named them "All." To reduce the effects of spuriously high or low dN/dS due to gappy sites, the correlation coefficients were recalculated for sites that are non-gap in at least four sequences, and for sites that are non-gap in at least six sequences. These datasets were named "4+" and "6+." Since there are a large number of sites that does not experience any nucleotide substitutions (their dN/dS is technically incalculable due to division by 0, only approximated using extrapolation from other sites), the correlation coefficients were also recalculated with all datasets after these constant sites were removed. These datasets were named "NC-All," "NC-4+," and "NC-6+." Therefore, with "All," "4+," "6+," "NC-All," "NC-4+," and "NC-6+," from which 6 Spearman coefficients were calculated for each alignment method.

Test simulation for determining the parameters for the main simulation

I simulated sequence evolution using INDELible (Fletcher and Yang 2009). To simulate coding sequences, INDELible evolves nucleotide sequences along the input tree based on a nucleotide substitution models. These substitutions are subject to selection as determined by dN/dS, randomly drawn from an input distribution for each site. Insertions and deletions, always multiples of three nucleotides, are independently modeled and have a uniform rate among sites; however, the total amount of indels is proportional to the branch length. To input the parameters, including tree length (divergence), gene-wise indel rate, gene size, and dN/dS based on those of real data, I need to know what input parameters would correspond to the output values of each gene. I started with a "test" simulation to establish the relationship between input parameters and values estimated from sequences. Both in the test simulation and the main simulation, the input tree and the lengths of its branches are scaled from the "reference tree" from UCSC genome browser (Figure 4.1, Kent et al. 2002).

My simulation included three parts of grid search. In part one, only the level of divergence was varied. The trees along which simulated sequences were evolved are produced by multiplying every branch of the "reference tree" by factors. These 1,000 factors range from 0.01 to 10 with increments of 0.01, and are hereafter called S_T (tree scaling factor). For each of the 1,000 trees, 5 replicates are produced. The indel rate (relative to substitution rate) is derived from Gerstein et al. (2003), using the rate of insertions and deletions that are a multiple of three. The relative insertion rate is

0.00354, and the relative deletion rate is 0.01357. Part two is a two-dimensional grid search that varies both the level of divergence and relative rate of indels. For the level of divergence, S_T takes ten values from 1 to 10. The input relative indel rates are produced by multiplying 0.00354 and 0.01357 with factors from 0.1 to 8 with increments of 0.1. These factors are called S_{RI} (relative indel rate scaling factor). I call their product ($S_T \times S_{RI}$) = S_{AI} (absolute indel rate scaling factor). For each of the 10×80 = 800 scenarios, 5 replicates are produced. A gamma distribution of site-wise dN/dS with a shape parameter of $\alpha = 0.5$ and mean = 0.17 (derived from the gene-wise mean of the whole real dataset) was used for both parts. To satisfy the INDELible input requirement, the distribution was binned into 50 bins between 0 and 1, 20 bins between 1 and 2 and 1 bin above 2. In each bin, the dN/dS value used was the median.

Part three dealt with the change of gene-wise dN/dS from input to output. In this part, S_T was set at 1.5 and S_{RI} was set at 0.6; they are close to the mean of real data and preliminary tests showed that dN/dS output is not sensitive to S_T or S_{RI} as long as S_T is not very small (e.g., below 0.5). The input mean dN/dS varies from 0.01 to 1 with increments of 0.01 (5 replicates for each dN/dS value), and the distribution is a gamma distribution with a shape parameter of $\alpha = 0.5$, and binned with the same method as the two previous tests (bins with a probability below 10^{-6} were discarded).

In all three test simulations, the distribution of indel size is a power-law distribution (Cartwright 2009) whose parameter is 1.8 and truncated at 40, and each replicate started with a coding sequence of 500 codons.

All simulated alignments from this test round were realigned with PROBCONS. Part one alignments were then used to produce maximum likelihood trees using RAxML (Stamatakis 2006) with the user tree option, and total tree lengths were calculated. The relationship between ML total tree length and S_T was obtained by a linear regression through origin. Part two alignments were processed through an in-house Perl script to determine the gene-wise deletion rate; the relationship between this deletion rate and S_{AI} is obtained by a linear regression through origin. The ratio between the alignment length and ancestral gene length was also mapped to S_{AI} with a linear regression through origin, because insertions can increase alignment lengths. Because a very large S_{AI} would cause the alignment to be very difficult and the relationships would be non-linear, as well as not occurring in real data, only test-simulated data with $S_{AI} <= 20$ were used. For part three, FUBAR was used to estimate gene-wise dN/dS, by dividing the mean dN by mean dS, using only "4+" sites. I did not use "NC-4+" because excluding substitution-free sites may cause a bias. A linear regression (not through origin) is used to estimate the relationship between input and output dN/dS. The procedures for the test round and regressions are illustrated in Figure 4.3A.

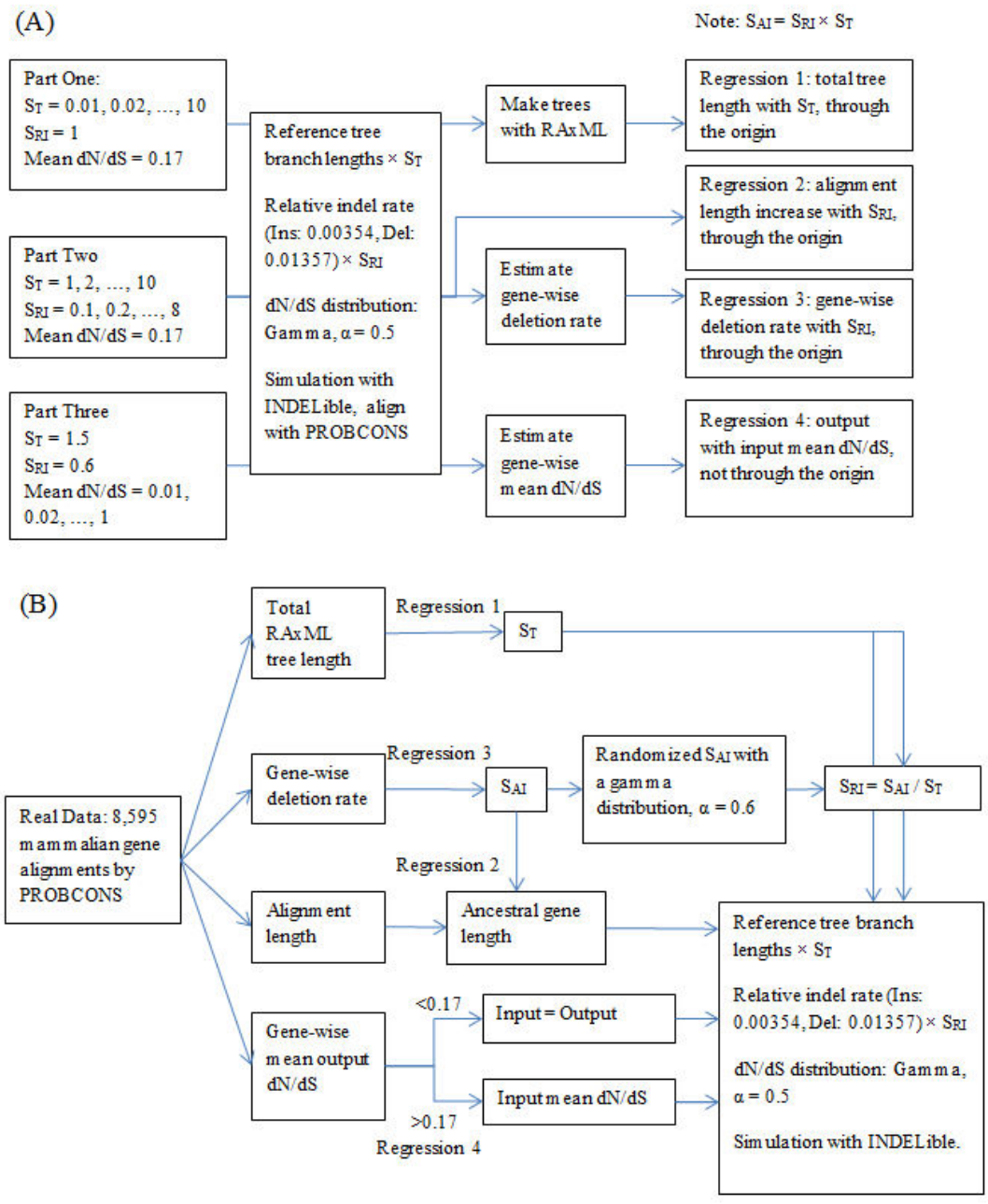

Figure 4.3 A flowchart describing the procedure I did to obtain the simulation parameters. (A) Three parts of the test round were used to determine how input tree length, indel rate, and dN/dS determine output tree length, deletion rate, alignment size, and dN/dS. (B) The formulas of input-output relations were then used to convert estimates from real data to input parameters of main simulation round.

