Experimental Phylogenetics: Generation of a Known Phylogeny

DAVID M. HILLIS, JAMES J. BULL, MARY E. WHITE, MARTY R. BADGETT, IAN J. MOLINEUX

Although methods of phylogenetic estimation are used routinely in comparative biology, direct tests of these methods are hampered by the lack of known phylogenies. Here a system based on serial propagation of bacteriophage T7 in the presence of a mutagen was used to create the first completely known phylogeny. Restriction-site maps of the terminal lineages were used to infer the evolutionary history of the experimental lines for comparison to the known history and actual ancestors. The five methods used to reconstruct branching pattern all predicted the correct topology but varied in their predictions of branch lengths; one method also predicts ancestral restriction maps and was found to be greater than 98 percent accurate.

THE DEVELOPMENT OVER THE PAST four decades of explicit methods for phylogenetic inference (*1*) has permitted biologists to reconstruct the broad outlines of evolutionary history and to interpret comparative biological studies within an evolutionary framework (*2*). However, evolutionary history usually cannot be observed directly, at least over the course of relevant magnitudes of change, so that assessment of phylogenetic methods has relied on numerical simulations. Although simulations have provided considerable insight into the effectiveness of various

Departments of Zoology and Microbiology, University of Texas, Austin, TX 78712.

phylogenetic algorithms, they are limited by an incomplete knowledge of biology: all models incorporate untested assumptions about evolutionary processes.

Direct tests of organismal phylogenetic histories are limited to a few studies of strains of laboratory animals (*3*) and plant cultivars (*4*). Even these cases have shortcomings: the organisms underwent little genetic differentiation, phylogenies were produced over the course of decades or centuries, and the histories are incompletely known. In contrast, viruses can be manipulated in the laboratory through thousands of generations per year, and mutation rates of viruses can easily be elevated through the use of mutagens, so that experimental studies of phylogenies with viruses should be feasible (*5, 6*). We report the creation of a known phylogeny of lineages derived from bacteriophage T7 and provide fine-scale restriction maps of the entire genomes of the experimental lineages, including the ancestors. Our purpose was to test the effectiveness of methods for inferring phylogeny and ancestral genetic character states by comparing the inferred evolutionary history against a known phylogeny and the true ancestors.

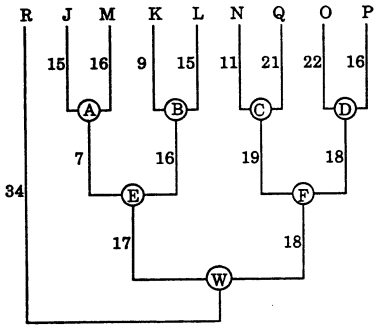

Fig. 1. True phylogeny for the experimental lineages of bacteriophage T7. The ancestors at each node are labeled with letters A through F and W (the latter represents wild-type T7). The numbers represent the number of restriction-site differences scored between the phages at each node of the phylogeny.

We chose to construct a symmetric phylogeny with equal distances among nodes (Fig. 1) (*7*). This topology (tree shape) has proven especially amenable to accurate reconstruction in theoretical studies (*8–10*) and may thus be regarded as a "null" model against which other topologies may be compared. We created a phylogeny of nine taxa (eight ingroup lineages and one outgroup lineage to root the tree) by serially propagating T7 phage in the presence of a mutagen and dividing the lineages at predetermined intervals (7); a clonal stock of wild-type T7 phage was the common ancestor of all lineages. There are 135,135 possible bifurcating trees for this many taxa, so the likelihood of inferring the correct phylogeny by chance alone is minimal. We compared the actual phylogeny (Fig. 1) to estimated phylogenies from five reconstruction methods; estimates were based on restriction-site maps produced for 34 restriction endonucleases in all terminal lineages (Figs. 2 and 3). To avoid bias, the actual phylogeny was unknown to the person mapping the restriction sites. We also produced restriction maps for the ancestral phage at each of the nodes of the true phylogeny (Fig. 3). Three aspects of the inferred phylogeny were compared to the actual phylogeny: branching topology, branch lengths, and ancestral states. The five methods of phylogenetic inference evaluated were parsimony (*12*), the Fitch-Margoliash method (*13*), the Cavalli-Sforza method (*14*), neighbor-joining (*15*), and the unweighted pair-group method of arithmetic averages (UPGMA) (*16*).

