

## Letter to the Editor

### Simple (Wrong) Models for Complex Trees: A Case from Retroviridae

David Posada and Keith A. Crandall

Department of Zoology, Brigham Young University

Although the use of adequate models of evolution should improve the accuracy of phylogenetic inference (Leitner, Kumar, and Albert 1997; Sullivan and Swoford 1997; Cunningham, Zhu, and Hillis 1998), computer simulation studies have shown that under certain circumstances, wrong models can recover the true tree with higher probability than the tree model employed to generate the data (Saitou and Nei 1987; Schöniger and von Haeseler 1993; Tajima and Takezaki 1994; Tateno, Takezaki, and Nei 1994; Yang 1997a; Takahashi and Nei 2000). These results have previously been attributed to the complexity of the topology estimation problem (Yang 1997a), but Bruno and Halpern (1999) have suggested that they could be better understood in terms of bias toward the true tree when the assumptions of the method are violated; in other words, getting the right answer for the wrong reason.

There is no evidence of whether this “phylogenetic bias” actually occurs with real data. While there are empirical examples of “phylogenetic robustness” (complex models produce no better results than simpler ones) (Russo, Takezaki, and Nei 1996), we are not aware of cases in which the use of simpler models leads to a better phylogenetic estimate. We report here on an empirical example in which only the use of simpler, wrong models of evolution leads to the estimation of trees that are in agreement with biochemical and immunological evidence and with previous phylogenetic studies.

DNA and amino acid sequences for the *gag*, *pol*, and *env* genes for 61–76 retroviral taxa were downloaded from GenBank. ClustalX (Thompson et al. 1997) was used for the alignment of the DNA and protein sequences. The best-fit model of DNA substitution for each data set was selected using Modeltest, version 1.05 (Posada and Crandall 1998). Neighbor-joining (NJ) trees (Saitou and Nei 1987) were constructed using uncorrected (NJ<sub>p</sub>) and maximum-likelihood (NJ<sub>ml</sub>) distances for DNA and using mean character (=uncorrected) (NJ<sub>mc</sub>) and maximum-likelihood (NJ<sub>mlprot</sub>) distances for proteins. A heuristic search was carried out to find the most parsimonious (MP) tree(s). Unweighted (MP<sub>dna</sub> and MP<sub>prot</sub>) and weighted (MP<sub>step</sub>, MP<sub>pars</sub>, and MP<sub>rice</sub>) MP trees were estimated. Three step matrices were used for the weighting. The DNASTEP matrix was constructed in MacClade, version 3.07 (Maddison and Maddison 1994), on the NJ tree estimated using maximum-likelihood (ML) distances under the best-fitting model of nucleotide substitution

(MP<sub>step</sub>). For protein sequences, two amino acid substitution matrices were used: protpars (Felsenstein 1991) (MP<sub>pars</sub>) and one nonmetric and asymmetrical step matrix estimated using the accepted mutation-parsimony method (AMP program; Rice, personal communication) (MP<sub>rice</sub>). Heuristic searches were performed in PAUP\* (Swofford 1998). The program GAML (Lewis 1998) was used to estimate ML trees for the DNA data sets. The method implemented in this program uses a genetic algorithm to increase search speed of the ML tree.

Sequences belonging to the same genus were aligned without much trouble. It was the alignment among the different genera that was problematic due to the high divergence among the sequences, especially for the DNA data sets. Numerous gaps were needed in each case to build the alignments, highlighting the importance of indels in the evolution of retroviruses. For the three genes, *env*, *gag*, and *pol*, the best-fitting model of evolution was the complex general time reversible model GTR+ $\Gamma$  (Rodríguez et al. 1990), with rate heterogeneity among sites. DNA and protein trees for the *pol* gene estimated under simpler models of evolution (MP<sub>dna</sub>, MP<sub>prot</sub>, NJ<sub>p</sub>, and NJ<sub>mc</sub>) presented almost all of the genera as monophyletic groups, while trees estimated under more complex models (NJ<sub>ml</sub>, NJ<sub>mlprot</sub>, MP<sub>pars</sub>, MP<sub>rice</sub>, and GAML [not shown]) showed less monophyletic groups (figs. 1 and 2). The MP<sub>step</sub> tree, which can be considered to use a model of medium complexity, recovered several monophyletic groups, although its topology was different from that of the MP<sub>dna</sub>, MP<sub>prot</sub>, NJ<sub>p</sub>, and NJ<sub>mc</sub> trees (Templeton test,  $P < 0.05$ ). For the *env* and *gag* genes, those trees estimated under simpler models again showed more monophyletic groups coincident with the current genera than those trees estimated under more complex models (trees not shown). NJ<sub>p</sub> trees were not significantly different (Templeton test,  $P > 0.05$ ) from MP trees. Topologies inferred from the three genes were different. In all cases, the nodal support was higher for the simpler models than for the complex models. This is not unexpected, as variances increase under complex models, and simpler models might be just overestimating the bootstrap values (Yang, Goldman, and Friday 1994).