After obtaining the mathematical relationship between simulation input and output, I applied them on real data to get the parameter used for the main round of simulation. For each of the 8,595 PROBCONS real gene alignments, the following information is collected: (1) the alignment length; (2) the total RAxML tree length; (3) gene-wise deletion rate; (4) gene-wise FUBAR-estimated dN/dS. The mean dN/dS are estimated by dividing the mean dN by mean dS, using only "4+" sites. Using (2) and (3), the S_T and S_{AI} corresponding to each gene is calculated. Using S_{AI} and (1), the "ancestral gene length" is calculated. Using (4) I calculated the input mean dN/dS; however if the regression formula (see Results) was used for the whole dataset, negative input mean dN/dS would appear in several genes, and there would be a strong bias in low-dN/dS genes. As a compromise, I only used the regression formula to calculate input dN/dS if the real data output is higher than 0.1704167; otherwise the input was the same number as output. The resulting polyline is continuous.

The formulas I used are as follows:

Total ML tree length = $S_T \times 0.4389$;

Gene-wise deletion rate = $S_{AI} \times 0.02108$;

(Alignment length / ancestral gene length) $- 1 = S_{AI} \times 0.01135$.

Output dN/dS = (Input mean dN/dS $\times$ 0.71511) + 0.04855 (if dN/dS > 0.1704167)

Output dN/dS = Input mean dN/dS (if dN/dS < 0.1704167)

Coding sequence simulation and analysis

For the main simulation round, 8,595 genes × 5 replicates were simulated with INDELible. For each gene, the ancestral gene length and level of divergence (achieved by multiplying the reference tree with S_T) were based on the values derived from the corresponding real gene. The distribution of dN/dS was a gamma distribution with a shape parameter of $\alpha = 0.5$ and mean = dN/dS calculated as per last paragraph, binned using the same methods as the test round (bins with a probability below 10^{-6} are discarded). Because I was studying the correlation between substitution rate (proportional to S_T) and absolute deletion rate (proportional to S_{AI}), I could not directly use the S_{AI} value estimated from the real gene. Instead, for each replicate, I drew S_{AI} from a gamma distribution with a shape parameter of $\alpha = 0.6$ and mean = 0.79 (the mean S_{AI} from the real data). This means the five replicates will have five different S_{AI}, while sharing other parameters. S_{RI} values (required calculate the parameters directly input) were then calculated by dividing S_{AI} with S_T. The whole process of deriving parameter for the simulation is also described via a flowchart in Figure 4.3B.

The simulated protein sequences were aligned with CLUSTALW, T-COFFEE, and PROBCONS, and then nucleotide alignments were threaded through the protein alignments. Together with the true alignment (as control for alignment error), their site-wise substitution measures were estimated with the FUBAR script by HyPhy.

Deletions were identified, and site-wise deletion rates were estimated with the exact same method as with real data.

To provide ranges of values with which the ones from real data can be compared with, bootstrap resampling was used. 1000 subsets of the simulated data were produced. In each subset, one random replicate was chosen from the five for each of the 8,595 "genes." Spearman coefficients were calculated for each subset. Each subset were used as-is, having sites with less than 4 or 6 non-gap characters removed, having sites without nucleotide substitution (constants) removed, or the combination of the them, similar to the procedures for real data ("All," "4+," "6+," "NC-All," "NC-4+," and "NC-6+"). For all the 6 Spearman coefficients, the mean, standard deviation, 2.5% and 97.5% quantiles were calculated. The corresponding real data value was compared to the range of simulated data values both with a Z-score and using quantiles. For the Z-score, the differences between the means of bootstrap values from simulated data and the values from real data were divided by the standard deviations of bootstrap values from simulated data; the resulting statistics were compared to a standard normal distribution and p-values were obtained.

Distribution of dN/dS in deleted sites

All real mammalian protein sites that have undergone at least one deletion in any lineage were extracted from the data set and their distributions of estimated dN/dS are computed. The distributions were compared with those from deletion-free sites to see

if any difference exists with chi-square tests. Effect sizes (Cohen's D, Cohen 1988) were calculated between dN/dS distributions in deletion and non-deletion codons. These analyses were done on "All" and "NC-4+" datasets only, because I thought the others would produce similar results and are thus unnecessary. These procedures were repeated for the simulated data. Similar to the previous section, one thousand bootstrap subsets were used, and the mean, standard deviation, 2.5%, and 97.5% quantiles were calculated. The corresponding real data values were compared to the range of simulated data values both with a Z-score (same formula as in the previous section) and using quantiles.

Analysis of gene-wise and within-gene correlations

For both real and simulated data, I calculated gene-wise dN, dS, dN/dS, and deletion rate. Gene-wise dN and dS are the mean of corresponding values of "4+" sites over the whole gene. I did not use "NC-4+" because excluding substitution-free sites may cause a bias. Gene-wise dN/dS is gene-wise dN divided by gene-wise dS. Gene-wise deletion rate is the total number of deletion divided by the total sum of weights of corresponding branch pairs from each used site.

I calculated the Spearman correlation between gene-level deletion rate and substitution measures in both real and simulated data, for each alignment method. Similar to previous sections, in the simulated data bootstrapping is used. Each subsample includes only one replicate for every simulated gene. The mean, standard deviation,

2.5%, and 97.5% quantiles were calculated. The corresponding real data values were compared to the range of simulated data values both with a Z-score and using quantiles.

I calculated within-gene Spearman correlation between deletion rates and substitution measures, using 466 real genes and 466 × 5 = 2,330 simulated genes that have the derived "ancestral gene length" longer than 1,500 amino acids. The correlation coefficients are calculated for both "all" and "NC-4+" datasets. Genes that do not have any deletions identified were removed from the data, while the rest were used to calculate the mean and standard deviation.

Table 4.1 gives an overview of my data.

Table 4.1 A summary of my data, both real and simulated.

	Real Data (PROBCONS)	Simulated Data (PROBCONS)	Simulated Data (TRUE)
Number of genes	8,595	42,975	42,975
Total alignment length (aa)	5675396	28458331	28465275
Proportion of substitution-free sites	0.3106	0.2266	0.2275
Number of deletions	50698	338330	340895
Mean deletion size (aa)	1.9591	1.9218	1.9274
Mean site-wise dN (sd)	0.3326 (0.9098)	0.2804 (0.5671)	0.2793 (0.5625)
Mean site-wise dS (sd)	1.9526 (2.6370)	1.5625 (1.8169)	1.5586 (1.8032)
Mean site-wise dN/dS (sd)	0.2687 (0.4891)	0.2705 (0.5630)	0.2705 (0.5654)
Mean gene-wise dN (sd)	0.3595 (0.2087)	0.2983 (0.1432)	0.2974 (0.1433)
Mean gene-wise dS (sd)	2.1206 (0.4309)	1.6833 (0.2955)	1.6789 (0.2936)
Mean gene-wise dN/dS (sd)	0.1702 (0.0940)	0.1790 (0.0884)	0.1789 (0.0888)
Mean gene-wise deletion rate (sd)	0.0166 (0.0222)	0.0186 (0.0258)	0.0189 (0.0266)

