Statistical Tests for Detecting Gene Conversion[1]

Stanley Sawyer
Department of Mathematics, Washington University, and
Department of Genetics, Washington University Medical School

March 13, 1989

Abstract

Statistical tests for detecting gene conversion are described for a sample of homologous DNA sequences. The tests are based on imbalances in the distribution of segments on which some pair of sequences agree. The methods automatically control for variable mutation rates along the genome, and do not depend on *a priori* choices of potentially monophyletic subsets of the sample. The tests show strong evidence for multiple intragenic conversion events at two loci in *E. coli*. The *gnd* locus in *E. coli* shows a highly significant excess of maximal segments of length 70–200bp, which suggests conversion events of that size. The data also indicate that the rate of these short conversion events might be of the order of the neutral mutation rate. There is also evidence for correlated mutation in adjacent codon positions. The same tests applied to a locus in an RNA virus were negative.

1. Introduction.

Recently developed statistical techniques for analyzing DNA sequences have shown strong evidence for intragenic recombination both in primates (Stephens 1985) and in bacteria (DuBose *et al.* 1988). In Stephens' (1985) test, a sample of homologous DNA sequences is partitioned into two subsets, and the collection of polymorphic sites which are consistent with this partition is considered. If the distribution of these sites is significantly nonuniform, it is inferred that one or more intragenic recombination events may have occurred involving the sequences in the sample. However, if the sample has more than three or four sequences, and the sequences are moderately or highly polymorphic, an appropriate partition may be hard to find. Also, partitions are treated

1. Key words: gene conversion, statistical test for gene conversion, conserved sequences, *Escherichia coli*.

Address for correspondence and reprints: Department of Mathematics, Washington University, St. Louis, MO 63130. This work was partially supported by National Science Foundation Grants DMS-8504315 and DMS-8802039.

individually, and statistical concerns about multiple comparisons may arise. For example, if there are $m = 8$ sequences, there are $2^{m-1} - 1 = 127$ possible partitions into two subsets, and some partitions may have significantly nonuniform distributions purely by chance. (Stephens' examples had $m = 3$ and 5.) Secondly, Stephens' procedure tests for a uniform distribution of polymorphic sites, but only corrects for mutational "hot" or "cold" spots along the genome by deleting particular segments with no polymorphic sites. If there is a moderate or high level of polymorphism, a more delicate method of controlling for variable mutation rates would be desirable.

Alternative procedures that are meant to address these concerns for gene conversion are described below. These tests, which are based on imbalances in the distribution of maximal segments on which some pair of sequences agree, were first applied to three samples of DNA sequences described in the literature: (i) seven strains of *E. coli* at the *gnd* locus (Sawyer *et al.*, 1987), (ii) eight strains of *E. coli* at the *phoA* locus (DuBose *et al.*, 1988), and (iii) 13 strains of human influenza A virus at the NS locus (Buonagurio *et al.*, 1986). Both *E. coli* loci show strong evidence for multiple intragenic conversion events. The *gnd* locus has a highly significant excess of maximal segments of length 70–200bp, which suggests conversion events of that size. The data also suggest that the rate of these short conversion events at a typical base may be greater than the base mutation rate. The *gnd* data set also shows a highly significant excess of maximal segments of length 2bp, which suggests a highly significant tendency for correlated neutral mutation at adjacent codon positions (see discussion below). The same tests applied to the human influenza data were negative. Also considered were (iv) simulated *gnd* data sets based on the consensus sequence and a plausible pedigree, both with and without gene conversion, and (v) a sample of nine cytochrome *c* bacterial protein sequences (Sneath *et al.*, 1975).

2. The Tests.

Consider a set of n aligned DNA sequences from the coding regions of genes or pseudogenes. Any base position that is not identical in all n DNA sequences, and that is not part of a codon position that encodes different amino acids, is called a *silent polymorphic site*. Consider only the silent polymorphic sites for the moment. If two sequences are compared, they will differ in a set of $d \leq s$ silent polymorphic sites, where s is the total number of silent polymorphic sites. These d discordant sites partition the genome into $d + 1$ subsets, which we will call the *fragments* determined by this partition. A *condensed fragment* is the set of silent polymorphic sites in a fragment. Thus the length of a condensed fragment is the number of

silent polymorphic sites either between two neighboring discordant sites, or else between a discordant site and one of the ends of the sequence. The sum of the lengths of the condensed fragments is $\sum x_i = s - d$, where x_i is the length of the i^{th} condensed fragment. We define the sum of the squares of the condensed fragment lengths SSCF as the sum of x_i^2 over all $d_k + 1$ fragments over all $n(n-1)/2$ pairs of sequences, where d_k is the number of discordant sites determined by the k^{th} pair of sequences. Similarly, we define MCF as the maximum of x_i for all such fragments for all pairs of sequences.

