eISSN: 09748369, www.biolmedonline.com

The impact of sequence parameter values on phylogenetic accuracy

Bhakti Dwivedi[1], *Sudhindra R Gadagkar[1,2]

[1]Department of Biology, University of Dayton, Dayton, OH 45469-2320, USA.
[2]Present Address: Central State University, Wilberforce OH 45384-1004, USA.

*Corresponding author: sgadagkar@centralstate.edu

Abstract
An accurately inferred phylogeny is important to the study of evolution. Factors affecting the accuracy of an inferred tree can be traced to several sequential steps leading to the inference of the phylogeny. We have examined here the impact of some features of nucleotide sequences in alignments on phylogenetic (topological) accuracy rather than any source of error during the process of sequence alignment or choice of the method of inference (as is usually done). Specifically, we have studied (using computer simulation) the implications of changing the values of the following five parameters, individually and in combination: sequence length (l), nucleotide substitution rate (r), nucleotide base composition (θ), the transition-transversion rate ratio (κ), and the substitution rate heterogeneity among the sites (α). An interesting, and unexpected, result was that κ has a strong positive relationship with phylogenetic accuracy, especially at high substitution rates. This simulation-based work has implications for empirical researchers in the field and should enable them to choose from among the multiple genes typically available today for a more accurate inference of the phylogeny being studied.

Keywords: Molecular evolution, phylogenetic inference, nucleotide substitution rate, transition-transversion rate ratio, phylogenetic accuracy, substitution saturation

Introduction

Phylogenetic reconstruction from an alignment of molecular sequences is the last in a series of consecutive steps that include obtaining the required sequences (either from DNA extracted from tissue or from a databank), aligning them, and employing a method of inference to reconstruct the phylogeny. Obviously, the correctness of the inferred phylogeny depends upon accuracy at each of these steps, and therefore, most studies that evaluate the determinants of the accuracy of phylogenetic inference have focused on understanding the contribution of these steps, particularly the latter two, to the accuracy of the inferred tree (Hillis 1995; Nei 1996; Takahashi and Nei 2000; Raghava *et al.* 2003; Huelsenbeck and Rannala 2004; Hall 2005; Rosenberg 2005a; Rosenberg 2005b; Ogden and Rosenberg 2006)

In the present study, however, rather than assess the contribution to phylogenetic accuracy of the three steps mentioned above, we have instead, focused on dissecting the features of the DNA sequences to determine the optimal combinations of sequence parameters that are associated with accurately inferred phylogenies. With the advances in sequencing technology already yielding DNA sequences for hundreds of taxonomic groups, and with the promise of much more to come with the popularization of next-generation sequencing (Mardis 2008) and metagenomics (Singh *et al.* 2009), several studies have focused on improving phylogenetic inference methods to handle large datasets (e.g., (Guindon and Gascuel 2003; Tamura, Nei and Kumar 2004). However, with the new technologies yielding large sets of diverse sequences and projects such as the Tree of Life (Maddison, Schulz and Maddison 2007) utilizing them to compare across extremely large times of divergence, it is also necessary to understand the effect of large differences in various parameters of these sequences on the accuracy of phylogenetic inference.

When an alignment of molecular sequences from different species is used to infer a phylogeny, what is actually being inferred is the evolutionary history of the sequences in the alignment, with the expectation that it accurately reflects the evolutionary history of the organisms whose sequences are in the alignment (Nei and Kumar 2000; Felsenstein 2003). However, since different genes can produce different evolutionary histories (trees) for a group of taxa (Nichols 2001; Gadagkar, Rosenberg and Kumar 2005), it is important to understand the individual and joint effects of the sequence

parameters on the accuracy of phylogenetic reconstruction.

DNA sequences can be characterized by summary statistics such as length and base composition. When two or more such sequences need to be compared to each other (as in an alignment prior to phylogenetic analysis) additional parameters come into play, such as the overall rate of nucleotide substitution (replacement of one nucleotide by another nucleotide), the ratio of two specific instantaneous rates of substitution: rate at which transitions (A↔G or C↔T) and transversions (all other changes) occur, and the rate variation among sites. These comprise some of the sequence parameters that are important for the accurate reconstruction of a phylogeny.

For example, the amount of evolution (substitution rate) among sequences is usually a deciding factor in reconstruction of a phylogeny, such that slow-evolving gene sequences (e.g., elongation factor-1α, small subunit ribosomal RNA) are used to infer the relationships among distantly related taxa (Regier and Shultz 1997; Struck *et al.* 2007), and fast-evolving sequences (such as animal mitochondrial genes, virus genomes, and the third codon position of protein coding genes) are used to infer the phylogeny among closely related organisms (Hillis *et al.* 1992; Ou *et al.* 1992; Yang 1996a, 1996b; Yoder, Vilgalys and Ruvolo 1996.) The reason fast-evolving genes cannot be used to infer phylogenies of distantly related taxa is of course, because higher evolutionary rates lead to multiple substitutions at the same site, and thus, a saturation of the phylogenetic signal, leading to incorrect tree reconstruction (Halanych and Robinson 1999; Xia 2000; Struck, Hessling and Purschke 2002; Struck *et al.* 2007). Therefore, different approaches have been developed to detect saturation in order to exclude entire genes or parts thereof from the phylogenetic analysis (Xia *et al.* 2003; Struck *et al.* 2008). For example, the ratio of the numbers of transitions to transversions plotted against the genetic distances (*p*) for all pairwise sequence comparisons in the alignment is a common test to determine whether sequences have experienced substitution saturation (Halanych and Robinson 1999; Struck, Hessling and Purschke 2002; Xia *et al.* 2003; Struck *et al.* 2007). In general, since the frequency of transitional substitutions is known to be higher than transversional substitutions in the genome (Wakeley 1996), this test detects saturation when the plot shows no further increase in the observed number of transitions despite increasing genetic distances. Thus, saturation of transitions at high levels of sequence divergence indicates saturation in the data.