Results

Correlation between deletion rates and substitution measures

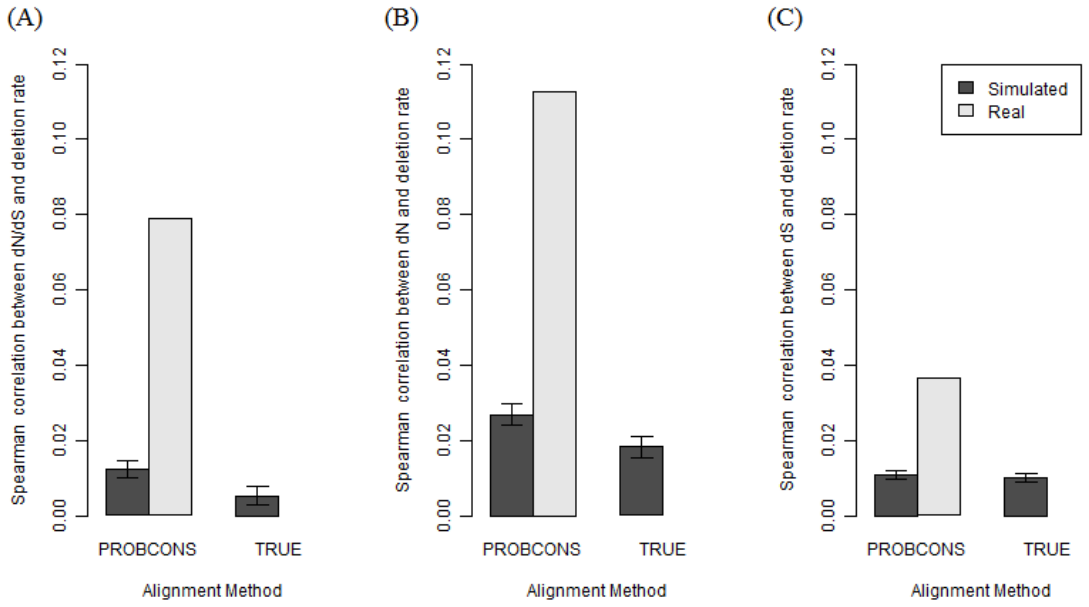

Figure 4.4 Site-wise Spearman correlation between deletion rate and substitution measures (A. dN/dS, B. dN, C. dS) in real and simulated data, complete dataset. For the simulated data, the shown value is the mean of 1,000 bootstrap re-samplings, and the error bar is 2.5% to 97.5% quantiles. I observe that in all three measures, real data produces a higher correlation than simulated data.

In Figure 4.4, the correlations between deletion rate and all three substitution measures (dN, dS, and dN/dS) are presented. In the dataset "All," a correlation of about $\rho \approx 0.08$ exists between dN/dS and deletion rate (Figure 4.4A), and $\rho \approx 0.11$ between dN and deletion rate (Figure 4.4B), estimated from the pooled sample of 8,595 real mammalian alignments. The differences between different alignment methods are minimal, therefore I only showed PROBCONS in the three algorithms (similar below). On the other hand, the corresponding correlation from simulated data is only $\rho \approx 0.01$

for dN/dS and $\rho \approx 0.03$ for dN. The real data are highly significantly different from these ($P < 0.0001$ in Z-test for all cases; all P values are from Z-tests in this section). For true simulated alignments, the correlation is well below the level of inferred alignments for dN and dN/dS. With dS (Figure 4.4C), the difference are much smaller, being $\rho \approx 0.04$ in real data and $\rho \approx 0.01$ for simulated data, but still significantly different from zero ($P < 0.0001$).

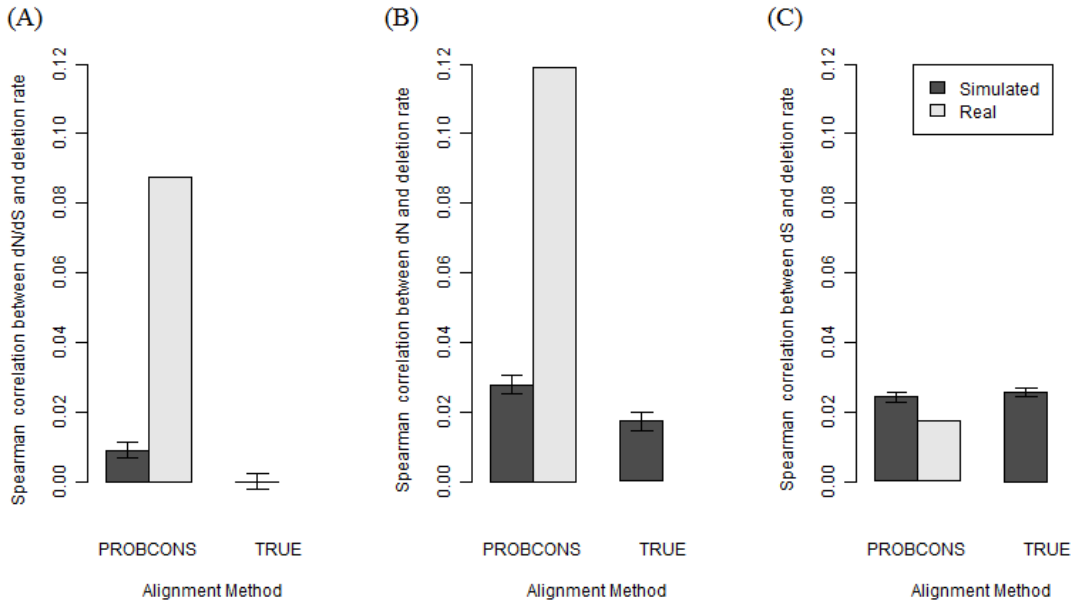

Figure 4.5 Site-wise Spearman correlation between deletion rate and substitution measures (A. dN/dS, B. dN, C. dS) in real and simulated data. Here I use the "NC-4+" dataset, from which substitution-free sites and sites that are gaps for more than five species are removed. For the simulated data, the shown value is the mean of 1,000 bootstrap re-samplings, and the error bar is 2.5% to 97.5% quantiles.

In the dataset "NC-4+," only sites that have at least one nucleotide substitution and are non-gaps in at least 4 species are used for the calculation. The patterns for dN and dN/dS have not changed (Figure 4.5A, 4.5B); the only visible difference is that the true alignment from simulated data has an extremely small correlation between dN/dS

and deletion rate ($\rho < 0.0001$), which is not significantly different from zero ($P = 0.4982$). On the other hand, the situation with dS has reversed (Figure 4.5C); here simulated data have a slightly higher correlation, $\rho \approx 0.025$ than real data, $\rho \approx 0.02$ ($P < 0.0001$). The other four datasets, "4+," "6+," "NC," and "NC-6+" presents similar results; it also appears that which one have a higher dS-deletion correlation depends on if substitution-free sites are removed. In datasets without substitution-free sites, the correlation between dN/dS and deletion rate in true simulated alignment is two magnitudes smaller than in datasets with substitution-free sites. It is likely that datasets without substitution-free sites are more reliable, because dN/dS of those sites are only estimated by extrapolation from other sites.

dN/dS distribution in codons that underwent deletion

Figure 4.6 shows the Cohen's D, a measure of differences in means of substitution measures between sites with and without deletion. It is another way to examine the relationship between deletion and substitution measures. Here the number of deletions does not matter; only presence or absence does. Biologically, if a site has a deletion, then it would be more "expendable;" another deletion is more likely to occur than by random. The observation from Cohen's D data is very similar to Spearman data: in both dN and dN/dS, real data gives a much higher effect than simulated data, with dN having slightly stronger signal; while in dS, the effect is much weaker. However, Cohen's D between deleted and non-deleted sites' dS are significantly different

between real and simulated data; simulated data have a higher Cohen's D in both "All" ($P < 0.0001$) and "NC-4+" ($P < 0.0001$) datasets.

