EVOLUTION

INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

PUBLISHED BY

THE SOCIETY FOR THE STUDY OF EVOLUTION

Vol. 47 **August, 1993** **No. 4**

Evolution, 47(4), 1993, pp. 993–1007

EXPERIMENTAL MOLECULAR EVOLUTION OF BACTERIOPHAGE T7

J. J. BULL, C. W. CUNNINGHAM, I. J. MOLINEUX,[1]
M. R. BADGETT, AND D. M. HILLIS
Department of Zoology, University of Texas, Austin, Texas 78712

Abstract.—We present an analysis of molecular evolution in a laboratory-generated phylogeny of the bacteriophage T7, a virus of 40 kilo-base pairs of double-stranded DNA. The known biology of T7 is used in concert with observed changes in restriction sites and in DNA sequences to produce a model of restriction-site convergence and divergence in the experimental lineages. During laboratory propagation in the presence of a mutagen, the phage lineages changed an estimated 0.5%–1.5% in base pairs; most change appears to have been G → A or C → T, presumably because of the mutagen employed. Some classes of restriction-site losses can be explained adequately as simple outcomes of random processes, given the mutation rate and the bias in mutation spectrum. However, some other classes of sites appear to have undergone accelerated rates of loss, as though the losses were selectively favored. Overall, the wealth of knowledge available for T7 biology contributes only modestly to these explanations of restriction-site evolution, but rates of restriction-site gains remain poorly explained, perhaps requiring an even deeper understanding of T7 genetics than was employed here. Having measured these properties of molecular evolution, we programmed computer simulations with the parameter estimates and pseudo-replicated the empirical study, thereby providing a data base for statistical evaluation of phylogeny reconstruction methods. By these criteria, replicates of the experimental phylogeny would be correctly reconstructed over 97% of the time for the three methods tested, but the methods differed significantly both in their ability to recover the correct topology and in their ability to predict branch lengths. More generally, the study illustrates how analyses of experimental evolution in bacteriophage can be exploited to reveal relationships between the basics of molecular evolution and abstract models of evolutionary processes.

Key words.—Bacteriophage, experimental molecular evolution, molecular systematics, parametric bootstrap, restriction-site evolution, selection, T7.

Received September 28, 1992. Accepted January 20, 1993.

Among the most rapidly expanding fields in evolutionary biology is molecular systematics, which attempts to reconstruct phylogenies of living (and, in some cases, extinct) taxa from DNA sequences or their encoded gene products (Nei 1987; Doolittle 1990; Hillis and Moritz 1990; Miyamoto and Cracraft 1991). From a cold and cruel perspective of the scientific method, the major weakness of this field is its difficulty in unambiguously falsifying hypotheses of phylogenetic relationships, and hence, of molecular evolution. In almost no cases is a phylogeny known a priori (Baum 1984; Fitch and Atchley 1987; Atchley and Fitch 1991), and thus reconstructions and models of molecular evolution that require knowledge of ancestry cannot be definitively tested. A common approach to circumvent some of these difficulties has been to generate pseudo-phylogenies with a computer and then to test methods of reconstruction against these pseudo-phylogenies (Li et al. 1987; Fitch and Ye 1991; Jin and Nei 1991; Nei 1991; Sidow and Wilson 1991). The advantage of this approach

[1] Department of Microbiology, University of Texas, Austin, Texas 78712 USA.

is that methods can be tested against millions of independent phylogenies, but a drawback is that the underlying models of molecular evolution must be assumed a priori and lack the complexities of natural biological evolution (Fitch and Atchley 1987; Nei 1987; Hillis et al. 1992).

An alternative approach is to develop a laboratory system in which lineages of rapidly evolving organisms are propagated so that the phylogenetic history is known. We have described an empirical system using the bacteriophage T7 which allowed us to create such phylogenies (White et al. 1991; Hillis et al. 1992). In this system, first described by Studier (1980), bacteriophage are grown in the presence of a mutagen that enhances the rate of base-pair substitutions. The phage are thus subjected to intense mutation pressure while maintaining selection for viability, and this rapid laboratory system may be regarded as an approximation to long-term evolution with low mutation rates. Because the underlying model of molecular evolution is biological, this system is expected to incorporate a level of complexity and reality not attainable in numerical simulations of evolution.

Our prior studies focused on a description of the experimental design, presentation of a set of data from one experimental phylogeny, and evaluation of five reconstruction methods applied to the data. The goal of the present study is to identify the properties of molecular evolution in this system as they pertain to phylogenetic history. Fulfillment of this goal broadens our understanding of molecular evolution in general and enables us to appreciate the complexity of evolutionary processes that must be addressed by phylogeny reconstruction methods. The measurement of evolutionary parameters has also enabled us to return to one goal of our previous study: computer simulations can be programmed with the estimates of evolutionary parameters and used to generate replicates of the empirical phylogeny, enabling further tests of reconstruction methods.

Perspective: The Utility of a Laboratory System

As a model of evolutionary history, the T7 system has an advantage over computer simulations because its molecular evolution is biological. Computer simulations assume simple models of evolution: typically random changes in nucleotide sequences being fixed stochastically, with constant probabilities. Variations in parameters through time, effects of selection, and many other potential complexities are omitted. An essential question is whether such omissions lead to serious misrepresentations of evolutionary history, and an important application of the T7 system is thus to examine the pitfalls of this assumed simplicity. There is an extensive background of genetic work on T7 that can be brought to bear on experimental studies: the wild-type sequence has been published, open reading frames in the sequence correspond to genes studied through classical means, and the biochemical functions of many gene products are understood as are the consequences of null mutations in those genes (Dunn and Studier 1983). One can thus conduct experiments on molecular evolution and then evaluate how many of these genetic details are required to account for the observations. Perhaps it will be found that the simplest models of molecular evolution are adequate, and the knowledge of T7 biology is superfluous. Alternatively, T7 biology may offer indispensable information about the course of molecular evolution. In the latter case, the study would provide motivation for striving to improve models applied to other systems as well.

The central goal of this paper is to apply experimental data obtained using T7 to a model of restriction-site evolution, that is, a model specifying the rates of site gains and losses that in turn reflect evolutionary history. To take advantage of the many levels of data and background information available in this study, our pursuit of this goal requires several steps: presentation of a basic model, assessment of change at the nucleotide level, and identification of genetic factors influencing rates of restriction-site evolution. Each of these steps is undertaken in a separate section below. Ultimately, estimation of the model's parameters enables us to program computer simulations to generate numerical replicates of the data (parametric bootstraps); these data are then applied to statistical evaluations of reconstruction methods.

Methods and Materials

Review of T7 Experimental Phylogeny.—T7 is an obligately lytic bacteriophage of 39,937 base pairs of double-stranded DNA whose complete sequence is known (Dunn and Studier 1983; Moffatt et al. 1984; Beck et al. 1989). From a combination of classical and molecular genetic methods, 51 genes have been identified. Approximately 8% of the total DNA sequence is noncoding (intergenic), but many of these inter-

genic regions are involved in vital functions such as gene regulation, ribosome binding, and so forth. The nucleotide composition of the + strand is 27% A, 23% C, 26% G, and 24% T, hence almost 50% GC for the double-stranded molecule.

The procedure for introducing mutations in this phage is to grow it in a bacterial culture containing the potent mutagen N-methyl-N′-nitro-N-nitrosoguanidine (NG). Phages in the resulting lysate are then used to infect a similar culture. The latent period of T7, which is normally 20 min at 37°C, increases somewhat in the presence of NG. Nevertheless, in a relatively short period, a stock of phage may be grown for hundreds of successive cycles in the presence of mutagen, potentially accumulating hundreds of mutations.