In the past years, the nucleotide sequences of a large number of retroviruses have been determined, and phylogenetic trees including several members of the family Retroviridae have been published (Doolittle et al. 1989, 1990; Xiong and Eickbush 1990; McClure 1993; Griffiths et al. 1997; Boeke and Stoye 1998; Vogt 1998; Coffin 1999). Current classification of retroviruses is based on this phylogenetic evidence and on morphology, biochemical properties, and range of hosts (Chiu et al. 1984; Coffin 1999). Today, we have a good idea of the “true” phylogenetic relationships among the main groups of retroviruses. When examining the results of

Key words: models of evolution, complex trees, phylogenetic bias, alignment, retrovirus, phylogeny.

Address for correspondence and reprints: David Posada, Department of Zoology, Brigham Young University, 574 Widtsoe Building, Provo, Utah 84602-5255. E-mail: dp47@email.byu.edu.

*Mol. Biol. Evol.* 18(2):271–275. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

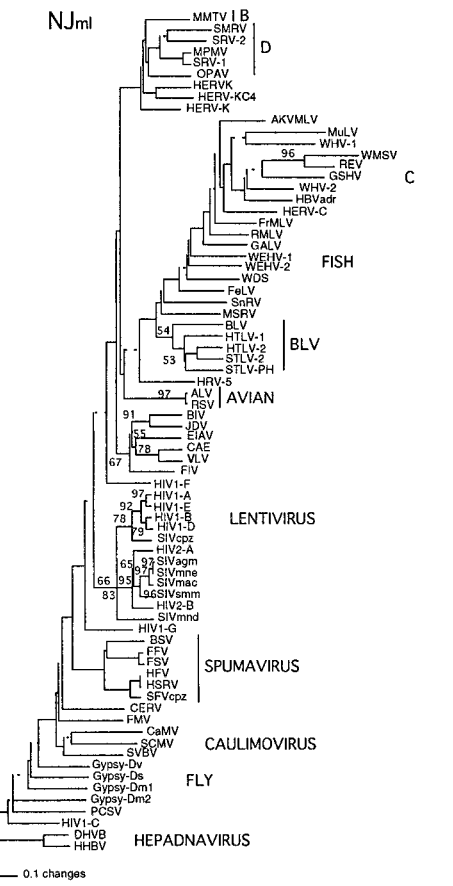
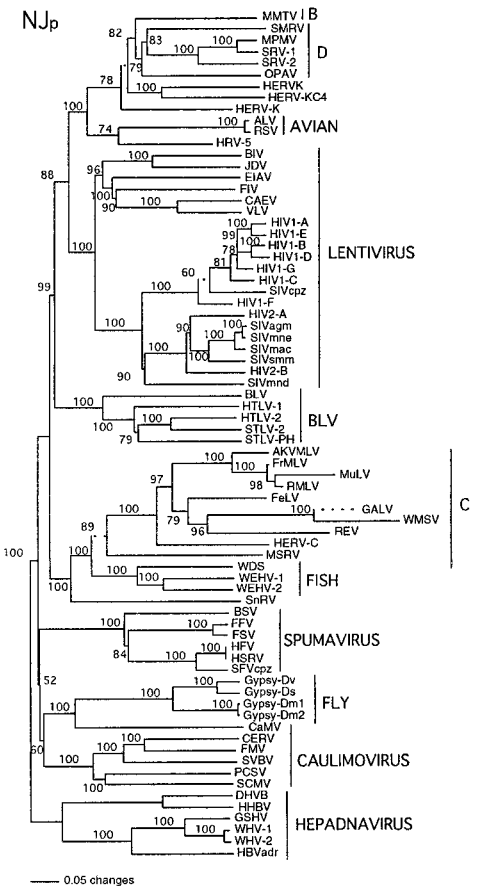
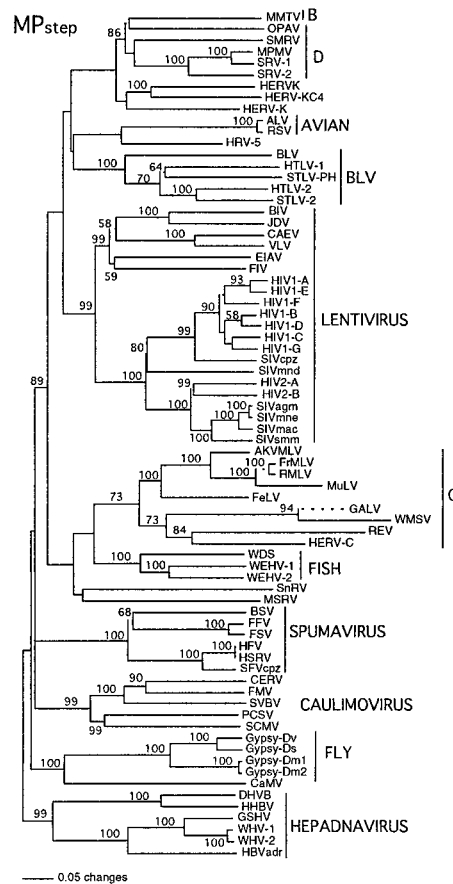