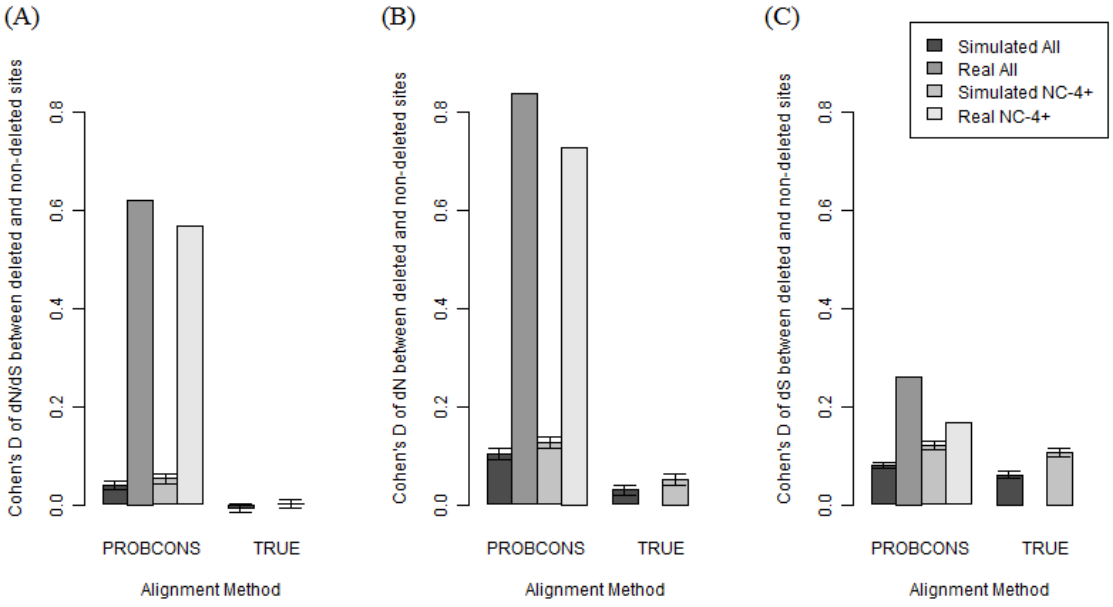

Figure 4.6 Effect size (Cohen's D) indicating the difference of (A) dN/dS, (B) dN, or (C) dS means between deleted and non-deleted sites. In each graph, both simulated and real data, and both "All" and "NC-4+" datasets are used. For the simulated data, the shown value is the mean of 1,000 bootstrap re-samplings, and the error bar is 2.5% to 97.5% quantiles.

To visualize the different distributions of dN/dS between deleted and non-deleted sites, I made bar graphs for real and simulated "NC-4+" data aligned with PROBCONS (Figure 4.7). In real data (Figure 4.7A), it is easy to see that non-deleted sites are twice likely to have dN/dS under 0.1 than deleted sites; while in all bins above 0.2, the fractions of deleted sites are much higher. On the contrary, in simulated data (Figure 4.7B), the fractions of deleted and non-deleted sites in each bin of dN/dS are very close. There are visible differences in the lowest (<0.1) and highest (>2) bins, but they are much less pronounced than in real data. Thus, I have shown that deleted sites

generally have a higher dN/dS (as well as dN) than non-deleted sites, and this effect is much higher in real data than in simulated data.

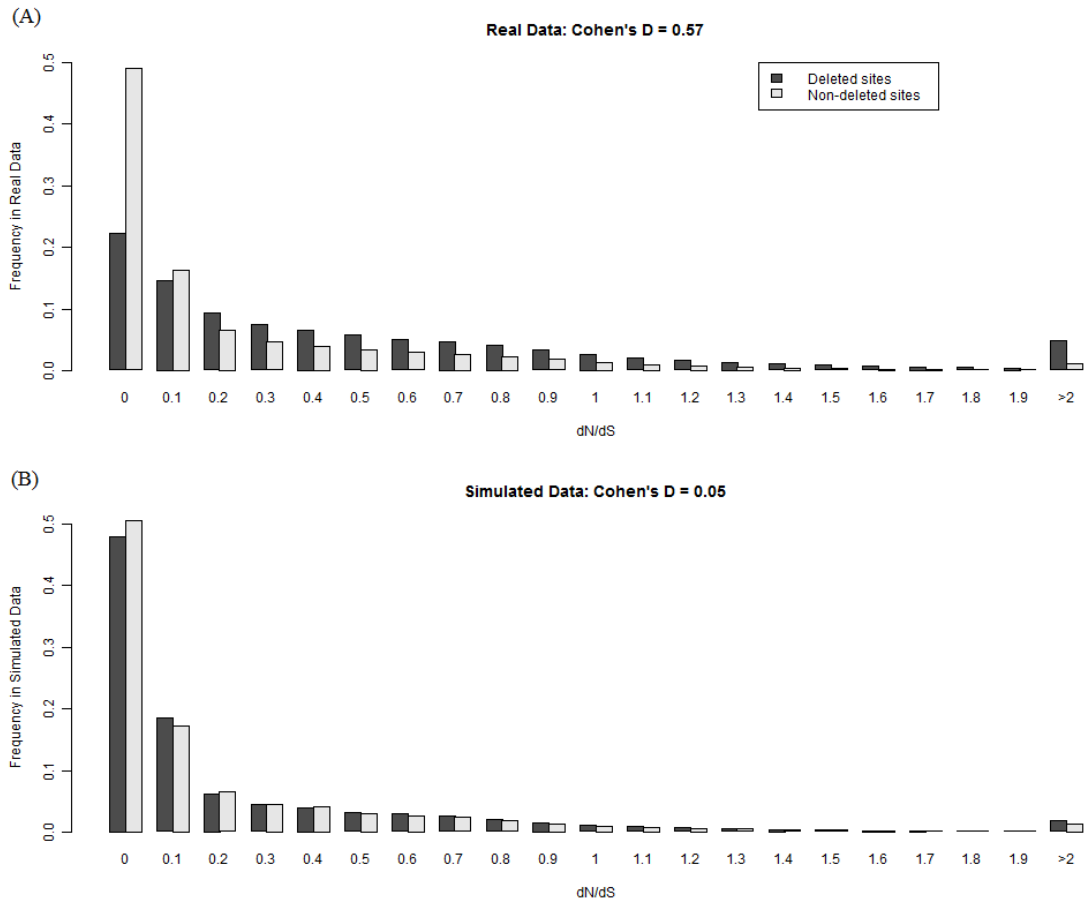

Figure 4.7 Histograms showing dN/dS distribution comparisons between sites with and without deletion, in both (A) real, (B) simulated data aligned with PROBCONS. It can be observed that the distributions are much more different in real data than in simulated data: the non-deleted sites have a heavier left tail, while the deleted sites have a heavier right tail.

Gene-wise correlation between deletion rates and substitution measures

Previously, I showed the correlation of deletion rate and substitution measures on a site level. However, there can be a large stochastic effect on the deletion or substitution of codons, which decreases the signal-to-noise ratio. To reduce the effects of noise, I looked at the same correlation at a gene level; using the same statistical method, with gene-averaged deletion rates and substitution measures instead of site-by-site. This may reduce the noise caused by stochasticity and put emphasis on differences among genes.

I estimated the dN and dS of each gene by taking a mean of all its sites. For deletion rate, I used a weighted mean that takes different total corresponding branch lengths among used sites into consideration. The dN/dS of the gene is calculated by dividing the mean dN by the mean dS ("ratio of mean" approach). When calculating the substitution measures over the whole gene, I only take "4+" sites into account.

Figure 4.8 shows the substitution measures plotted against deletion rates in real and simulated data, aligned by PROBCONS. The real-data plots have a top-right leaning, with the most high-deletion incidents appearing in the right side of the bulk of dots. On the contrary, in the simulated-data plots, the shape outline of the dots skews to the left – similar to the skew in dN/dS distribution among genes itself.

