J Mol Evol (2002) 54:396–402
DOI: 10.1007/s00239-001-0034-9

JOURNAL OF MOLECULAR EVOLUTION

The Effect of Recombination on the Accuracy of Phylogeny Estimation

David Posada, Keith A. Crandall

Department of Zoology, Brigham Young University, Provo, UT 84602, USA

Received: 12 March 2001 / Accepted: 28 September 2001

Abstract. Phylogenetic studies based on DNA sequences typically ignore the potential occurrence of recombination, which may produce different alignment regions with different evolutionary histories. Traditional phylogenetic methods assume that a single history underlies the data. If recombination is present, can we expect the inferred phylogeny to represent any of the underlying evolutionary histories? We examined this question by applying traditional phylogenetic reconstruction methods to simulated recombinant sequence alignments. The effect of recombination on phylogeny estimation depended on the relatedness of the sequences involved in the recombinational event and on the extent of the different regions with different phylogenetic histories. Given the topologies examined here, when the recombinational event was ancient, or when recombination occurred between closely related taxa, one of the two phylogenies underlying the data was generally inferred. In this scenario, the evolutionary history corresponding to the *majority* of the positions in the alignment was generally recovered. Very different results were obtained when recombination occurred recently among divergent taxa. In this case, when the recombinational breakpoint divided the alignment in two regions of similar length, a phylogeny that was different from any of the true phylogenies underlying the data was inferred.

Key words: Recombination — Reticulate evolution — Mosaic genes — Phylogeny estimation — Accuracy — Phylogenetic simulations

Correspondence to: David Posada, Variagenics, 60 Hampshire Street, Cambridge, MA 02139, USA; *email:* dposada@variagenics.com

Introduction

While there are many examples of the use of phylogenies, these applications often rely on accurate estimates of phylogenetic relationships. Traditional methods of phylogeny estimation, such as maximum parsimony (MP), minimum evolution (ME), and maximum likelihood (ML), assume that a single evolutionary history underlies the sample of sequences under study. However, different regions of an alignment can have different evolutionary histories due to processes such as crossing-over, gene conversion, horizontal transfer, and hybridization (hereafter generally called recombination) (Sneath 1975). In those studies that have explored the possibility of recombination, it has been found to have a significant impact on the conclusions (e.g., Drouin et al. 1999; Holmes et al. 1999; Robertson et al. 1995; Sanderson and Doyle 1992; Zhou et al. 1997). In practice, recombination is ignored and its possible consequences neglected.

Only a few studies have dealt with the effects of recombination in a phylogenetic context. Wiens (1998) carried out computer simulations to understand the effect of combining data sets with different histories (gene trees) when the goal is to estimate the species tree (i.e., only one of the gene trees is correct). He concluded that the combined analysis of genes with different histories might diminish the chances of recovering the species tree. Recently, Schierup and Hein (2000) characterized some of the consequences of ignoring recombination when using phylogenies to make demographical, chronological, or substitutional inferences. However, an interesting and largely unaddressed question is how recombination affects the "accuracy" of phylogenetic

inference. Intuition suggests that recombination will confound methods of phylogeny estimation, but in what fashion? Recombination could lead to the estimation of trees that do not reflect any of the true histories. Alternatively, phylogenetic methods might simply find the most frequent history in the alignment.

Here we performed computer simulations to characterize the effect of ignoring the presence of recombination on the "accuracy" of phylogenetic reconstruction.

Methods

The methodology proposed proceeds by the following steps.

1. Simulate recombinant data sets under two model trees (or one model tree in the case of no recombination).
2. Apply traditional methods of phylogeny estimation, which assume a single evolutionary history, to the recombinant data sets.
3. Compare the estimated phylogeny with the two model trees using two criteria: recovery of the model trees and percentage of clades in the models trees recovered.

Simulation of Recombination

There are at least two general approaches to simulating recombinant sequence alignments. The first strategy is a time-forward approach in which a population of chromosomes is evolved from the past to the present by introducing a series of recombination events and mutations each generation. However, the phylogeny of the sample is not known until the simulation is finished, and therefore this approach does not allow for the use of particular phylogenies or for the positioning of the recombination event in a specific place in the phylogeny.

