Exceptional Convergent Evolution in a Virus

J. J. Bull,*,† M. R. Badgett,* H. A. Wichman,‡ J. P. Huelsenbeck,§ D. M. Hillis,*,† A. Gulati,* C. Ho* and I. J. Molineux†,**

**Department of Zoology, **Department of Microbiology, †Institute of Cellular and Molecular Biology, University of Texas, Austin, Texas 78712, ‡Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844 and §Department of Integrative Biology, University of California, Berkeley, California 94702*

ABSTRACT

Replicate lineages of the bacteriophage ϕX 174 adapted to growth at high temperature on either of two hosts exhibited high rates of identical, independent substitutions. Typically, a dozen or more substitutions accumulated in the 5.4-kilobase genome during propagation. Across the entire data set of nine lineages, 119 independent substitutions occurred at 68 nucleotide sites. Over half of these substitutions, accounting for one third of the sites, were identical with substitutions in other lineages. Some convergent substitutions were specific to the host used for phage propagation, but others occurred across both hosts. Continued adaptation of an evolved phage at high temperature, but on the other host, led to additional changes that included reversions of previous substitutions. Phylogenetic reconstruction using the complete genome sequence not only failed to recover the correct evolutionary history because of these convergent changes, but the true history was rejected as being a significantly inferior fit to the data. Replicate lineages subjected to similar environmental challenges showed similar rates of substitution and similar rates of fitness improvement across corresponding times of adaptation. Substitution rates and fitness improvements were higher during the initial period of adaptation than during a later period, except when the host was changed.

MOLECULAR sequences are routinely used to infer evolutionary histories across the entire spectrum of life. Underlying these uses of molecular data is the assumption that convergent evolution is statistically unimportant—that identical, independent changes in different lineages are not common enough to obscure the true historical signal (DOOLITTLE 1994; HEDGES and MAXSON 1996). However, the impact of convergent molecular evolution on phylogenetic reconstruction is difficult to evaluate because the true phylogeny is almost never known and thus cannot be used to identify errors in the phylogeny estimated from molecular data. This potential effect on phylogenetic reconstruction is complicated by the fact that not enough is known about the evolutionary causes of convergence to predict where it might be a problem.

A rarity of convergence in molecular evolution is not obviously controversial, as it can be reconciled with both the selectionist and neutralist paradigms. Under neutrality, most substitutions are stochastic, so convergence is merely accidental. If instead, most substitutions are considered adaptive, convergence is likewise unexpected unless organisms face the same selective environment, and then only if they have few alternative pathways of adaptation to that environment.

Experimental systems provide a direct way of assessing the magnitude of convergent evolution. Most obviously, they avoid the circularity of using inferred phylogenies to infer properties of molecular evolution that themselves influenced the reconstruction. Beginning and end points of the evolutionary process can be assessed directly to avoid uncertainty about which changes occurred. Here we adapted the bacteriophage ϕX 174 to grow at high temperature in a chemostat on either of two hosts. Our objective was to assess patterns of molecular evolution during adaptation to a generalized form of environmental selection (as opposed to a drug or specific host mutation). The data provide rates of substitution during adaptation, as well as rates of convergence, and allow us to assess their impact on phylogenetic reconstruction.

Corresponding author: J.J. Bull, Department of Zoology, University of Texas Austin, TX 78712. E-mail: bull@bull.zo.utexas.edu. Inquiries regarding this article should be addressed to individual authors.

MATERIALS AND METHODS

Phage and hosts: A wild-type isolate of ϕX 174 was used as the ancestor for these studies. The complete genome sequence of this isolate, designated ancestor A, was found to differ from SANGER *et al.*'s (1977; GenBank V01128) viral strand sequence of ϕX 174 at five positions: 4784 (we observed T), 4518 (A), 2811 (T), 1650 (G), and 833 (A). The hosts used for phage propagation were *Escherichia coli* C and the Type I restrictionless (*hsd*), ϕX 174^S *Salmonella typhimurium* LT2 strain IJ750 [*xyl-404 metA22 metE551 galE719 trpD2 ilv-452 hsdLT6 hsdSA29 hsdSB121 fla-66 rpsL120* H1-b H2-e *nix*] [provided by (as MS3849) M. M. SUSSKIND].

Chemostat: A chemostat consisting of two 100 × 15 mm glass test tubes was used to select phage populations (DYKHUIZEN 1993). LB broth with 0.005% antifoam B (Sigma A5757) was pumped continuously into the first tube, which maintained host cells at a density suitable for phage growth; phage were absent from this "cell" tube. Liquid from the cell tube was drawn continuously into the second tube, which contained phage (the "phage" tube). The volume of liquid in the phage tube was maintained at 1–2 ml, with a flow through rate of 6–10 ml/hr; thus the volume of the phage

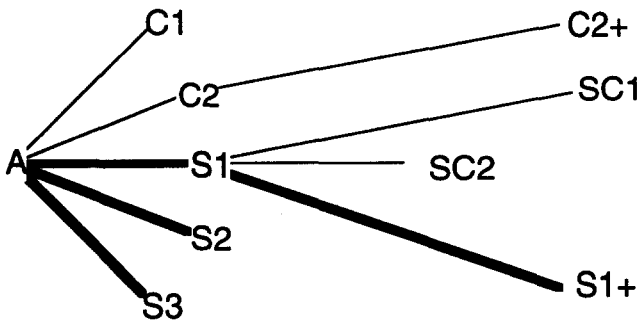

FIGURE 1.—Evolutionary history of the ϕX 174 lineages. Branch lengths reflect the duration of propagation (all lineages were grown for either 11 or 22 days); letters indicate isolates. Hosts were either *S. typhimurium* (thick lines) or *E. coli* C (thin lines).

tube was replaced on the order of 100 times per day. Both tubes were aerated by bubbling with filtered air. Both tubes were immersed in a water bath whose temperature was maintained 0.7° above the designated chemostat temperature, to compensate for the observed cooling effect of the aeration. Chemostat tubes, cells, and media were changed at least every 3rd day or as needed. Samples from the phage tube were extracted over chloroform and used to reinoculate the phage tube of the chemostat when the tubes were changed.

Adaptation to temperature and lineage structure: Chemostat populations of the phage were adapted gradually to growth at high temperature (43.5°). Starting with ancestor A, the chemostat population of phage (generally in excess of 10^7 phage) was grown at 38° for 1 day, and the temperature was elevated 1° daily until completing 1 day at 43°. The temperature was finally elevated to and maintained at 43.5° for an additional 5 days, yielding a total propagation period of 11 days. From the 11th day population, one or more clonal phage isolates were obtained; fitnesses of these isolates were determined and an isolate was sequenced completely. Five independent lineages were propagated in this manner, all starting with separate samples of ancestor A (Figure 1). Two of these five independent lineages were propagated on *E. coli* C throughout the process (designated C1 and C2), and three were propagated on *S. typhimurium* (designated S1, S2, S3); we designate these five lineages as the primary lineages. C1, C2, S1, and S2 were all initiated from the same clonal stock of A, whereas S3 was initiated from a separate plaque of A (not sequenced).