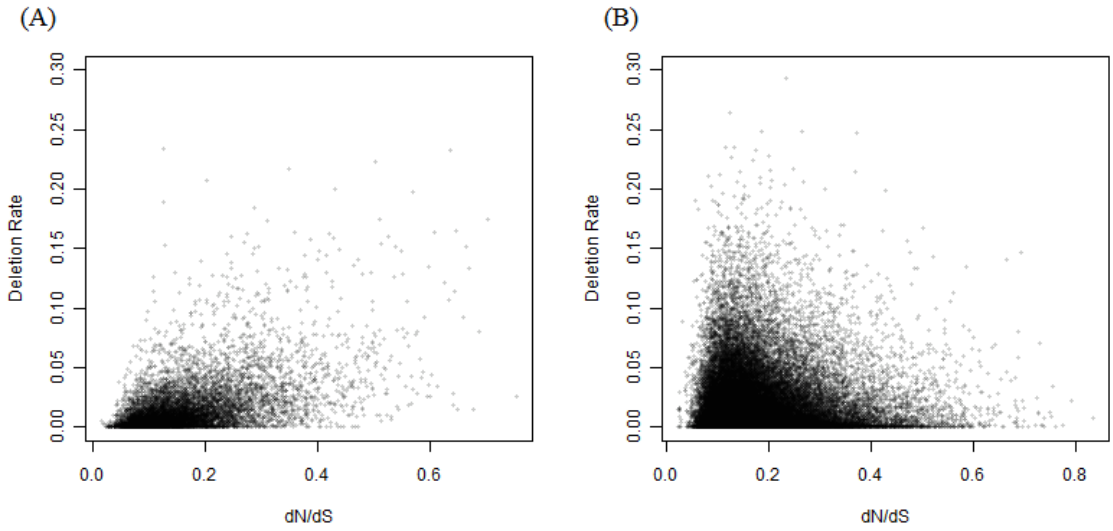

Figure 4.8 Gene-wise deletion rates plotted against dN/dS, in both (A) real and (B) simulated data.

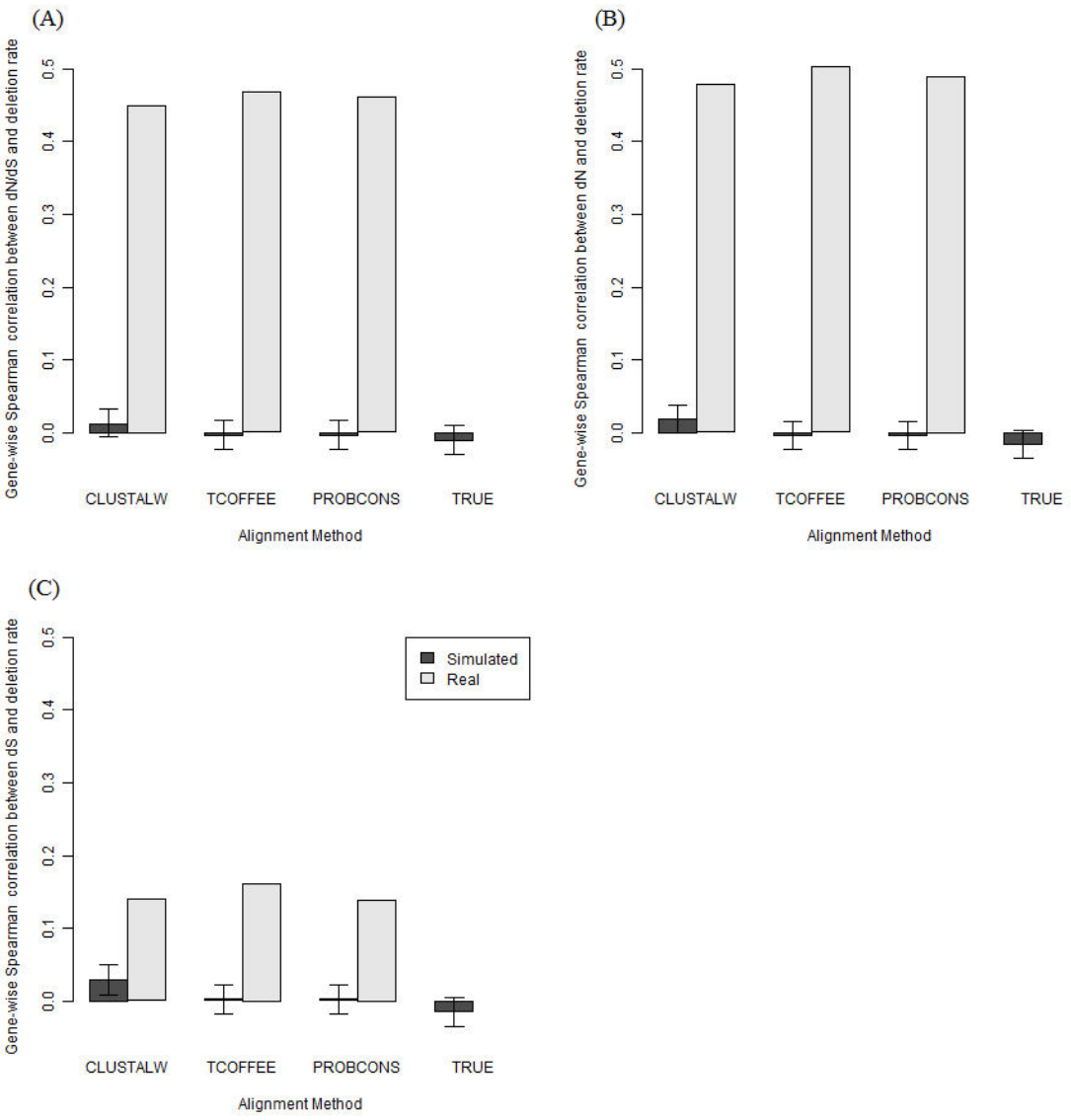

Figure 4.9 Gene-wise Spearman correlations between deletion rate and substitution measures (A. dN/dS, B. dN, C. dS). In both dN/dS and dN, the correlation in real data is very high (≈0.45) compared to simulated data (<0.05); the difference is much less pronounced in dS. For the simulated data, the shown value is the mean of 1,000 bootstrap re-samplings, and the error bar is 2.5% to 97.5% quantiles.

Figure 4.9 further supports it by showing the Spearman correlation coefficients using the same data. In all cases, the correlations in real data are significantly higher than in simulated data ($P < 0.0001$). However, the amount of correlation in real data clearly

depends on what substitution measure is used. For both dN and dN/dS, ρ is as high as 0.4 to 0.5; but for dS, ρ is close to 0.15. At the same time, all simulated data with the exception of dN and dS in CLUSTALW alignments show correlations that are not significantly different from zero; even dN in CLUSTALW is only marginally significant. For dN, dS as well as dN/dS, true simulated alignments give a weakly negative correlation.

Within-gene correlation between deletion rates and substitution measures

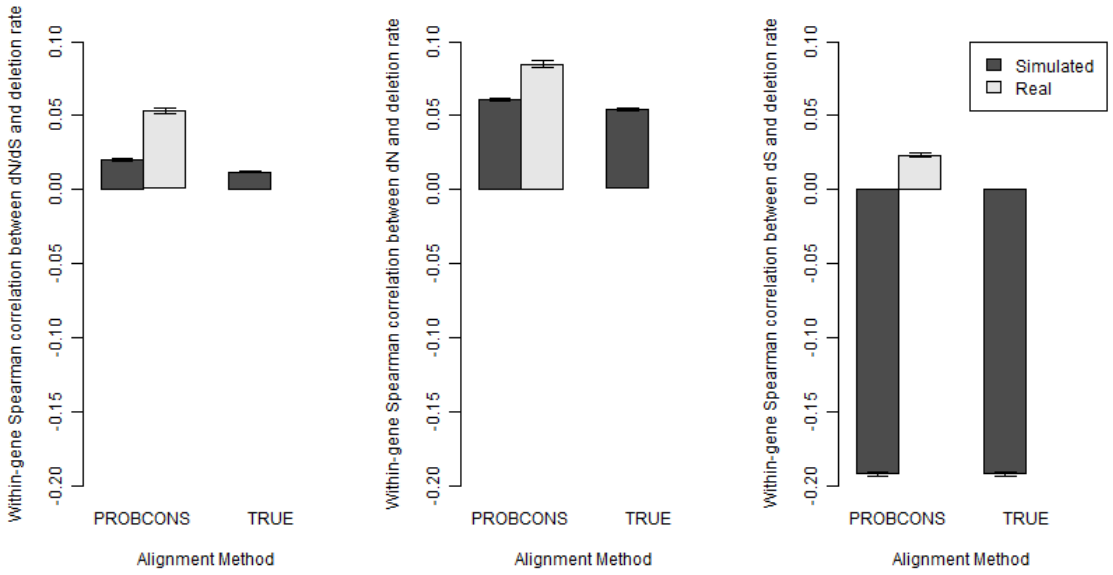

Figure 4.10 Within-gene site-wise Spearman correlation between deletion rate and substitution measures (A. dN/dS, B. dN, C. dS) in real and simulated data. Here I use the "NC-4+" dataset, from which substitution-free sites and sites that are gaps for more than five species are removed. The column shown is the mean over all genes that have at least one deletion and estimated ancestral gene size over 1,500 aa, and the error bar is the standard error.