All methods predicted the correct branching order of the known phylogeny, but no method predicted the actual branch lengths for every branch (*17*). To compare the five methods for their ability to predict branch lengths, the correlation between observed and predicted branch lengths was calculated for each method. These five correlations were significantly heterogeneous, with parsimony yielding the highest value and UPGMA yielding the lowest value (*18*). The UPGMA method is known to be sensitive to unequal rates of change (*1*), and the number of changes per branch was quite variable in the true phylogeny [although the number of changes per ingroup branch is not significantly heterogeneous from the expectation under a Poisson distribution; test from (*19*)].

The experimental system also enabled us to determine ancestral states directly. Of the methods tested, only parsimony makes predictions about ancestral character states (parsimony may be used to optimize states onto phylogenies inferred by other methods, but, for these data, all methods estimated the same branching pattern). In comparing inferred ancestral states to the actual ancestral states, three outcomes are possible: the ancestral states may be (i) correctly inferred, (ii) incorrectly inferred, or (iii) ambiguous (when more than one character optimization is possible). For 202 variable sites assayed in each of seven ancestors, parsimony correctly inferred 1369 states (97.3%), incorrectly inferred 18 states (1.3%), and was ambiguous about 20 states (1.4%). Seven states (of four sites) could not be observed in some ancestors because they fell under deletion mutations (Fig. 3). If the 91 wild-type states that were invariant in all lineages are included, the inferred restriction maps are an average of 98.6% identical to the actual maps (with either delayed or accelerated