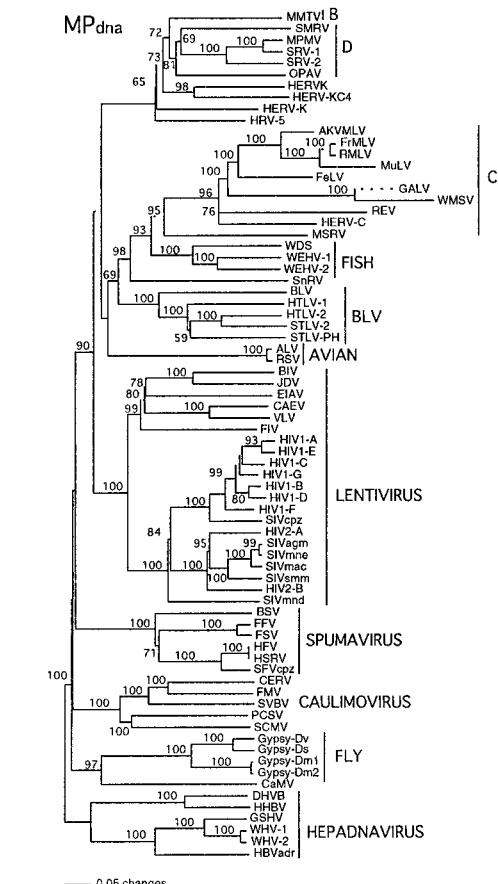




FIG. 2.—Pol protein trees: maximum-parsimony tree (MP<sub>prot</sub>), maximum-parsimony tree weighted by the protpars matrix (MP<sub>pars</sub>) (Felsenstein 1991), maximum-parsimony tree weighted by a nonmetric and asymmetrical step matrix estimated with the program AMP (Rice, personal communication) (MP<sub>rice</sub>), neighbor-joining tree from mean character distances (NJ<sub>mc</sub>), and neighbor-joining tree from maximum-likelihood distances (NJ<sub>mlprot</sub>). To obtain this last tree, a protein rate matrix was estimated using the reversible model for protein evolution (Yang, Nielsen, and Masami 1998); this matrix was used to obtain pairwise maximum-likelihood distances according to Goldman and Yang's (1994) codon model implemented in PAML, version 2.0 (Yang 1997b). The neighbor-joining tree was estimated from the resulting distance matrix. Support of the nodes trees was evaluated using the bootstrap technique (Felsenstein 1985) with 1,000 replicates (100 replicates for the unweighted parsimony tree), except for the NJ<sub>mlprot</sub> tree, for which the amount of computation would be unpractical.