"Extended" lineages were propagated at 43.5° from two of these evolved isolates (Figure 1). Isolate C2 was the ancestor for the C2+ lineage, propagated on *E. coli* an additional 22 days. Isolate S1 was the ancestor for three extended lineages: S1+, grown 22 days on *S. typhimurium*; SC1, grown 22 days on *E. coli*; and SC2, grown 11 days on *E. coli*. (S1+ and SC1 were propagated from the original isolated plaque of S1; SC2 was propagated from a separate plaque of S1). The lineages leading from A to SC1 and SC2 thus experienced both *E. coli* and *S. typhimurium* as hosts. The complete genomes of isolates from each of these extended lineages were also sequenced, and the fitnesses of the isolates were determined.

Fitnesses were determined for the isolates and often also for the culture from which the isolate was obtained. The isolate and whole-culture values were indistinguishable except for lineage S2; the fitness of the first isolate chosen was vastly below that of the culture's fitness, so a second isolate, whose fitness was indistinguishable from that of the whole culture, was then chosen to represent this lineage.

Fitness assays: Viral fitness was estimated as the $\log_2$ of the increase in phage concentration per hour. Hosts were grown at 42° with shaking in LB to a density of $\sim 10^8$ cells per ml. At this point (t_0), phage were added to a concentration of $10-10^3$/ml, the mixture was grown for 1 hr with shaking, and phage titers were then determined by plating. A temperature of 42° was chosen for these assays, compromising between the need to use a high enough temperature to inhibit growth of the ancestral phage but not so high that the host would be severely impaired. Fitness was sensitive to the host used in the assay and, where fitnesses are presented, the host used is indicated. Fitnesses are quantified as $\log_2$ of the ratio of phage concentration at 60 min divided by the phage concentration at t_0. These numbers can thus be thought of as the number of doublings of phage concentration per hour; this is a simple measure of the growth capacity of an organism and does not assume reproduction by fission: 100 progeny represents 6.6 doublings ($2^{6.6}$), for example.

Sequencing: Sequences were determined by the dideoxy chain termination method using the single-stranded viral DNA as template. Most sequences were obtained from a core facility at the University of Texas using an ABI 377 automated sequencer, although some of the earlier sequences were obtained by manual sequencing. Sequences were compared to the complement of the published viral strand sequence (SANGER *et al.* 1977), and any deviations from this published sequence were verified for authenticity. Ambiguities were resequenced, if necessary using a different primer or different DNA extraction.

Protection against contamination: The conclusions of this study are sensitive to contamination: contamination can create identical changes in otherwise unrelated lineages and thus give the false impression of convergence. To minimize the possibility of contamination, we took the precaution of adapting lineages one at a time or, when two chemostats were running concurrently, the two chemostats were isolated in different rooms and used different hosts. In a further attempt to evaluate the role of contamination in generating convergence in our study, the S3 lineage was evolved at the University of Idaho (by H.A.W.), whereas all other lineages were evolved in Texas. The levels of convergence between S3 and the other *S. typhimurium*-evolved lineages were similar, so we do not consider contamination a plausible explanation for the results.

Reconstruction: Phylogeny was estimated using the maximum likelihood criterion assuming a HKY85+Γ model of DNA substitution (HASEGAWA *et al.* 1985; YANG 1993, 1994). This model of substitution allows for unequal base composition, a transition:transversion rate bias, and among-site rate variation. Two analyses were performed, using different numbers of taxa. The first analysis included complete genome sequences from the six isolates A, C1, C2, S1, S2, and S3. The second analysis used all 10 isolates but omitted from all isolates the 27 bases that were deleted in lineage SC1.

The impact of convergence was analyzed using a likelihood ratio test. Under the null hypothesis, the true tree was assumed. The likelihood (L_0) of this tree was calculated while forcing the interior branches to equal zero, to represent a star phylogeny rather than a strictly bifurcating one. (Our maximization of the likelihood when internal branches are set to zero differs from FELSENSTEIN's 1993 formulation of this test.) The program PAML 1.2 was used to calculate the likelihood under the null hypothesis, which required treating some taxa as true ancestors (YANG 1996a). The likelihood under the alternative hypothesis (L_1) is more general in that it allows interior branches to be independently estimated. The alternative hypothesis introduces, respectively, four and eight additional parameters for the first and second analyses. PAUP* version 4.0.d53 (SWOFFORD 1997) was used to calcu-

late likelihoods under the alternative hypothesis. The likelihood ratio

$$\Lambda = \max[L_0]/\max[L_1]$$

was calculated for both analyses. Because the null hypothesis is a special case of the alternative hypothesis, $-2\log(\Lambda)$ should be asymptotically distributed as a chi-square with 4 and 8 d.f. for the first and second analyses, respectively.

An alternative means of determining the distribution of $-2\log\Lambda$ under the null hypothesis is Monte Carlo simulation (or parametric bootstrapping). The parametric bootstrap procedure takes maximum likelihood estimates of parameters under the null hypothesis and constructs many simulated replicates of the same size as the original. For each simulated data set, $-2\log\Lambda$ is calculated. The proportion of time that the observed value of $-2\log\Lambda$ is observed or exceeded in the simulated replicates represents the significance level of the test. We used the parametric bootstrap procedure to determine the significance level for both analyses (a chi-square distribution poorly fits the distribution determined using parametric bootstrapping).

Terminology: We observed many parallel substitutions (LI 1997): the same nucleotide substitution (*e.g.*, A→G) at the same site occurring in two or more lineages. We also observed reversions from a derived state back to the wild-type, ancestral state. The combined magnitudes of these types of substitutions confounded phylogenetic reconstruction. We collectively refer to these parallelisms and reversions as convergences, after DOOLITTLE (1994), because they have the effect of causing lineages to appear more closely related than they are in reality. This use of "convergence" is broader than that of many authors.

Parallel substitution rates: Rates of substitutions were estimated for the five primary lineages in a way that accounts for parallel substitutions and, in particular, in a way that allows extrapolation to an arbitrary number of lineages. This model assumes that similarities between replicate lines arise by chance. Consider one site in the genome: let c be the probability of a substitution existing at that site in the isolate taken at the end of the propagation. Since we have five independent lineages, the six terms in the expansion

$$[c + (1 - c)]^5 = c^5 + 5c^4(1 - c) + 10c^3(1 - c)^2 + 10c^2(1 - c)^3 + 5c(1 - c)^4 + (1 - c)^5,$$

respectively, give the probability of observing substitutions in 5, 4, . . . , 1, and 0 of the five lineages at this site. If c is taken as a per-site constant rate over a genome of N sites, then Nc^5 is the expected number of sites showing the same substitution in all five lineages, and so on; for an arbitrarily chosen pair of lineages, the value c^2 is the expected fraction of sites with parallel changes.

The model may be extended to include invariant sites—sites unable to change (as opposed to sites able to change but which did not). Let $1 - \nu$ be the fraction of sites that are treated as invariant; the probability of observing a site with no substitutions in any of the five lineages is then

$$P(0) = (1 - \nu) + \nu(1 - c)^5,$$

and the probability of observing a site substituted in just one of the five lineages is

$$P(1) = \nu 5c(1 - c)^4,$$

and so on. If n_i is the number of sites observed to be substituted in i of the five lineages, maximization of the log of the likelihood

$$k\Pi P(i)^{n_i}$$

over ν and c will yield maximum likelihood estimates of both parameters. Using the data for the five primary lineages in Table 1 and a genome size of 5386 bases, estimation of these two parameters was done numerically, using a systematic search of the parameter space, starting with a large grid size and then progressively reducing the grid size around the successive optima.