To estimate the significance of SSCF and MCF, we compare the scores for the observed sequences to those for artificial data sets obtained from 10,000 random permutations of the orders of the s silent polymorphic sites. The data are now viewed as an $n \times s$ matrix of letters corresponding to bases; permutations correspond to rearranging the columns of this matrix. Before permuting the s sites (or columns), each site is assigned to one of four classes according to whether its codon is (i) two-fold degenerate, (ii) three-fold degenerate, (iii) four-fold degenerate, or (iv) mixed, the last for leucine and arginine codons (Li *et al.*, 1985; Sawyer *et al.*, 1987). Serine codons are assigned to class i or iii according to whether the codon is AGY or UCN. Sites in classes i–iii are third-position sites in otherwise nondegenerate codon positions. Permutation of the sites is constrained so that sites of a given class can only be assigned to site positions of the same class in the original data set. Thus, for each pair of sequences, the number of discordant sites d_k and the sum of the fragment lengths $\sum x_i = s - d_k$ is preserved by the permutations, although $\sum x_i^2$ will generally vary. P-values for SSCF and MCF are defined as the proportion of permuted data sets that have SSCF or MCF scores (respectively) that are greater than or equal to the original score.

A theoretical justification for these tests is as follows. If there has been no gene conversion since the most recent common ancestor of the sequences, the distribution of bases at silent polymorphic sites would have been determined by independent neutral mutation within the same pedigree at all sites. The distribution of the bases at a silent polymorphic site given the original base would depend on the degree of degeneracy at the site, but should otherwise be more-or-less independent of position. These distributions should be preserved by the permutations described above. A gene conversion event between two sequences in the sample would result in a segment of bases in which the two sequences agree, and may produce an unusually long fragment. Given a fixed number of lengths $\{x_i\}$ with $\sum x_i$ held constant, the sum $\sum x_i^2$ is minimized when the $\{x_i\}$ are equal. Thus a gene conversion event should tend to increase SSCF and MCF. Note that attempting to control for end effects by ignoring the end fragments of the sequenced region would cause $\sum x_i$ to be variable, so that the meaning of $\sum x_i^2$ as a heterogeneity measure would be lost. The

resulting P-values should be less significant. In fact, this was observed for the two *E. coli* data sets.

By construction, SSCF and MCF are not influenced by hot and cold spots in the mutation rate if the mutation rate is the same across the sequences, since monomorphic sites are excluded. The tests also control for strain-dependent mutation rates, each of which is constant along the sequence. However, if a subset of the strains has a depressed mutation rate in part of the sequenced region, fragments containing mutational cold spots in pairs of strains could mimic gene conversion events. Biological mechanisms by which this could happen are not clear, however, and we do not feel that this is the cause of the very significant values of SSCF encountered below.

It is useful to define two variants of SSCF and MCF which do not control for mutational hot and cold spots, but which would have greater power for detecting large conversion events if the mutation rate is constant. For each pair of sequences, consider the set of d discordant silent polymorphic sites described above which partitions the sequence into $d+1$ fragments. If y_i is the length of the i^{th} fragment in the original data set (i.e., the length of the *uncondensed* fragment), then $\sum y_i = L - d$, where L is the total number of sites. We define SSUF as the sum of y_i^2 for all uncondensed fragments for all pairs of sequences, and similarly MUF is the maximum y_i for all fragments for all pairs of sequences. In determining the significance of SSUF and MUF, permutations of the data (now viewed as an $n \times L$ matrix of bases) map silent sites onto silent sites of the same class as before, with monomorphic sites and sites within amino-acid polymorphic codon positions held fixed. For each pair of sequences, these permutations preserve the number d of discordant silent polymorphic sites and the sum $\sum y_i$ for that pair of sequences, but $\sum y_i^2$ will generally vary. P-values for SSUF and MUF are defined as above. The statistics SSUF and MUF give alternative measures of the size of pairwise conserved segments. In one example below, SSUF is significant ($P = 0.0133$) but SSCF is not significant ($P = 0.2163$).

Permuting all silent polymorphic sites, rather than permuting within each degeneracy class, typically gave essentially the same results. Permuting all polymorphic sites in the *E. coli* data sets, including those in amino-acid variable codon positions, and using arbitrary polymorphic sites as potential fragment boundaries, typically gave less significant results. This may indicate that the amino-acid polymorphisms are younger on the average than the events causing significance for silent polymorphic sites. Parallel selection could in principle produce concordant sequences of amino-acid varying sites that would mimic gene conversion; the *E. coli* data sets have relatively few amino-acid polymorphisms and this was not a factor. The P-values for SSCF and SSUF were much more significant than those for MCF and MUF in the *E. coli* data

sets, while they were comparable in the protein data set. The difference between the two cases may depend on the size of the fragments that cause high values for the statistics. Alternately, MCF and MUF may simply put too much weight on the most recently diverged pairs of sequences in these data sets. Tests based on SSCF are probably the most powerful for detecting pairwise conserved segments that are short with respect to an appropriate measure of polymorphism.