Most studies that have investigated the influence of sequence parameter values on phylogenetic accuracy have varied one sequence parameter at a time, and therefore, have failed to record the influence of the interaction among the parameters, the inference methods, and any other factors considered, such as the topology of the model tree. Therefore, we simulated evolution along different model trees (Fig. 1) to generate non-coding DNA sequences, varying different parameters across wide ranges (that are, however, biologically realistic), and compared the performance of different inference methods in reconstructing the phylogeny. During the simulation process, we systematically varied the following sequence parameters: sequence length (*l*), overall rate of nucleotide substitution (*r*), nucleotide base-composition (θ), transition-transversion rate ratio (κ), and the heterogeneity of substitution rates among sites (α), in order to study their individual and joint effects on the accuracy of phylogenetic tree reconstruction, using the following inference methods: Neighbor-Joining (NJ), Maximum Parsimony (MP) and Likelihood-based methods (ML and PhyML). In addition, we simulated evolution along tree topologies of different size, shape, and relative branch lengths.

While most of our results agree with those in the literature, there is one notable exception, namely, those involving κ, which shows a positive relationship with phylogenetic accuracy, thus appear to contradict previous studies (Yang 1998; Rosenberg and Kumar 2003). Our results are, however, consistent for all the tree topologies examined, regardless of the values of the other parameters, although differences exist in the extent of accuracy achieved. The positive relationship between the value of κ and phylogenetic accuracy is stronger when the evolutionary rate, *r*, is high and thus, contributing to saturation of the pairwise genetic distances among the sequences in the alignment. Our results also showed that all the four inference methods performed equally well under substantial saturation (high *r* and high κ), while there were significant differences in accuracy among them at high *r* and low κ. Our findings suggest that highly divergent datasets are still usable, as the phylogenetic information is often not completely lost, and may be

retrieved using sites that have experienced more transversions.

Materials and Methods

Computer Simulation

Nucleotide sequence alignments were generated using the computer program Dawg version 1.1.2 (Cartwright 2005) for four ultrametric, 16-taxon topologies (Fig. 1), obtained from Ogden and Rosenberg (2006). Simulations were also performed using non-ultrametric 16-taxon topologies as in Fig. 1 (not shown). DNA evolution was simulated using only nucleotide substitution events under the HKY model (Hasegawa, Kishino and Yano 1985), while systematically varying the following sequence parameters in a fully factorial manner: sequence length (l), nucleotide base frequencies (expressed as the G+C content, θ), rate of nucleotide substitution (r) as a multiplier, the transition to transversion rate ratio (κ), and the shape parameter (α) of the gamma distribution that describes the rate variation among-sites. The values of the sequence parameters used in the simulations are given in Table 1. All other options in the simulation program were set to default during simulation.

While the ranges of these parameters have been deliberately kept rather large in order to understand the full scope of their influence on phylogenetic accuracy, they are not unrealistic, and have in fact been obtained empirically from mammalian genes and used in earlier studies (Rosenberg and Kumar 2003; Gadagkar and Kumar 2005). In particular, such ranges, particularly for r and κ, are seen in mitochondrial genes (Yoder, Vilgalys and Ruvolo 1996; Yang 1996c; Yang 1998; Yang and Yoder 1999), nuclear non-coding introns (Saitou and Ueda 1994), and even in some nuclear genes (Rosenberg and Kumar 2003). Each sequence parameter combination (total 576 "genes") was used for the simulations along the four model trees, and 100 replicates were obtained for each gene, thus producing 230,400 datasets.

Phylogenetic tree reconstruction

After the simulations were done, the sequence alignments obtained were subjected to phylogenetic inference using Neighbor-Joining (NJ), Maximum Parsimony (MP), and Maximum likelihood (ML) methods as implemented in PAUP* version 4.0 b10 (Swofford 2003). In addition, likelihood analysis was also done using PhyML version 2.4.4 (Guindon and Gascuel 2003) because of its

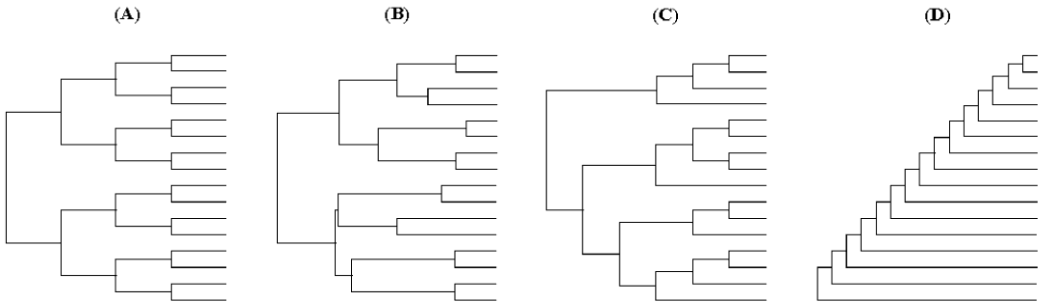

Figure 1. The model trees. The four ultrametric 16-taxon topologies obtained from Ogden and Rosenberg (2006) used as model trees for the simulations of DNA evolution, shown with relative branch lengths: (A) Balanced tree with equal branch lengths, (B) Balanced tree with random branch lengths, (C) Random tree generated under a pure birth (Yule) process, and (D) Pectinate tree . During simulation, the total number of substitutions to be made for a given branch, for a given parameter-combination, was obtained as the product of the branch length in the model tree, the rate multiplier and the sequence length. Values from the latter two parameters were obtained from Table 1.

Table 1. **The sequence parameter values used in the simulations.** The sequence length, l, is measured as the number of nucleotides, the nucleotide substitution rate, r is a multiplier, that, when multiplied by a given branch length in the model tree and the sequence length, yields the expected number of substitutions to be introduced in that branch during simulation. The nucleotide base frequencies are expressed in terms of G+C content, θ. κ represents the transition to transversion rate ratio, and the shape parameter, α specifies the extent of rate heterogeneity among sites.

Sequence Parameters	Values
Sequence length, l	500, 2500
G+C content, θ	0.20, 0.50, 0.80
Transition-transversion rate ratio, κ	1, 5, 10, 20
Gamma distribution shape parameter, α	0.1, 0.5, 5.0, $\propto$ (*infinity*)
Rate of nucleotide substitution, r	0.025, 0.05, 0.1, 0.2, 0.4, 0.8

speed and efficiency in comparison to ML. HKY (Hasegawa, Kishino and Yano 1985) pair-wise distance estimates were used for the NJ analyses. In PhyML, the initial tree was built using BIONJ (Guindon and Gascuel 2003). The parameters of the HKY substitution model (the four base frequencies and the transition/transversion rate ratio) along with the proportion of invariable sites and the gamma distribution shape parameter were estimated from the simulated data using PAUP*. For the MP and ML analyses, a heuristic search was done using the stepwise addition algorithm for the provisional tree and subsequent branch swapping was done using the Nearest-Neighbor Interchange (NNI) method. When multiple trees were recovered, a strict consensus of these trees was taken to produce a single tree. All other settings were set to default in PAUP*, and PhyML program.