Figure 4.10 shows the within-gene correlation for both real and simulated data. I only took genes that have an estimated ancestral length over 1,500 aa, because with smaller genes, stochastic effects will be too strong. In dN/dS, the real data gives a slightly higher correlation compared to the simulated data ($\rho \approx 0.05$ compared to $\rho \approx 0.02$), though not to the level of entire genome site-wise correlation. dN is similar; but in dS the results are surprising. In simulated data, there is a very strong negative correlation between dS and deletion rates within genes ($\rho \approx -0.2$), which is not observed in real data. It is most likely that this is due to some feature in the simulation algorithm.

Discussion

Implications on protein sequence evolution

This study shows that there is indeed a positive correlation between the probability a codon being deleted and its dN/dS value, indicating similarity in patterns of purifying selection against deletion and amino acid replacement. This can be interpreted as that both replacement and deletion can damage the function of an amino acid residue in the protein; thus reducing the fitness of individuals bearing such mutation. However, this site-wise correlation is very weak, on the order of $\rho \approx 0.1$; therefore it would be rather difficult to predict one kind of selection from the other. In other words, selection against deletion is not completely consistent from selection against replacement.

One reason of the low correlation may be the existence of "indifferent DNA" (Graur et al. 2013, 2015). Indifferent DNA refers to sequences that are subject to strong purifying selection against deletions but not substitutions, due to its functionality relying more on the length than the exact sequences. For example, fourfold degenerate (synonymous) nucleotide sites in proteins can be any of the four nucleotides, but deletion of that nucleotide causes a frameshift; a promoter motif needs to have a certain distance from the start codon, thus the sequence between them can freely change as long as the length is kept. In my case, it is possible that some amino acids are required for the protein structure but do not actively bind anything. Some functional sites can also have a range of choices of amino acids, due to similar biochemical properties of these amino acids.

In both the Spearman correlation and distribution comparison, dN also produces a correlation to deletion rates, at a level similar to dN/dS. On the other hand, such correlation is very weak when dS is used, even undistinguishable from simulated data in some cases. If I compare the substitution measures to biological processes, dN/dS is an indicator of selection, dS of mutation, and dN is a combination of both. Thus, it is reasonable to suggest that this correlation between deletion and substitution is largely due to a similar selection scheme instead of correlated mutation rates, though mutational effects cannot be ruled out.

This study on the correlation between substitutions and indels is the first one that involves genomic protein-coding genes, and includes both site-wise and gene-wise

analyses. Using the deletion rate inferred from multiple sequence alignments instead of data on genetic diseases (Miller et al. 2007) made the rate estimation across multiple species rather than human-specific. Alignment-derived deletion rates are also available as long as the genomes of these species are annotated, while disease-derived rates are limited to clinical data and lethal sites are excluded. However, due to alignment errors and partial sequences in some species, alignment-derived deletion rates are less reliable. Nevertheless, I think that I have taken precautions for these disadvantages, respectively by use of simulation and datasets "4+"/"6+."

The potential non-independence between selection against substitutions and deletions can also be relevant in studies involving simulated sequence evolution. In protein simulation, the algorithm writer must decide whether to account for this correlation. For example, INDELible, one of the most comprehensive and frequently used simulation programs, does not allow variation of indel rates along the sequence (Fletcher and Yang 2009). On the other hand, programs like SIMPROT (Pang et al. 2005) implements an algorithm that chooses indel positions relative to their substitution rates. ROSE (Stoye et al. 1998) and indel-Seq-Gen (Strope et al. 2009) limit indels to less conserved regions of sequences.

Difference between site-wise and gene-wise analyses

Site-wise and gene-wise analyses on evolutionary parameters often yield different results (e.g., Wang et al. 2013). Here, I have shown that the Spearman correlation

between dN, dS as well as dN/dS and deletion rate are more than 4 times higher in gene-wise comparisons than in site-wise comparisons. dN/dS values vary in a much larger range in site-wise than gene-wise analyses (Lindblad-Toh et al. 2011). It is possible for sites (individual amino acids) to undergo positive selection, which yields a dN/dS above 1, while this is rare for a whole gene because a protein's basic structure need to be kept consistent for it to function. Therefore, a site-wise study can provide a higher resolution on the selection schemes on the coding part of genomes. On the other hand, site-wise studies suffer from a low sample size for each data point, and thus larger random error.

Although the most likely reason for a much higher gene-wise correlation is that the stochasticity of both substitution and indel rates on the site level causes a low signal-to-noise ratio, it is also possible that between-gene differences contribute more to the overall correlation than within-gene differences. Since traditional methods for distinguishing within-category and between-category effects such as ANCOVA are not applicable due to the non-normal distribution of deletion rates, I looked at the within-gene correlation between deletion rates and substitution measures in individual genes. Since the within-gene Spearman's ρ is smaller than that from the total dataset, it is safe to say that the within-gene effects are smaller than the between-gene effects. On a population genetics level, it is possible that 1) selection is similar within one gene because the deleterious effects of mutations are similar, and 2) mutation rates within one gene (a short region in the genome) do not have much variation.

Artifactual correlation caused by alignment errors

Aside from the biological insights into protein sequence evolution, this study also provides information about consequences of alignment errors. There is no pre-determined correlation between indels and dN/dS in the simulated sequences; thus all estimated correlation is due to artifacts. The correlation between dN/dS and deletion in true alignments of simulated sequences is undistinguishable from zero, which confirmed this point. The same correlations estimated from inferred alignment, on the other hand, are consistently higher than zero. Since the only difference between them is the presence of alignment error, I can conclude that the small correlation observed in simulated reconstructed alignments is caused by alignment errors.

On the other hand, there is a correlation between dN and deletion as well as dS and deletion in simulated sequences that cannot be explained by alignment errors. This phenomenon appears in both true and inferred alignments, and in both site-wise and gene-wise analyses. Since dN, dS, and deletion rate are all indicators of total evolutionary change along the whole tree, the most likely explanation is different rates of evolution (tree length) among different genes.

Multiple sequence alignment is a mathematically difficult (NP-complete) problem; an optimal solution, though theoretically exists, is impossible to implement due to the time needed is immense. All current multiple sequence alignment algorithms use heuristic methods. These algorithms typically produce alignments that are shorter than the true alignment due to preferring mismatches over gaps, and gives mathematically

optimal placements while the real process is sub- or co-optimal (Landan and Graur 2008, Landan and Graur 2009). Regions that are rich in insertions and deletions are difficult to align due to co-optimal placement of gaps; thus putting gaps and mismatches together more often than it should be.

Different alignment algorithms produce a similar level of correlation between dN/dS and deletion rate, despite PROBCONS and T-COFFEE being substantially more accurate than CLUSTALW (e.g., Thompson et al. 2011). Therefore, although the causal relationship between alignment error and Spearman coefficient is uncontested (because true alignments have minimal Spearman coefficients), the amount of alignment error is not a good indicator of Spearman coefficient. I suggest that a large part of the correlation is caused by the most difficult parts of the alignment such as co-optimal and sub-optimal sites (Landan and Graur 2009); even the best alignment algorithm is not able to accurately resolve them. In other words, the artifact is caused by the shared errors of different alignment algorithms.

Phase-1 and Phase-2 deletions

A phase-1 or phase-2 codon deletion (deletions that only partially encompass the first and the last codon involved) can cause an amino acid mismatch without nucleotide substitutions. They are also called non-conservative deletions because they do not conserve the undeleted amino acids (de la Chaux et al. 2007). However, past studies demonstrated that such events are rare. In a study on pairwise indel event between

mouse and rat, 12% of indels found are non-conservative, in contrast with a simulation expectation of 29% (Taylor et al. 2004); another study (de la Chaux et al. 2007) gave an even lower estimate that 4% of all deletions are non-conservative from 3-primate alignments.