The definition of fragments given above were designed to detect *internal* gene conversion events; i.e., gene conversion between two sequences in the sample. An alternative definition of *outer fragment* can be made for detecting gene conversion (or reciprocal recombination) from strains outside the sample. See Section 8 below. The outer fragment analogs of SSCF, SSUF, MCF, and MUF were essentially nonsignificant for the two *E. coli* data sets, but not for the protein data set. This may be due to an intrinsic lack of power of the outer fragment analogs of SSCF, etc., or perhaps it is because the samples are so broadly representative of *E. coli* strains in general that significant outer fragments could not be produced by conversion from an outside strain. DuBose *et al.* (1988) suggest that an analog of the notion of reproductively isolated groups for eukaryotes might be groups of bacteria that are isolated with respect to exchanging small segments of DNA. Statistical tests based on outer fragments might be used to test the extent of such groups.

3. The *gnd* Locus in *E. coli*.

Sawyer *et al.* (1987) describe a data set consisting of a 768-base segment from the reading frame of the *gnd* locus in 7 strains of *E. coli*. Out of 256 codon positions, 12 are amino-acid polymorphic. Of the 274 potentially degenerate sites within the remaining amino-acid monomorphic codon positions, 81 are silent polymorphic. Of the 81 sites, 44 are singly polymorphic (i.e., having one base in one strain and a second base in the remaining strains), 19 sites are at 2-fold degenerate sites, 5 sites are third-position sites within isoleucine codon positions (i.e., are 3-fold degenerate), 42 sites are 4-fold degenerate, and the remaining 15 are at sites of irregular degeneracy. As in Sawyer *et al.* (1987), isoleucine was considered 2-fold degenerate for *gnd* since the ATA codon did not occur, so that 24 sites were considered 2-fold degenerate.

The results of the tests described in Section 2 are given in Table 1. Although both SS statistics are highly significant, only one uncondensed fragment of 200bp is statistically significant by itself as measured by the distribution of simulated values of MCF or MUF (Table 2). If this particular fragment were excluded, both SS statistics would still be highly significant. Note that Table 1 also gives the distance between the observed value and the mean of

the permuted scores in terms of the standard deviation of the permuted statistics. While this distribution is probably not normal, the fact that the observed value of SSCF for the *gnd* data set was 7.21 standard deviations above the mean is consistent with the fact that the largest of 10,000 simulated values of SSCF was still well below the observed value.

Table 1. The *gnd* locus in *E. coli*[a]

Statistic	Observed Score	P-value[d]	S.D.[b,c] above Mean[b]	S.D.[b] of scores
SSCF:	5213	0	7.21	203.0
MCF:	22	0.0040	3.98	2.3
SSUF:	822020	0	7.10	23896.2
MUF:	200	0.0348	2.36	24.6

[a] *Permuting 81 silent polymorphic sites in 7 strains 10,000 times.*

[b] *Mean and S.D. are the mean and standard deviation of the scores for 10,000 random permutations of the data set.*

[c] *(Observed score – Mean)/S.D.*

[d] *The relative number of permuted data sets with scores greater than or equal to the observed score.*

Thus the seven strains appear to have undergone a large amount of between-strain gene conversion within the *gnd* locus, although it may not be possible to say exactly where. A comparison of the actual fragment lengths with the simulated fragment lengths show that the data set has a highly significant excess of fragments of length 70bp or longer ($P < 10^{-3}$). Specifically, 7.1% of the observed fragments have length 70bp or longer but only 4.8% of the simulated fragments, amounting to a relative increase of approximately 50%. Potentially, any or all of these 51 observed fragments could have been caused by gene conversion. DuBose *et al.* (1988) discuss the biological mechanisms which might generate gene exchanges of this size.

The four largest condensed and uncondensed fragments are given in Table 2, along with P-values for the fragment lengths as determined by the permuted MCF and MUF scores. Note that none of the fragments overlap either end of the 768bp sequenced region.