Assessing phylogenetic accuracy

The accuracy of the phylogenetic trees inferred was measured as the percentage of internal branches (or nodes) reconstructed correctly in the inferred tree, P_C, obtained as $P_C = \left[1 - \frac{d_T}{(2m-6)}\right]100$, where m is the number of sequences in the phylogeny (16) and d_T is the topological distance between the inferred tree and model tree (Robinson and Foulds 1981; Penny and Hendy 1985). P_C values were averaged over all the (100) replicates for each parameter combination, to give $\bar{P}_C$ and is expressed in percent. For example, 60 percent accuracy means 60 percent of the internal branches are reconstructed correctly in the reconstructed (or inferred tree) when compared to the model tree.

Results

Overall Performance

We first examine the overall accuracy, $\bar{P}_C$ of each inference method for the lowest and highest values of each parameter (Table 2). It can be seen that there is a large difference in accuracy between the two extreme values considered in this study for some parameters, and not for others. Furthermore, some inference methods appear to show a greater difference than others. When the accuracy is compared between the lowest and highest parameter values, the inference methods, in general, show an increase in accuracy for l, the sequence length, a decrease for r, the substitution rate, very slight to no change for θ, the base composition, an increase for κ, the transition-transversion rate ratio, and a decrease for α, the shape parameter, with the greatest difference in accuracy seen in the case of r, and the least in the case of θ (although a G+C content of 0.50 shows a slightly higher accuracy when compared to the two extreme values of 0.20 and 0.80; not shown). In general, most of these

results are not novel, and have been shown before, although perhaps not in such detail. However, what is surprising is the behavior of κ, the transition-transversion rate ratio. This parameter has seldom been the focus in the literature, but available studies have generally attributed a negative relationship between phylogenetic accuracy and the value of κ (Yang 1998; Rosenberg and Kumar 2003). The results of this study, on the other hand, show that the marginal effect of κ, when averaged over the other parameters, has a positive relationship with accuracy. This is dealt with at length below.

Among the inference methods, NJ results show the greatest difference between the lowest and highest values of each of the parameters, on average, while all the other methods show comparable values of accuracy between them. NJ also shows the least values for $\overline{P}_C$, whether at the lowest or highest parameter values, when compared to the other methods, all of which were somewhat comparable. As far as the topologies are concerned, the highest accuracy is seen for the balanced tree with equal branch lengths (Balanced tree A), followed closely by the random tree topology. Balanced tree B (with unequal branch lengths) has a lower accuracy, in general, than the above two topologies, while the pectinate tree does very poorly. These results are consistent among the inference methods.

The general trend in the association of the sequence parameters with phylogenetic accuracy is the same (as explained above) among all the inference methods and model tree topologies, except when the MP method is used on the datasets obtained from the pectinate tree. This shows a slight decrease in accuracy with increase in κ. The pectinate tree shape also appears more sensitive to changes in the values of sequence parameters and shows greater differences in accuracy for the lowest and highest parameter values, regardless of the inference method examined. However, the ML methods (PhyML and ML) yield more accurate results than MP or NJ from the pectinate tree datasets.

Table 2. Marginal effects of low and high values of sequence parameters on phylogenetic accuracy. The overall accuracy, $\overline{P}_C$, of the four phylogenetic methods (PhyML, ML, MP, and NJ) for each model tree at the lowest and highest values of sequence length (*l*), substitution rate (*r*), G+C content (θ), transition-transversion rate ratio (κ), and the shape parameter (α), when averaged over all the values of the other parameters (see text).

		Sequence Parameters									
		l		***r***		**θ**		**κ**		**α**	
Method	**Model Tree**	**Min.**	**Max.**	**Min.**	**Max.**	**Min.**	**Max.**	**Min.**	**Max.**	**Min.**	**Max.**
PhyML	Balanced A	96.10	96.91	99.46	85.40	96.47	96.43	91.67	99.42	98.29	94.68
	Balanced B	87.62	91.93	93.46	78.64	89.61	89.71	85.44	92.53	89.44	89.42
	Random	94.36	96.72	98.62	84.00	95.48	95.50	90.75	98.36	96.10	95.00
	Pectinate	68.44	89.64	83.43	70.05	78.45	78.30	78.63	78.00	64.23	86.60
ML	Balanced A	94.45	97.87	99.12	84.58	95.69	95.71	91.58	99.08	97.19	94.80
	Balanced B	87.65	93.82	94.64	77.91	90.11	90.22	86.32	93.59	90.65	90.10
	Random	91.26	97.36	98.06	80.53	93.67	93.78	89.39	97.59	93.92	93.53
	Pectinate	66.03	90.06	86.63	53.67	76.59	76.69	72.07	82.97	68.80	81.20
MP	Balanced A	93.65	98.96	98.78	86.38	96.41	96.37	93.27	98.00	94.12	97.06
	Balanced B	85.51	92.29	93.05	78.24	89.11	89.20	86.41	90.48	85.88	90.28
	Random	91.72	98.60	97.81	85.78	95.06	95.04	92.94	96.29	91.51	96.57
	Pectinate	56.99	79.50	80.22	47.96	67.05	66.89	69.48	65.56	54.26	75.03

NJ	Balanced A	88.13	94.22	98.74	68.30	90.52	90.57	83.15	98.02	94.40	87.79
	Balanced B	80.95	89.28	92.93	62.97	84.40	84.45	77.77	91.36	86.49	83.19
	Random	85.03	93.98	97.48	65.65	88.73	88.81	81.77	96.11	90.07	87.71
	Pectinate	58.17	83.32	81.84	42.83	69.21	69.20	63.68	76.65	56.99	76.76

Since the effect of κ on phylogenetic accuracy is contrary to generally held views, it warrants closer scrutiny. After studying its marginal effects for each inference method and for each topology when taken over all the other parameter values (Table 2), we next studied the interaction between κ and each of the other parameters, taken one at a time (while averaging over the rest of the other parameters), while noting differences among the model trees and inference methods as well.. Finally, we explain the behavior of κ from the perspective of substitution saturation.