The size of the evolving T7 population may be varied simply by adjusting the number of phage particles in a lysate that is used to initiate the succeeding infection. Alternatively, the phage may be plated on semisolid medium to generate plaques, each of which is derived from a single phage. Plating is done in the absence of mutagen; thus, a clonal stock of phage is present in each plaque, which usually contains 10^7–10^8 particles. These clonal stocks can then be studied directly or can be used to continue a lineage.

In the study of Hillis et al. (1992), an experimental phylogeny was constructed that consisted of nine terminal lineages (eight ingroup, one outgroup), whose common ancestor was a clonal isolate of a wild-type stock of T7 (fig. 1). Beginning with this wild-type isolate, the ingroup taxa were created as follows: two independent lineages were propagated for 40 mutagenic cycles (referred to as the two primary branches). After 40 cycles, phage were plated, and a single plaque isolate was chosen from each lineage; two independent lineages were then initiated from each isolate, thus creating four lineages (four secondary branches). Each of these four lineages was propagated as before, and after 40 additional cycles of mutagensis, each was split again, creating eight terminal branches. Forty cycles after this last split (120 cycles from the wild-type ancestor), propagation of the eight lineages was stopped. The outgroup was also propagated from wild-type, but passaged for only 105 mutagenic cycles.

Plaque isolates were obtained and used to continue the lineages after every five mutagenic cycles on the ingroup branches and after every cycle on the outgroup branch. The presumed advantage of plaque purification is twofold. First, it

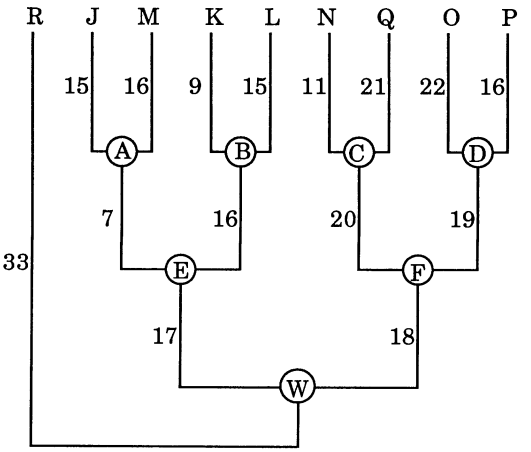

FIG. 1. Model of the experimental phylogeny using T7. Nodes are indicated by letters, and lineages between adjacent nodes are referred to as branches. Node R constitutes the single outgroup taxon, J–P the ingroup taxa (A–F are nodes of the ingroup as well). Each ingroup branch consisted of 40 mutagenic cycles; the outgroup branch was 105 mutagenic cycles. Numbers adjacent to branches represent the numbers of changes in restriction sites and deletions that were detected between the two ends of the branches; these numbers correct some of the branch lengths presented in Hillis et al. (1992).

reduces the effective population size of phage, so that natural selection is greatly weakened (and random drift is conversely strengthened) as a factor in determining whether a mutation becomes incorporated in the T7 lineage. As a consequence, the rate of fixation of deleterious mutations should be enhanced over that which would occur in the absence of plaque purification. Since most mutations in this system are likely to be either neutral or somewhat deleterious, plaque purification should thereby enhance the overall rate of nucleotide substitution. (Based on the crude assays of time from infection to lysis and plaque size, the fitness of every mutated lineage declined substantially during the course of propagation.) The second motivation for plaque purification applies only at the branch points of the phylogeny, where it ensures that all descendants in a lineage have a single, defined ancestor. Further details of the methods of propagation of T7 were provided previously (White et al. 1991; Hillis et al. 1992).

Nucleotide Sequences.—Sequences of the terminal lineages and ancestors were obtained for nucleotide positions 36,014–36,464 and 36,622–36,837. In choosing this region for sequencing, we were motivated to study a region that was

likely to evolve at a high rate but which was also sufficiently important to the phage that deletions would not be favored (a comparison of published sequences for T7 and for the closely related phage T3 was used as the basis for this inference). PCR was then used to amplify the region between nucleotide positions 35,941 and 36,905 from the phage genome, and DNA sequences were obtained from parts of the amplified product.

The Evolutionary Parameters Model

The data set from Hillis et al. (1992) consisted of 199 variable restriction sites throughout the genome and three deletions that arose in one specific region. Thus, the experimental phylogeny presented a large set of characters. However, the retrieval of evolutionary history, which is the objective of the systematist, depends more on the rates of change at individual sites than on the overall amount of change. In this section, we introduce a simple model of evolution for application to the experimental phylogeny. This model is of a single two-state character (states 0 or 1, such as the absence or presence of a restriction site), with probabilities of change in each direction. The step undertaken here is to derive a model that defines the parameters of interest and reveals how they are related to the observations. Application to the data will be undertaken in a later section.

Consider a specific restriction enzyme and an arbitrary site in the sequence of the phage ancestor of our phylogeny (wild-type). The site will either be recognized (cleaved) by the restriction enzyme or not. Whichever the ancestral state (cleaved or not cleaved), we define a *conversion* as a change from the ancestral state to the alternative state and a *reversion* as a change from the converted state back to the ancestral state. Conversions and reversions are measured only at the endpoints of branches; hence changes that arise and are lost between the endpoints of a branch are not recorded.

To associate these concepts with symbols, we use the following definitions:

c = probability per branch that the site undergoes a conversion

r = probability per branch that a formerly converted site reverts to its ancestral state

$P(i, j)$ = probability that the site experiences i conversions and j reversions in the phylogeny

$n(i, j)$ = the number of times the site is observed to undergo i conversions and j reversions

Our analysis must be confined to the ingroup part of the phylogeny, because internal nodes are not characterized for the outgroup, and there are consequently no branches within the outgroup lineage that are similar in length to the ingroup branches. Note that c and r are per-site probabilities. They have no bearing on the numbers of sites likely to be converted elsewhere in the phage genome, rather they determine the rates of evolution at one site.

The procedure adopted in estimating the two parameters (c and r) is to calculate a priori likelihoods of the observations as a function of the parameters and to find values of the parameters that maximize these likelihoods. The likelihood function depends on $P(i, j)$, which in turn depend on c and r. We begin by illustrating how to calculate $P(i, j)$ as functions of c and r. The simplest case is that in which no mutations occur, $P(0, 0)$. As there are 14 ingroup branches, and in this case no conversions occur on any of these branches,

$$P(0, 0) = (1 - c)^{14} \tag{1}$$

A slightly more complicated case is that of a single conversion, $P(1, 0)$. This probability is the sum of three terms. One term represents the probability that the site is converted along a primary branch:

$$[2c(1 - c)][(1 - r)^2(1 - c)^2] \cdot [(1 - r)^4(1 - c)^4] \tag{2a}$$

The quantity in the first pair of brackets is the probability that the site is converted on either of the two primary branches. The quantity in the second pair of brackets is the probability that no further changes occur on any of the four secondary branches, and the quantity in the third pair of brackets is the probability that no changes occur on any of the eight terminal branches. $P(1, 0)$ also includes the probability that the conversion occurs on a secondary branch:

$$[(1 - c)^2][4c(1 - c)^3][(1 - c)^6(1 - r)^2] \tag{2b}$$

Finally, $P(1, 0)$ includes the probability that the conversion occurs on a terminal branch:

$$[(1 - c)^2][(1 - c)^4][8c(1 - c)^7] \tag{2c}$$

The probability $P(1, 0)$ is therefore the sum of probabilities (2a), (2b),and (2c). The probabilities for any $P(i, j)$ may be calculated similarly,

although the number of terms typically exceeds three.