A similar maximum likelihood estimation procedure was done for T7 sequence data (1091 sites from HILLIS *et al.* 1994), using the equations of BULL *et al.* (1993) but introducing an invariant-sites variable ν as above. However, for comparison to the ϕX 174 data, the value of c from the T7 study needs to be corrected upward by ~3.76, because each T7 branch constituted only about 266 doublings, whereas a ϕX 174 lineage experienced closer to 1000 doublings.

RESULTS

We independently adapted five "primary" lines from the same ancestor to the chemostat (on either of two hosts); an additional four lines were grown further from two isolates obtained from the primary lineages. The high temperature chemostat represents a specific and no doubt novel environment for the phage (GODSON 1978; DOWELL 1980). Our interest lies in both the fitness improvement that occurred during growth (reproduction) in this environment as well as in the nucleotide-level changes which accompanied the fitness changes.

Fitness: Fitness improved substantially in all evolved lines (Figure 2). The initial fitness of ancestor A was low on both hosts at 42°, whereas, by the 11th day, fitnesses of isolates within each primary lineage had improved at least 4000-fold (~12 doublings/hr). The 22-day extended lineages S1+ and C2+ showed slight but significant improvements over their respective ancestors (in t-tests, fitnesses of isolates S1+ and C2+ are significantly higher than the fitnesses of S1 and C2, respectively; $P < 0.05$). The average fitness increase along the two primary lineages A→S1 and A→C2 is significantly greater than the average increase from S1→S1+ and C2→C2+, indicating a slowdown in the rate of fitness improvement ($P < 0.01$, even ignoring the fact that these extensions were conducted for 22 instead of 11 days). Extended lines SC1 and SC2 showed major fitness increases over their ancestor S1 on host *E. coli*, but these extensions were carried out on a new host. It appears that fitness improvement is greater early than late except when the phage is placed in a new environment.

Substitutions—overall rates and convergences: Rates of nucleotide substitution showed a pattern similar to rates of fitness change (Table 1). All isolates differed from A by at least 12 and up to 26 substitutions, and a 27-base deletion was observed in SC1. [The same deletion was observed previously and did not appear to have an obvious effect on viral replication (MÜLLER and WELLS 1980), but caused a 50% increase in eclipse rate (INCARDONA and MÜLLER 1985).] On a per-day basis, the number of substitutions was relatively uniform

TABLE 1

Nucleotide substitutions in the experimental lineages

Site	No.	Gene	Protein	C1	C2	C2+	SC1	SC2	S1	S1+	S2	S3
3932	1	—	—	a	a	a	G	a	a	a	a	a
3974	1	—	—	c	c	c	T	c	c	c	c	c
4110	4	A	m	c	c	T	T	T	T	T	T	T
4122	3	A	m	A	A	A	A	A	A	A	g	g
4168	2	A	m	a	a	G	G	G	G	G	a	a
4420	3	A	m	a	a	a	A	C	C	C	a	C
4623	2	A, A*	m, m	g	A	G	g	g	g	g	g	g
4657	1	A, A*	m, m	a	a	G	a	a	a	a	a	a
4700	2	A, A*	m, m	t	t	t	T	G	G	G	t	t
4751	1	A, A*	s, s	G	a	a	a	a	a	a	a	a
4804	1	A, A*	m	c	c	c	T	c	c	c	c	c
4805	1	A, A*	s, s	a	a	a	C	a	a	a	a	a
5185	1	A, A*, B	m, m, s	T	c	c	c	c	c	c	c	c
5268	1	A, A*, B	m, m, m	a	a	a	a	a	a	a	C	a
5365	1	A, A*, B	m, m, s	t	t	t	C	C	C	C	t	t
22	1	A, A*, B	s, s, m	g	g	g	T	g	g	g	g	g
28	1	A, A*, B	s, s, m	g	A	A	g	g	g	g	g	g
31	2	A, A*, B	s, s, m	G	a	a	G	a	a	a	a	a
319	4	C	m	T	g	g	G	T	T	T	g	T
323	1	C	m	a	a	a	G	a	a	a	a	a
384	1	C	m	g	g	g	T	g	g	g	g	g
396	1	D	m	c	c	c	c	G	c	c	c	c
500	4	D	s	g	g	g	G	A	A	A	A	T
533	1	D	s	c	c	c	c	T	c	c	c	c
563	1	D	s	c	c	c	c	c	c	T	c	c
756	4	D, E	m, s	C	C	C	C	C	C	C	C	t
795	2	D, E	m, s	a	a	a	G	G	a	a	a	a
872	1	J	m	g	T	T	g	g	g	g	g	g
903	1	J	m	g	g	g	g	g	g	g	g	A
905	2	J	m	g	g	g	A	g	g	g	A	g
986	1	—	—	a	a	a	D	G	G	G	a	a
1025	3	F	m	A	g	g	A	A	A	A	g	A
1111	1	F	s	c	c	T	c	c	c	c	c	c
1159	1	F	s	t	t	t	t	t	t	C	t	t
1180	1	F	s	t	t	t	t	t	t	t	t	C
1250	1	F	m	a	a	a	a	a	a	a	G	a
1305	4	F	m	g	g	g	G	G	A	A	A	g
1533	2	F	m	c	T	T	c	c	c	c	T	c
1565	2	F	m	a	a	a	G	a	a	G	a	a
1613	2	F	m	a	a	a	T	T	T	T	a	T
1618	1	F	m	c	c	c	c	c	c	c	c	G
1702	1	F	s	t	t	t	t	t	t	C	t	t
1727	4	F	m	T	T	T	c	T	c	c	T	c
1809	1	F	m	a	a	a	a	a	a	G	a	a
2009	5	F	m	t	t	t	T	T	A	A	A	A
2167	5	F	m	t	t	t	T	T	G	G	G	G
2410	1	G	m	a	a	a	a	a	a	a	a	G
2591	2	G	m	C	t	t	t	C	t	t	t	t
2594	1	G	m	a	a	G	a	a	a	a	a	a
2949	1	H	m	g	g	g	A	g	g	g	g	g
2964	1	H	m	g	g	T	g	g	g	g	g	g
2979	1	H	m	g	A	A	g	g	g	g	g	g
2980	4	H	m	T	T	T	c	T	c	c	T	c
3011	1	H	m	a	a	a	a	a	a	T	a	a
3013	1	H	m	c	c	c	T	c	c	c	c	c
3042	1	H	m	c	c	c	T	c	c	c	c	c
3061	1	H	m	c	c	c	T	c	c	c	c	c
3062	1	H	m	t	t	t	G	t	t	t	t	t
3111	1	H	m	A	g	g	g	g	g	g	g	g

TABLE 1

Continued

Site	No.	Gene	Protein	C1	C2	C2+	SC1	SC2	S1	S1+	S2	S3
3120	3	H	m	c	T	T	c	T	c	c	T	c
3166	1	H	m	c	c	c	c	c	c	c	c	T
3336	1	H	m	g	g	A	g	g	g	g	g	g
3337	2	H	m	c	c	c	T	T	c	c	c	c
3378	4	H	m	C	C	C	C	C	a	a	a	a
3584	1	H	m	t	t	t	t	G	t	t	t	t
3826	1	H	m	a	a	a	G	a	a	a	a	a
3847	1	H	m	c	A	A	c	c	c	c	c	c
3873	1	H	m	g	g	g	g	A	g	g	g	g

Nucleotide substitutions are given for the ϕX174 viral strand. The first column identifies the nucleotide position, as defined by SANGER *et al.* (1977). Differences between the sequence of the ancestor and the published sequence are indicated in METHODS AND MATERIALS and are not listed here. The second column provides the number of independent changes observed at the site, the third column indicates the gene, and the fourth column whether the substitution is missense or silent in each gene. Columns 5–13 give the bases at each position in the nine different isolates. Lower case bases are those of the wild-type sequence. Where changes occurred during the evolution of a lineage, bases are underscored. For extended lineages (C2+, SC1, SC2, and S1+), bases that were substitutions in their primary-lineage ancestor and were maintained during the extension are shown in capital Roman letters, and new changes (including changes back to the ancestral type) are underscored. A D is shown for position 986 in SC1 because that position was lost as part of a deletion of bases 965–991. Deleted bases are considered uninformative and thus not listed as substitutions. No other deletions were observed, and this deletion was not observed in other isolates.

across the five primary lineages and even on the extended lineages SC1 and SC2, ranging from 12 to 15 per 11 days. Extended lineages S1+ and C2+ showed a significantly lower rate of substitution from that of the primary lineages (six and seven changes per 22 days, $P < 0.01$).

