Letter to the Editor

Recombination and the Molecular Clock

Mikkel H. Schierup and Jotun Hein

Department of Ecology and Genetics, University of Aarhus, Denmark

Much recent work focuses on the study of sequence evolution of fast-evolving RNA-based viruses such as HIV, influenza, and foot-and-mouth disease. From both an evolutionary and an epidemiological point of view, it is of interest to know whether sequences evolve according to a molecular clock or whether evolutionary rates vary among evolutionary lineages or over time. Present phylogenetic analyses based on the likelihood ratio test (Felsenstein 1981; Huelsenbeck and Rannala 1997) often reject a molecular clock (Elena, Gonzalez-candelas, and Moya 1992; Holmes, Pybus, and Harvey 1999, Kelsey, Crandall, and Voevodin 1999), which, in turn, is often taken as evidence for rate variation caused by varying selection pressures (Holmes, Pybus, and Harvey 1999). However, many viruses readily recombine (Robertson et al. 1995; Holmes, Worobey, and Rambaut 1999; Santti et al. 1999), which implies that no single phylogenetic tree describes the genealogy of the sampled sequences. We have found that very small levels of recombination invalidate the likelihood ratio test of the molecular clock.

Data sets were simulated under the coalescent model with recombination (Hudson 1983) with the scaled recombination rate $\rho = 4Nr$, where N is the effective population size and r is the recombination rate per gene per generation. For each value of ρ, 1,600–2,000 replicates were used. The simulation program was written in C and can be accessed through http://www.daimi.au.dk/~compbio/. To mimic an average viral data set, we simulated 1,000-bp sequences evolving according to a Jukes-Cantor model of substitution with constant rate (i.e., a molecular clock) and an average distance between sequences of 20% divergence. From the simulated data sets, the maximum-likelihood values of the most likely phylogenies with and without the assumption of a molecular clock were compared using the DNAml and DNAmlk programs of PHYLIP (Felsenstein 1995), assuming the Jukes-Cantor model under which the sequences were simulated. We restricted analysis to cases in which the two methods returned the same tree topology. This was done because use of the χ^2 distribution for likelihood ratio tests assumes that hypotheses are nested (Huelsenbeck and Rannala 1997; Whelan and Goldman 1999). In this case, $\delta = -2\Delta\ln(\text{likelihood})$ is approximately χ^2 distributed with $n - 2$ degrees of freedom, where n is the number of sequences (Felsenstein 1981).

Table 1 shows for 10 sequences the percentage of cases in which the molecular clock is rejected using a $\chi^2(8)$ distribution at the 0.1%, 1%, and 5% levels for different rates of the population recombination rate ρ. Also shown is the mean of δ, which is expected to be 8 for a $\chi^2(8)$ distribution. The three percentiles and the mean δ value for $\rho = 0$ are in good agreement with the $\chi^2(8)$ distribution. However, even low levels of recombination cause a large proportion of false rejections of the molecular clock, and when $\rho > 8$, the clock is rejected in almost all cases. When ρ approaches infinity, all sequences are expected to be equidistant and a molecular clock should reappear, but no sign of this is observed even for our largest value of ρ, 64. Conditioning on the observed number of recombinations in the sequences, we found that the likelihood of rejecting the molecular clock exceeds 50% when the total number of recombinations in the history of the 10 sequences exceeds 6. We emphasize that six recombination events in many cases would not be detectable in data sets. The last column of table 1 shows the percentage of cases in which the same topologies were found with DNAml and DNAmlk. This percentage decreases with increasing ρ values, because recombination affects the topology. We also analyzed the remaining cases in which different topologies were found by forcing DNAmlk to use the same topology as that found by DNAml. This led to an even higher percentage of rejections of the clock when recombination was present; thus, the results of table 1 are an underestimate of the effect of recombination.

We argue that recombination is the simplest explanation for the lack of a molecular clock in many data sets of viruses. For example, the recombination rate for HIV 1 is likely to be higher than even the largest value used here. Thus, dating the origin of the HIV 1 pandemic from an early (1959) sequence (Korber, Theiler, and Wolinsky 1998; Zhu et al. 1998) may yield misleading results. The implications may extend to the human mitochondrial data, where evidence for recombination was reported recently (Awadalla, Eyre-Walker, and Smith 1999).