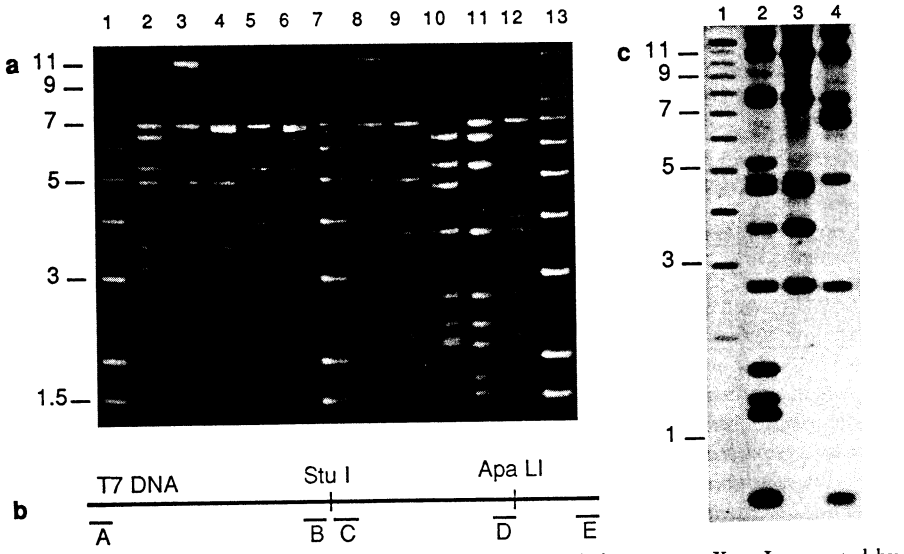

Fig. 2. (**a**) DNA from each terminal lineage cleaved with the restriction enzyme Xmn I, separated by electrophoresis on 0.8% wedge-shaped agarose gels and stained with ethidium bromide. Lanes 1, 7, and 13 contain size standards (sizes shown in kilobase pairs); lane 2 contains wild-type T7 DNA; lanes 3 through 6 contain DNA from lineages R, Q, P, and O, respectively; lanes 8 through 12 contain DNA from lineages N, M, L, K, and J, respectively. (**b**) Strategy for mapping restriction site variation. Viral DNA was cleaved with the restriction enzyme Stu I (which has a single recognition site in T7 that remained invariant in all lineages) and then partially digested with the restriction enzyme to be mapped. After electrophoresis on 0.8% wedge-shaped agarose gels, Southern blots were successively hybridized with four oligonucleotides that matched sequences located at each end of T7 DNA and on either side of the Stu I fragment (A through C and E). To improve resolution of the map for several restriction sites, we cleaved with the restriction enzyme Apa LI (which also has an invariant recognition site in all lineages), partially digested with the enzyme to be mapped, and hybridized the Southern blot with an oligonucleotide located at position D. (**c**) An example of an autoradiogram produced as described in (b), probed with oligonucleotide B. Lane 1 contains size standards (indicated in kilobase pairs); lanes 2 through 4 contain DNA from lineages O, P, and Q partially cleaved with Sau 3A1 (an isoschizomer of Mbo I.