Another general strategy to simulate recombinant sequence alignments is the genealogical or phylogenetic simulations. In this approach, each site in a sample of DNA sequences is evolved upon a phylogeny that can change in different regions of the alignment. Note that the phylogenetic approach attempts to simulate the result, but not the process, of recombination in terms of a sequence alignment. The set of phylogenies underlying the recombinant alignment can be generated at random using the coalescent with recombination (Hudson 1983). However, in the coalescent process it is difficult to have control over the exact position of the recombination event in the history of the sample or the shape and branch lengths of the phylogenetic tree(s).

Here we used an alternative phylogenetic approach to the coalescent with recombination, using fixed arbitrary topologies for different regions, allowing control of when, where, and between which sequences the recombination events happen (Grassly and Holmes 1997; McGuire et al. 1997; Worobey 2001; Worobey et al. 1999). To simulate a recombinant alignment, a breakpoint partitioning the alignment in two regions (left region and right region) was arbitrarily selected. Nucleotides at each side of the breakpoint were evolved under two model trees (left-side tree and right-side tree) (Fig. 1). For example, a 75% recombinational breakpoint in a 1000-character alignment implies that the left region, including sites 1–750, was evolved on the left-side tree, while the right region, including sites 751–1000, was evolved on the right-side tree. It should be noted that the actual location of the sites evolved under the left-side or under the right-side tree, which in turn determines the number of physical breakpoints along the alignment, does not matter, as phylogenetic methods assume independence of sites. Four arbitrary breakpoints were simulated, 50, 75, 90, and 100%. Note that a 100% breakpoint implies no recombination.

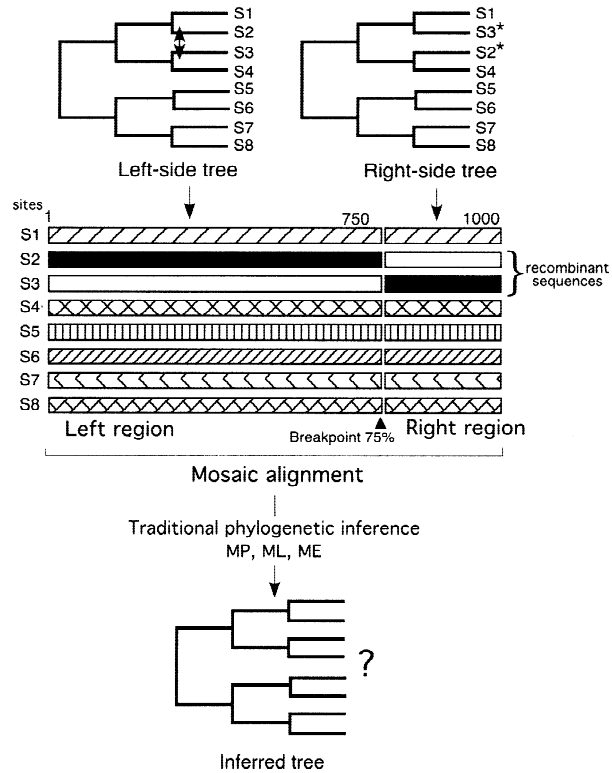

Fig. 1. Simulation of recombination and phylogenetic inference from mosaic alignments. To simulate alignments that are mosaics of two histories as a result of a recombinational event, different regions of the alignment (left region and right region) were evolved under different tree topologies (left-side tree and right-side tree). The boundary between the left region and the right region defines the recombinational breakpoint. Here a reciprocal exchange is represented, but nonreciprocal exchanges were also performed. Traditional phylogenetic inference was performed from the mosaic alignments ignoring the presence of recombination (i.e., assuming that there is only one tree underlying the data).