More complex substitution models than used here including rate variation over the sequence are expected to increase the likelihood of rejecting the molecular clock in most cases; thus, our estimates are conservative. We conclude that methods of testing the molecular clock that incorporate recombination or are independent of recombination would be very desirable.

Key words: virus, molecular clock, likelihood ratio test, recombination, mtDNA.

Address for correspondence and reprints: Mikkel Heide Schierup, Department of Ecology and Genetics, University of Aarhus, Building 540, Ny Munkegade, DK-8000 Aarhus C., Denmark. E-mail: mikkel.schierup@biology.au.dk.

Mol. Biol. Evol. 17(10):1578–1579. 2000

Acknowledgments

We thank Thomas Christensen for programming assistance and Xavier Vekemans, Roald Forsberg, and two

Table 1
Recombination and Probability of Rejecting the Molecular Clock

ρ[a]	Expected No. of Recombinations[b]	Rejection at 5% Level (%)	Rejection at 1% Level (%)	Rejection at 0.1% Level (%)	Average δ[c]	Probability of Same Topology[d] (%)
0	0	6.5	1.4	0.2	8.5	72.7
1	2.83	33.9	25.2	19.6	22.9	68.1
2	5.66	48.3	39.2	32.0	30.3	62.6
8	22.63	82.4	75.2	68.7	64.2	50.5
64	181.05	98.0	95.3	91.0	76.3	32.8

[a] The combination rate ρ is scaled as the number of recombinations in a gene per $2N$ generations, where N is the effective diploid population size.

[b] The number of recombinations expected in the phylogenetic history of 10 sequences is given by $\rho \sum_{i=1}^{n-1} 1/i$.

[c] Maximum likelihood estimates were obtained using the DNAml and DNAmlk programs of PHYLIP with the GLOBAL option (Felsenstein 1995). Average δ denotes twice the difference in log likelihood.

[d] The probability that DNAml and DNAmlk return a tree with the same topology.

anonymous reviewers for comments on the manuscript. This study was supported by grant 9701412 from the Danish Natural Sciences Research Council and by BRICS, Center of the Danish National Research Foundation.

LITERATURE CITED

AWADALLA, P., A. EYRE-WALKER, and J. M. SMITH. 1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science **286**:2524–2525.

ELENA, S. F., F. GONZALEZCANDELAS, and A. MOYA. 1992. Does the Vp1 gene of foot-and-mouth-disease virus behave as a molecular clock. J. Mol. Evol. **35**:223–229.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA-sequences—a maximum-likelihood approach. J. Mol. Evol. **17**:368–376.

———. 1995. PHYLIP (phylogeny inference package). Version 3.572. Distributed by the author, Department of Genetics, University of Washington, Seattle.

HOLMES, E. C., O. G. PYBUS, and P. H. HARVEY. 1999. The molecular population dynamics of HIV-1. Pp. 177–207 *in* K. A. CRANDALL, ed. The evolution of HIV. Johns Hopkins University Press, Baltimore, Md.

HOLMES, E. C., M. WOROBEY, and A. RAMBAUT. 1999. Phylogenetic evidence for recombination in dengue virus. Mol. Biol. Evol. **16**:405–409.

HUDSON, R. R. 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23**:183–201.

HUELSENBECK, J. P., and B. RANNALA. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276**:227–232.

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.

KORBER, B., J. THEILER, and S. WOLINSKY. 1998. Limitations of a molecular clock applied to considerations of the origin of HIV-1. Science **280**:1868–1871.

ROBERTSON, D. L., P. M. SHARP, F. E. MCCUTCHAN, and B. H. HAHN. 1995. Recombination in HIV-1. Nature **374**:124–126.

SANTTI, J., T. HYYPIA, L. KINNUNEN, and M. SALMINEN. 1999. Evidence of recombination among enteroviruses. J. Virol. **73**:8741–8749.

WHELAN, S., and N. GOLDMAN. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenies. Mol. Biol. Evol. **16**:1292–1299.

ZHU, T. F., B. T. KORBER, A. J. NAHMIAS, E. HOOPER, P. M. SHARP, and D. D. HO. 1998. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature **391**:594–597.

KEITH CRANDALL, reviewing editor

Accepted June 2, 2000