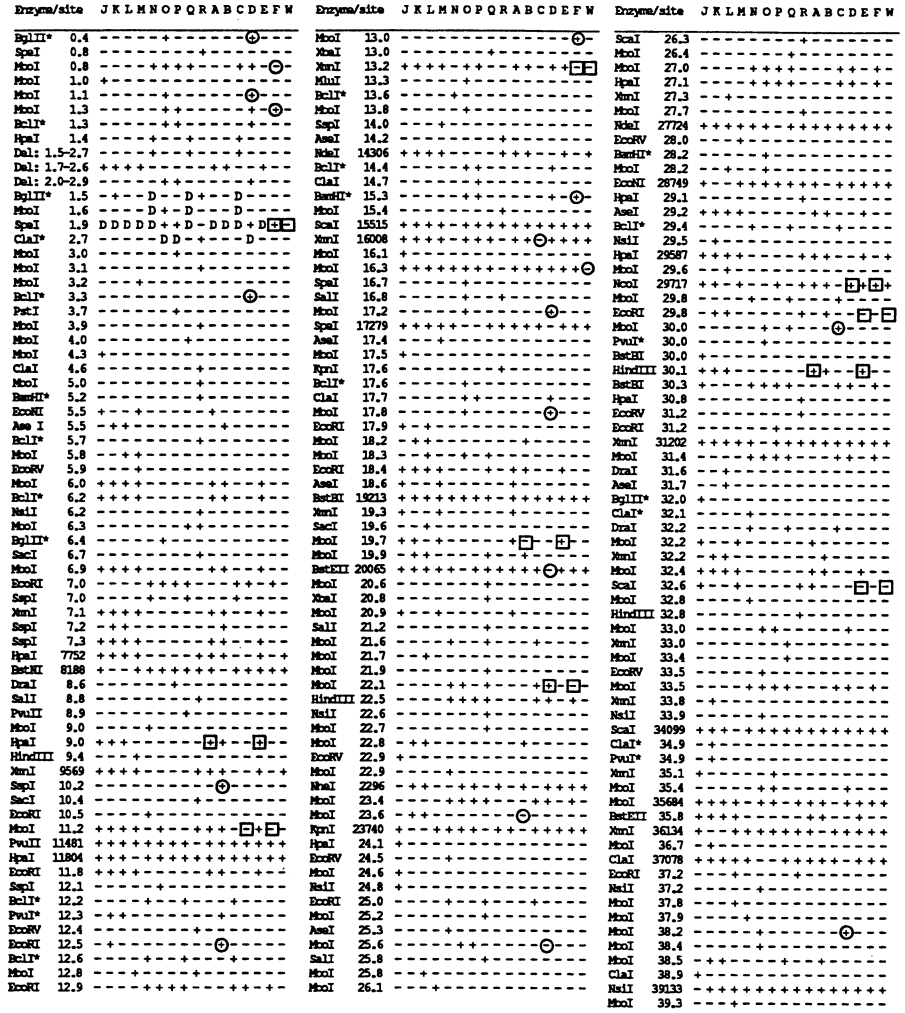

Fig. 3. Variable restriction enzyme cleavage sites and deletions of the terminal lineages (J through R) and ancestral nodes (A through F) compared to the wild-type T7 genome (W). Wild-type sites for these enzymes that are not listed are found among all the lineages. Sites with positions shown without decimals are wild-type sites that have been lost in some of the lineages; their positions were determined from the complete sequence of T7 DNA (*11*). The remaining site locations were mapped as shown in Fig. 2; these positions are indicated in kilobases. Sites marked with an asterisk are also Mbo I sites; they are listed only once unless multiple changes produced differences between Mbo I and the larger recognition sequence in some of the taxa (for example, the Mbo I and Bam HI sites at position 5.2). Sites in ancestors that were inferred incorrectly by parsimony analysis of the terminal lineages are circled; those for which inference was ambiguous are surrounded by squares. A letter "D" indicates that the site occurs in a region that has been deleted in the respective lineage; these sites were coded as unknown in the analyses.

character optimization used) (*1*). Three of the incorrectly inferred states involved sites that were found only in the ancestral lineages, hence would not have been detected as variable characters if ancestors were unavailable (as is typically the case in phylogenetic studies).

The results of this study directly support the legitimacy of methods for phylogenetic estimation, not only with regard to reconstructing branching relationships, but also branch lengths and ancestral genotypes. Perhaps more importantly, they point the direction to a field of research in which methods of reconstruction can be tested against various known phylogenies of real organisms differing in topological and other evolutionary characteristics in the same fashion that tests have been conducted with simulated, theoretical phylogenies. Experimental phylogenetics is not a substitute for numerical studies, nor is it likely that laboratory phylogenies will ever display the full complexity of phylogenies produced over long-term evolution, but such studies will fill an important void in the science of phylogenetic reconstruction.