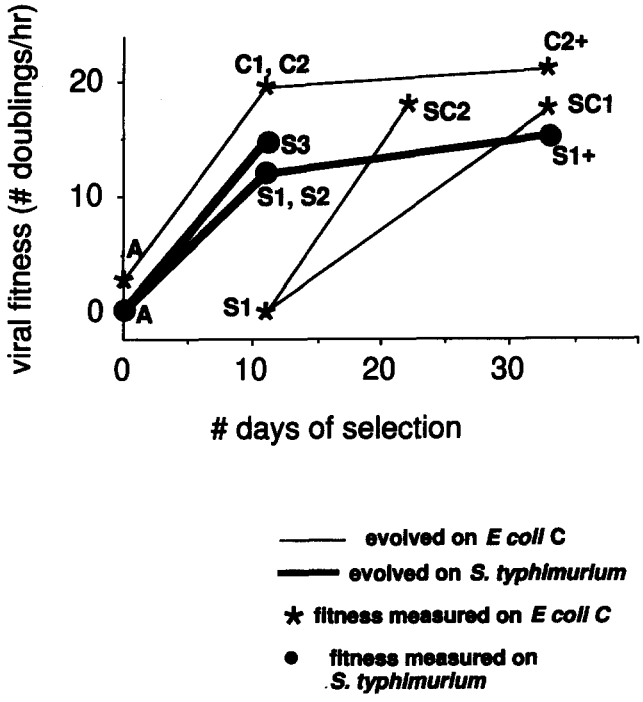

FIGURE 2.—ϕX 174 fitness at 42°. Dots and stars give average fitnesses of the isolates from different lineages. Some points represent two isolates, where the two values are too similar to be viewed separately. *, Fitness assays on host *E. coli* C; ●, fitness assays on *S. typhimurium*. Lines connect ancestors with descendants: heavy lines, growth on *S. typhimurium*; thin lines, growth on *E. coli*. The isolate S1 had very different fitnesses on the two hosts.

Despite the low per-genome rate of substitution (0.2–0.5%), many of the same substitutions occurred independently in two or more lineages (Tables 1 and 2). In addition, some of the changes that accumulated in the primary lineages were reversed in extended lineages, further contributing to the convergence (Table 2 deals specifically with convergences).

These levels of convergence are difficult to reconcile with random models of substitution, which assume that convergence is coincidental because of high rates of substitution. Consider the five primary lineages: 5348 positions in the genome experienced no substitutions, and {22, 7, 8, 1, 0} sites showed substitutions in {1, 2, 3, 4, 5} lineages, respectively. The maximum likelihood analysis of these observations yielded the extreme result that 99% of sites in the genome were estimated as invariant (unable to change); the probability of change at the remaining 1% of sites was estimated as 0.264 per lineage. This estimation procedure assumes that parallelism is coincidental, so it can only accommodate the high rates of parallelism we observed by treating most sites as invariant; a high rate of change at the remaining few sites then boosts the expected rate of coincidental parallelism. However, even this extreme fit is inadequate: a chi-square test comparing observations to expectations rejects the model, but only marginally so (expectations: 5293.3, 26.5, 17.5, 5.8, 0.9, 0.1 *vs.* observed 5348, 22, 7, 8, 1, 0). It should be emphasized that this model only addresses the sites at which substitutions occurred; the fact that, with one exception, all sites experiencing substitutions in more than one lineage were also substituted for the same base further

TABLE 2

Convergent substitutions

Isolate	C1	C2	C2+	SC1	SC2	S1	S1+	S2	S3	Unique
C1	[12]	5	5	5	8	4	4	3	2	3
C2		[12]	(11)	3	6	2	2	5	0	5
C2+			[18]	5	8	4	4	6	1	5
SC1				[26]	(10)	(7)	(8)	3	3	13
SC2					[23]	(12)	(12)	6	5	4
S1						[15]	(15)	6	7	3
S1+							[21]	6	7	5
S2								[13]	3	2
S3									[13]	6
Extension	**C1**	**C2**	**C2+**	**SC1**	**SC2**	**S1**	**S1+**	**S2**	**S3**	**Rev**
C2 → C2+	0		[8]	3	3	3	3	2	2	1
S1 → SC1	8	8	8	[26]	6		1	4	2	7
S1 → SC2	7	7	7	6	[14]		0	3	1	3
S1 → S1+	0	0	0	1	0		[6]	0	0	0

Top: Matrix of the number of changes found in one isolate that were also found in the other isolate (the matrix is necessarily symmetric). All changes are assessed relative to ancestor A, so for example, substitutions that occurred in S1 but were reversed during the extension from S1 to SC1 are not listed for SC1. Numbers on the diagonals (in square brackets ([ ]) give the number of substitutions observed in each isolate. Numbers in parentheses indicate that one of the isolates shares an intermediate ancestor with the other isolate, hence some or all of the substitutions in common are due to the common ancestry. Unique changes are those observed in only one of the lineages (hence do not contribute to convergence; the same unique substitution is present in multiple isolates if preserved throughout an extended lineage). For isolates from extended lineages (C2+, SC1, SC2, and S1+), the number of unique changes given are those arising during the extended phase (*e.g.*, for isolate SC1, unique changes are counted only during the phase from S1 → SC1), because the number of unique changes during the A → S1 phase can be obtained under S1. Reversions are not counted in the unique category.

Bottom: convergences occurring between the four extended lineages and all of the isolates. The left end of each row indicates the extended lineage, the top of the column indicates the isolate. The entry in each cell of the table is the number of substitutions (including reversions) occurring during the extension which matched the sequence of the isolate at the top of the column; changes already present in the primary-lineage ancestor are not counted. For example, eight of the substitutions during S1 → SC1 gave rise to the same base as found in C1. The last column (Rev) indicates the number of reversions occurring during the extension: several changes that accumulated during the primary lineage changed back to the wild-type allele during the extension. Cells are left blank if the isolate is the ancestor for the lineage. As above, numbers in parentheses indicate a common ancestry; numbers in square brackets indicate the number of substitutions that occurred on the lineage.

contributes to the deviation between this process and randomness.