Interaction of κ with each of the other sequence parameters

κ *and l*: When *l* and κ are varied simultaneously, the difference in $\bar{P}_C$ is almost non-existent between the two values of sequence length, whereas the effect of changes in the value of κ is quite obvious, with the accuracy of almost 80 percent when κ = 1 and reaching practically 100 percent when κ = 20, in the case of balanced and random tree topologies (not shown). The pectinate tree also shows an increase in accuracy with increase in kappa (except when inferred using MP method), although the extent of accuracy obtained is the least in comparison to other tree shapes for all inference methods, and irrespective of the sequence length (*l* = 500 or *l* = 2500). The percent increase in accuracy for each increase in κ was maximum in the case of NJ, improving the accuracy by about 10-20%.

κ *and* θ: No significant trends were observed between Kappa and the base composition (not shown). The accuracy improved with increase in κ, irrespective of the GC content (not shown). No significant differences were observed among the GC content values, G+C = 0.50 yields a slightly better accuracy than extreme G+C values (0.20 or 0.80). The results were consistent among all the inference methods and tree shapes, except again for MP in the case of the pectinate tree, where a slight decrease in accuracy was seen with increase in kappa at all GC content values (not shown).

κ *and r*: The interaction of κ with *r*, shown in Figure 2, is the most important in affecting the accuracy of phylogenetic inference because an increase in the value of *r* adds to the effect that κ has on accuracy. Taken individually, phylogenetic accuracy decreases with increase in evolutionary rate and increases with increases in Kappa, regardless of the inference method or tree topology (Figure 2). However, when a high *r* is coupled with a high Kappa (e.g., when κ = 20 and r ≥ 0.4), this results in an increased accuracy of about 90 percent or greater. The decline in accuracy is seen mainly when $r \geq 0.2$, and $\kappa \leq 10$. At low substitution rates ($r \leq 0.2$), the accuracy is almost close to 100 percent at all kappa values. The extent of accuracy achieved for each interaction of *r* and κ, differs among the phylogenetic methods (in the order, PhyML > ML > MP > NJ), and with the topology of the model tree (Balanced A > Random > Balanced B > Pectinate tree). Interestingly, when κ is high (κ = 20), all the inference methods, and tree shapes perform equally well and lead to a similar improved accuracy. The only exception is that when $r \leq 0.4$, the accuracy decreases slightly (~ 5%) with increase in Kappa, for MP trees associated with the pectinate topology (Figure 2D). When balanced topology A was used, the percent increase in accuracy at high rates ($r \geq 0.4$) from κ = 1 to κ = 20 was high, but differed considerably among the inference methods: 30% in ML, 35% in PhyML, 70% in ML, and 90% in NJ (Figure 2A). In case of pectinate tree, the accuracy from κ = 1 to κ = 20 decreases by 3% at low rates ($r \leq 0.05$), and at higher rates ($r \geq 0.4$) eventually increases by almost 50% (Figure 2D).

κ *and α*: The among-site rate variation, measured as alpha (α) is an important determinant of phylogenetic accuracy (Yang 1996c). In our study, the results reported above regarding the interaction between Kappa and substitution rate hold true only when the rate is homogenous across the sites (α = infinity).

When the among-sites rate is heterogeneous (α = 0.1) the results are relatively constant across all rates, irrespective of the value of Kappa (Figure 3). Figure 3 shows the influence of α for each combination of substitution rate and kappa, for the inference method PhyML, under the four model tree topologies. When α = infinity (Fig. 3A), the accuracy is low when Kappa is low and *r* is high. However, as the value of Kappa increases, the accuracy dramatically increases. On the other hand, when the rate is heterogeneous among sites, in particular, when only fewer sites experience evolutionary changes ($\alpha \leq 0.5$) the accuracy is high to begin,