REFERENCES AND NOTES

1. D. L. Swofford and G. J. Olsen, in *Molecular Systematics*, D. M. Hillis and C. Moritz, Eds. (Sinauer, Sunderland, MA, 1990), pp. 411–501.
2. N. Eldredge and J. Cracraft, *Phylogenetic Patterns and the Evolutionary Process* (Columbia Univ. Press, New York, 1980); J. Felsenstein, *Am. Nat.* **125**, 1 (1985); D. R. Brooks and D. A. McLennan, *Phylogeny, Ecology, and Behavior* (Univ. of Chicago Press, Chicago, 1991); P. H. Harvey and M. D. Pagel, *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, Oxford, 1991).
3. W. M. Fitch and W. R. Atchley, *Science* **228**, 1169 (1985); in *Molecules and Morphology in Evolution: Conflict or Compromise?*, C. Patterson, Ed. (Cambridge Univ. Press, Cambridge, 1987), pp. 203–216; W. R. Atchley and W. M. Fitch, *Science* **254**, 554 (1991).
4. B. R. Baum, in *Cladistics: Perspectives on the Reconstruction of Evolutionary History*, T. Duncan and T. F. Stuessy, Eds. (Columbia Univ. Press, New York, 1984), pp. 192–220.
5. F. W. Studier, in *Genes, Cells, and Behavior: A View of Biology Fifty Years Later*, N. H. Horowitz and E. Hutchings, Jr., Eds. (Freeman, San Francisco, 1980), pp. 72–78.
6. M. E. White, J. J. Bull, I. J. Molineux, D. M. Hillis, in *Proceedings of the Fourth International Congress of Systematic and Evolutionary Biology*, E. Dudley, Ed. (Dioscorides Press, Portland, OR, 1991), pp. 935–943.
7. T7 phage was grown in 1-ml cultures of *Escherichia coli* strain W3110 in the presence of the mutagen *N*-methyl-*N'*-nitro-*N'*-nitrosoguanidine (20 μg/ml). After lysis, a new culture was infected with 10 μl of the lysate. After every five serial lysates, phage were plated on agar and a single plaque was randomly selected for further propagation. The distance between each node of the ingroup phylogeny (lineages J through P and their ancestors) was equal to 40 serial lysates; each lysate is the consequence of two to three bursts (generations) of phage. The outgroup lineage (R) was derived from wild-type T7 by a similar protocol (*6*). To guard against contamination among lineages, we assayed DNA from each isolated plaque with restriction enzymes whose restriction patterns evolve rapidly in this system (either Mbo I or a combination of Hae II, Hpa I, Kpn I, and Spe I) and looked for massive convergence of the restriction patterns. Two cases of contamination were detected by this procedure; after verifying the contamination with additional restriction digests, we regrew the lineages from the last contamination-free ancestor.
8. J. Felsenstein, *Syst. Zool.* **27**, 401 (1978).
9. J. P. Huelsenbeck, *ibid.* **40**, 257 (1991).
10. J. A. Lake, *Mol. Biol. Evol.* **4**, 167 (1987).
11. J. J. Dunn and F. W. Studier, *J. Mol. Biol.* **166**, 477 (1983).
12. D. L. Swofford, *Phylogenetic Analysis Using Parsimony* (Illinois Natural History Survey, Champaign, 1990).
13. W. M. Fitch and E. Margoliash, *Science* **155**, 279 (1967).
14. L. L. Cavalli-Sforza and A. W. F. Edwards, *Am. J. Hum. Genet.* **19**, 233 (1967).
15. N. Saitou and M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987).
16. R. R. Sokal and C. D. Michener, *Univ. Kansas Sci. Bull.* **38**, 1409 (1958).
17. For the parsimony reconstruction, bootstrap confidence intervals were 100% for all nodes except A (92%; refer to Fig. 1 for this node), based on 1000 replications; the consistency index for informative characters was 0.750.
18. Correlations between the actual and predicted branch lengths were 0.91 for parsimony (average of accelerated and delayed character transformation), 0.89 for the Fitch-Margoliash method, 0.88 for the Cavalli-Sforza method, 0.88 for the neighbor-joining method, and 0.82 for the UPGMA method. As the estimates of branch lengths by these methods may violate the assumptions underlying parametric statistics, we tested the possibility of statistically significant differences among the five correlations in the following fashion. For each branch *i*, each of the five estimates of that branch was assigned at random and without replacement to one of the five reconstruction methods; the resulting matrix was thus a shuffling of the original matrix, with the constraint that an estimate of branch *i* was always assigned as an estimate of branch *i*, even though the method to which it was assigned varied. Correlations between actual and estimated branch lengths were computed for each of the methods in this randomized matrix, and these data were recorded; the process was repeated 10,000 times. The heterogeneity of corre-

lations from the actual estimates was found to be significantly greater than that of the randomized data ($P = 0.017$), with the correlation for UPGMA being significantly less than the randomized value minimum ($P < 0.011$). The heterogeneity of correlations among the set of four methods with UPGMA removed is no longer significantly large, but the correlation for parsimony is larger than the maximum of randomized values at $P = 0.05$. By these criteria, UPGMA appears to be significantly worse than the other methods, and there is some evidence that parsimony is superior.

19. G. W. Snedecor and W. G. Cochran, *Statistical Methods* (Iowa State Univ. Press, Ames, ed. 7, 1980).

20. Supported by the NSF (to D.M.H.) and the NIH (to I.J.M.). We thank F. W. Studier, I. Tessman, and T. Kunkel for advice on mutagenesis in phage.

8 August 1991; accepted 22 November 1991