The likelihood of a set of observations is

$$L(c, r) = K \prod_{i,j} P(i, j)^{n(i,j)}, \quad (3)$$

where the product is extended over all observed classes, and the constant K is independent of c and r. $P(i, j)$ varies according to c and r, so $L(\cdot)$ will also vary with c and r, and values of these parameters that maximize the likelihood are chosen as the estimates. Since $[P(i, j)]^0 \equiv 1$ regardless of c and r [for $P(i, j) > 0$], individual terms of $P(i, j)$ need to be calculated explicitly as functions of c and r only if $n(i, j) > 0$. For a small data set such as that of Hillis et al. (1992), the maximum-likelihood estimates may be calculated to an acceptable precision from an exhaustive numerical search of the feasible parameter space.

To this point, the model has been developed for a single, arbitrary site in the phage, and it is at this juncture that the application becomes complicated. With data from a very large number of replicate phylogenies, the likelihood in equation (3) would be applied separately to each defined site in the phage. Then, $n(i, j)$ would be the number of phylogenies observed with i conversions and j reversions for that site. With only one phylogeny, however, each site is represented by just one observation, and application of this model to individual sites is virtually meaningless. We therefore group sites into sets and apply the model collectively to these sets, in essence treating them as though every member of the set shares the same values of c and r. Then n (i, j) is the number of sites in the set with i conversions and j reversions. Although c and r are undoubtedly unique for each site, some sites may have sufficiently similar c and r values to be regarded as homogeneous, and statistical tests may be applied post hoc to evaluate the legitimacy of this assumption.

This need to partition restriction sites according to different evolutionary characteristics requires that application of the model be postponed for a better understanding of restriction-site evolution in this system. The next few sections thus explore the molecular basis of nucleotide changes in the T7 system and bring that information to bear on restriction-site evolution.

Nucleotide Data

Table 1 lists the base-pair changes observed and branches of their origins for the 665 bases assayed. A total of 21 changes were observed, 19 of which were G → A or C → T transitions. NG is known to cause G·C → A·T transitions preferentially in DNA (Horsfall et al. 1990). It is, of course, not possible to distinguish between G → A and C → T *mutations* in double-stranded DNA after one or more rounds of replication, because a G → A mutation is replicated in the other strand as a C → T change, but this ambiguity is not important here. The observed rate of change per base is just under 0.002 per ingroup branch (18 changes per 14 lineages per 665 bases), but as nearly all changes are confined to just two bases, the rate per G/C is approximately 0.004. The rate per ingroup lineage would be nearly three times these values because each ingroup terminal taxon is three branch lengths removed from the wild-type ancestor.

The background information on T7 genetics

Table 1. Base-pair changes in the T7 phylogeny. Nucleotide positions are from Dunn and Studier (1983). The "mutation" column refers to the change as observed in the sense strand of T7 DNA. The column "AA impact" denotes whether the nucleotide change resides within a gene and whether changes within genes affect the deduced amino-acid sequence. Three differences were detected between the sequence of our wild-type T7 DNA and that determined by Dunn and Studier (1983): C instead of T at positions 36163 and 36461, and G instead of A at 36388. When the changes in this table are added to the branch lengths (numbers of restriction-site changes) in figure 1, there is no significant heterogeneity in branch lengths in the ingroup.

Branch	Position	Mutation	AA impact	Gene	Codon position
A-M	36122	C → T	silent	17	3
	36281	G → A	silent	17	3
	36288	G → A	intergenic		
	36378	G → A	silent	17.5	3
	36441	C → T	silent	17.5	3
B-L	36176	C → T	silent	17	3
C-N	36689	C → T	Ala → Val	18	2
D-O	36137	C → T	silent	17	3
	36441	C → T	silent	17.5	3
	36716	G → A	Ser → Asn	18	2
D-P	36745	C → T	silent	18	1
F-C	36175	C → T	Ala → Val	17	2
E-B	36142	T → G	Phe → Cys	17	2
W-E	36398	C → T	Thr → Met	17.5	2
	36705	G → A	silent	18	3
	36826	C → T	intergenic		
W-F	36412	G → A	Val → Ile	17.5	1
	36818	G → A	silent	18	3
W-R	36321	T → G	intergenic		
	36395	G → A	Gly → Glu	17.5	2
	36647	C → T	silent	18	3

enables us to consider some properties of selection on the evolution of base-pair change even in this small sample. The sequenced regions overlap with the coding regions for T7 genes *17* (nucleotide positions 34,623–36,284), *17.5* (36,343–36,546), and *18* (36,552–36,821). Gene *17* codes for a tail fiber protein, gene *17.5* for a lysis function, and gene *18* for a maturation function in DNA packaging (Dunn and Studier 1983). Genes *17* and *18* are essential for viability, whereas gene *17.5* is nonessential or only conditionally essential for producing viable phage progeny. Of the 21 mutations, 3 occurred in intergenic regions, 11 of the 18 changes in coding regions comprise silent substitutions, and the other 8 are missense mutations. For G → A and C → T mutations in the sequenced portions of these genes, the expected proportion of silent substitutions is only 0.36 if changes occur randomly, versus the observed 0.61, and the excess of silent substitutions in the data is marginally significant ($0.03 < P < 0.05$, omitting intergenic changes and transversions, based on a Fisher's exact test of the totals). Missense mutations occurred in all three genes, and the proportion of silent substitutions to missense mutations is not appreciably different across the three genes, even though only two of them are essential for progeny phage production. These observations are not implausible, as not all misense mutations are deleterious, and a nonessential gene may nonetheless confer a fitness advantage to the phage despite that fact that the phage can reproduce without it.

The sequence data provide two potentially important insights in explaining restriction-site evolution. First, most mutations are G → A or C → T. The rate of loss of restriction sites may thus increase with the number of G and C bases in the recognition sequence, and the rate of gain may increase with the number of A and T bases. Second, nucleotide changes in coding regions depend on the impact of the mutation on the amino-acid composition of the protein. If this latter effect is widespread and strong, rates of restriction-site evolution may vary greatly between sites in T7 DNA, depending on how the changes affect underlying genes.

Mutational Bias Explains Restriction-Site Losses

The question at hand is whether the nucleotide data provide unique insights to restriction-site evolution in this system. We begin by considering restriction sites present in the wild-type ancestor of the phylogeny and the rates at which they were lost. The sequence data presented above revealed that most nucleotide-level changes were G → A or C → T, so it is a straightforward prediction that the rate of loss of a restriction site should increase with the number of G and C bases in the recognition sequence: a recognition sequence with four G or C bases can be lost with a mutation at any of the four G/C positions, whereas a recognition sequence lacking G and C cannot be lost with such mutations. The expectation is met: 0 of 22 sites lacking G/C were lost, 16 of 65 sites with 2 G/C were lost at least once in the phylogeny, and 7 of 14 sites with 4 G/C were lost. The probability of observing heterogeneity this extreme under the null model (no effect of G/C content) is less than 0.002 (χ^2_2), so we reject the null model in favor of the alternative that the G/C content of a restriction-enzyme recognition sequence influences the probability of site loss.

