JOURNAL OF BACTERIOLOGY, Nov. 1992, p. 6886–6895
0021-9193/92/216886-10$02.00/0

Vol. 174, No. 21

Evolutionary Genetics of the Proline Permease Gene (*putP*) and the Control Region of the Proline Utilization Operon in Populations of *Salmonella* and *Escherichia coli*

KIMBERLYN NELSON* AND ROBERT K. SELANDER

Institute of Molecular Evolutionary Genetics, Mueller Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802

Received 16 June 1992/Accepted 27 August 1992

Virtually complete sequences (1,467 bp) of the proline permease gene (*putP*) and complete sequences (416 to 422 bp) of the control region of the proline utilization operon were determined for 16 strains of *Salmonella*, representing all eight subspecies, and 13 strains of *Escherichia coli* recovered from natural populations. Strains of *Salmonella* and *E. coli* differed, on average, at 16.3% of *putP* nucleotide sites and 17.5% of control region sites; the average difference between strains was much larger for *Salmonella* strains (4.6% of *putP* sites and 3.4% of control region sites) than for *E. coli* (2.4 and 0.9%, respectively). There was no difference in the distribution of polymorphic amino acid positions between the membrane-spanning and loop regions of the permease molecule, and rates of synonymous nucleotide substitution were virtually the same for the two domains. Statistical analysis yielded evidence of three probable cases of intragenic recombination, including the acquisition of a large segment of *putP* by strains of *Salmonella* subspecies VII from an unidentified source, the exchange of a 21-bp segment between two strains of *E. coli*, and the acquisition by one strain of *E. coli* of a cluster of 14 unique polymorphic control region sites from an unknown donor. An evolutionary tree for the *putP* and control region sequences was generally concordant with a tree for the *gapA* gene and a tree based on multilocus enzyme electrophoresis, thus providing evidence that for neither gene nor for enzyme genes in general has recombination occurred at rates sufficiently high or over regions sufficiently large to completely obscure phylogenetic relationships dependent on mutational divergence. It is suggested that the recombination rate varies among genes in relation to functional type, being highest for genes encoding cell surface and other proteins for which there is an adaptive advantage in structural diversity.

With the objective of understanding the evolutionary mechanisms that generate genotypic diversity and determine genetic population structure in bacteria, we are currently examining a representative sample of strains of *Salmonella* and *Escherichia coli* for nucleotide sequence variation in several chromosomal genes encoding proteins that serve a variety of cellular functions. Earlier, we reported the results of an analysis of *gapA*, the structural gene of the soluble glycolytic enzyme glyceraldehyde-3-phosphate dehydrogenase (32), and here we present a similar analysis of the proline permease gene (*putP*) and the associated control region of the proline utilization (*put*) operon (23).

The *put* operon, which is located at 22 and 23 min on the *Salmonella* and *E. coli* chromosomes, respectively (2, 42), consists of two divergently transcribed structural genes, *putP* and *putA*, separated by the *put* control region, a segment of approximately 420 bp containing regulatory elements. Proline permease is an integral membrane-spanning protein that mediates transport of proline into the cell, where it is degraded to glutamate for use as a nitrogen or carbon source by the action of a bifunctional (oxidase-dehydrogenase) enzyme encoded by *putA*. Expression of *putP* is tightly regulated in relation to the endogenous concentration of proline (7, 34).

Because activity of the *put* operon is not required for cell survival and growth, except when proline is the sole nitrogen source (23), we were interested in determining whether *putP* is subject to less severe constraints on nucleotide substitution than *gapA*, which is evolutionarily conserved in both prokaryotes and eukaryotes (8). There is a marked difference between these genes in synonymous codon usage as measured by the codon adaptation index (CAI) (51), with the highly expressed *gapA* gene having an average index of 0.81 and *putP* having an average index of 0.33. A comparison of the published sequences of *putP* and the *put* control region in *Salmonella* serovar Typhimurium laboratory strain LT2 (12, 26) and *E. coli* laboratory strain K-12 (27, 28) showed a 17% nucleotide difference, compared with a 6% difference for *gapA* (32), but nothing concerning the extent of diversity in *putP* within either species has been reported. The proline permeases of *Salmonella* strains and *E. coli* are also of interest because they can be structurally subdivided by hydropathy analysis and turn potential predictions into membrane-spanning regions, cytoplasmic or periplasmic loop regions, and a tail region (26, 27), thus providing an opportunity to study interdomain variation in rates of nucleotide and amino acid substitution.