The results were almost identical if all silent sites were randomly permuted, rather than just within degeneracy classes. However, randomly permuting all polymorphic sites instead of silent polymorphic sites in the definitions of SSCF, MCF, SSUF, and MUF (and using all polymorphic sites as potential fragment boundaries) gave less significant results even though more polymorphic sites were involved (Table 3). With 95% confidence, at least half

Table 2. The largest fragments in *gnd*

Condensed fragments with P-values:			
K/6: 364–563 (22)	[0.004]	5/8: 229–401 (15)	[0.195]
6/8: 151–335 (16)	[0.117]	K/8: 184–338 (14)	[0.313]
Uncondensed fragments with P-values:			
K/6: 364–563 (200)	[0.035]	5/8: 229–401 (173)	[0.119]
6/8: 151–335 (185)	[0.059]	K/8: 184–338 (155)	[0.256]

NOTE — *'K/6: 364–563' denotes a fragment with boundaries determined by the sequences 'K' and '6' (see text) and base range 364–563 in the original data set. The first position in the aligned region has position 1. The number in parentheses is the condensed or uncondensed fragment length. The number in brackets is the relative number of permuted data sets (see Table 1) whose MCF or MUF score respectively is greater than or equal to the observed fragment length.*

of the amino-acid variation in this data set is known to be due to deleterious variants (Sawyer *et al.*, 1987). Perhaps the decreased significance of Table 3 is caused by the selectively deleterious polymorphisms being younger than the gene conversion events.

Table 3. All polymorphic sites in *gnd*[a]

Statistic	Observed Score	P-value	S.D. above Mean	S.D. of scores
SSCF:	5894	0.0002	4.81	233.2
MCF:	19	0.0252	2.58	2.3
SSUF:	655292	0	5.23	20330.6
MUF:	170	0.0207	2.72	18.0

[a] *Permuting 100 polymorphic sites in 7 strains 10,000 times. See Table 1 for definitions.*

If only the 44 singly polymorphic silent sites are used (Table 4), SSCF and SSUF are just barely significant ($0.01 < P < 0.02$ for both), and MCF and MUF have $P \geq 50\%$. In contrast, if the 37 multiply polymorphic silent sites are used, SSCF and SSUF remain highly significant ($P < 10^{-3}$ for both; see Table 4). The decreased significance for singly polymorphic silent sites may be an indication that these polymorphisms are of comparable age to the conversion events. The latter could imply that the rate of gene conversion involving segments in the range 70–200bp is of the same order of magnitude

as the neutral substitution rate per base. This implication is also consistent with the simulation results in Section 6 below.

Table 4. Silent polymorphic sites in *gnd*

Singly poly.[a]*:*	Observed Score	P-value	S.D. above Mean	S.D. of scores
SSCF:	4638	0.0155	2.57	355.4
MCF:	15	0.7467	-0.57	3.2
SSUF:	1940880	0.0116	2.77	102556.3
MUF:	290	0.5336	-0.26	47.3
Multiply polymorphic sites[b]*:*				
SSCF:	1107	0.0004	4.95	53.73
MCF:	9	0.2216	0.94	1.55
SSUF:	1307486	0.0001	4.79	39013.62
MUF:	242	0.3822	0.20	32.81

[a] *Permuting 44 singly polymorphic silent sites 10,000 times. See Table 1 for definitions.*

[b] *Permuting 37 multiply polymorphic silent sites 10,000 times.*

There was a significantly large number of uncondensed fragments of length 2bp, corresponding to third-position sites in adjacent codon positions which both differ in some pair of strains ($P \approx 0.0035$: 18% of observed uncondensed fragments as opposed to 14% for the simulated distribution; these were the only uncondensed fragments of length 3 or less in either the observed or the permuted data sets). There was also an excess of condensed fragments of length 0 ($P \approx 0.0067$: 47% as opposed to 42%), corresponding to an excess of observed DNA segments that are monomorphic in amino-acid synonymous positions but bounded by sites at which two strains simultaneously differ. The *E. coli phoA* data set also shows an excess of empty condensed fragments ($P \approx 0.0001$; see below). These observations suggest the possibility of mutation caused by "templating by local DNA sequences" which "can account for simultaneous multiple mutations" (Golding and Glickman, 1985, 1986; see also Koch, 1971; Milkman and Crawford, 1983; Powers and Smithies, 1986). Gene conversion is an example of such templating. The excess of short fragments in the *E. coli* data sets are not artifacts of the method, since neither the human influenza data set, the protein data set, nor the mutation-only simulated data set of Section 6 below had significant excesses of short condensed fragments. There were also highly significant excesses of short condensed fragments in the two *E. coli* data sets if all silent sites were permuted, rather than within degeneracy classes. However, the simulation results in Section 6 below show

that an excess of short fragments can result from a pattern of gene conversion of segments of length 50–200bp, so that these short-range correlations might be viewed as additional evidence for the presence of gene conversions perhaps in this size range.