The next issue is whether the probability of site loss is determined adequately by the number of G and C bases in the recognition sequence, or whether the impact of the mutation on T7 coding regions also needs to be considered. Each G (C) in a wild-type restriction site was thus evaluated as to whether a G → A (C → T) change would occur in an intergenic region, would generate a silent substitution in a coding region, or would cause a missense mutation. Because recognition sequences for the enzymes studied contained either 0, 2, or 4 G or C bases, each site (i) was classified according to the number of intergenic (G_i), silent (S_i), and missense and nonsense (M_i) substitutions by which G → A and C → T mutations could individually prevent cleavage of the site (table 2). These data were then fit to the following model:

Define R_i as the probability that site i is retained throughout the phylogeny, and let

$$R_i = k(1 - s)^{S_i}(1 - m)^{M_i}(1 - g)^{G_i}, \qquad (4)$$

where k is a constant ($0 < k \leq 1$), and s is the probability that a G → A or C → T conversion occurs somewhere in the phylogeny (causing loss of the site), the impact of which would be a silent substitution. Similarly, m is the probability for a missense substitution, and g for an intergenic substitution.

TABLE 2. Restriction enzymes and T7 characteristics. Categories are as follows. Wt sites: number of wild-type restriction sites whose mutation to a noncleaved state would have been detectable in our assays (hence less than or equal to the number of wild-type restriction sites). GC status: characteristics of wild-type sites according to the number of G → A and C → T mutations that could cause loss of the site, plus the consequences of those mutations on coding sequences (key is listed at the end of this paragraph). Losses: the number of assayable wild-type sites observed to be lost at least once in the phylogeny; key is listed below. S: number of 1-off sites for which a site gain would have resulted in a silent substitution in a known gene (coding regions from Dunn and Studier 1983). M: number of 1-off sites for which a site gain would have resulted in an amino-acid change in a known gene. G: number of 1-off sites in intergenic regions. Gains: number of sites in the molecule not cleaved in wild-type but which evolved to be cleaved at some later point in the phylogeny. Multiple gains at the same site (convergences) are listed as a single gain. Key for "GC status" and "Losses" categories: (in each of the following triplets, the first entry is the number of G → A or C → T mutations that would cause loss of the restriction site with a silent substitution in a T7 gene, the second entry is the number corresponding to missense mutations, and the third entry is the number corresponding to mutations in intergenic regions) A = (0,0,0), B = (1,1,0), C = (0,1,1), D = (2,0,0), E = (0,2,0), F = (0,0,2), G = (0,4,0), H = (2,2,0), I = (1,3,0).

				1-Off sites			
Enzyme	Wt sites	GC status	Losses	S	M	G	Gains
*Apa*LI	1	1I	0	2	6	2	0
*Ase*I	10	10A	0	31	17	8	7
*Bam*HI	0	—	—	9	4	1	3
*Bcl*I	1	1B	0	12	44	1	10
*Bgl*II	1	1B	0	5	15	5	4
*Bst*BI	5	4B, 1E	1B	8	21	1	2
*Bst*EII	1	1I	1I	17	58	9	1
*Bst*NI	1	1I	1I	6	1	2	0
*Cla*I	3	1B, 2E	1E	20	24	1	7
*Dra*I	6	6A	0	27	16	13	3
*Eco*NI	1	1H	1H	2	5	1	1
*Eco*RI	0	—	—	34	35	3	11
*Eco*RV	0	—	—	25	46	3	7
*Hind*III	0	—	—	19	23	1	4
*Hpa*I	17	15B, 1E, 1F	3B	12	49	9	6
*Kpn*I	3	3I	1I	4	21	1	1
*Mbo*I	4	2B, 1D, 1E	1D	141	233	42	88
*Mlu*I	1	1H	0	2	7	0	1
*Nco*I	1	1I	1I	7	15	3	0
*Nde*I	7	5E, 2C	1E, 1C	10	21	1	0
*Nhe*I	1	1I	1I	14	23	0	0
*Nsi*I	7	5E, 2F	1F	17	15	3	6
*Pst*I	0	—	—	5	5	1	1
*Pvu*I	0	—	—	5	12	0	3
*Pvu*II	3	3G	1G	13	4	0	1
*Sac*I	0	—	—	1	0	0	3
*Sal*I	0	—	—	1	9	0	4
*Sca*I	3	2D, 1E	2D	9	48	0	2
*Spe*I	2	1B, 1E	1E	6	51	4	3
*Ssp*I	6	6A	0	24	17	8	6
*Stu*I	1	1I	0	3	10	5	0
*Xba*I	3	1B, 1E, 1F	0	20	24	3	2
*Xho*I	0	—	—	1	3	0	0
*Xmn*I	12	1B, 6D, 5E	1B, 2D, 1E	10	16	1	8

The likelihood of a set of observations is thus

$$L(k, s, m, g) = \kappa \prod_i R_i^{r_i}(1 - R_i)^{1-r_i}, \quad (5)$$

where κ is a constant and r_i is unity if the site was retained throughout the phylogeny, zero if lost.

The four parameters were estimated from the data by maximum likelihood using a comprehensive, systematic search of the parameter space ($\hat{k} = 0.955$, $\hat{s} = 0.23$, $\hat{m} = 0.07$, $\hat{g} = 0.16$). The parameterized model was then used to predict the number of losses in each category and was found to offer an acceptable fit to the data by a χ^2 criterion. However, a model constrained so that $s = m = g \equiv x$ was also found to provide

an acceptable fit to the data, a fit that was not significantly worse than the fit provided by the full model ($\hat{x} = 0.120$, $\hat{k} = 0.955$ for the full phylogeny; respective values for the ingroup taxa alone are 0.1056 and 0.955). Thus, this simpler model of two parameters is preferred over the four-parameter model, and the success of the two-parameter model indicates that the probability of site loss is adequately approximated merely by counting the number of G and C bases in the recognition sequence regardless of their effects on T7 genes. The rate of change per G/C calculated from this model is approximately 1.1% per ingroup branch (3% per lineage), three times the rate calculated from the nucleotide data (calculated as $1 - [\hat{k}(1 - \hat{x})]^{1/14}$). We do not attach significance to the difference between these estimates, given the limited sampling base for each set of data; in particular, both sets have been obtained from too limited a set of genes/sites for us to be confident that either represents the average for the entire molecule.

Whereas the G/C bias provides a useful supplement to our understanding of wild-type site losses, it cannot account for the two cases in which a lost site was regained. If a site loss stems from a G → A or C → T change, regain of the site must result from the reverse transition. The fact that even two reversions were observed for such a rare class of mutation suggests that the reversions may have been strongly selected, but there are insufficient data to warrant serious evaluation of this possibility. We thus emerge from this part of the analysis with three partitions among restriction sites present in wild-type: sites whose recognition sequences contain zero, two, or four G or C bases. For example, all wild-type *Ase*I, *Dra*I, and *Ssp*I sites are lumped in the same partition, because the recognition sequences for all three enzymes lack G and C entirely.

Site Gains

Most of the data presented in Hillis et al. are gains of new restriction sites rather than losses of ancestral sites. In this section we will attempt to analyze site gains along the lines of our preceding analysis of site losses, but one difference is immediately apparent. With site losses, it was possible to begin with an a priori knowledge of all sites at which changes might occur, whereas for site gains, there is no exact counterpart. We therefore consider whether we can identify a set of defined positions to which all gains are confined.