The topological differences between the left-side tree and the right-side tree defined the phylogenetic position of the recombinational event: recent or ancient, among closely related or distant taxa, and whether the recombinational exchange was reciprocal or nonreciprocal (Fig. 2). For each set of parameters, 100 eight-taxon sets of aligned DNA sequences were evolved according to the HKY model of nucleotide substitution (Hasegawa et al. 1985) on each of the two model trees. The base frequencies used were arbitrarily set to 0.1, 0.2, 0.3, and 0.4, for A, C, G, and T, respectively, and the transition/transversion ratio was set to 2. The number of sites simulated was 100, 500, 1000, 3000, and 5000. Sequences were evolved under two substitution rates (expected number of substitutions per nucleotide from the root to the tip of the tree), 0.6 and 0.3. Three tree shapes were also explored: unbalanced, intermediate, and balanced. The nonreciprocal simulations were designed after the results from the reciprocal simulations were obtained. Given that in the reciprocal simulations the number of sites did not influence the results, only 100, 500, and 1000 sites were used in the nonreciprocal simulations.

Phylogeny Estimation

Phylogenetic trees were estimated from the whole alignments, and therefore, ignoring the presence of recombination. MP, ME, and ML

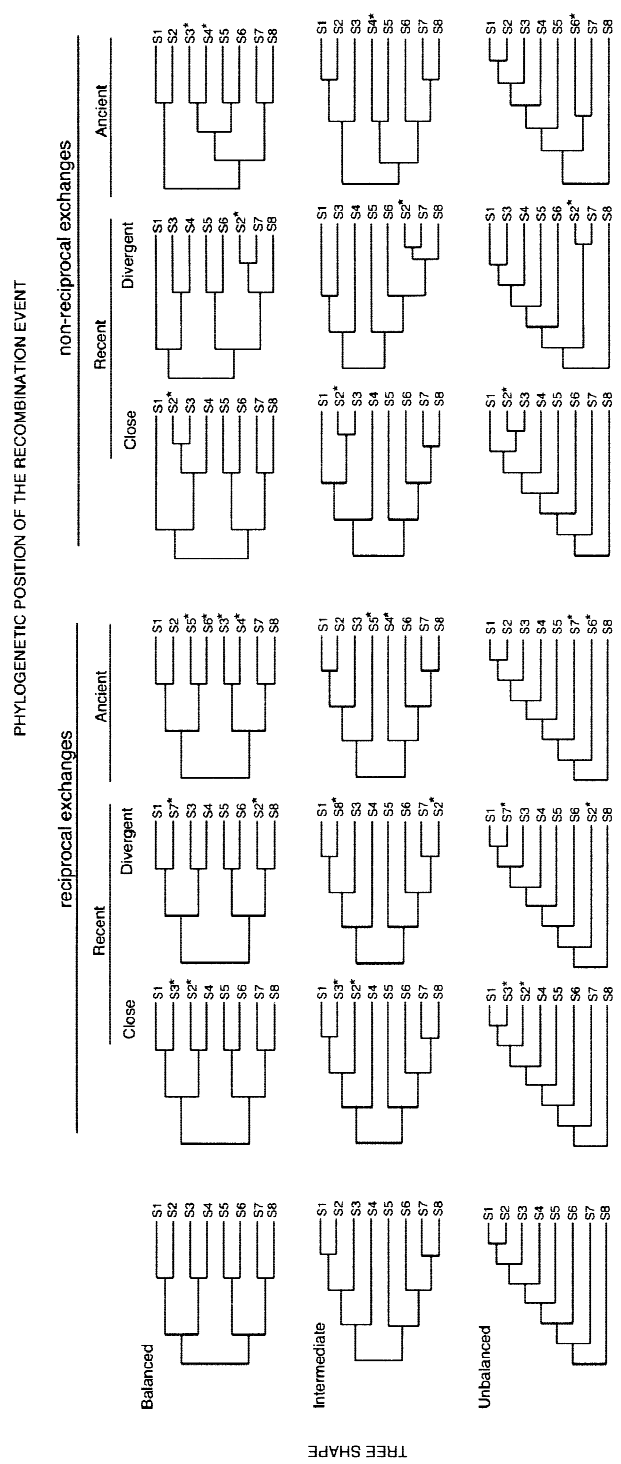

Fig. 2. Topological definition of the recombinational events. The difference in topology between the left-side tree and the right-side tree defines the phylogenetic position of the recombinational event and whether the exchange of sequence fragment was reciprocal or nonreciprocal. Different right-side topologies were used to imitate a recent recombination between close or divergent taxa or an ancient recombination. Different tree shapes—balanced, intermediate, and unbalanced—were also investigated. *Asterisks* denote recombinant taxa.