Unfortunately, with the simulation and alignment methods I used, I could not account for the effects for such deletions, nor could I mimic them by simulation. Nevertheless, the mismatch caused by non-conservative deletions usually does not happen in the same site as the gap. For example, if ACGCAT (Thr-His) became A---AT (Asn), the Asn residue will be aligned into one of the sites, while the gap occupies the other. The elevated dN/dS would thus only occur in the non-gap site. It is possible that the presence of such a mismatch complicates the alignment process and attracts other alignment errors, but I am not able to quantify this effect.

Caveats and future directions

In my study, the simulation part was used as a negative control. In other words, it was used as a baseline when indel rates and dN/dS are independent from each other. I suggest that in future studies, a positive control can be implemented. If a simulation includes a correlation between indel and substitution models (or even perfectly linearly correlated rates), I could see how the results would compare to the real data. After all, even if the input indel and replacement rates are perfectly linear to each other, the site-wise correlation would still not be one because of stochastic effects.

In the phylogenetic tree used in this study, I put the horse (Perissodactyla) and the dog (Carnivora) together as sister groups, while the cow (Cetartiodactyla) is a sister group for the horse+dog clade. This hypothesis, named Pegasoferae, is supported by a phylogenetic study using molecular data (Nishihara et al. 2006). However, the evolutionary relationship among horse, dog, and cow is still under debate. A rival hypothesis groups the horse and the cow together (Perissodactyla + Cetartiodactyla = Euungulata), to the exclusion of the dog (Prasad et al. 2008). I reasoned that in both hypotheses, the branch separating two of them from the third is very short, and this controversy would have a minimal effect on the estimation of evolutionary parameters. Therefore, I have arbitrarily chosen the Pegasoferae hypothesis. It may be a good idea to check if the choice of phylogenetic tree will affect the result in the future.

Conclusion

This study has demonstrated that in the evolution of mammalian proteins, the selection regimes on amino acid replacement and on short deletions are weakly correlated to each other. Codons that are less likely to undergo nonsynonymous substitutions are statistically also less likely to be deleted. However, in practice this correlation can be overestimated due to the effects of alignment errors.

Chapter Five: Summary

As a highly automatized procedure for molecular evolutionary studies, multiple sequence alignment is frequently scrutinized for its accuracy but also often ignored and deprioritized in actual evolutionary studies. Many biologists like to focus on biological questions like what kind of selection has occurred in which genes, and when did two species diverge, rather than paying attention to the computational processes and how errors may undermine the credibility of the final results. High-accuracy alignment algorithms like PROBCONS (Do et al. 2005) and T-COFFEE (Notredame et al. 2000) have been published more than a decade ago, but many researchers still default to CLUSTALW (Thompson et al. 1994) without knowing better. It is difficult to estimate how many inaccurate studies litter the literature due to faulty methodology; how many false positives have been reported in studies positive selection (Markova-Raina and Petrov 2011).

In my dissertation, I examined the effects of alignment errors and the choice of alignment- and alignment-refining algorithms on evolutionary analyses. Particularly, I focused on studies downstream of the phylogenetic reconstruction, such as branch-length estimation (when the topology is fixed by the user tree) and selection-pattern analyses.

In Chapter Two, I established a method to compare the branch lengths of two phylogenetic trees with identical topology. This measure, named Normalized Tree Distance (NTD), is completely unaffected by the scales of both of the compared trees.

Using mammalian protein-coding sequences, I showed that NTD has a log-normal distribution when a large number of genes are used.

In Chapter Three, I studied the effects of choice of alignment- and alignment-refining algorithm on the accuracy of branch-length estimates. The measure I described in the previous chapter was used to quantify the accuracy. I identified the choice of algorithms as well as four variables of the evolutionary scenario to have significant effects on branch-length accuracy; there were clearly predictable differences between branch lengths estimated from "good" and "bad" alignments. I also discovered the alignment-refining algorithm T-COFFEE (whose evaluation model was used as a criterion for filtering) provides the strongest improvement in branch-length accuracy across multiple evolutionary scenarios. However, the optimal alignment algorithm (out of the seven I studied) depends chiefly on the input tree topology.

In Chapter Four, I checked how inferred alignments differ from true (simulated) alignments in analyses of selection patterns. I chose the correlation between purifying selection on deletion and amino acid replacement, an understudied topic in molecular evolution, as an example. Using a combination of real and simulated data, I have shown that while there is a real correlation (that is biologically explainable), alignment error can complicate the results by introducing artifacts into the correlation.

I believe there is still much to do on the topic of alignment accuracy affecting molecular evolutionary studies, especially quantifying the effects of low-quality alignment tools, such as CLUSTALW, on downstream analyses. In particular,

estimations that are sensitive to false positives, such as positive selection, hybridization, horizontal gene transfers, need to be scrutinized. When there is sufficient proof that alignment algorithm choice is vital to their accuracy, researchers would have stronger incentive to pay attention to this step during real-data studies. It may be also worthy to reexamine some milestone macroevolutionary studies in the recent past, to see if the usage of high-accuracy alignment algorithms combined with alignment refining can improve the results.

References

Ahola V, Aittokallio T, Vihinen M, Uusipaikka E. 2006. A statistical score for assessing the quality of multiple sequence alignments. *BMC Bioinform.* 7:484.

Brown JM., Hedtke SM., Lemmon AR., Lemmon EM. 2010. When Trees Grow Too Long: Investigating the Causes of Highly Inaccurate Bayesian Branch-Length Estimates. *Syst. Biol.* 59(2):145–161.

Cantarel BL, Morrison HG, Pearson W. 2006. Exploring the Relationship between Sequence Similarity and Accurate Phylogenetic Trees. *Mol. Biol. Evol.* 23(11):2090–2100.

Cartwright R. 2009. Problems and Solutions for Estimating Indel Rates and Length Distributions. *Mol. Biol. Evol.* 26(2):473–480.

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

Chang J-M, Tommaso PD, Notredame C. 2014. TCS: A New Multiple Sequence Alignment Reliability Measure to Estimate Alignment Accuracy and Improve Phylogenetic Tree Reconstruction. *Mol. Biol. Evol.* 31(6):1625–1637.

Chen J-Q, Wu Y, Yang H, Bergelson J, Kreitman M, Tian D. 2009. Variation in the Ratio of Nucleotide Substitution and Indel Rates across Genomes in Mammals and Bacteria. *Mol. Biol. Evol.* 26(7):1523–1531.

Cohen J. 1988. Statistical Power Analysis for the Behavioral Sciences (second ed.). Lawrence Erlbaum Associates. 67.

de la Chaux N, Messer PW, Arndt PF. 2007. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evolutionary Biology*, 7:19.

Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* 15:330–340.

Douzery EJP, Scornavacca C, Romiguier J, Belkhir K, Galtier N, Delsuc F, Ranwez V. 2014. OrthoMaM v8: A Database of Orthologous Exons and Coding Sequences for Comparative Genomics in Mammals. *Mol. Biol. Evol.* 31:1923–1928.

Edgar RC, Batzoglou S. 2006. Multiple sequence alignment. *Current Opinion in Structural Biology* 16:368–373.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* 32:1792–1797.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evol.* 63(1):1–19.

Felsenstein J. 1985. Phylogenies and the comparative method. *The American Naturalist* 125(1):1–15.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle. Available from evolution.genetics.washington.edu/phylip.html

Fitzpatrick DA, Logue ME, Stajich JE, Butler G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol. Biol.* 6:99.

Fletcher W, Yang Z. 2009. INDELible: A Flexible Simulator of Biological Sequence Evolution. *Mol. Biol. Evol.*26(8):1879–1888.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates C, Fairley S, Fitzgerald S, et al. 2011. Ensembl 2011. *Nucl. Acids Res.* 39:D800–D806.

Gamble T, Berendzen PB, Shaffer B, Starkey DE, Simons AM. 2008. Species limits and phylogeography of North American cricket frogs (Acris: Hylidae). *Mol. Phylogenet. Evol.* 48:112–125.

García-Pereira MJ, Caballero A, Quesada H. 2011. The relative contribution of band number to phylogenetic accuracy in AFLP data sets. *J. Evol. Biol.* 24:2346–2356.

Göker M, Scheuner C, Klenk H, Stielow JB, Menzel W. 2011. Codivergence of mycoviruses with their hosts. *PLoS One*. 6:e22252.

Graur D, Zheng Y, Azevedo RBR. 2015. An evolutionary classification of genomic function. *Genome Biol. Evol.* 7(3):642–645.