The strong bias in mutation spectrum suggests one possibility: sequences that are a single G → A or C → T mutation from becoming a recognition sequence (1-off sites) for a particular restriction enzyme may account for nearly all gains. For example, the sequence GATC is cleaved by *Mbo*I and its isoschizomers. The sequence GGTC is one G → A mutation from GATC, and GACC is one C → T mutation from GATC. Might the set of GGTC and GACC sequences in T7 constitute a complete set of the sites at which all *Mbo*I gains occur? As *Mbo*I gains are otherwise possible only for changes other than G → A or C → T (which are uncommon) or for double mutations (which should be rare), these 1-off sites offer the best hope of identifying a set of T7 positions at which site gains may be expected.

Restriction-site mapping does not yield a precise location of each restriction site, so it is not possible to directly assess whether each gain occurred at a 1-off site. However, because of the large number of sites gained, it is possible to extract the relevant information statistically, with a model in which numbers of 1-off sites are used to predict numbers of site gains. Consequently, T7 was searched for the number of 1-off sites for each of the 34 enzymes used in the study. Each 1-off site was further classified according to whether it would cause a silent, missense (plus nonsense), or intergenic substitution (table 2). The data were analyzed with multiple regression for the following model:

$$\text{Number of sites gained} = c_0 + c_1 S + c_2 M + c_3 G + \text{error}, \quad (6)$$

where S, M, and G were the numbers of silent, missense (plus nonsense), and intergenic 1-off sites, respectively. Owing to multiple regression's sensitivity to extreme correlations among the independent variables, 1-off sites for *Mbo*I were excluded from the analysis, because they contributed exceptionally high values for S, M, and G: the three correlations ranged from 0.88 to 0.92 when *Mbo*I 1-off sites were included, but only 0.33 to 0.52 when *Mbo*I was excluded. The most important conclusions follow:

1. S is the best single predictor.—The regression with just S (M and G omitted, or $c_2 = c_3 = 0$) explained 37% of the variance ($\hat{c}_0 = 0.89$, $\hat{c}_1 = 0.20$, $P < 0.001$), whereas the regression with just M explained only 15% of the variance ($P < 0.03$), and the regression with just G explained less than 4% of the variance ($P < 0.29$).

2. M and G do not offer a significant improve-

ment.—The full model in (6) explained only 5% additional variance above that explained by S alone, and the increase afforded by both M and G or by either alone is not statistically significant.

3. The effects of S, M, and G are significantly heterogeneous.—The full model in (6) explains significantly more variance than does the reduced model requiring that S, M, and G all have the same effect ($c_1 = c_2 = c_3$), although only marginally so ($P < 0.05$).

4. The regression model does not adequately explain all the variance.—By a χ^2 criterion, the fitted full model leaves a significant amount of variance in site gains unexplained. This result is not surprising, as more than half the variance remains unexplained by the model. Table 2 even reveals one enzyme (*Sac*I) for which the number of sites gained actually exceeds the total number of 1-off sites.

We conclude that the rate of gain in restriction sites is likely affected by both the G/C bias in mutation spectrum and also by the impact of the mutation on T7 proteins (hence T7 biology is important), but the models do not adequately account for all of the observed site gains. This leaves us with two problems, (1) how to adapt the evolutionary parameters model to the case in which there is no a priori set of T7 sites at which gains may be expected, and (2) how to partition site gains into meaningful categories when applying them to the model. The first of these challenges has a straightforward solution. In the context of the evolutionary parameters model, it means that we cannot observe the class $P(0, 0)$ and the maximum likelihood formula in (3) must be corrected to avoid this bias:

$$L^*(c, r) = K \prod_{i,j \neq 0,0} \left\{ \frac{P(i, j)}{1 - [1 - P(0, 0)]^{14}} \right\}^{n(i,j)}. \quad (7)$$

Once a maximum-likelihood estimate $\hat{c}$ is obtained from (7), an effective number of "presites" present in wild-type may be estimated simply as N_e:

$$N_e = \frac{N_o}{1 - (1 - \hat{c})^{14}}, \quad (8)$$

where N_o is the observed number of sites gained in the phylogeny. This effective number of sites estimates the total number of sites in the phage that share the same c values, even though no conversions were observed at some of those sites in our single experimental phylogeny. In this sense, N_o is a sample of the larger set of sites (N_e) at which gains would be observed in an infinite number of replicates.

The second concern—that is, partitioning site gains into classes sharing similar conversion and reversion rates—is more difficult to address in a meaningful way. For the most part, each site gain (conversion) is likely due to a single base-pair change, and the conversion probability for that site is independent of the number of other sites in T7 DNA at which gains are observed. So there is no basis for partitioning enzymes according to the number of positions in T7 DNA at which gains are observed. Likewise, there is no basis for grouping enzymes according to the G/C content of the recognition sequence, since each site gain likely has a unique possible origin (given the low mutation rate). One may indeed choose to partition among enzymes with different G/C content on the grounds that *reversions* will be differentially affected, but reversions were so uncommon and heterogeneous in the data that this hypothesis lacks support.

Only one partition of the site-gain data was made: *Mbo*I sites were distinguished from all others. This distinction was made for two reasons. First, there were many *Mbo*I site gains, enabling meaningful estimates—nearly half of the site gains in the phylogeny were *Mbo*I sites. Second, the recognition sequence of *Mbo*I is GATC, a sequence which has important regulatory functions in the host including being the recognition sequence for the host's *dam* methylation system. Wild-type T7 DNA possesses an overwhelming deficiency in *Mbo*I sites—only six sites are present versus the nearly 160 expected in a random sequence of 40 kbp—and it is plausible that gains of GATC sequences in T7 carry special fitness consequences that do not accrue to other enzymes. If so, then the true evolutionary parameters for gained *Mbo*I sites may differ systematically from those of other enzymes.

Conversion and Reversion Rate Estimates in the Evolutionary Parameters Model

Once the partitions of restriction-site data were chosen, estimates of c and r (conversions and reversions) were obtained for each partition (table 3). The c and r values have been converted into probabilities of site losses and gains: for sites present in wild-type (class I), the probability of loss is the probability of conversion, and the probability of gain is the probability of reversion;

TABLE 3. Estimates from the evolutionary parameters model. Probabilities are calculated per site per ingroup branch. We have translated the probabilities of site conversions and reversions from the text as their corresponding probabilities of gains and losses: a loss means that a site formerly cleaved by a restriction enzyme changed to one that is not cleaved, and a gain is the reverse process. (There is an obvious order of precedence here, as a site absent from wild-type must be gained before it can be lost, and so forth.) N_o is the observed number of sites at which gains and/or losses were observed in the ingroup (excluding changes observed only in the outgroup), and N_e is the effective number of sites as defined in the text. (For class I sites, N_e was calculated directly from the number of known restriction sites with different G/C characteristics in wild-type T7; see table 2.) Numbers in parentheses are approximate 95% confidence intervals obtained by (nonparametric) bootstrapping 1000 replicates (subject to the discreteness of the distribution of bootstrap samples) except for class Ia. Confidence limits for class Ia losses are based on a binomial test; no estimate of (re)gain probabilities is provided, because in the absence of any losses, there is no opportunity to observe gains. Otherwise, nonparametric bootstrap trials were conducted separately for the class I and class II sites. For the class Ib (Ic) sites, we sampled with replacement among all Ib (Ic) sites and calculated gain and loss probabilities on each of the bootstrap samples. The lower confidence limit for the probability of a class Ib and Ic gain is zero, reflecting the fact that only 1 of the sites in each class observed to converse experienced a reversion, hence the bootstrap samples often failed to include any reversions (yielding an estimate of zero). For the class II sites at which conversions were observed, we randomly sampled among the sites and then partitioned them into IIa and IIb sites before computing probabilities of gains and losses. Some class IIb enzymes cleave sites that are also cut by *Mbo*I. For example, the internal 4 bases in the recognition sequences of *Bam*HI, *Bgl*II, *Pvu*II, and *Bcl*I are GATC (sites cut by these enzymes and some other sites were highlighted with an asterisk in fig. 3 of Hillis et al.). These sites were classified here as class IIa sites unless the non-*Mbo* enzyme revealed a different history of conversion and reversion at the site than did *Mbo*I, in which case the site was listed as both class IIa and IIb according to the different patterns (only three such sites).