searches were performed, and the best tree found under each criterion was recorded. For the ME and ML estimation, analyses were conducted using the true model, HKY; a simpler—and wrong—model, Jukes and Cantor (1969) (JC); and an unnecessarily general, overparameterized model, the general time-reversible (GTR) model (Tavaré 1986). PAUP* (Swofford 1998) was used for all analyses. All of the phylogenetic reconstruction methods and models of nucleotide substitution used here are described by Swofford et al. (1996).

Phylogeny Reconstruction Evaluation

Each recombinant alignment was simulated under two trees. Clearly, no single tree can therefore be considered to be an accurate reflection of the true evolutionary history. In this study there are two model trees for each data set, a left-side tree and a right-side tree (see Fig. 1). We therefore recorded the number of times the inferred tree matched either the left- or the right-side tree, as well as the proportion of clades in the

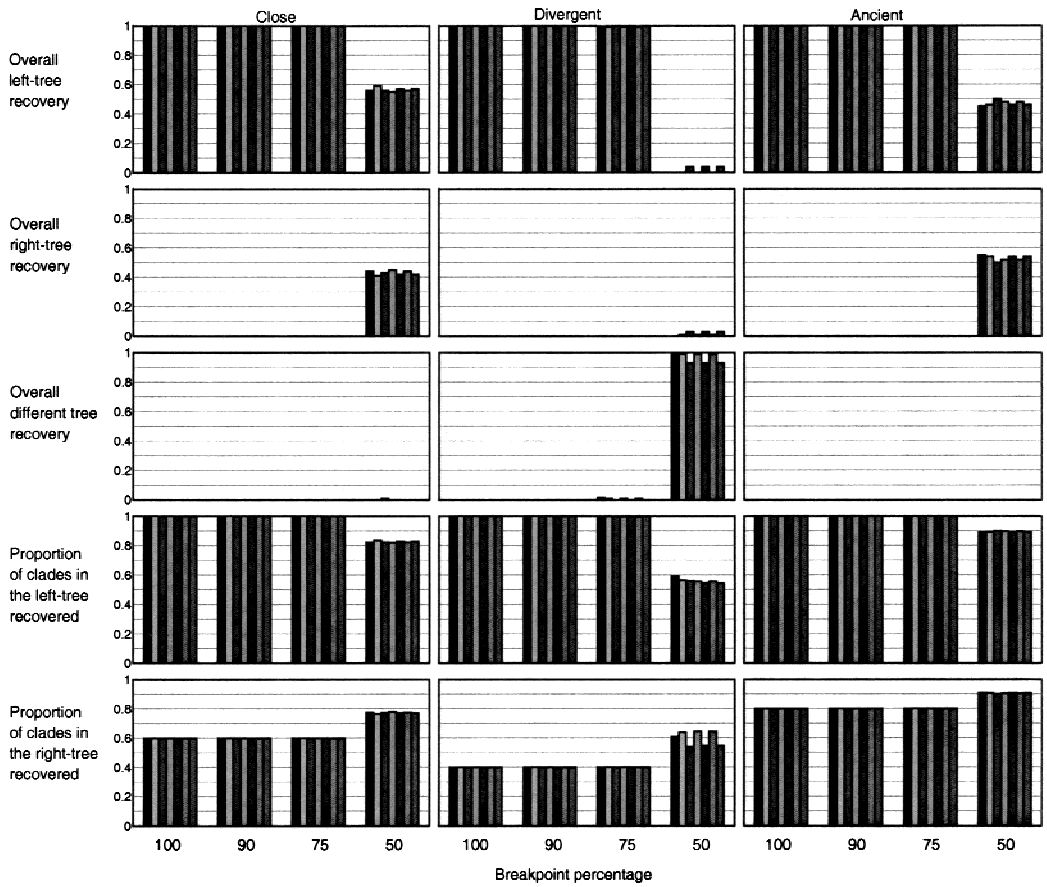

Fig. 3. Overall topology and clade recovery from balanced trees when the exchanges are reciprocal. The probability of recovering the left-side tree, the right side-tree, a different tree, a clade in the left-side tree, and a clade in the right-side tree is plotted for 1000-bp sequences. Each bar is the average of 100 replicates. The breakpoint percentage indicates the fraction of the alignment evolved under the left-side tree (i.e., the relative size of the left region) (see Fig. 1). The phylogenetic criteria and substitution models used were MP (■), ML-JC (▩), ME-JC (▩), ML-HKY (▩), ME-HKY (▩), ML-GTR (▩), and ME-CTR (□).

inferred tree present in the left- or right-side tree. This allows for a direct comparison of two intuitively alternative results: recovery of the "highest proportion" tree versus selection of a tree unrelated to either model tree.