Although some convergence was observed between most pairs of lineages, the highest levels of convergence were typically observed among lineages grown on the same host. An interesting comparison is between the SC lineages and the primary lineages. Using S1 as ancestor and *E. coli* as the new host, SC1 was propagated for 22 days, and SC2 was propagated for 11 days. Eight of 26 substitutions during this extension rendered SC1 more like C1 and C2; seven of 14 rendered SC2 more like C1 and C2. Most of the convergences between SC1 and C1 / C2 were reversals of substitutions found in the S1 ancestor, whereas most of those between SC2 and C1 / C2 were parallelisms. The other two extended lineages (C2+, S1+) show yet a different pattern: none of the changes in either of these extended lineages were convergent with changes in the primary lineages grown on the same host (although a few convergences with other lineages are apparent). Without replicates of these extended lineages, the rate of convergence after the initial period of adaptation cannot be determined because changes favored in the later stages of propagation may be different than in the early stages.

A subtle point in these studies is the level of independence between replicates. By starting C1, C2, S1, and S2 from the same lysate of A, the course of evolution in these four lineages could have been determined by features common to this lysate. As the lysate was itself grown from an isolated plaque, there would have been limited opportunity for much variation to accumulate. Nevertheless, it is conceivable that certain combinations of mutations could have been created in linkage disequilibrium and thus influenced some convergences. The facts that S3 was grown from its own isolated plaque of A and that SC1 and SC2 were grown from separate isolated plaques of S1, combined with the high levels of convergence between these lineages and the (other) primary lineages, indicates that convergence was not an artifact of peculiarities inherent to the A lysate.

Erroneous phylogenetic reconstruction: The convergent substitutions interfered with an accurate reconstruction of the true history of these lineages. Considering first the six-taxon tree that contains A and the five

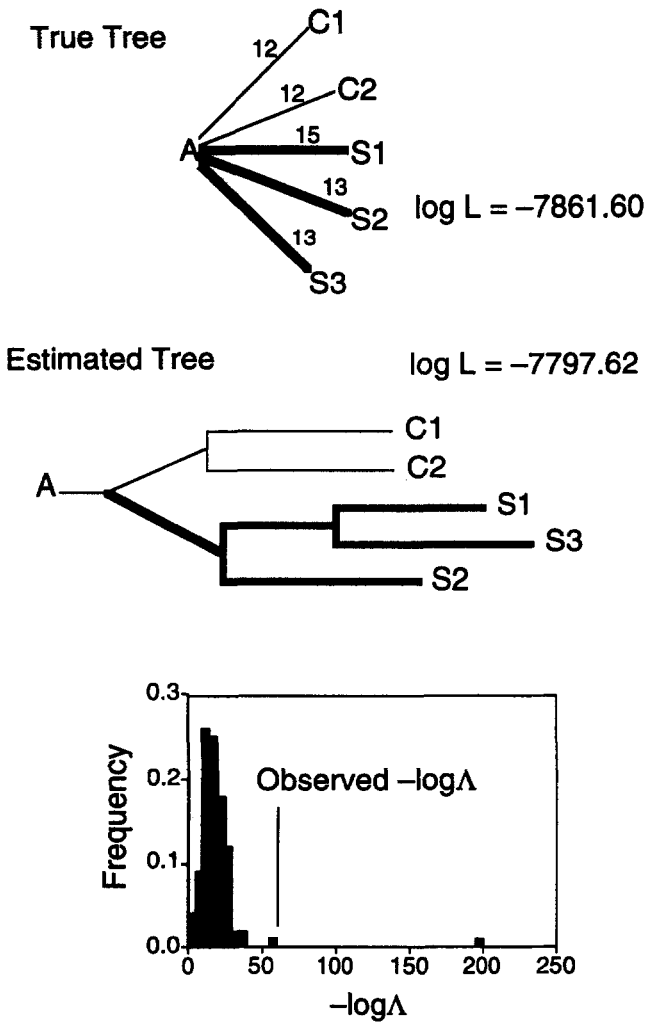

FIGURE 3.—(Top) True history of the five primary ϕX 174 lineages relative to ancestor A. All lineages were grown for 11 days. The numbers on each branch represent the numbers of substitutions occurring throughout the 5386-base genome along each lineage. Log likelihood is for the maximum likelihood reconstruction when the data are forced to obey the true topology. (Middle) Estimated topology and its log likelihood. (Bottom) The distribution of the likelihood ratio statistic, which compares the fit of the data to the true tree *vs.* the fit of the data to the best-fit tree. The observed ratio lies outside 99% of the distribution, indicating that the true tree is not an acceptable fit to the data.

primary lineages (S1, S2, S3, C1, C2), the maximum likelihood reconstruction united all five evolved lineages on a branch from ancestor A, as if all five shared a common ancestor more recently than A (Figure 3). The tree further divided the three *S. typhimurium*-evolved lineages (S1, S2, S3) from the two *E. coli*-evolved lineages (C1, C2). Although some structure is to be expected when using small samples to reconstruct star phylogenies, the strong clustering of phage isolates evolved on the same host suggests more than just sampling error. This intuition is supported statistically: the likelihood for the true tree was significantly worse than for the estimated tree, indicating that the true tree is rejected (at $P \sim 0.01$, Figure 3). This erroneous reconstruction is not simply due to our incorporation of ancestor A into the analysis (the ancestor would typically be absent in nonexperimental studies): reconstruction without ancestor A also yields an estimate that is significantly different from a star phylogeny ($P < 0.01$; analysis not shown).

The estimated changes were mapped on the reconstruction in Figure 3 to consider how much of the parallelism would be detected when the true phylogeny is not known. When a reconstruction is compromised by parallelism, the parallelism should be underestimated because it is often interpreted as common ancestry. A total of 26 parallel substitutions were observed across 16 sites in the five primary lineages (a single parallelism occurs with the second independent origin of a change, a second parallelism with the third independent origin, and so on). Only nine of the 26 (35%) parallelisms were detected in the reconstruction, at nine of the 16 sites. The underestimation is thus appreciable.

One indication of the high rate of parallel change that does not require knowledge of the true phylogeny is that the shape parameter α was estimated to be $\sim$0.01. This low value is the model's representation of a genome in which most change is restricted to a few sites, which themselves have high rates of convergence (as was observed in the likelihood analysis above that estimated 99% of the sites as invariant).

Reconstruction of the full set of 10 taxa is likewise compromised (Figure 4). As with the six-taxon tree, the true phylogeny is rejected at the 1% level. Extended lineages S1+ and C2+ are correctly united with their respective ancestors S1 and C2, but the lineages SC1 and SC2 are shifted from their immediate ancestor S1 toward the ancestor A, reflecting the extensive reversions and parallelisms that rendered them similar to C1 and C2. Another feature of this reconstruction is that S2 is united with C1 and C2, in contrast to its grouping in the six-taxon phylogeny. This shift presumably reflects the several parallelisms between S2 and C1/C2 on the one hand and others between S2 and S1/S3 (Table 2). Although not shown in our presentation, the placement of SC1 is also sensitive to which taxa are included: SC1 groups with C1 and C2 when the lineage SC2 is omitted from the analysis. We speculate that these major shifts in taxon clustering arise because the taxon contains several convergences with both sides of the estimated tree: fewer independent changes are required by joining the taxon to either side of the tree than by placing it intermediate, near ancestor A.