Restriction-site class	Probability of gain	Probability of loss	N_o	N_e
Present in wild-type				
Ia (0 G,C)	—	0 (0.0, 0.001)	0	22
Ib (2 G,C)	0.045 (0.0, 0.141)	0.017 (0.010, 0.026)	15	65
Ic (4 G,C)	0.084 (0.0, 0.239)	0.049 (0.016, 0.098)	6	14
Absent in wild-type				
IIa (*Mbo*I)	0.033 (0.017, 0.052)	0.108 (0.052, 0.180)	80	213
IIb (non-*Mbo*)	0.013 (0.003, 0.026)	0.046 (0.008, 0.103)	73	436

these interpretations are reversed for sites absent in wild-type (class II). Confidence intervals of some estimates are large, but several conclusions do emerge:

1. The probability of class Ic losses is significantly greater than the probability of class Ib losses, which in turn is significantly greater than the probability of class Ia losses. This result is expected from the foregoing analysis of wild-type site losses, due to the different G/C content of the class Ia, Ib, and Ic enzymes. (Although confidence intervals for some of these probabilities overlap, a direct test of the difference by a nonparametric bootstrap analysis revealed statistical significance.)

2. The probability of class IIa (*Mbo*I) loss is significantly greater than the probability of class Ib loss. Both classes of enzymes have the same G/C content in their recognition sequences, so we must find some other basis for the difference. The difference may result from selection: class Ib losses constitute deviations from the wild-type sequence and may thus be disadvantageous, whereas losses of *Mbo*I gains may be advantageous: some are reversions to the wild-type sequence and, even for losses that do not revert to wild-type, they at least destroy a GATC sequence (recall that GATC is notably underrepresented in the wild-type sequence, suggesting that it is disadvantageous to T7). In line with this hypothesis, it is not difficult to understand that deleterious *Mbo*I sites could have been gained despite their disadvantage, as lineages were propagated from single plaques every 5 cycles.

3. The probability of a class IIa gain (*Mbo*I) is significantly greater than the probability of a class IIb gain. (Again, confidence intervals overlap, but a direct test of the difference reveals significance.) Most site gains probably result from single base-pair changes from the wild-type sequence, so the rates should be similar for both IIa and IIb enzymes. Why the difference? One possibility is that we have overestimated the rate of convergent *Mbo*I gains. The estimated rate of site gains will be inflated by errors in mapping that conservatively assign homology to neighboring but distinct sites. *Mbo*I may be more prone to this error than other enzymes because gained

*Mbo*I sites are more densely distributed on the genetic map of the experimental lineages. At the time this idea was proposed, we had no direct evidence for or against this possibility; subsequently, sequences obtained by CWC have revealed three clustered *Mbo*I sites that had been scored by us as only one site (position 0.8 in Hillis et al. 1992). So the higher apparent rate for *Mbo*I gains may indeed be an artifact of their high density and our conservative scoring procedure.

4. The probability of a class IIa (*Mbo*I) loss is significantly greater than the probability of a class IIa gain. This result may be anticipated from the simple fact that each gain likely has a single possible origin, whereas a loss has two possible origins (the G and C in the recognition sequence GATC). However, the excess rate of IIa losses is also consistent with point (2) above, in which we suggested that *Mbo*I site gains are at a selective disadvantage.

From these maximum-likelihood estimates, we calculated expected numbers of observations in each class and compared them to the actual $n(i, j)$. No significant heterogeneity was detected for any of the classes, so the null model of homogeneity of c and r within each of these classes is supported. This procedure of testing for heterogeneity can be extended to other subsets of the data (e.g., the numbers of site gains on primary, secondary, or terminal branches, and so on); we have performed a few such additional tests but again found no evidence for rejecting the null model.

The values in table 3 enable estimates of the per-base rate of evolution in the phylogeny, assuming that the values from these few restriction sites apply to other positions in the molecule as well (at least to G/C bases). For example, the loss rate for wild-type sites with 2 G/C in the recognition sequence was estimated as 0.017. Given that nearly all mutations appear to have targeted G or C, the rate of change per G/C is half 0.017, or 0.0085. The rate for class Ic is only slightly higher, at 0.012. (Above, the estimate obtained for the combined loss data was 0.011, using a slightly different model.) For sites absent from wild-type, the (per-site) probability of gain should reflect the per-base probability of change, which again likely applies only to G and C bases. However, as we have reason to doubt the validity of the *Mbo*I estimate, only the non-*Mbo*I estimate is honored here (0.013). Thus, estimates from the restriction site data are similar to each other (approximately 1% per G/C per ingroup lineage), but they are two- to threefold the value estimated from the nucleotide data (0.004 for G/C).

Extending the Data: Parametric Bootstraps

The objective of experimental molecular evolution as it applies to molecular systematics is to provide known phylogenies of biological taxa that enable direct tests of phylogenetic methods. Ideally, we would prefer to use empirical data sets at all levels of method evaluation—not just a few token data sets to demonstrate that reconstruction methods do indeed sometimes recover the correct topology, but sufficient data sets to indicate how often the methods are likely to succeed as well as to detect subtle differences in success rates among the methods. The dilemma that faces experimental phylogenetics is that a small number of empirical data sets provides little statistical power in evaluating methods, yet each data set requires months or years of laboratory work. It is thus desirable to have some means of maximizing the utility of individual data sets.

A common statistical approach in overcoming the limitations of a single data set is to resample the original data set many times and analyze each subsample as a different data set (techniques known as bootstrapping and jackknifing). The main drawback of this approach is that the different subsamples are not independent of each other, hence various biases arise in the distribution generated from the subsamples (e.g., Efron 1979, 1987; Nei 1991; Hillis and Bull 1993). An alternative approach is suggested by the analysis underlying table 3: a single data set is used to parameterize a model. In turn, this parameterized model is used to generate new, independent data sets. Conversion and reversion rate estimates are then obtained from each simulated data set, yielding a distribution of estimates that can be subjected to statistical tests. This approach is known as "parametric bootstrapping" (Efron 1985; Felsenstein 1988). Despite the similarity in name, it is fundamentally different from the traditional bootstrapping approach of merely resampling data (known as nonparametric bootstrapping).

To illustrate this approach, we return to the analysis of reconstruction methods in Hillis et al. (1992). In that study, we applied five reconstruction methods to the single, empirical data set. All methods yielded the correct topology, so

no differences were evident among the methods, and there was thus no basis for distinguishing this aspect of the methods' performances. We would like to know how often the methods are expected to recover the correct topology and whether they differ in this ability. This question can be addressed with parametric bootstrapping. Three methods of phylogenetic reconstruction will be compared: parsimony, neighbor-joining, and UPGMA. The latter two methods are known as distance methods because they estimate ancestry from pairwise estimates of genetic distances among taxa, whereas parsimony predicts relationships by minimizing numbers of evolutionary changes separating taxa.