Results

Results were very consistent across the different lengths of the alignments and mutation rates. Different tree shapes or phylogenetic methods gave different quantitative results in some cases. Figures 3 and 4 represent typical phylogenetic estimation outcomes for the reciprocal and nonreciprocal recombination events, respectively. Exceptions to the patterns in these figures are emphasized in the text. Basically, three situations were observed.

(a) The Topology Recovered Was Always (Or Almost Always) One of the Model Trees. This result was obtained with 90 or 75% breakpoints across all conditions with a few exceptions (see b, below). In this case, the exact model tree recovered corresponded to the one responsible for most of the sites in the alignment (i.e., the left-side tree). This result was also obtained with 50% breakpoints when the recombinational events were ancient, or when the recombinational events occurred between closely related sequences (except for nonreciprocal exchanges in balanced trees; see b). For the 50%-breakpoints, either each model tree was recovered half of the time (all reciprocal exchanges, and close nonreciprocal exchanges in unbalanced trees), or the left-side tree was recovered most of the time (ancient nonreciprocal exchanges in balanced or intermediate tree trees), or the right-side tree was recovered most of the time (close nonreciprocal exchanges in intermediate trees, or ancient nonreciprocal exchanges in unbalanced trees).

(b) The Topology Recovered Was In Some Cases Slightly Different From Any of the Model Trees. This result was observed in some 75% breakpoints when recombination occurred between divergent sequences. This applied to all phylogenetic methods for reciprocal exchanges in unbalanced trees (20–30% of the time), but