The clustering described by the excess of short fragments does not show up for traditional measures of polymorphism along the genome. For example, the observed positions of silent codon polymorphisms are not significantly clustered in the *gnd* data set in comparison with random permutations of codon positions for (i) the number of contiguous groups of two or more silent polymorphic codon positions (a statistic used by Golding and Glickman, 1986), (ii) the sum of the lengths squared of contiguous groups, or (iii) the maximum length of a contiguous group. These codon locations are also nonsignificant for the chi-square runs test used by Brown and Clegg (1983) and Maeda *et al.* (1988). Similarly, the events polymorphic/monomorphic for sites are not significantly autocorrelated along the *gnd* sequenced region (Sawyer *et al.*, 1987). Thus tests based on the pairwise fragments defined above may have greater statistical power for detecting simultaneous local multiple mutations than tests based on contiguous polymorphic positions.

4. The *phoA* Locus in *E. coli*.

DuBose *et al.* (1988) describe DNA sequence data for nine strains of *E. coli* at the *phoA* locus, for which the largest open reading frame has 1413 bases. Two of the nine strains were identical on this reading frame and were combined, leaving eight strains. Out of 471 codon positions in the reading frame, 10 were amino-acid polymorphic. Of the 518 potentially degenerate sites within the remaining amino-acid monomorphic codon positions, 61 are silent polymorphic. Of the 61 sites, 28 are singly polymorphic, 15 sites were 2-fold degenerate, 2 sites were 3-fold degenerate, 35 sites were 4-fold degenerate, and 9 sites were within leucine or arginine codons. Isoleucine was treated as 3-fold degenerate since the ATA codon appeared 12 times in the data set for *phoA*. The same analysis as in Table 1 led to the data in Tables 5–6. The results for SSCF and SSUF in Table 5 give strong evidence for intragenic conversion at the *phoA* locus among these strains. The same conclusion had been obtained previously by DuBose *et al.* (1988) using seven of the nine strains. The four largest condensed and uncondensed fragments are given in Table 6, along with P-values for the fragment lengths as determined by the permuted MCF and MUF scores. Note that while the observed values of SSCF and SSUF are highly significant ($P < 10^{-3}$), no individual fragment is significant at the 5% level (Table 6). Note that, in contrast with Table 2, three of the 5 distinct fragments in Table 6 overlap ends of the aligned region, suggesting the possibility of

gene conversion events extending outside the region (DuBose *et al.*, 1988). The *phoA* data showed a highly significant excess of condensed fragments of length 0 ($P \approx 0.0001$; 53% of observed condensed fragments as opposed to 42% of simulated fragments). There was no significant excess of uncondensed fragments of length 2, although uncondensed fragments of lengths 5 and 8 corresponding to two or three codon positions were weakly significant ($0.03 < P < 0.05$). The lower significance for short uncondensed fragments in *phoA* in comparison with *gnd* may be due to the lower level of polymorphism in the *phoA* data set. There was also an excess of uncondensed fragments of 190bp or longer ($P \approx 0.03$, with 5.9% of the observed uncondensed fragments of that length in comparison with 4.3% of the simulated distribution).

Table 5. The *phoA* locus in *E. coli*[a]

Statistic	Observed Score	P-value	S.D. above Mean	S.D. of scores
SSCF:	10820	0	5.36	661.5
MCF:	60	0.0574	1.68	8.4
SSUF:	7846182	0.0003	4.43	407499.3
MUF:	1281	0.1418	1.30	203.3

[a] *Permuting 61 silent polymorphic sites in 8 strains 10,000 times. See Table 1 for definitions.*

Table 6. The largest fragments in *phoA*

Condensed fragments with P-values:					
K/3:	133–1413	(60) [0.057]	4/5:	778–1145	(23) [1.000]
5/7:	796–1352	(34) [0.945]	1/7:	1–776	(21) [1.000]
Uncondensed fragments with P-values:					
K/3:	133–1413	(1281) [0.142]	5/7:	796–1352	(557) [1.000]
1/7:	1–776	(776) [0.915]	1/4:	1–548	(548) [1.000]

NOTE — *See Table 2 for explanation of notation.*

5. Human Influenza A Virus.

Buonagurio *et al.* (1986) give DNA sequence data for 890 bases in the NS gene of 15 strains of human influenza type A viruses. The largest open reading frame of the assumed ancestral strain A/WSN/33 consisted of 678 bases; two A/Houston viruses agreed on this segment and were combined. Strain USSR/90/77 was left out because it did not differ from Maryland/2/80

at silent sites. Since these sequences are from RNA viruses, the mutation rate is expected to be much higher, and any effects of gene conversion should be more difficult to detect. Also, the virus strains were gathered at different times over a period of 53 years, which would make between-strain conversion events more difficult although not impossible. The analysis of Section 2 was carried out for the 13 strains on the 678-base reading frame (Table 7). The data set in Table 7 had more polymorphic sites in amino-acid variable codon positions than silent polymorphic sites, so that results are presented for permutations of all polymorphic sites (with polymorphic sites used to define fragments) as well as for silent sites.