One thousand data sets were simulated according to the evolutionary parameters model using the parameter estimates in table 3 and the topology in fig. 1 (the outgroup was treated as an independent three-branch lineage whose evolutionary rates were the same as for the outgroup lineages). Neighbor-joining succeeded in predicting a single tree with the correct topology in 991 of the trials, parsimony in 978 of the trials, and UPGMA in 973 of the trials. (In 15 of its 22 "incorrect" reconstructions, parsimony produced two best-fit trees, one of which was the correct one.) The most compelling conclusion is that all three methods are remarkably successful. We thus infer that all three methods would usually predict the correct topology in actual repetitions of the T7 study. Furthermore, these small differences are statistically heterogeneous for these sample sizes, so we also infer that the methods would exhibit consistent differences over a large number of replicates of the original study.

These simulations can also be used to evaluate branch length predictions. In our original study (Hillis et al. 1992), we noted that the correlation between actual and predicted branch lengths was highest for parsimony (0.89), intermediate for rate-insensitive distance methods such as neighbor-joining (0.86), and lowest for UPGMA (0.80). A somewhat unorthodox statistical approach was applied to the data, which suggested that these differences were statistically significant. Would the methods differ consistently in their ability to predict branch lengths under replications of the experimental phylogeny? One hundred parametric bootstrap simulations were conducted, and the correlation between predicted and actual branch length calculated for each method. Correlations were highest for parsimony (mean of 0.94), next highest for neighbor-joining (0.92), and lowest for UPGMA (0.87). The simulations thus produced the same ranking as the experimental data. These differences were also statistically significant: correlations were higher for parsimony over neighbor-joining in 98 of the trials, for parsimony over UPGMA in 95 of the 97 trials in which UPGMA predicted the correct tree, and for neighbor-joining over UPGMA in 83 of the 97 trials for which UPGMA predicted the correct tree.

Discussion

An earlier paper described an in vitro system using bacteriophage T7 as an experimental model of molecular evolution. That study constructed an experimental phylogeny and presented restriction site data with the explicit purpose of evaluating methods of phylogeny reconstruction. The objective here has been instead to describe molecular evolution in that system. A specific goal was to produce a model describing the rates at which restriction sites were gained and lost, these rates determining the rates of molecular divergence and convergence in the experimental phylogeny (the evolutionary parameters model).

The evolutionary parameters model describes rates of restriction site loss and gain for the T7 experimental phylogeny. To assist in explaining those rates, we used the following information:

1. The T7 wild-type DNA sequence, which provides the numbers and positions of existing restriction sites as well as positions that can mutate to become restriction sites with a single base-pair change (1-off sites)
2. The T7 coding regions, whereby we can assess the impact of a base-pair change on T7 genes (classified here as intergenic mutations, mutations causing a missense mutation, or mutations causing a silent substitution)
3. Limited DNA sequences of the experimental phylogeny, indicating that most base-pair changes were G → A or C → T
4. Recognition sequences of the restriction enzymes used

The main objective of this model was to quantify rates of restriction-site gains and losses along the branches of the experimental phylogeny, and more specifically, to determine whether those rates were affected by the data in 1–4. Specifically, do different classes of restriction sites evolve at different rates? And how much genetic detail is warranted in describing molecular evolution in this system?

The results can be summarized as follows:

1. Sites present in the wild-type sequence were lost at different rates, according to the number of G/C bases in the recognition sequences; no significant improvement in the model was afforded by the impact of these mutations on T7 genes. At present, therefore, the model of wild-type site loss lacks intricate detail about T7 biology.

2. Gains of new restriction sites are more difficult to explain. The total number of sites gained in T7 DNA varied widely among enzymes, a phenomenon explained only partly by differences in the numbers of 1-off sites and the impact of those changes on T7 genes. In addition, the rate of convergent site gains was higher for *Mbo*I than for other enzymes, and we do not necessarily have a satisfactory explanation for this difference.

3. Subsequent losses of new *Mbo*I sites occurred at a significantly higher rate than did losses of wild-type sites with similar G/C content. Selection for the wild-type sequence is implicated.

The results thus reveal various levels of complexity in restriction site evolution in the T7 experimental phylogeny. Some aspects of the data seem to be adequately represented by purely random processes (e.g., relative rates of wild-type site loss, given the G/C bias in mutations), whereas other dimensions of the data reflect additional complications. The complexity required to explain these results is greater than that assumed in simple models of restriction site evolution (e.g., Li 1986), but it is not nearly as great as we might have expected. Of course, a more detailed understanding of T7 biology might well lead to further revelations of complexity—a deeper understanding of the relationship between amino acid sequence and protein function would undoubtedly enable us to subdivide missense mutations into classes of different selective impact to yield a meaningful improvement in explanatory power—and further revelations of complexity.

Parametric Bootstrap

A major objective of studies in experimental molecular evolution is to provide empirical data sets for unambiguous tests of evolutionary models. The experimental approach is labor-intensive, however, and it is not feasible to generate empirical data with near the ease of computer simulations. Estimation of evolutionary rates from one data set enables the investigator to reap the benefits of both approaches: simulations parameterized with the estimates from the empirical data can then be used to generate thousands of independent pseudoreplicates bearing characteristics of the empirical study. They thereby enable statistical tests not possible by merely resampling the original data (nonparametric bootstrapping).

The parametric bootstrap offers at least three applications:

1. Replications of data to evaluate methods of analysis, as here (Felsenstein 1988)
2. Tests of phylogenetic estimates (Felsenstein 1988); when a phylogenetic reconstruction is attempted, evolutionary parameters may be estimated from the best-fit tree and used to program replicates to evaluate "confidence" in the reconstruction
3. Extensions of data to new designs; for example, the observed evolutionary parameters could be extrapolated to new topologies and longer-term studies

The first two of these applications provide direct parallels to traditional (nonparametric) bootstrapping, and it may seem that the parametric bootstrap offers no advantages. But there is an important distinction: nonparametric bootstrapping is subject to various biases that the parametric bootstrap avoids. For example, suppose that a single sample of data is obtained, and the estimate for some statistic is $\hat{X}$. Using nonparametric bootstrapping, we would reject the hypothesis that the true population value of X is zero only if 97.5% of the bootstrap replicates of $\hat{X}$ were less than zero or if 97.5% of them were greater than zero (when using a two-tailed test): because of the bias inherent in the resampling process, nonparametric bootstrapping is used to obtain estimates of the variance but not the mean.

The parametric bootstrap works in the following way. The data are fit to some underlying model; this model is parameterized with the estimates and used to generate additional data sets. Pseudoreplicate estimates of $\hat{X}$ are obtained from these samples and may be used to test any of various hypotheses. In this case, we have a much more powerful statistical basis for testing hypotheses, because each replicate can be treated as an independent data set. For example, we would reject the hypothesis that the "true" pseudo-replicate value of X is zero if the numbers of positive and negative trial values merely differed significantly from 1:1. However, the models be-

ing tested differ in the two cases: the model underlying a parametric bootstrap is an a priori model whose few parameters are derived from the data, whereas the model underlying nonparametric bootstrapping is merely a sampling of the original data set.

In a comparison of three reconstruction methods (neighbor-joining, parsimony, and UPGMA), parametric bootstrapping established two results. First, each method predicted the correct phylogeny in at least 97% of the trials; successful reconstruction would therefore be expected in most empirical replicates of the T7 phylogeny. Second, the methods revealed significant differences in ability to predict branch lengths, with parsimony outperforming neighbor-joining, and both of these methods outperforming UPGMA.