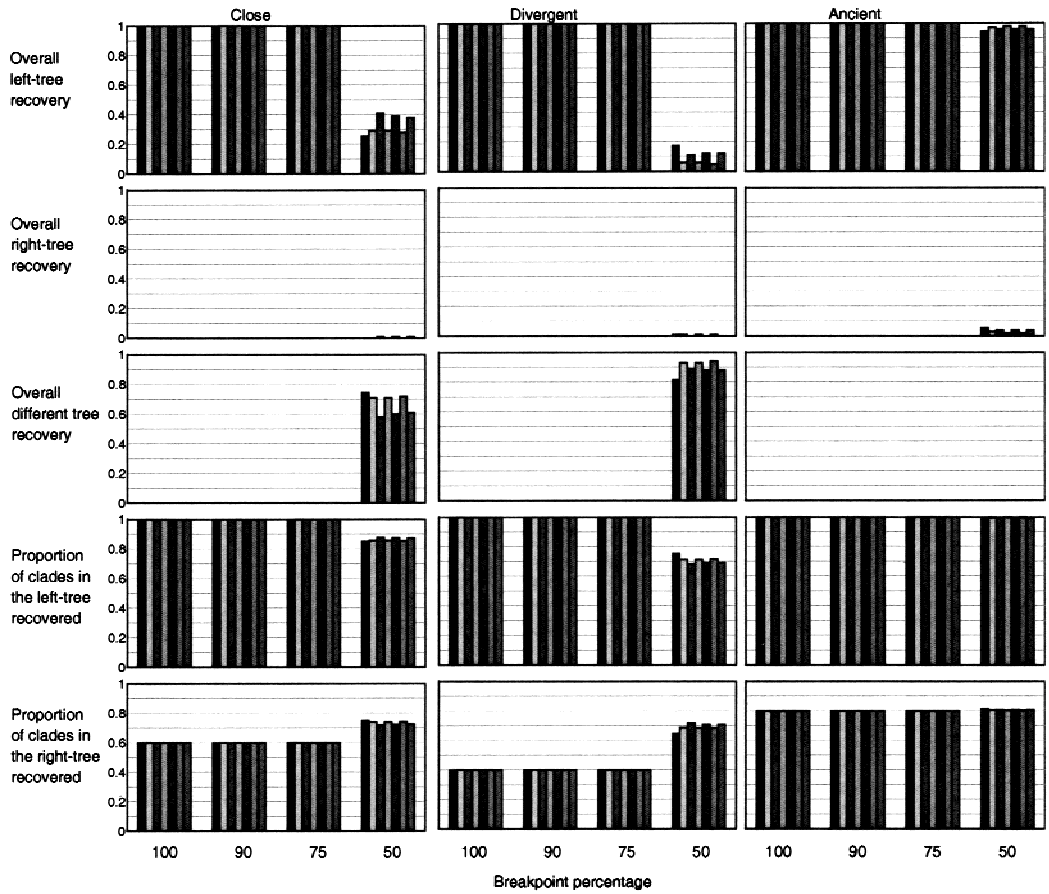

Fig. 4. Overall topology and clade recovery from balanced trees when the exchanges are nonreciprocal. The probability of recovering the left-side tree, the right side-tree, a different tree, a clade in the left-side tree, and a clade in the right-side tree is plotted for 1000-bp sequences. Each bar is the average of 100 replicates. The breakpoint percentage indicates the fraction of the alignment evolved under the left-side tree (i.e., the relative size of the left region) (see Fig. 1). The phylogenetic criteria and substitution models used were MP (■), ML-JC (▒), ME-JC (▒), ML-HKY (▒), ME-HKY (▒), ML-GTR (▒), and ME-GTR (□).

only for ME for reciprocal exchanges in intermediate trees (25% of the time) and for nonreciprocal exchanges in intermediate or unbalanced trees (65–70% of the time). With 50% breakpoints, this outcome was observed only when the exchanges were nonreciprocal, for close events in balanced trees (around 70%—ME, only 60%—of the time), and for divergent events in unbalanced trees (around 60% of the time). In all cases, the tree inferred was similar to one of the model trees, recovering 80–95% of their clades.

(c) The Topology Recovered Was Always (Or Almost Always) Very Different From Any of the Model Trees. This result was observed only with 50% breakpoints in recombinational events involving divergent sequences. For reciprocal exchanges this was observed for all tree shapes. The trees obtained were very different from any of the model trees, recovering only 30% (intermediate or unbalanced trees; 15% for ME for intermediate trees-) or 55–60% (balanced trees) of the model tree clades. For nonreciprocal exchanges, this outcome was observed only for balanced and intermediate trees. The trees obtained were less different from any of the model trees than in the reciprocal case, recovering only 70% (intermediate trees) or 40–60% (intermediate trees) of their clades.

In general, different tree shapes did not have a significant effect on whole phylogeny or individual clade recovery when the exchanges were reciprocal, but they did in some cases when the exchanges were nonreciprocal (see above). Increasing the number of characters did not improve the ability of the different phylogenetic methods to recover one of the model trees in the presence of recombination relative to the absence of recombination. Indeed, recovering the model trees was more difficult with fewer characters (100 characters), in both the presence and the absence of recombination. The different mutation rates used in the simulations did not have an overall effect on the ability to recover the model trees. The impact of recombination was similar for all phylogenetic optimality criteria, but in some specific cases (outlined above) the ME criterion inferred a different tree

when the other criteria recovered one of the model trees. MP recovered the model trees more often than the other criteria with small data sets (100 characters). The model of nucleotide substitution implemented did not seem to have an effect on any result.