←

FIG. 1.—Pol DNA trees: maximum-parsimony tree (MP<sub>dna</sub>) (random addition, tree bisection-reconnection [TBR], 1,000 replicates), maximum-parsimony tree weighted by an estimated step matrix (MP<sub>step</sub>) (random addition, TBR, 100 replicates), neighbor-joining tree from p-distances (NJ<sub>p</sub>), and neighbor-joining tree from maximum-likelihood distances (NJ<sub>ml</sub>) estimated under the best-fit model of DNA substitution (GTR+Γ). Support of the nodes was evaluated using the bootstrap technique (Felsenstein 1985) with 1,000 replicates.

the present study, only those trees estimated according to simple, likely wrong, models of evolution agree with current evidence. In most of the reconstructed trees, different genera appear as monophyletic groups. These groups have normally high bootstrap values indicating that, given the data sets at hand, we can be confident in the nodes defining these clusters. When more complex, more realistic, models of evolution are employed, fewer genera are recovered as monophyletic, the level of support is lower, and the topologies are very different from the assumed "known" trees.

Phylogenetic bias, by which "incorrect" models can give "correct" answers, has been identified in simulation studies. Why this bias occurs is a question that remains unsolved. Here, we report an empirical example of this bias. The causes of it are probably complex and dependent on several factors. One possible factor contributing to the bias is most likely a problematic alignment, in which sequences belonging to the same group (genus) are easily aligned, whereas the opposite is true for sequences belonging to different groups. Complex models might be confounded when trying to extract information from the bad intragroup sequence alignment, while simpler models use basically the observed patterns. This would warrant a word of caution for the estimation of phylogenies from highly divergent data sets. It would be interesting to test whether this bias appears when analyzing highly divergent sequences. It should be noted that the three genes studied might not be giving independent evidence for the bias. Most likely, they share common causes responsible for it, probably related to the high level of divergence.

The use of particular models of nucleotide substitution may change the results of an analysis (Leitner, Kumar, and Albert 1997; Sullivan and Swofford 1997; Cunningham, Zhu, and Hillis 1998; Kelsey, Crandall, and Voevodin 1999). In general, phylogenetic methods may be less accurate or may be inconsistent when the model of evolution assumed is wrong (Felsenstein 1978; Huelsenbeck and Hillis 1993; Penny et al. 1994; Bruno and Halpern 1999). Statistical procedures exist to select the best-fit mode at hand (Fratl et al. 1997; Huelsenbeck and Crandall 1997; Posada and Crandall 1998), and we strongly encourage their use in phylogenetic reconstruction. However, there are unusual cases in which wrong models can lead to a better answer, and here we report such an example. Identifying these few exceptions and their causes will help us to understand the role of models in phylogenetic inference. Indeed, the conclusions of this paper are based on two main assumptions: (1) that there is a good knowledge of how the "true tree" of the family Retroviridae should look, and (2) that the processes governing the evolution of retroviruses are complex and are not well described by simple models of evolution. We think that these assumptions are easily justified.

### Acknowledgments

We thank Dr. Ziheng Yang for implementing the estimation of amino acid likelihood pairwise distances

in PAML and for guidance in their calculation. Dave Swofford kindly provided beta versions of PAUP\*. Two anonymous reviewers made helpful suggestions. This work was supported by the Alfred P. Sloan Foundation and the NIH R01-HD34350-01A1 HIV grant. NEXUS alignments are available on request from the authors.