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. 2013. On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biol. Evol.* 5(3):578–590.

Hall BG. 2005. Comparison of the Accuracies of Several Phylogenetic Methods Using Protein and DNA Sequences. *Mol. Biol. Evol.* 22:792–802.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22(2):160–174.

Jordan G, Goldman N. 2012. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* 29(4):1125–1139.

Jukes TH, Kimura M. 1984. Evolutionary constraints and the neutral theory. *J. Mol. Evol.* 21(1):90–92.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucl. Acids Res.* 33:511–518.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief *Bioinform.* 9(4):286–298.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res.* 12(6):996–1006.

Knowles LL, Lanier HC, Klimov PB, He Q. 2012. Full modeling versus summarizing gene-tree uncertainty: Method choice and species-tree accuracy. *Mol. Phylogenet. Evol.* 65:501–509.

Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Wägele JW, Misof B. 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front. Zool.* 7:10.

Kuhner MK, Felsenstein J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol.* 11(3):459–468.

Landan G. 2005. Ph.D. Dissertation. Tel-Aviv University.

Landan G, Graur D. 2007. Heads or Tails: a simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.* 24:1380–1383.

Landan G, Graur D. 2008. Local reliability measures from sets of co-optimal multiple sequence alignments. *Pac. Symp. Biocomput.* 13:15–24.

Landan G, Graur D. 2009. Characterization of pairwise and multiple sequence alignment errors. *Gene.* 441(1–2):141–7.

Leaché AD, Mulcahy DG. 2007. Phylogeny, divergence times and species limits of spiny lizards (Sceloporus magister species group) in western North American deserts and Baja California. *Mol. Ecol.* 16:5216–5233.

Lepage T., Bryant D., Philippe H., Lartillot N. 2007. A General Comparison of Relaxed Molecular Clock Models. *Mol. Biol. Evol.* 24(12):2669–2680.

Li W-H., Tanimura M. 1987. The molecular clock runs more slowly in man than in apes and monkeys. *Nature* 326:93–96.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478: 476–482.

Lovell SC, Robertson DL, 2010. An Integrated View of Molecular Coevolution in Protein–Protein Interactions. *Mol. Biol. Evol.* 27:2567–2575.

Markova-Raina P, Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Res.* 21:863–874.

Marshall DC, Simon C, Buckley TR. 2006. Accurate Branch Length Estimation in Partitioned Bayesian Analyses Requires Accommodation of Among-Partition Rate Variation and Attention to Branch Length Priors. *Syst. Biol.* 55(6):993–1003.

Marshall DC. 2010. Cryptic Failure of Partitioned Bayesian Phylogenetic Analyses: Lost in the Land of Long Trees. *Syst. Biol.* 59(1):108–117.

Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, et al. 2007. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.* 17(12):1797–1808.

Misof B, Misof K. 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Syst. Biol.* 58(1):21–34.

Miyata T, Yasunaga T, Nishida T. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. *PNAS* 77(12):7328–7332.

Morrison DA. 2009. Why would phylogeneticists ignore computerized sequence alignment? *Syst. Biol.* 58: 150-158.

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Pond SLK, Scheffler K. 2013. FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Mol. Biol. Evol.* 30 (5):1196–1205.

Nishihara H, Hasegawa M, Okada N. 2006. Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. *PNAS* 103: 9929–9934.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302(1):205–217.

Nuin PAS, Wang Z, Tillier ERM. 2006. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinform.* 7:471.

Nye TMW, Lio P, Gilks WR, 2006. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinform.* 22:117–119.

Ogden TH, Rosenberg MS. 2006. Multiple Sequence Alignment Accuracy and Phylogenetic Inference. *Syst. Biol.* 55(2): 314–328.

Pang A, Smith AD, Nuin PAS, Tillier ERM. 2005. SIMPROT: Using an empirically determined indel distribution in simulations of protein evolution. *BMC Bioinform.* 6:236.

Pazos F, Juan D, Izarzugaza JMG, Leon E, Valencia A. 2008. Prediction of Protein Interaction Based on Similarity of Phylogenetic Trees. *Methods in Mol. Biol.* 484:523–535.

Privman E, Penn O, Pupko T. 2012. Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol. Biol. Evol.* 29:1–5.

Penn O, Privman E, Landan G, Graur D, Pupko T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol. Biol. Evol.* 27:1759–1767.

Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Quéinnec E, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. *Current Biology* 19:706–712.

Prasad AB, Allard MW, NISC Comparative Sequencing Program, Green EE. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol. Biol. Evol.* 25:1795–1808.

Regier JC, Shultz JW, Ganley ARD, Hussey A, Shi D, Ball B, Zwick A, Stajich JE, Cummings MP, Martin JW, Cunningham CW. 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst. Biol.* 57: 920–938.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math. Biosciences* 53:131–147.

Rodríguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Löffelhardt W, Bohnert HJ, Philippe H, Lang BF.. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Current Biology* 15:1325–1330.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinform.* 19(12):1572–1574.

Rosa ML, Fiannaca A, Rizzo R, Urso A. 2013. A Study of Compression–Based Methods for the Analysis of Barcode Sequences. *Comput. Intell. Methods for Bioinform. and Biostat. Lect. Notes in Comput. Sci.* 7845:105–116.

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

Sanderson MJ. 2002. Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. *Mol. Biol. Evol.* 19(1):101–109.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol. Evol.* 1:114–118.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539.

Smith SA, Donoghue MJ. 2008. Rates of Molecular Evolution Are Linked to Life History in Flowering Plants. *Science* 322:86–89.

Soria-Carrasco V, Talavera G, Igea J, Castresana J. 2007. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. *Bioinform.* 23:2954–2956.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinform.* 22(21):2688–2690.

Stoye J, Ever D, Meyer F. 1998. Rose: generating sequence families. *Bioinform.* 14(2):157–163.

Strope CL, Abel K, Scott SD, Moriyama EN. 2009. Biological Sequence Simulation for Testing Complex Evolutionary Hypotheses: indel-Seq-Gen Version 2.0. *Mol. Biol. Evol.* 26(11):2581–2593.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* 30(12):2725–2729.

Taylor MS, Ponting CP, Copley RR. 2004. Occurrence and Consequences of Coding Sequence Insertions and Deletions in Mammalian Genomes. *Genome Res.* 14(4):555–566.

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

Thompson JD, Linard B, Lecompte O, Poch O. 2011. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One* 6:e18093

von Reumont BM, Meusemann K, Szucsich NU, Dell'Ampio E, Gowri-Shankar V, Bartel D, Simon S, Letsch HO, Stocsits RR, Luan Y, et al. 2009. Can comprehensive background knowledge be incorporated into substitution models to improve

phylogenetic analyses? A case study on major arthropod relationships. *BMC Evol. Biol.* 9:119.

Wang H, Susko E, Roger AJ. 2013. The Site-Wise Log-Likelihood Score is a Good Predictor of Genes under Positive Selection. *J Mol. Evol.* 76:280–294.

Wang L, Jiang T. 1994. On the complexity of multiple sequence alignment. *J. Comput. Biol.* 1:337–348.

Wang L-S, Leebens-Mack J, Wall PK, Beckmann K, dePamphilis CW, Warnow T. 2011. The impact of multiple protein sequence alignment on phylogenetic estimation. *IEEE/ACM Transactions on Comp. Biol. and Bioinform.* 8: 1108–1119.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.

Wu M, Chatterji S, Eisen JA. 2012. Accounting for alignment uncertainty in phylogenomics. *PLoS ONE* 7:e30288.

Yamamoto S, Kasai H, Arnold DL, Jackson RW, Vivian A, Harayama S. 2000. Phylogeny of the genus Pseudomonas: intrageneric structure reconstructed from the nucleotide sequences of gyrB and rpoD genes. *Microbiology* 146:2385–2394.

Zhang Z, Gerstein M. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucl. Acids Res.* 31(18):5338–5348.

Zhou X, Xu S, Xu J, Chen B, Zhou K, Yang G. 2012. Phylogenomic Analysis Resolves the Interordinal Relationships and Rapid Diversification of the Laurasiatherian Mammals. *Syst. Biol.* 61:150–164.