Discussion

Clearly, these results pertain to the particular topologies used in the simulations. Parameter space could have been more thoroughly explored by generating multiple random recombinant genealogies, for example, using the coalescent with recombination (see above). However, in such case it would be much more complex to track simultaneously the combined effect of the number of true histories in the data, different tree shapes, number of recombination events, phylogenetic position of the recombinational events, and breakpoint percentages. By using fixed model topologies, with one recombination event defining two true histories, we simplified the problem to allow for precise control over these variables. Our objective was not to evaluate all possible scenarios but, rather, to answer formally the question of whether recombination may confound phylogenetic estimation and to explore some situations under which this effect might vary.

The mechanics of these simulations do not mimic "conventional mechanisms of recombination," which would be more closely resembled by a classical forward simulation. That was never the intention. What matters is that the result of both forward and genealogical approaches is exactly the same, a mosaic sample (alignment) of sequences where different sites can have different phylogenetic histories, and that such mosaic alignments resemble very well a real sample of sequences where homologous exchange has occurred. Indeed, phylogenetic incongruence is a natural way to look for recombination (Sneath et al. 1975).

In consequence, there are many biological scenarios under which this study might be relevant. For example, considering that no gene boundary is defined in the simulations, one could think of the alignments produced here as products of homologous intragenic recombination. In such a case, the nonreciprocal exchanges imitate the product of eukaryote crossing-over when both recombinant products, but not the parents, remain in the sample. While this situation might be uncommon, but possible in nature, it is the worst-case scenario. The samples produced with nonreciprocal exchanges might be similar to those produced in nature by gene conversion, by crossing-over when only one recombinant product remains in the sample, or by other nonreciprocal exchanging mechanisms typical of virus and bacteria. Again, because no gene boundaries are defined, one could think of the samples simulated here as combined alignments of two genes, delimited by the recombinational breakpoint, with two histories, i.e., intergenic recombination. Such data sets are likely to be produced by biological mechanisms such as horizontal gene transfer. Furthermore, we could also consider that one of the genes (the left region) represents the species tree, and the other gene (the right region) represents a different history resulting from a recombinational event like those already mentioned, but also from hybridization, lineage sorting, or gene duplication followed by loss (Maddison 1997). In this scenario, the recovery of the left-side tree would indicate how often the species tree is inferred from combined data sets with mixed phylogenetic signals (Wiens 1998). Because each nucleotide position is considered independent in phylogenetic estimation, the "gene boundaries" are irrelevant to our specific question of how recombination influences our ability to reconstruct an evolutionary history accurately.

Indeed simulation studies often make simplifying assumptions. A small, constant, number of taxa was used in our studies. Likewise, sequence evolution was simplistic in that a molecular clock was assumed and a simple stochastic model of nucleotide substitution was used that clearly does not capture the full complexity of sequence evolution (e.g., codon position, structural constraints, etc.).

There are two intuitive predictions regarding the impact of recombination on phylogeny reconstruction. The first prediction suggests that recombination may confound phylogenetic inference. The second prediction is that the effect of a single recombination event on phylogeny reconstruction is often mild in the sense that the tree responsible for the majority of the alignment is recovered. Here we have characterized some situations in which these predictions hold true. In these simulations the confounding effect of recombination was evident when the taxa involved were divergent and the recombinational breakpoint divided the sequences in half, especially when the exchange was reciprocal. On the other hand, the most common history was generally recovered when recombination events were ancient or involved closely related sequences (with one exception). There were also cases where the recombinational breakpoint divided the sequences in half and one of the model trees was recovered most of the time, probably because of inherent phylogenetic methods bias towards certain topologies. Of course, an accurate history of the mosaic sequences cannot be estimated by traditional phylogenetic methods because the true history would represent the mosaic relationships in a way that cannot be represented by a single nonreticulate tree. Instead, network approaches to estimating genealogical relationships are a sensible alternative to traditional methods when recombination is suspected or identified among sequences in an alignment (for a review see Posada and Crandall 2001).