### LITERATURE CITED

- BOEKE, J. D., and J. P. STOYE. 1998. Retrotransposons, endogenous retroviruses and the evolution of retroelements. Pp. 343–436 in J. M. COFFIN, S. H. HUGHES, and H. E. VARMUS, eds. *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
- BRUNO, W. J., and A. L. HALPERN. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.* **16**:564–566.
- CHIU, I.-M., R. CALLAHAN, S. R. TRONICK, J. SCHLOM, and S. A. AARONSON. 1984. Major *pol* gene progenitors in the evolution of oncoviruses. *Science* **223**:364–370.
- COFFIN, J. M. 1999. Molecular biology of HIV. Pp. 3–40 in K. A. CRANDALL, ed. *Evolution of HIV*. Johns Hopkins University Press, Baltimore, Md.
- CUNNINGHAM, C. W., H. ZHU, and D. M. HILLIS. 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution* **52**:978–987.
- DOOLITTLE, R. F., D.-F. FENG, M. S. JOHNSON, and M. A. MCCLURE. 1989. Origins and evolutionary relationships of retrovirus. *Q. Rev. Biol.* **64**:1–30.
- DOOLITTLE, R. F., D. F. FENG, M. A. MCCLURE, and M. S. JOHNSON. 1990. Retrovirus phylogeny and evolution. *Curr. Top. Microbiol. Immunol.* **157**:1–18.
- FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**:401–410.
- . 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**:783–791.
- . 1991. PHYLIP (phylogenetic inference package). Version 3.4. Distributed by the author, Department of Genetics, University of Washington, Seattle.
- FRATL, F., C. SIMON, J. SULLIVAN, and D. L. SWOFFORD. 1997. Gene evolution and phylogeny of the mitochondrial cytochrome oxidase gene in Collembola. *J. Mol. Evol.* **44**:145–158.
- GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**:725–736.
- GRIFFITHS, D. J., P. J. W. VENABLES, R. A. WEISS, and M. T. BOYD. 1997. A novel exogenous retrovirus sequence identified in humans. *J. Virol.* **71**:2866–2872.
- HUELSENBECK, J. P., and K. A. CRANDALL. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* **28**:437–466.
- HUELSENBECK, J. P., and D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* **42**:247–264.
- KELSEY, C. R., K. A. CRANDALL, and A. F. VOEVODIN. 1999. Different models, different trees: the geographic origin of PTLV-I. *Mol. Phylogenet. Evol.* **13**:336–347.
- LEITNER, T., S. KUMAR, and J. ALBERT. 1997. Tempo and mode of nucleotide substitutions in *gag* and *env* gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J. Virol.* **71**:4761–4770.

- LEWIS, P. O. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Mol. Biol. Evol.* **15**:277–283.
- MADDISON, W. P., and D. R. MADDISON. 1994. *MacClade: analysis of phylogeny and character evolution*. Sinauer, Sunderland, Mass.
- MCCLURE, M. 1993. Evolutionary history of reverse transcriptase. Pp. 425–444 in A. M. SKALKKA and S. P. GOFF, eds. *Reverse transcriptase*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
- PENNY, D., P. J. LOCKHART, M. A. STEEL, and M. D. HENDY. 1994. The role of models in reconstructing evolutionary trees. Pp. 211–230 in R. W. SCOTLAND, D. J. SIEBERT, and D. M. WILLIAMS, eds. *Models in phylogenetic reconstruction*. Clarendon Press, Oxford, England.
- POSADA, D., and K. A. CRANDALL. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**:817–818.
- RODRÍGUEZ, F., J. F. OLIVER, A. MARÍN, and J. R. MEDINA. 1990. The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**:485–501.
- RUSSO, C. A. M., N. TAKEZAKI, and M. NEI. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.* **13**:525–536.
- SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406–425.
- SCHÖNIGER, M., and A. VON HAESELER. 1993. A simple method to improve the reliability of tree reconstructions. *Mol. Biol. Evol.* **10**:471–483.
- SULLIVAN, J., and D. L. SWOFFORD. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenies. *J. Mamm. Evol.* **4**:77–86.
- SWOFFORD, D. L. 1998. *PAUP\*: phylogenetic analysis using parsimony and other methods*. Version 4.0 beta. Sinauer, Sunderland, Mass.
- TAJIMA, F., and N. TAKEZAKI. 1994. Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. *Mol. Biol. Evol.* **11**:278–286.
- TAKAHASHI, K., and M. NEI. 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. *Mol. Biol. Evol.* **17**:1251–1258.
- TATENO, Y., N. TAKEZAKI, and M. NEI. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* **11**:261–277.
- THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN, and D. G. HIGGINS. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **24**:4876–4882.
- VOGT, V. M. 1998. Retroviral virions and genomes. Pp. 27–70 in J. M. COFFIN, S. H. HUGHES, and H. E. VARMUS, eds. *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
- XIONG, Y., and T. H. EICKBUSH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**:3353–3362.
- YANG, Z. 1997a. How often do wrong models produce better phylogenies? *Mol. Biol. Evol.* **14**:105–108.
- . 1997b. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**:555–556.
- YANG, Z., N. GOLDMAN, and A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11**:316–324.
- YANG, Z., R. NIELSEN, and H. MASAMI. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**:1600–1611.

ROSS CROZIER, reviewing editor

Accepted October 18, 2000