Table 7. Human influenza A virus

Silent sites[a]*:*	Observed Score	P-value	S.D. above Mean	S.D. of scores
SSCF:	19639	0.6245	-0.38	1288.2
MCF:	27	0.9276	-1.14	6.2
SSUF:	5423277	0.0588	1.69	253618.0
MUF:	321	0.9577	-1.45	64.8
All sites[b]*:*				
SSCF:	55003	0.2163	0.72	2708.6
MCF:	40	0.5832	-0.27	7.5
SSUF:	2495073	0.0133	2.51	96611.0
MUF:	254	0.3248	0.32	40.4

[a] *Permuting 52 silent polymorphic sites in 13 strains 10,000 times. See Table 1 for definitions.*

[b] *Permuting 120 polymorphic sites in 13 strains 10,000 times.*

As Table 7 indicates, none of the statistics are significant for permutations of silent polymorphic sites, although SSUF is significant for permutations of all polymorphic sites. In contrast to the *E. coli* data, there was no excess (or deficit) of either long or short fragments with respect to permutations of either all silent or all polymorphic sites.

6. Two Simulated *gnd* Data Sets.

Simulated *gnd* data sets were constructed in order to get a quantitative idea of the relative sizes of mutation and gene conversion rates. With sites as traits, there are many equally parsimonious pedigrees with the sequences in the *gnd* data set as end nodes (DuBose, personal communication). In

these pedigrees, the links leading up to the current nodes average about half the distance to the common ancestor. A typical 7-strain pedigree with this property was selected. For convenience, current nodes were taken to be the same distance from the common ancestor, and mutation and gene conversion rates in the pedigree links were assumed proportional to distance. The *gnd* consensus sequence was taken as the common ancestor. Potential mutation times were chosen at random in the time interval since the common ancestor with weights proportion to the number of extant strains, and considered in time order. Each potential mutation was assigned at random to a degenerate site in an extant strain, with a random distinct base as product. Potential transversion mutations were then rejected with probability 50%, as well as any mutation that changed an encoded amino acid. These rules were introduced so that simulated data sets would have, on the average, approximately the same relative number of two-fold and four-fold degenerate sites as the observed data set. Surviving mutations were propagated up the pedigree. The expected number of potential mutations needed to produce 81 silent polymorphic sites was computed to be 216. A random data set was then constructed with 216 potential mutations and no gene conversions, and led to the results in Table 8. Note that none of the statistics SSCF, MCF, SSUF, and MUF have significant observed values. This simulated data set also did not have a significant excess of either long or short condensed or uncondensed fragments.

Table 8. A simulated *gnd* data set. Mutation only[a]

Statistic	Observed Score	P-value	S.D. above Mean	S.D. of scores
SSCF:	6008	0.2719	0.54	413.57
MCF:	17	0.7002	-0.48	3.80
SSUF:	850780	0.5083	-0.11	45199.92
MUF:	176	0.7723	-0.75	34.93

[a] *Permuting 78 polymorphic sites in 7 strains 10,000 times. See Table 1 for definitions.*

Gene conversion between strains in the sample (or their ancestors) was modeled by randomly choosing 80 gene conversion times in the time interval since the common ancestor with weights $n*(n-1)$, where n is the number of extant strains. Gene conversion times and potential mutation times were sorted together and considered in time order. Each gene conversion was assumed to map a random segment of length 50–200bp from a randomly chosen extant strain to the same positions in a second randomly chosen extant strain. This amount of gene conversion drastically reduced the amount of polymorphism

generated by 216 potential mutations. It was found by experimentation that approximately 414 potential mutations were needed to generate numbers of silent polymorphic sites in the range 60–99. A random data set was generated with 414 potential mutations and 80 gene conversions under these conditions, and led to the results in Table 9. Note that the P-values of SSCF, MCF, SSUF, and MUF are quite close to those of Table 1. As in Table 4, using only the 45 singly polymorphic silent sites in the data set of Table 9 led to marginally significant results for SSCF and SSUF ($0.01 < P < 0.04$), while the 38 multiply polymorphic silent sites led to much more significant results than in Table 9 (SSCF and SSUF were 12.28 and 10.60 standard deviations above the mean, respectively). Simulated data sets constructed with 60 gene conversions of 50–200bp, or 130 gene conversions of 35–150bp, had less significant (although still significant) P-values. The simulated data set of Table 9 had a weakly significant excess of uncondensed fragments of length 2 ($P \approx 0.03$: $P \approx 0.06$ for length 5), and a highly significant excess of uncondensed fragments of 114–275bp ($P < 10^{-3}$; the longest uncondensed fragment was 275bp). There was also a highly significant excess of empty condensed fragments ($P \approx 0.0008$). Thus the excess of short fragments in the *E. coli* data sets could have been caused by gene conversion of segments of length 50–200bp.