Fitness basis of substitutions: The fitness increases can be collectively attributed to the observed substitutions, since the entire genome was sequenced, but we do not know which substitutions or what fraction of them improved fitness. At one extreme, all substitutions could be equally beneficial. At the other extreme, the fitness improvement could be due primarly to just one

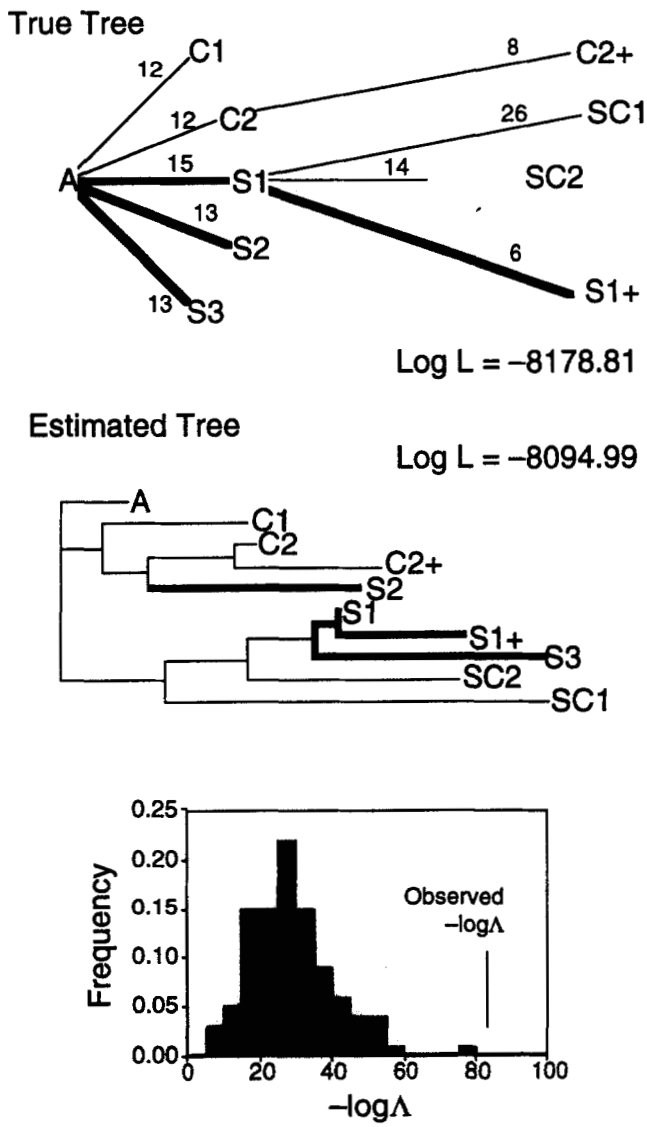

FIGURE 4.—(Top) True history of the nine ϕX 174 lineages relative to ancestor A. Lineages were grown for 11 or 22 days. The numbers on each branch represent the numbers of substitutions occurring throughout the 5386-base genome along each lineage. Log likelihood is for the maximum likelihood reconstruction when the data are forced to obey the true topology. (Middle) Estimated topology and its log likelihood. (Bottom) The distribution of the likelihood ratio statistic, which compares the fit of the data to the true tree *vs.* the fit of the data to the best-fit tree. Again, the likelihood ratio test rejects the true tree at the 1% level.

of the substitutions in each primary lineage and to an additional substitution in each extended lineage, other substitutions being neutral or of weak effect. The accumulation of substitutions with small fitness effects is plausible: assuming 1000 doublings of phage concentration during the 11-day selection, a mutant with a 1% fitness advantage per doubling period could have increased to a frequency of 0.5 from an initial frequency of one in a million (using formulae modified from CROW and KIMURA 1970, p. 192). This magnitude of fitness effect would be indistinguishable from neutrality in our assays. The only observation that bears directly on these possibilities comes from an isolate obtained in a preliminary study, which indicates that some substitutions can have a large effect: an isolate with only two substitutions had a fitness on *E. coli* of ~14, showing more than half the fitness improvement seen in C1 and C2.

Convergence may indicate the beneficial effects of some substitutions. Mutation pressure alone is unlikely to cause a meaningful increase in substitution frequency in our design. At 1000 doublings, a mutation rate of 10^{-5} per doubling would lead to a final frequency of only 0.01, not enough to observe in multiple lineages. Consequently, the plausible alternative is selection.

A substitution process driven by selection is also consistent with the observed excess of missense substitutions. The ratio of missense to silent changes is significantly higher than expected for this spectrum of mutations ($P < 0.01$, expectations calculated empirically by randomly assigning the observed nucleotide changes across all sites in the genome, comparing the number of substitutions silent in all reading frames to substitutions causing missense changes in at least one reading frame). The inference of selection from an excess of missense changes is indirect, however, and may not provide a true measure of the fraction of sites that were selected. Furthermore, silent substitutions may not be neutral: position 500 evolved silent substitutions in all three *S. typhimurium* lineages and reverted to the ancestral state in one of the SC lineages.

DISCUSSION

Our study may be the first to show that an erroneous phylogenetic reconstruction results from convergent molecular evolution (DOOLITTLE 1994; HEDGES and MAXSON 1996; but see BEINTEMA *et al.* 1977). Two manifestations of the problems caused by convergence were as follows: (i) the number of changes since common ancestry of the lineages was underestimated, and (ii) lineages grown on the same host were grouped together in the reconstruction. The errors in reconstruction were not an artifact of sampling just a small portion of the genome, because the entire genome sequence was used. Furthermore, the estimated histories were statistically significant deviations from the true histories.

How general is a high rate of convergence? Virtually all phylogenetic trees based on molecular data reveal homoplasy, so some level of convergence is common if not universal (PEACOCK and BOULTER 1975; BEINTEMA *et al.* 1977; ROMERO-HERRERA *et al.* 1979; SANDERSON and DONOGHUE 1996; WELLS 1996). Should we then extrapolate from the present study and conclude that convergence commonly causes mistakes in reconstruction? Not yet. First, a ubiquity of convergence is not, by itself, a problem for reconstruction. A certain amount

of convergence is expected simply under random substitutions, and reconstruction methods accommodate this type of convergence. Second, reconstruction of a star phylogeny, as here, is especially sensitive to convergence because there is no historical signal to dampen its effects. Third, the ϕX 174 data appear to have an atypically high rate of convergence, because the estimate of the shape parameter (α) was near zero (0.01); estimates from other data are typically 10 times higher (YANG 1996b).

Inspection of the design used in these experiments provides further grounds for doubting the generality of the findings to other systems. Three properties of the design may have led to the high rate of convergence, all of which are atypical of what is thought to apply to most organisms:

1. Replicate lineages were exposed to nearly identical selective environments.
2. Strong, mass selection was operating on very large populations.
3. The replicate lineages started with virtually identical genomes.

We can offer plausibile arguments for why all three properties might contribute to convergence: (i) similar environments create similar selective pressures, which in turn favor similar adaptations (a hypothesis consistent with the observed phylogenetic clustering of lineages grown on the same host); (ii) the array of beneficial mutations with large effects, which will be those favored under strong selection, is much more limited than those with small effects; and (iii) the fitness effect of a mutation is genome-dependent, so when the same genome is subjected to the same selection, the same mutations will be favored. The appeal of the ϕX 174 experimental system is that simple experiments can be conducted to test the role of each of these factors in promoting convergence. In the meantime, observations from other experimental studies are illuminating.

