

How many alignment gaps are too many for accurate phylogenetic inference? A simulation study

Bhakti Dwivedi¹ and Sudhindra R. Gadagkar^{1*}

¹ Department of Biology, University of Dayton, 300 College Park Ave., Dayton, OH 45469-2320

Molecular sequences evolve by substitution and insertion/deletion (indel) events. While substitutions change the composition of the sequences, indels increase (decrease) the length of a given sequence compared to its ortholog in another species. When the sequences for a given gene from multiple species are to be used in comparative sequence analysis, these indels need to be reconstructed to ensure an accurate alignment of the sequences. A reconstruction of the evolutionary relationships among the species (phylogenetic inference) requires such an alignment of the sequences since the relationships are determined by comparing the state at each site among the sequences. Clearly, the presence of a gap in a given sequence (representing an indel) will render such a comparison difficult. Yet, gaps are an integral part of a sequence alignment. Therefore, this project was undertaken to determine the effects of the type and number of alignment gaps on phylogenetic accuracy. A computer program was used to simulate evolution along a model tree (a published 66-species mammalian tree) for many genes. The simulation process was begun with a random sequence, which was then evolved along the model tree using only substitutions, and the resulting 66 terminal sequences were obtained. Two types of gaps were then introduced into these sequences to simulate indels: (1) at individual sites scattered randomly across the entire alignment (“spotty” gaps), and (2) at multiple contiguous sites (“chunky” gaps) in one randomly selected sequence in the alignment. The total number of sites with gaps ranged from 10% to 90% of the gene length. Phylogenetic analysis was done using the Neighbor-Joining (NJ) method on the following datasets: “OS” – the original sequences with no gaps, “IS” – the same sequences with gaps but with the most likely state at each gap inferred using the parsimony principle, “PD” – sites deleted in pair-wise sequence comparisons during phylogenetic analysis only if one or both sequences showed a gap at that site, and “CD” – sites deleted completely from the entire alignment even if a single sequence contained a gap at that site. Phylogenetic accuracy was determined by the number of incorrectly inferred branches as compared to the model tree. The results show that when gaps were present (spotty or chunky), as expected, longer gene sequences were affected less than shorter sequences in terms of phylogenetic accuracy, but only in CD analyses, and with phylogenetic accuracy significantly reduced only after the number of gaps exceeded ~25% of the gene length. A surprising result was that the deletion of sites with gaps in PD analyses had little impact in reducing the accuracy of phylogenetic inference when compared to the corresponding IS or OS analyses, even when the gaps constituted up to 90% of the gene length. Furthermore, this was true whether the gaps were spotty or chunky. These results suggest that for accurate phylogenetic inference using NJ, gaps in sequence alignments are best handled by inferring the state based on the parsimony principle or by pair-wise deletion of the affected sites, irrespective of the number of gaps in the alignment.

* Corresponding author: gadagkar@notes.udayton.edu