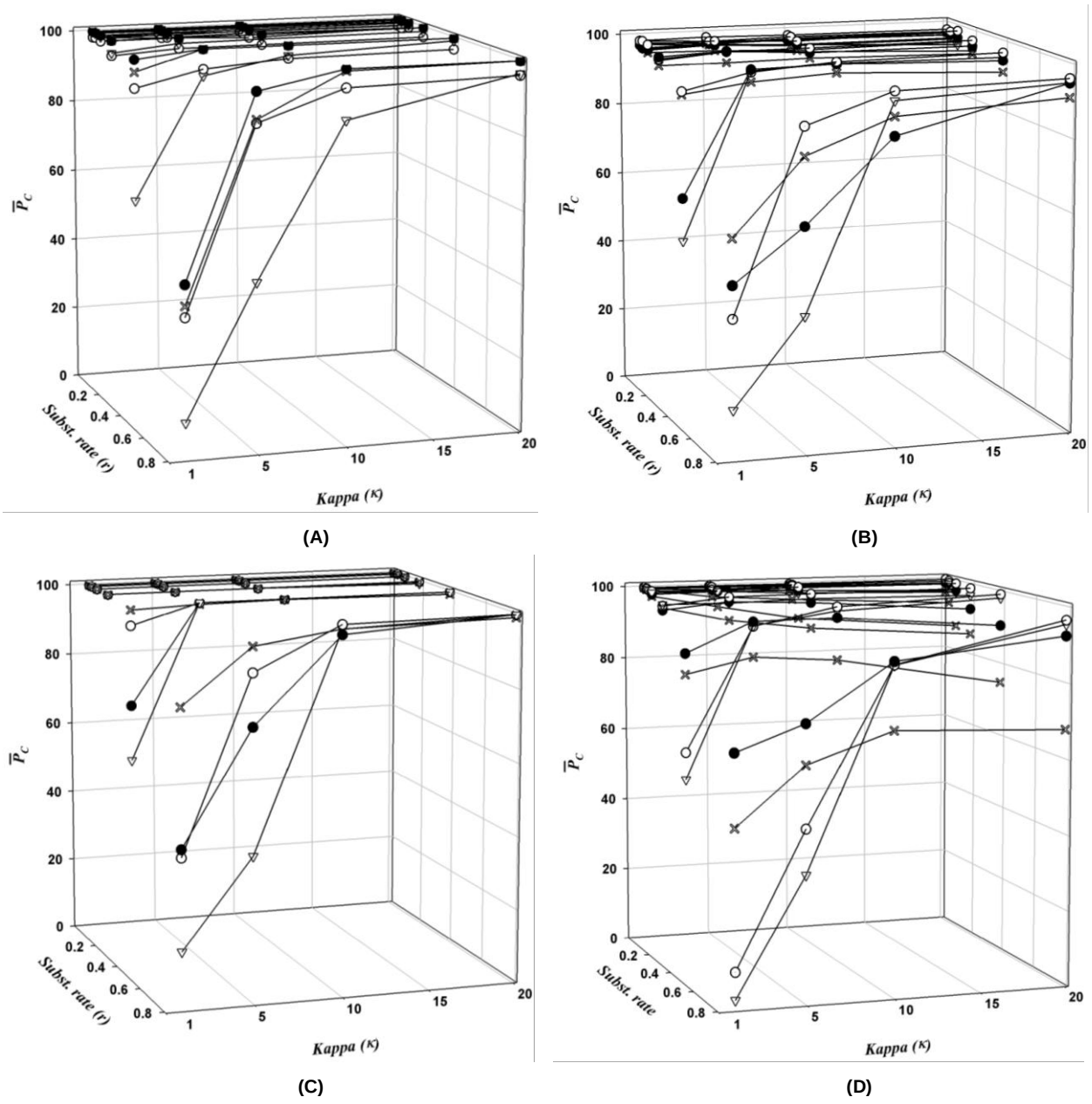

Figure 2. Joint effects of kappa and substitution rates on the phylogenetic accuracy. The phylogenetic accuracy ($\overline{P}_C$) plotted against substitution rate, *r*, and transition-transversion rate ratio, κ, for the four model tree topologies of Fig. 1: (A) Balanced 'A', (B) Balanced 'B', (C) Random, and (D) Pectinate. $\overline{P}_C$ for the four inference methods, PhyML, ML, MP, and NJ, is represented by the following symbols: filled circle, open circle, cross, and open triangle, respectively. Each point in the graph represents an average over all the replicates for *l*=2500, θ =0.50, and α = infinity.

and, as a consequence, does not change much with increase in Kappa, even at higher rates (Figure 3B). κ = 20; not shown), irrespective of the model tree. In PhyML or ML and MP the percent increase in accuracy from κ = 1 to κ = 20 is approximately 60% and 40%, respectively. This percent increase however, is different for Balanced B topology and Pectinate tree shape (e.g., see Fig. 3 for PhyML).

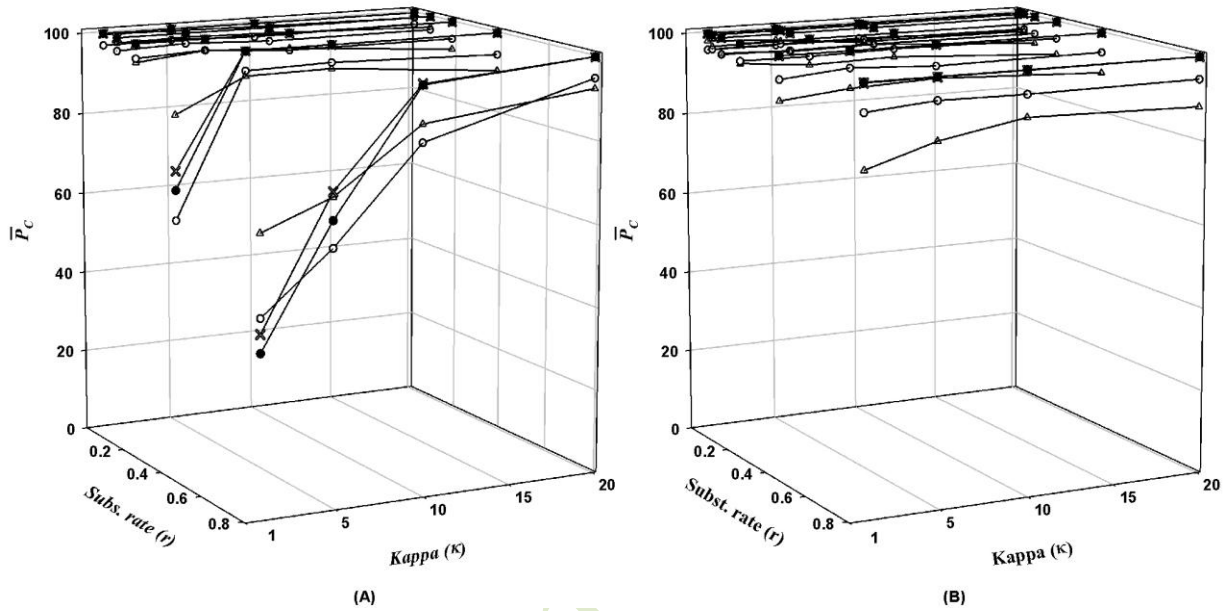

Figure 3. Combined effect of Kappa and Substitution rate on phylogenetic accuracy, for the PhyML method. The results are shown for two values of the shape parameter: (A) α = infinity (rate homogeneity among sites), and (B) α = 0.1 (rate heterogeneity among sites). The phylogenetic accuracy for the four tree topologies (Fig 1), Balanced 'A', Balanced 'B', Random, and Pectinate, is represented by the following symbols: filled circle, open circle, cross, and triangle, respectively. Each point in the graph represents an average over all the replicates for l=2500, and θ=0.50.

Nucleotide Substitution Saturation and Role of Transversions

Distantly diverged gene sequences (or sequences with high substitution rates) often experience substantial substitution saturation, especially in the third codon position of protein-coding genes. This saturation can misrepresent the phylogenetic information contained in the sequences, leading to incorrect phylogenetic inference. Some simple ways of handling such sequences include avoiding sequences with pair-wise evolutionary distances larger than 1 (Nei and Kumar 2000) and plotting either number of substitutions or the transition to transversion ratios against a corrected genetic distance (Xia and Xie 2001). A more sophisticated method is the entropy-based test of substitution saturation as implemented in DAMBE (Xia *et al.* 2003). This test, when used on our datasets, suggested significant saturation in the sequences obtained from simulations with higher evolutionary rates ($r \geq 0.4$), regardless of the tree topology. The effect was more profound in dataset with homogenous rate distribution, irrespective of kappa and other parameters. At high substitution rates, and under rate homogeneity, it is likely that each site in the sequence will be "hit" multiple times. If the transition rate is much higher than transversion rate (as in κ=20), than it is more likely for one transition to be followed by another transition simply because another transition is more likely than a transversion. Under these circumstances, a high transition to transversion rate ratio should lead to saturation of transitions more than transversions. On the other hand, a