A striking comparison to this ϕX 174 work is provided by a study that used a clonal isolate of phage T7 to create nine lineages (HILLIS *et al.* 1992, 1994). The comparable rate of substitution per site per lineage from that propagation was 0.05, *vs.* 0.264 from the present study. The expected fraction of parallel substitutions in a pair of lineages is thus 28 times higher from this ϕX 174 study than from the T7 study. Consistent with this comparison, the T7 phylogeny was correctly estimated by all six attempted reconstruction methods, and, furthermore, Monte Carlo sampling methods showed it to be very robust to correct reconstruction (BULL *et al.* 1993). Of course, recovery of the correct T7 phylogeny was also facilitated by the structure of the phylogeny, not just by the lower rate of convergence.

The T7 design shared properties 1 and 3 with the ϕX 174 design, but it differed in property 2, because the T7 lineages were bottlenecked to a single plaque every five cycles, and each terminal isolate was a full 24 bottlenecks removed from the phylogeny's deepest ancestor. This periodic reduction in population size grossly interferes with adaptation. T7 was grown in the presence of a powerful mutagen, perhaps enhancing the maladaptive effect of the bottlenecks. A large fraction of substitutions in the T7 study may thus have been irrelevant or antagonistic to adaptation; fitnesses actually decreased during the propagation, as judged by plaque sizes and other indirect measures. If convergence is driven by strong selection in large populations, then the T7 design would not be expected to show a high level of convergence.

A second study of T7 bolsters this view. YIN (1993) observed different paths of restriction site evolution between different plaques and even within a plaque, when the virus was grown on a host that supplied or obviated functions normally expressed by the phage. The changes in restriction site pattern were due chiefly to deletions that included the phage genes made redundant by the host. The virus followed many paths of adaptation, hence, it failed to show strong convergence at the level of specific nucleotide changes. Although YIN's (1993) study encouraged T7 to adapt to a common host environment, the spatial structure of the experimental design prevented population-wide competition, thereby preserving variety rather than selecting for the fittest phage, violating property 2 above. The role of spatial structure in facilitating the evolution of diversity was also recognized in another experimental microbial system (KORONA *et al.* 1994).

Finally, CUNNINGHAM *et al.* (1997) reported multiple, independent origins of deletions and stop codons in experimental lineages of T7 subjected to mass selection with infrequent bottlenecks. The selective factor in that study was grossly similar to that in the YIN (1993) study (loss of a T7 function was beneficial to phage growth), except that selection was applied to large populations without spatial structure. Thus, in comparing these three T7 studies, the nature of selection (property 2) may explain the different levels of convergence.

Despite our hesitance in regarding these ϕX 174 results as typical of other systems, one should not be too quick to dismiss convergence as a problem in phylogenetic reconstruction. Indeed, the phenomenon known as "long branch attraction" results when high levels of substitution create enough convergence to impair reconstruction (FELSENSTEIN 1978). The fact that two thirds of parallelisms in the ϕX 174 primary lineages would have been misdiagnosed had we not known their true history cautions that the extent of convergence in other systems may be underestimated whenever the true history is unknown. High rates of parallel substitutions and back mutations were proposed in some early phylogenetic work for which molecular data were mapped on phylogenies derived from nonmolecular data (BEINTEMA *et al.* 1977; ROMERO-HERRERA *et al.*

1979). Convergence may be especially prone to cause problems when few informative characters separate taxa.

The ϕX 174 results may even find direct relevance in special cases, as in the evolution of drug resistance by viruses and bacteria (FUNATSU and WITTMANN 1972; VIZAN *et al.* 1991; YOSHIDA *et al.* 1993; HEISIG 1995; MOLLA *et al.* 1996; WEISBLUM 1995). In these cases, the major factor promoting convergence may be the extreme, common selection generated by the toxic effects of the drug. In an interesting application of convergence, MOLLA *et al.* (1996) used parallelisms among human immunodeficiency virus (HIV) isolates to identify substitutions that were likely to be (and later were shown to be) important in drug resistance. The present case even extends those precedents in that the adaptation has occurred to broad environmental agents (temperature, host, and perhaps media) rather than to a specific challenge such as a drug, and convergence is genome-wide rather than confined to a single gene.

No easy patterns: Even in our simple system, the patterns of convergence are not entirely obvious. As an example, SC1 and SC2 were both extended on host *E. coli* from ancestor S1, but SC1 was extended for twice as long as SC2. SC1 is reconstructed as though it is more closely related to C1 or C2 than is SC2, as if the longer extension time allowed more convergence with C1 and C2 (recall that C1 and C2 were both grown on host *E. coli*). Yet Table 2 (top) shows the unexpected result that isolate SC2 is actually more convergent than SC1 with C1 and C2. Furthermore, most of the convergence between SC1 and C1/C2 arose from reversions in SC1, whereas parallelisms accounted for most of the convergence between SC2 and C1/C2. There is, therefore, no clear pattern that can be used to generalize these data to numerical simulations.

Fitness evolution: ϕX 174 fitness, measured as growth rate, increased by an order of magnitude in all lineages derived from low-fitness ancestors. We do not know the distribution of fitness effects for the 12 or more substitutions that were observed in each isolate, but some of the fitness effects must have been substantial (even the average fitness improvement per substitution is substantial for the primary lineages). One intriguing question is whether the level of convergence differs between substitutions with large effects and those with small effects. It is commonly argued that beneficial mutations with small effects are much more common than those with large effects (reviewed by LANDE 1983); however, substitutions with large effects will increase much faster than those with small effects, unless pleiotropic effects are so deleterious as to offset any net advantage. As a consequence of both points, there should be greater and more rapid convergence for mutations with large effects, an argument that underlies property 2 above. One could test this prediction in the ϕX 174 system by growing replicate lineages from an ancestor that was already well adapted to the chemstat (*e.g.*, replicate the extended lineages S1+ and C2+).

These simple expectations about convergence are complicated by any antagonistic interactions among substitutions, whereby either substitution alone is beneficial, but together they are not. In the absence of interactions, convergence should approach 100% for substitutions offering fitness gains of more than a few percent. No substitutions in our study occurred in all lineages, which might be regarded as evidence of fitness interactions. However, because only one isolate was sampled per lineage, the present data may underestimate the true level of convergence, because other changes could be present but not fixed in each lineage. Also, seven sites show a strict pattern of a substitution present in all *S. typhimurium* lines or all *E. coli* lines, so it may indeed be the case that some substitutions are universally beneficial for growth on a particular host.

For most lineages, evolution of higher fitness could have been in response to any combination of host, temperature, and chemostat conditions. The comparison of lineage S1+ with SC1 and with SC2 holds constant all of these factors except host, because all three of these extended lineages started with the *S. typhimurium*-adapted, high-temperature-adapted ancestor S1. S1+ accumulated only six new substitutions during 22 days, whereas SC1 accumulated 26 new substitutions over 22 days and SC2 accumulated 14 substitutions in 11 days. Thus, most of the changes that accumulated during the propagations of SC1 and SC2 are likely driven by host switching.

Furthermore, some changes may have nothing to do with host, chemostat, or temperature *per se*. They may merely accommodate earlier substitutions. For example, the initial substitutions conferring *E. coli* resistance to the antibiotic streptomycin caused a fitness drop in the absence of the drug; subsequent changes enabled the bacterium to recover by modulating the pleiotropic effects of the initial resistance (SCHRAG and PERROT 1996). A compelling example of this "recovery" phenomenon was observed for HIV resistance to a protease inhibitor. In the initial selection, several mutations accumulated in the presence of the drug. In a subsequent experiment, the drug was withdrawn after the first mutations appeared, yet the virus proceeded to accumulate the additional substitutions anyway (BORMAN *et al.* 1996). Similar recovery may explain substitutions that accompanied HIV drug resistance evolution but that were located outside active sites of the drug target (MOLLA *et al.* 1996). The additional experiments required to tease apart the component causes of convergent substitutions in ϕX 174 are feasible.