Prospectus

This paper has offered a study of molecular evolution that combines biological and genetic details in a phylogenetic context. The analysis enjoys the advantage that the phylogeny is known a priori, so it becomes possible to calculate evolutionary parameters in the absence of uncertainties about phylogeny. The experimental organism is not of special interest by itself, so the value of the study must rest on its generality to other systems. Yet, generalities are not immediately apparent precisely because of the incorporation of genetic detail. The irony is that, by increasing the level of molecular resolution, we have discovered features that render the experiment unique, hence less general. The utility of the study is further compromised by our use of a design to enhance the rate of mutation at the expense of phage fitness.

On further reflection, however, there are two respects in which the T7 system may generalize. First the bias in mutation spectrum induced by the mutagen (nitrosoguanidine) bears a surprising resemblence to the mutations observed in some other taxa. A strong G → A and C → T bias has been observed in eukaryotic pseudogenes, presumably reflecting a similar bias in the mutation spectrum (Gojobori et al. 1982; Li et al. 1984); likewise, the predominant class of mutation in the HIV virus is G → A (Vartanian et al. 1991; Moriyama et al. 1991). Second, the T7 system serves as a model for any other system with unique features. As more is discovered about the molecular genetic basis of change in different organisms, the notion of a uniform process of molecular evolution applying to most taxa is vanishing rapidly. Every taxon has its own evolutionary history as regards mutations, population structure, and selection. A goal of evolutionary biology is to discover these unusual features of a taxon's evolutionary history, which in most cases is possible only through inference based on phyletic comparisons. An experimental phylogeny offers an uncompromising proving ground for methods that purport to be able to recover the various nuances of evolution history.

Molecular analysis of experimental evolution appears to be a feasible endeavor. The accumulated rate of base-pair change achieved in this study (1%–3% of G/C bases from wild-type) was adequate to achieve meaningful levels of restriction site change and to compare different models of molecular evolution with respect to the role of selection. The potential exists for achieving rates five times those observed here and to provide empirical tests of many compelling questions about evolution at the molecular and higher levels.

Acknowledgments

We thank C. Pease for advice on the maximum-likelihood model, E. Holmes for references, and M. Riley and R. Lenski for comments on the manuscript. J. Huelsenbeck discovered the references on parametric bootstrapping and thereby avoided our inventing a new term for it. This work and C.W.C. were supported by National Science Foundation grant DEB 9106746 to D.M.H., J.J.B., and I.J.M.

Literature Cited

Atchley, W. R., and W. M. Fitch. 1991. Gene trees and the origins of inbred strains of mice. Science 254:554–558.

Baum, B. R. 1984. Application of compatibility and parsimony methods at the infraspecific, specific, and generic levels in *Poaceae*. Pp. 192–220 *in* T. Duncan and T. F. Stuessy, eds. Cladistics: perspectives on the reconstruction of evolutionary history. Columbia University Press, New York.

Beck, P. J., S. Gonzalez, C. L. Ward, and I. J. Molineux. 1989. Sequence of bacteriophage T3 DNA from gene 2.5 through 9. Journal of Molecular Biology 210:687–701.

Doolittle, R. F. 1990. Molecular evolution: computer analysis of protein and nucleic acid sequences. Methods in enzymology, vol. 183. Academic Press, New York.

Dunn, J. J. and F. W. Studier. 1983. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of genetic elements. Journal of Molecular Biology 166:477–535.

Efron, B. 1979. Bootstrapping methods: another look at the jackknife. Annals of Statistics 7:1–26.

———. 1985. Bootstrap confidence intervals for a class of parametric problems. Biometrika 72:45–58.

———. 1987. Better bootstrap confidence intervals. Journal of the American Statistical Association 82: 171–185.

Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. Annual Review of Genetics 22:521–565.

Fitch, W. M., and W. R. Atchley. 1987. Divergence in inbred strains of mice: a comparison of three different types of data. Pp. 203–216 *in* C. Patterson, ed. Molecules and morphology in evolution: conflict or compromise? Cambridge University Press, Cambridge.

Fitch, W. M., and J. Ye. 1991. Weighted parsimony: Does it work? Pp. 147–154 *in* M. M. Miyamoto and J. Cracraft 1991.

Gojobori, T., W.-H. Li, and D. Graur. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. Journal of Molecular Evolution 18: 360–369.

Hillis, D. M., and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology 42:182–192.

Hillis, D. M., and C. Moritz. 1990. Molecular systematics. Sinauer, Sunderland, Mass.

Hillis, D. M., J. J. Bull, M. E. White, M. R. Badgett, and I. J. Molineux. 1992. Experimental phylogenetics: generation of a known phylogeny. Science 255:589–592.

Horsfall, M. J., A. J. E. Gordon, P. A. Burns, M. Zielenska, G. M. E. Vander Vliet, and B. W. Glickman. 1990. Mutational specificity of alkylating agents and the influence of DNA repair. Environmental Molecular Mutagenesis 15:107–122.

Jin, L., and M. Nei. 1991. Relative efficiencies of the maximum-parsimony and distance-matrix methods of phylogeny construction for restriction data. Molecular Biology and Evolution 8:356–365.

Li, W.-H. 1986. Evolutionary change of restriction cleavage sites and phylogenetic inference. Genetics 113:187–213.

Li, W.-H., J. Sourdis, and P. M. Sharp. 1987. Reconstruction of phylogenetic trees and estimation of divergence times under nonconstant rates of evolution. Cold Spring Harbor Symposium on Quantitative Biology 52:847–856.

Li, W.-H., C.-I. Wu, and C.-C. Luo. 1984. Nonrandomness of point mutations as reflected in nucleotide substitution in pseudogenes and its evolutionary implications. Journal of Molecular Evolution 21:58–71.

Miyamoto, M. M., and J. Cracraft, eds. 1991. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

Moffat, B. A., J. J. Dunn, and F. W. Studier. 1984. Nucleotide sequence of the gene for bacteriophage T7 RNA polymerase. Journal of Molecular Biology 173:265–269.

Moriyama, E. N., Y. Ina, K. Ikeo, N. Shimizu, and T. Gojobori. 1991. Mutation pattern of human immunodeficiency virus genes. Journal of Molecular Evolution 32:360–363.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

———. 1991. Relative efficiencies of different tree-making methods for molecular data. Pp. 90–128 *in* M. M. Miyamoto and J. Cracraft 1991.

Sidow, A., and A. C. Wilson. 1991. Compositional statistics evaluated by computer simulations. Pp. 129–146 *in* M. M. Miyamoto and J. Cracraft 1991.

Studier, W. F. 1980. The last of the T phages. Pp. 72–78 *in* N. H. Horowitz and E. Hutchings, Jr., eds. Genes, cells, and behavior: a view of biology fifty years later. W. H. Freeman, San Francisco.

Vartanian, J.-P., A. Meyerhans, B. Asjo, and S. Wain-Hobson. 1991. Selection, recombination, and G → A hypermutation of human immunodeficiency virus type 1 genomes. Journal of Virology 65:1779–1788.

White, M. E., J. J. Bull, I. J. Molineux, and D. M. Hillis. 1991. Pp. 935–943 *in* E. Dudley, ed. Proceedings of the Fourth International Congress of Systematic and Evolutionary Biology. Dioscorides Press, Portland, Oreg.