transition to transversion rate ratio of one (κ=1) means the two rates are the same. The latter situation has a relatively higher likelihood of saturation (when compared to κ=20) of both the substitution types, although now transversions are more likely than transitions (twice as likely, to be exact). We believe that, at high rates ($r \geq 0.4$), the improved accuracy (≥ 90%) at high kappa (κ=20) in comparison to the accuracy (≤ 40%) at low kappa (κ=1), is because there is little saturation of transversions at high kappa and saturation of both types at low kappa. Phylogenetic trees inferred from the transition-only and transversion-only sites confirmed this expectation. Phylogenetic trees reconstructed from transversion-only sites (for dataset with r = 0.8, κ =20, α = infinity) yielded trees with most of the of the internal branches correct (≥ 90), while trees inferred using the same dataset but from transition-only sites failed to infer almost any of the internal nodes correctly (and showed an accuracy close to zero), showing saturation in the transitions. However, for datasets with r = 0.8, κ =1, and α = infinity, neither transition-only nor transversion-only sites gave accurate trees, suggesting that both transitions and transversions had undergone saturation). These results hold for all the tree shapes investigated in this study.

Discussion

We have presented here the results of a simulation-based study undertaken to investigate the influence of the following sequence parameters: sequence length (l), nucleotide substitution rate (r), base-composition (θ), transition-transversion rate ratio (κ) and the shape parameter (α) (that specifies the extent of heterogeneity in the substitution rate across sites), individually and jointly, on the accuracy of phylogenetic reconstruction by four inference methods: Neighbor-Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML and PhyML) methods. The model trees used were four 16-taxon tree topologies (Figure 1). All five parameters were varied and combined in a factorial manner (Table 1) and 100 replicates were generated for each of the 576 parameter combinations, for each model tree. The accuracy of phylogenetic reconstruction was measured as the number of correctly inferred internal branches, determined based on the topological distances (Robinson and Foulds 1981; Penny and Hendy 1985), and averaged over the 100 replicates for a given parameter combination, and finally expressed as $\overline{P}_C$, the average percent of correct branches.

The marginal effects of l, r, θ, and α held no surprises in their general trends, although the results were not entirely devoid of interest (Table 2). Clearly, however, the most interesting parameter was κ, the transition-transversion rate ratio. In particular, this parameter was influenced with the changing values of r. When $r \leq 0.2$ the phylogenetic accuracy is almost 100%, regardless of the kappa. When rate is high $r \geq 0.4$, however, the accuracy is clearly affected and is correlated to the value of kappa. Of the four inference methods, NJ appears particularly prone to the impact of the combination of these two parameters (Fig. 2). Overall, when the accuracy is assessed for all kappa at high substitution rate (r = 0.8) the likelihood methods outperform MP, which outperforms the NJ method. This trend is common to the balanced trees. For the random tree shape, MP is better than the PhyML and ML (at low kappa), and under pectinate tree topology, PhyML outperforms MP, which in turn, does better than ML, and which in turn, is better than NJ (Figure 2). Also, when the rate is heterogeneous among sites, PhyML performs the best and MP the worst, overall (data not shown). When $\kappa \geq 10$ all methods perform equally well, with high level of accuracy. These results are for Balanced A topology (Figure 1A) in particular. Comparison among the model tree topologies shows no difference in the trend observed for kappa (see Figure 2A-C) except for the pectinate tree (Figure 2D).

We also determined whether the effect of kappa is limited to the specifics of the model tree used in the study. Therefore, we examined factors such as the number of taxa and ultrametric vs. non-ultrametric trees to confirm the consistency in the effect of kappa on the phylogenetic accuracy. The results were similar (not shown).

Here, it seems that high accuracy at high kappa is a result of the phylogenetic signal being present in the transversional substitutions. At high kappa, relying only on transitions results in an incorrect phylogeny and at low kappa both transitions and transversions contribute to incorrect phylogeny, while an inference done using only the transversion events at high kappa almost always produced an accurate tree. This is because of saturation of transitions and transversions at low kappa, (when the rate is

high), and saturation of transition events relative to other mutations at high kappa.

Application to Phylogenetic Tree Reconstruction using Real Data

DNA sequences comprising vertebrate mitochondrial (mtDNA) COI sequences from *Masturus lanceolatus* (sunfish), *Homo sapiens* (human), *Bos taurus* (cow), *Balaenoptera musculus* (blue whale), *Pongo pygmaeus* (Bornean orangutan), *Pan troglodytes* (chimpanzee), *Gallus gallus* (chicken), and *Alligator mississippiensis* (American alligator) were obtained based on Xia (2000). As in all protein-coding gene sequences, the third codon position is the most variable and the second is the most conserved. The substitution saturation test of Xia *et al.* (2003) for the first, second and third codon positions of the mtDNA sequences indicated that the third codon positions had experienced substitution saturation and as a consequence, were not likely to be useful for phylogenetic inference. The third codon positions in vertebrate mitochondrial COI sequence evolves extremely fast, and exhibits a high kappa, κ ~ 50 and alpha, α = 0.70 (moderate rate homogeneity). Despite the signs of saturation in the third codon position (Xia and Xie 2001; Xia *et al.* 2003), phylogenetic analysis with this data resulted in a tree that was almost congruent to the (first + second) codon positions that are believed to be conserved and expected to produce an accurate tree (Figure 4 a&b). When sites with transitions and transversions were analyzed separately for the third codon position, the two trees were significantly different from each other. In fact, the tree inferred from the transitions-only sites of the third codon positions (Figure 4c) is absurd, whereas the phylogenetic tree with the transversions-only sites of the third codon positions (Figure 4d) showed exactly the same topology as in Figures 4a and 4b. The transversions-only sites of the third codon position also resulted in the correct grouping of two taxa, *Bos taurus* and *Balaenoptera musculus* which are not grouped in the tree inferred from all sites of the third codon position (Figure 4b). Analyzing the transitions and transversions separately in this manner shows that transversions contain stronger phylogenetic signal than transitions, and are capable of masking the distorted signals coming from the saturated transitions sites that may be misleading. This however, may be applicable only under certain conditions (as in this simulation based study or the empirical example used), for instance, high transition to transversion rate ratio, increased evolutionary rate, less rate heterogeneity among sites. Thus, even though a sequence (non-coding or coding) or subsets of a sequence have undergone substantial substitution saturation, it does not mean that it cannot be used for phylogenetic inference.. The phylogenetic signal is likely to be present in the sites that have experienced transversions, which may be useful in the inference of an accurate phylogeny.