We thank IRWIN TESSMAN and BEN FANE for advice on ϕX 174 biology. M. BRAUER, W. CRILL, M. HOLDER, S. MESSENGER, and M. RAIN offered comments on the manuscript. This project was funded by grants DEB 95-08987 and DEB 92-20520 from the National Science

Foundation and by grant MCB 95-06954 from the National Institutes of Health. J.J.B. was supported as the J. F. Miescher Regents Professor and D.M.H. as the A. Roark Centennial Professor.

LITERATURE CITED

BEINTEMA, J. J., W. GAASTRA, J. A. LENSTRA, G. W. WELLING and W. M. FITCH, 1977 The molecular evolution of pancreatic ribonuclease. J. Mol. Evol. **10:** 49–71.

BORMAN, A. M., S. PAULOUS and F. CLAVEL, 1996 Resistance of human immunodeficiency virus type 1 to protease inhibitors: selection of resistance mutations in the presence and absence of the drug. J. Gen. Virol. **77:** 419–426.

BULL, J. J., C. W. CUNNINGHAM, I. J. MOLINEUX, M. R. BADGETT and D. M. HILLIS, 1993 Experimental molecular evolution of bacteriophage T7. Evolution **47:** 993–1007.

CROW, J. F., and KIMURA, M., 1970 *An Introduction to Population Genetics Theory*. Harper and Row, New York.

CUNNINGHAM, C. W., K. JENG, J. HUSTI, M. BADGETT, I. J. MOLINEUX *et al.*, 1997 Parallel molecular evolution of deletions and nonsense mutations in bacteriophage T7. Mol. Biol. Evol. **14:** 113–116.

DOOLITTLE, R. F., 1994 Convergent evolution: the need to be explicit. Trends Biochem. Sci. **19:** 15–18.

DOWELL, C. E., 1980 Growth of bacteriophage ϕX 174 at elevated temperatures. J. Gen. Virol. **49:** 41–50.

DYKHUIZEN, D., 1993 Chemostats used for studying natural selection and adaptive evolution. Methods Enzymol. **224:** 613–631.

FELSENSTEIN, J., 1978 Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27:** 401–410.

FELSENSTEIN, J. 1993 PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

FUNATSU, G., and H. G. WITTMANN, 1972 Ribosomal proteins. 33. Location of amino-acid replacements in protein S12 isolated from *Escherichia coli* mutants resistant to streptomycin. J. Mol. Biol. **68:** 547–550.

GODSON, G. N., 1978 The other isomeric phages, pp. 103–112 in *The Single-Stranded DNA Phages*, edited by D. T. DENHARDT, D. DRESSLER and D. S. RAY. Cold Spring Harbor Laboratory, New York.

HASEGAWA, M., H. KISHINO and T. A. YANO, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

HEDGES, S.B., and L. R. MAXSON, 1996 Molecules and morphology in amniote phylogeny. Mol. Phylogenet. Evol. 6:312–314.

HEISIG, P., 1995 Genetic evidence for a role of parC mutations in development of high-level fluoroquinoline resistance in *Escherichia coli*. Antimicrob. Agents Chemother. **40:** 879–885.

HILLIS, D. M., J. J. BULL, M. E. WHITE, M. R. BADGETT and I. J. MOLINEUX, 1992 Experimental phylogenetics: laboratory construction of a known phylogeny. Science **255:** 589–592.

HILLIS, D. M., J. P. HUELSENBECK and C. W. CUNNINGHAM, 1994 Application and accuracy of molecular phylogenies. Science **264:** 671–677.

INCARDONA, N. L., and U. R. MÜLLER, 1985 Eclipse kinetics as a probe of quaternary structure in bacteriophage ϕX 174. J. Mol. Biol. **181:** 479–486.

KORONA, R., C. H. NAKATSU, L. J. FORNEY and R. LENSKI, 1994 Evidence for multiple adaptive peaks from populations of bacteria evolving in a structured habitat. Proc. Natl. Acad. Sci. USA **91:** 9037–9041.

LANDE R., 1983 The response to selection on major and minor mutations affecting a metrical trait. Heredity **50:** 47–65.

LI, W-H., 1997 *Molecular Evolution*. Sinauer, Sunderland, MA.

MOLLA, A., M. KORNEYEVA, Q. GAO, S. VASAVANONDA, P. J. SCHIPPER *et al.*, 1996 Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. Nature Med. **2:** 760–766.

MÜLLER, U. R., and R. D. WELLS, 1980 Intercistronic regions in ϕX 174 DNA II. Biochemical and biological analysis of mutants with altered intercistronic regions between genes J and F. J. Mol. Biol. **141:** 25–41.

PEACOCK, D., and D. BOULTER, 1975 Use of amino acid sequence data in phylogeny and evaluation of methods using computer simulation. J. Mol. Biol. **95:** 513–527.

ROMERO-HERRERA, A. E., N. LIESKA, M. GOODMAN and E. L. SIMONS, 1979 The use of amino acid sequence analysis in assessing evolution. Biochimie **61:** 767–779.

SANDERSON, M. J., and M. J. DONOGHUE, 1996 The relationship between homoplasy and confidence in a phylogenetic tree, pp. 67–89 in *Homoplasy: The Recurrence of Similarity in Evolution*, edited by M. J. SANDERSON and L. HUFFORD. Academic Press, New York.

SANGER, F., G. M. AIR, B. G. BARRELL, N. L. BROWN, A. R. COULSON *et al.*, 1977 Nucleotide sequence of bacteriophage ϕX 174 DNA. Nature **265:** 687–695.

SCHRAG, S. J., and V. PERROT, 1996 Reducing antibiotic resistance. Nature **381:** 120–121.

SWOFFORD, D. L., 1997 PAUP*: Phylogenetic Analysis Using Parsimony (and other methods), version 4.0. Sinauer Associates, Sunderland, MA.

VIZAN, J. L., C. HERNANDEZ-CHICO, I. DEL CASTILLO and F. MORENO, 1991 The peptide antibiotic microcin B17 induces double-strand cleavage of DNA mediated by *E. coli* DNA gyrase. EMBO J. **10:** 467–476.

WEISBLUM, B., 1995 Erythromycin resistance by ribosome modification. Antimicrob. Agents Chemother. **39:** 577–585.

WELLS, R. S., 1996 Excessive homoplasy in an evolutionarily constrained protein. Proc. R. Soc. Lond. B **263:** 393–400.

YANG, Z, 1993 Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10:** 1396–1401.

YANG, Z, 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39:** 306–314.

YANG, Z., 1996a Phylogenetic Analysis by Maximum Likelihood (PAML), version 1.2. Department of Integrative Biology, University of California, Berkeley.

YANG, Z., 1996b Among site rate variation and its impact on phylogenetic analysis. TREE **11:** 367–372.

YIN, J., 1993 Evolution of bacteriophage T7 in a growing plaque. J. Bacteriol. **175:** 1272–1277.

YOSHIDA, H., M. NAKAMURA, M. BOGAKI, H. ITO, T. KOJIMA *et al.* 1993 Mechanism of action of quinolones against *Escherichia coli* DNA gyrase. Antimicrob. Agents Chemother. **37:** 839–845.

Communicating editor: W-H. LI