Table 9. A simulated *gnd* data set with 80 gene conversions[a]

Statistic	Observed Score	P-value	S.D. above Mean	S.D. of scores
SSCF:	6306	0	7.46	259.18
MCF:	30	0.0011	5.01	3.00
SSUF:	893223	0	6.71	29992.34
MUF:	275	0.0017	4.40	25.95

[a] *Permuting 83 silent polymorphic sites in 7 strains 10,000 times. See Table 1 for definitions.*

Tables 8–9, and the excess of short fragments in the data set of Table 9, suggest that gene conversion may be the cause of the very significant P-values in Tables 1 and 5 for the *E. coli* data sets. If so, the relative rates of gene conversion and the sizes of the segments involved may not be radically different from those of Table 9. The simulated data set had $414/274 \approx 1.51$ potential mutations per silent site and $80 * ((50 + 200)/2)/768 \approx 13.0$ gene conversions affecting a typical site. Thus, short-segment gene conversion may be 8–10 times more common per base than mutation. However, confidence intervals about these estimates should be considered quite large. The effect of the simulated gene conversion on the amount of surviving polymorphism is also

interesting. Approximately twice the neutral mutation rate was required to generate the same number of neutral polymorphic sites with this amount of gene conversion. Also, the random data set of Table 9, in addition to having 83 silent polymorphic sites, had 64 monomorphic silent sites with mutations since the common ancestor. Random data sets generated with 414 potential mutations and no gene conversion typically had 0–7 monomorphic silent sites with ancestral mutation. Thus about as much neutral polymorphism may have been destroyed by gene conversion in the observed *gnd* data set as has survived.

The observed *gnd* data set has 11 silent polymorphic sites in which 3 or more bases occur in the polymorphism. Random data sets generated under the conditions of Tables 8–9 tended to have half that many or fewer, which may be an indication that gene conversion from strains outside the sample may have significant effects. Gene conversion from outside strains was crudely modeled by choosing times at random since the common ancestor weighted with the number of extant strains, then choosing a random segment of length 50–200bp in a strain existing at that time, and then treating all sites that were silent polymorphic at that time in that segment as potential mutation sites in the sense described above. This treats an external gene conversion as a block of simultaneous neutral mutations occurring at silent polymorphic sites. A third random data set was constructed with 340 potential mutations, 60 internal gene conversion events of 50–200bp, and 20 external gene conversion events of the same length. This data set had 82 polymorphic sites, of which 15 had 3 or more bases in the seven strains, and had P-values for SSCF and SSUF that were more significant than those in Table 9. Simulations with more than 20 external gene conversions (modeled in this manner) with 70–90 silent polymorphic sites tended to have many more complex polymorphisms than the observed data set. However, this simulation should be repeated with a more realistic model of external gene conversion.

7. Protein Sequences in Bacteria.

Sneath *et al.* (1975, Table 2) have protein sequences for cytochrome *c*-551 for nine *Pseudomonas* and *Azotobacter* bacterial strains. The sequences have 82 amino-acid positions, of which 52 are polymorphic and 16 are singly polymorphic. Sneath *et al.* (1975) analyze these data for conserved sequences by constructing tables of pairwise incompatibility indices between pairs of positions in the protein. These tables provide qualitative information about conserved sequences, but do not easily lend themselves to tests of statistical significance. The tests of Section 2 applied to these data give overall measures of the size of conserved segments between pairs of strains, along with P-values

for the significance of these measures. Biological mechanisms that could cause gene conversion events between bacteria this distantly related are apparently not known. Thus conserved sequences between pairs of strains are more likely to be due to selection than to gene conversion.

The results of the tests of Section 2 for permutation of all polymorphic amino-acid sites in these protein sequences show a weak tendency for conserved segments (Table 10). Table 10 also provides an example in which the P-values for MCF and MUF are small but not markedly different from those for SSCF and SSUF. An analysis of the fragment lengths shows no excess (or deficit) of short condensed or uncondensed fragments, but a weak tendency for an excess of uncondensed fragments of 9–16 codon positions ($0.02 < P < 0.05$).