The results in this study show that substitution saturation itself may not be a problem, as it may be dealt with in terms of whether it is the transition or transversions sites that have undergone saturation. We have shown with simulated and empirical datasets that a correct phylogeny can be obtained from saturated datasets when there is saturation among transversions in the sequences.

These results also emphasize the importance of the substitution rate in impacting the accuracy of phylogenetic inference. A low substitution rate usually ensures accurate phylogenetic inference, irrespective of the values of the other parameters used in this study. The results also establish optimal values of parameter combinations for accurate phylogenetic inference, under the different simulation conditions employed. It is expected that these results will be useful in studies that do phylogenetic analyses with empirical DNA sequences.

Conclusions

This study has established optimal values of five sequence parameters, singly and in combination, for improving the accuracy of phylogenetic inference, under varying conditions of model tree topology and inference method. An important conclusion of this study is that substitution saturation need not render a dataset unsuitable for phylogenetic analysis. In addition, the results here suggest values of the transition-transversion rate ratio and the evolutionary rate that result in saturation of the signal from the transitions in the data, but where the transversions still carry sufficient signal to offset the distorted signal from the transitions, so as to yield accurately inferred trees.

Acknowledgements

We thank Ohio Supercomputer Resources (OSC) for providing the computational resources for the bulk of the analysis in this study. This

research was supported by means of research funds from the University of Dayton: start-up funds for SRG and summer research funding from the Graduate School for BD.

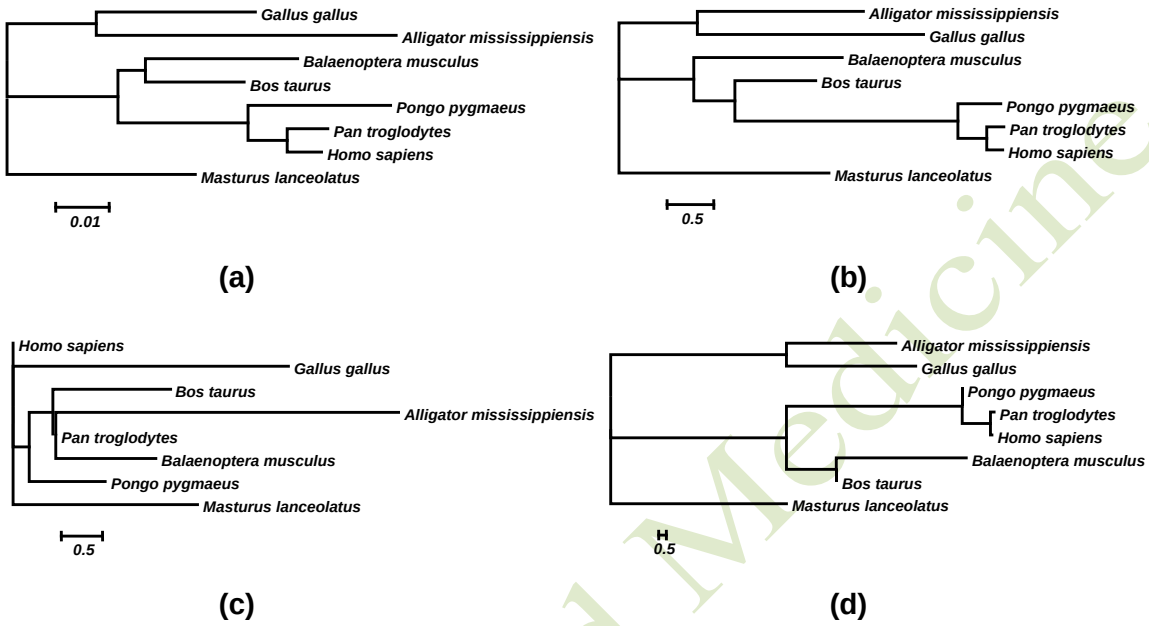

Figure 4. Phylogenetic tree reconstruction based on the vertebrate mitochondrial COI sequences. Maximum Likelihood tree were reconstructed using DNAML program (Felsenstein 1989) with default options, as implemented in DAMBE (Xia and Xie 2001) based on the vertebrate mitochondrial COI sequences using: (a) First and Second codon positions, (b) Third codon positions, (c) Transition-only sites of the third codon positions, and (d) Transversion-only sites of the third codon positions.

References

Cartwright, R. A. (2005). "DNA assembly with gaps (Dawg): simulating sequence evolution." Bioinformatics 21: 31-38.

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

Felsenstein, J. (2003). Inferring phylogenies. Sunderland, MA, Sinauer Associates Inc.

Gadagkar, S. R. and Kumar, S. (2005). "Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous." Molecular Biology and Evolution 22(11): 2139-2141.

Gadagkar, S. R., Rosenberg, M. S. and Kumar, S. (2005). "Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree." Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 304B(1): 64-74.

Guindon, S. and Gascuel, O. (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Systematic Biology 52(5): 696-704.

Halanych, K. M. and Robinson, T. J. (1999). "Multiple substitutions affect the phylogenetic utility of cytochrome b and 12S rDNA data: Examining a rapid radiation in Leporid (Lagomorpha) evolution." Journal of Molecular Evolution 48(3): 369-379.

Hall, B. G. (2005). "Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences (vol 22, pg 792, 2005)." Molecular Biology and Evolution 22(4): 1160-1160.

Hasegawa, M., Kishino, H. and Yano, T. A. (1985). "Dating of the Human Ape Splitting by a Molecular Clock of Mitochondrial-DNA." Journal of Molecular Evolution 22(2): 160-174.

Hillis, D. M. (1995). "Approaches for Assessing Phylogenetic Accuracy." Systematic Biology 44(1): 3-16.

Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. R. and Molineux, I. J. (1992). "Experimental Phylogenetics - Generation of a Known Phylogeny." Science 255(5044): 589-592.