That recombination affects phylogenetic inference is a popular concept. Surprisingly, no study has been explicitly aimed to test this idea formally or to characterize such an effect (but see Wiens 1998). This study shows that in some cases a phylogeny obtained from recombinant data is very different from any of the true histories underlying the data. Inferences based on such phylogenies might be very misleading (see also Schierup and Hein 2000). Some caution might be required in cases where recombination is likely to occur. Such data sets might be scanned for the presence of recombination prior to phylogenetic analysis. There are many methods to detect recombination from DNA sequences, and their performance has only recently been evaluated (Brown et al. 2001; Wiuf et al. 2001; Posada, submitted; Posada and Crandall 2002).

Acknowledgments. David Swofford provided essential assistance with the simulations. We want to thank David Hillis, Jack Sites, Mike Whiting, Marcos Pérez-Losada, and Andrew Rambaut for reading and commenting on the manuscript. Martin Kreitman and two anonymous reviewers provided very helpful suggestions. This work was supported by a BYU Graduate Studies Award (D.P.), NSF DEB 9974124 (D.L.S.), NSF DEB 0073154, NIH Grant RO1-HD33982, and the Alfred P. Sloan Foundation (K.A.C.).

References

Brown CJ, Garner EC, Dunker KA, Joyce P (2001) The power to detect recombination using the coalescent. Mol Biol Evol 18:1421–1424

Drouin G, Prat F, Ell M, Paul Clark GD (1999) Detecting and characterizing gene conversions between multigene family members. Mol Biol Evol 16:1639–1390

Grassly NC, Holmes EC (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol Biol Evol 14:239–247

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

Holmes EC, Urwin R, Maiden MCJ (1999) The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis.* Mol Biol Evol 16:741–749

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23:183–201

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HM (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–132

Maddison W (1997) Gene trees in species trees. Syst Biol 46:523–536

McGuire G, Wright F, Prentice MJ (1997) A graphical method for detecting recombination in phylogenetic data sets. Mol Biol Evol 14:1125–1131

Posada D, Crandall KA (2001) Intraspecific gene genealogies: Trees grafting into networks. Trends Ecol Evol 16:37–45

Posada D, Crandall KA (2002) Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc Nat Acad Sci USA (in press)

Robertson DL, Sharp PM, McCutchan FE, Hahn BH (1995) Recombination in HIV-1. Nature 374:124–126

Sanderson MJ, Doyle JJ (1992) Reconstruction of organismal and gene phylogenies from data on multigene families: Concerted evolution, homoplasy, and confidence. Syst Biol 41:4–17

Schierup ME, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. Genetics 156:879–891

Sneath PHA (1975) Cladistic representation of reticulate evolution. Syst Zool 24:360–368

Sneath PHA, Sackin MJ, Ambler RP (1975) Detecting evolutionary incompatibilities from protein sequences. Syst Zool 24:311–322

Swofford DL (1998) PAUP* Phylogenetic analysis using parsimony and other methods. Sinauer Associates, Sunderland, MA

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic Inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematics. Sinauer Associates, Sunderland, MA, pp 407–514

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura RM (ed) Some mathematical questions in biology—DNA sequence analysis. Am Math Soc, Providence, RI, pp 57–86

Wiens JJ (1998) Combining data sets with different phylogenetic histories. Syst Biol 47:568–581

Wiuf C, Christensen T, Hein J (2001) A simulation study of the reliability of recombination detection methods. Mol Biol Evol 18:1929–1939

Worobey M (2001) A novel approach to detecting and measuring recombination: New insights into evolution in viruses, bacteria, and mitochondria. Mol Biol Evol 18:1425–1434

Worobey M, Rambaut A, Holmes EC (1999) Widespread intraserotype recombination in natural populations of dengue virus. Proc Natl Acad Sci USA 96:7352–7357

Zhou J, Bowler LD, Spratt BG (1997) Interspecies recombination, and phylogenetic distortions, within the glutamine synthetase and shikimate dehydrogenase genes of *Neisseria meningitidis* and commensal *Neisseria* species. Mol Microbiol 23:799–812