Table 10. Cytochrome *c*-551[a]

Statistic	Observed Score	P-value	S.D. above Mean	S.D. of scores
SSCF:	6656	0.0966	1.36	611.66
MCF:	50	0.0776	1.55	6.91
SSUF:	20306	0.0647	1.64	1627.02
MUF:	80	0.0579	1.52	11.98

[a] *Permuting 52 polymorphic amino acid positions in 9 strains 10,000 times. See Table 1 for definitions.*

8. A Test for External Gene Conversion.

The definition of fragments given above were designed to detect gene conversion in which both source and target sequences were in the sample. An alternative definition of *outer fragments* can be formulated for detecting gene conversion or reciprocal recombination from outside the sample. Given an individual sequence, consider the set of e silent polymorphic sites at which that sequence agrees with at least one other sequence in the sample. These e sites define a partition of the sequence into $e + 1$ (*uncondensed*) *outer fragments*. Thus an outer fragment consists of the silent polymorphic sites at which the given sequence has a unique base in the sample, together with adjacent monomorphic sites and sites within amino-acid polymorphic codon positions, that are bounded by either two of the e sites or one of the e sites and a sequence end. A *condensed outer fragment* is the set of silent polymorphic sites in an uncondensed outer fragment. If one of the strains in the sample has undergone gene conversion from or reciprocal recombination with a strain which has had a very different history from the strains in the sample (but which nevertheless codes for the same protein), then a large outer fragment could result.

No outer fragment for either of the two *E. coli* data sets contained more than 3 silent polymorphic sites. The outer fragment analogs of SSCF, SSUF, MCF, and MUF were essentially nonsignificant not only for the two *E. coli* data sets, but also for the three simulated *gnd* data sets of Section 6, even though the third simulated *gnd* data set was explicitly constructed to model external gene conversion. Thus it appears that the outer fragment analogs of these statistics are not powerful enough to detect external gene conversion at the rate they occur in *gnd* and *phoA*.

However, the outer fragment analogs of SSCF and MCF were significant for the protein data set of Table 10 ($0.01 < P < 0.05$ in both cases), although SSUF and MUF were not significant. The P-values of the four statistics in Table 10 are of comparable size, in contrast to most of the *E. coli* tests in which the SS statistics are much more significant than the M statistics. These observations together indicate that the shape of the pairwise conserved segments in the protein data is of a different qualitative character than in the *E. coli* data sets. The nature of this qualitative difference is an interesting open question.

Acknowledgments

We would like to thank R. DuBose, D. Dykhuizen, and D. Hartl for many helpful discussions, and for access to their data and ms. before publication. We would also like to thank W. Fitch and J. Felsenstein for helpful comments which have lead to many improvements in the manuscript. This work was partially supported by National Science Foundation Grants DMS-8504315 and DMS-8802039.

LITERATURE CITED

1. Brown, A. and M. Clegg. 1983. Analysis of variation in related DNA sequences. Pp. 107–132 *in* B. Weir, ed. Statistical analysis of DNA sequence data. Marcel Dekker, New York.

2. Buonagurio, D., S. Nakada, J. Parvin, M. Krystal, P. Palese, and W. Fitch. 1986. Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene. Science **232:**980–982.

3. DuBose, R., D. Dykhuizen, and D. Hartl. 1988. Genetic exchange among natural isolates of bacteria: Recombination within the *phoA* locus of *Escherichia coli*. Proc. Nat. Acad. Sci. USA **85:**7036–7040.

4. Golding, G. B. and B. Glickman. 1985. Sequence-directed mutagenesis: Evidence from a phylogenetic history of human α-interferon genes. Proc. Nat. Acad. Sci. USA **82:**8577–8581.

5. Golding, G. B. and B. Glickman. 1986. Evidence for local DNA influences on patterns of substitutions in the human α-interferon gene family. Can. J. Genet. Cytol. **28:**483–496.

6. Koch, R.. 1971. The influence of neighboring base pairs upon base-pair substitution mutation rates. Proc. Nat. Acad. Sci. USA **68:**773–776.

7. Li, W.-H., C.-I Wu, and C.-C. Luo. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2:**150–174.

8. Maeda, N., C.-I. Wu, J. Bliska, and J. Reneke. 1988. Molecular evolution of intergenic DNA in higher primates: pattern of DNA changes, molecular clock, and evolution of repetitive sequences. Mol. Biol. Evol. **5:**1–20.

9. Milkman, R. and I. Crawford. 1983. Clustered third-base substitutions among wild strains of *E. coli*. Science **221:**378–380.

10. Powers, P. and O. Smithies. 1986. Short gene conversions in the human fetal globin gene region: a by-product of chromosome pairing during meiosis? Genetics **112:**343–358.

11. Sawyer, S., D. Dykhuizen, and D. Hartl. 1987. Confidence interval for the number of selectively neutral amino acid polymorphisms. Proc. Nat. Acad. Sci. USA **84:**6225–6228.

12. Sneath, P., M. Sackin, and R. Ambler. 1975. Detecting evolutionary incompatibilities from protein sequences. Systematic Zoology **24:**311-332.

13. Stephens, J. C.. 1985. Statistical methods of DNA sequence analysis: Detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2:**539–556.