Huelsenbeck, J. P. and Rannala, B. (2004). "Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models." Systematic Biology 53(6): 904-913.

Maddison, D. R., Schulz, K. S. and Maddison, W. P. (2007). "The Tree of Life Web Project." Zootaxa(1668): 19-40.

Mardis, E. R. (2008). "Next-generation DNA sequencing methods." Annual Review of Genomics and Human Genetics 9: 387-402.

Nei, M. (1996). "Phylogenetic analysis in molecular evolutionary genetics." Annual Review of Genetics 30: 371-403.

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

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

Nichols, R. (2001). "Gene trees and species trees are not the same." Trends in Ecology and Evolution 16(7): 358-364.

Ogden, T. H. and Rosenberg, M. S. (2006). "Multiple sequence alignment accuracy and phylogenetic inference." Systematic Biology 55(2): 314-328.

Ou, C. Y., Ciesielski, C. A., Myers, G., Bandea, C. I., Luo, C. C., Korber, B. T. M., Mullins, J. I., Schochetman, G., Berkelman, R. L., Economou, A. N., Witte, J. J., Furman, L. J., Satten, G. A., Macinnes, K. A., Curran, J. W. and Jaffe, H. W. (1992). "Molecular Epidemiology of HIV Transmission in a Dental Practice." Science 256(5060): 1165-1171.

Penny, D. and Hendy, M. D. (1985). "The Use of Tree Comparison Metrics." Systematic Zoology 34(1): 75-82.

Raghava, G. P. S., Searle, S. M. J., Audley, P. C., Barber, J. D. and Barton, G. J. (2003). "OXBench: A benchmark for evaluation of protein multiple sequence alignment accuracy." BMC Bioinformatics 4.

Regier, J. C. and Shultz, J. W. (1997). "Molecular phylogeny of the major arthropod groups indicates polyphyly of Crustaceans and a new hypothesis for the origin of hexapods." Molecular Biology and Evolution 14(9): 902-913.

Robinson, D. F. and Foulds, L. R. (1981). "Comparison of Phylogenetic Trees." Mathematical Biosciences 53(1-2): 131-147.

Rosenberg, M. S. (2005a). "Evolutionary distance estimation and fidelity of pair wise sequence alignment." Bmc Bioinformatics 6.

Rosenberg, M. S. (2005b). "Multiple sequence alignment accuracy and evolutionary distance estimation." Bmc Bioinformatics 6.

Rosenberg, M. S. and Kumar, S. (2003). "Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference." Molecular Biology and Evolution 20(4): 610-621.

Saitou, N. and Ueda, S. (1994). "Evolutionary Rates of Insertion and Deletion in Noncoding Nucleotide-Sequences of Primates." Molecular Biology and Evolution 11(3): 504-512.

Singh, J., Behal, A., Singla, N., Joshi, A., Birbian, N., Singh, S., Bali, V. and Batra, N. (2009). "Metagenomics: Concept, methodology, ecological inference and recent advances." Biotechnology Journal.

Struck, T., Hessling, R. and Purschke, G. (2002). "The phylogenetic position of the Aeolosomatidae and Parergodrilidae, two enigmatic oligochaete-like taxa of the 'Polychaeta', based on molecular data from 18S rDNA sequences." Journal of Zoological Systematics and Evolutionary Research 40(3): 155-163.

Struck, T. H., Nesnidal, M. P., Purschke, G. and Halanych, K. M. (2008). "Detecting possibly saturated positions in 18S and 28S sequences and their influence on phylogenetic reconstruction of Annelida (Lophotrochozoa)." Molecular Phylogenetics and Evolution 48(2): 628-645.

Struck, T. H., Schult, N., Kusen, T., Hickman, E., Bleidorn, C., McHugh, D. and Halanych, K. M. (2007). "Annelid phylogeny and the status of Sipuncula and Echiura." BMC Evolutionary Biology 7.

Swofford, D. L. (2003). PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.0b10. Sunderland, MA, Sinauer Associates Inc.

Takahashi, K. and Nei, M. (2000). "Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of

sequences are used." Molecular Biology and Evolution 17(8): 1251-1258.

Tamura, K., Nei, M. and Kumar, S. (2004). "Prospects for inferring very large phylogenies by using the neighbor-joining method." Proceedings of the National Academy of Sciences of the United States of America 101(30): 11030-11035.

Wakeley, J. (1996). "The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance." TREE 11: 158-163.

Xia, X. (2000). Chapter 20-Data Analysis in Molecular Biology and Evolution. Boston, Kluwer Academic.

Xia, X. (2000). Data Analysis in Molecular Biology and Evolution. Boston, Kluwer Academic Publishers.

Xia, X. and Xie, Z. (2001). "DAMBE: Software package for data analysis in molecular biology and evolution." Journal of Heredity 92(4): 371-373.

Xia, X., Xie, Z., Salemi, M., Chen, L. and Wang, Y. (2003). "An index of substitution saturation and its application." Molecular Phylogenetics and Evolution 26: 1-7.

Xia, X. H. (1998). "The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes." Molecular Biology and Evolution 15(3): 336-344.

Xia, X. H., Xie, Z., Salemi, M., Chen, L. and Wang, Y. (2003). "An index of substitution saturation and its application." Molecular Phylogenetics and Evolution 26(1): 1-7.

Yang, Z. H. (1996a). "Maximum-likelihood models for combined analyses of multiple sequence data." Journal of Molecular Evolution 42(5): 587-596.

Yang, Z. H. (1996b). "Phylogenetic analysis using parsimony and likelihood methods." Journal of Molecular Evolution 42(2): 294-307.

Yang, Z. H. (1996c). "Among-site rate variation and its impact on phylogenetic analyses." Trends in Ecology & Evolution 11(9): 367-372.

Yang, Z. H. (1998). "On the best evolutionary rate for phylogenetic analysis." Systematic Biology 47(1): 125-133.

Yang, Z. H. and Yoder, A. D. (1999). "Estimation of the Transition/Transversion Rate Bias and Species Sampling." Journal of Molecular Evolution 48: 274-283.

Yoder, A. D., Vilgalys, R. D. and Ruvolo, M. (1996). "Molecular evolutionary dynamics of cytochrome *b* in strepsirrhine primates: The phylogenetic significance of third codon position transversions." Molecular Biology and Evolution 13: 1339-1350.