MATERIALS AND METHODS

Bacterial strains. A sample of 16 strains of *Salmonella*, including two representatives of each of the eight currently recognized subspecies (21, 45), was selected from *Salmonella* Reference Collection C (6), as follows: subspecies I, strains S3333 (serovar Typhi) and S4194 (serovar Typhimurium); subspecies II, S2985 and S2993; subspecies IIIa, S2980 and S2983; subspecies IIIb, S2978 and S2979; subspecies IV, S3015 and S3027; subspecies V, S3041 and S3044;

* Corresponding author.

subspecies VI, S2995 and S3057; and subspecies VII, S3013 and S3014.

From the *Escherichia coli* Reference Collection (ECOR) (33) and the research reference collection of T. S. Whittam, we selected 12 strains as follows: EC10 and EC14 (ECOR group A); EC32, EC58, and EC70 (ECOR group B1); EC52 and EC64 (ECOR group B2); EC40 (ECOR group D); and E3406, E2666-74, E830587, and E851819 (T. S. Whittam).

This is the same set of strains previously analyzed for sequence variation in *gapA* (32), with the exception that one *E. coli* strain (A8190) was not included in the sample studied here.

For comparative purposes, a partial sequence of *putP* and the *put* control region of *Klebsiella pneumoniae* ATCC 13883 was also determined.

PCR amplification and nucleotide sequencing. From each strain, we extracted total DNA (59) and amplified a 1,890-bp fragment containing 1,467 bp (97%) of the 1,512-bp *putP* gene and the entire control region (416 to 422 bp) of the *put* operon by polymerase chain reaction (PCR) (40). Oligonucleotide primers for amplification of this segment were designed from the published sequences of *Salmonella* serovar Typhimurium LT2 (12, 26) and *E. coli* K-12 (27, 28), as follows: the 5′ primer was 5′-ACCCCCCATGGTGGTGGT TCCCAT-3′, and the 3′ primer was 5′-TGACGGCGGA GCGGAATGATAATG-3′.

Single-stranded DNA was generated by the λ exonuclease procedure (14), and the resulting template was sequenced by the dideoxynucleotide chain termination method with Sequenase (United States Biochemical, Cleveland, Ohio). Both orientations of the 1,890-bp segment were sequenced by the use of additional pairs of internal primers, and the data were assembled and edited with the SEQMAN and SEQMANED programs (DNASTAR, Madison, Wis.).

Nucleotide sequence accession numbers. The sequences reported here have been assigned GenBank accession numbers L01132 to L01159.

RESULTS

Sequence variation in *putP*. Among the *putP* genes of the 16 *Salmonella* strains studied, there were 216 polymorphic sites in the 1,467-bp segment sequenced (Fig. 1 and Table 1). The sequences of pairs of strains differed on average at 4.6% of nucleotide sites and 1.3% of amino acid positions. Most of the variation is attributable to differences between the subspecies (Table 1), with the sequences of strains of the same subspecies being only slightly different or, in the case of subspecies VII, identical (Fig. 1).

Variation among the 12 *putP* sequences of strains of *E. coli* was about half that seen in strains of *Salmonella* (Table 1 and Fig. 2); there were 108 polymorphic nucleotide sites, and the sequences of pairs of strains differed, on average, at 2.4% of nucleotide sites and 0.3% of amino acid positions.

There were 370 polymorphic nucleotide sites among the combined sample of 28 sequences, and the genes of *Salmonella* strains and *E. coli* differed on average at 16.3% of sites. In all, 43 amino acid positions were variable (Table 1), with an average species difference of 5.5%, which includes 20 positions at which all strains of *Salmonella* differed from all strains of *E. coli* (Fig. 3).

Sequence variation in the *put* operon control region. The control region varied in length from 416 to 422 bp among strains of *Salmonella* and *E. coli*. Length variation among strains of either organism involved primarily single-nucleotide insertions or deletions, and to align the two groups of sequences, several additional small insertions and deletions were required, with a resulting total aligned length of 434 bp.

In the *Salmonella* strains, the control region varied in length from 416 to 422 bp, with an average of 420 bp (Fig. 1). The sequences of both strains of subspecies II had a deletion of one copy of the tandem repeat 5′-TAAA-3′ that occurs just upstream of the Shine-Dalgarno sequence for *putP* (Fig. 1). Both strains of subspecies IIIa had a single-base deletion between the Shine-Dalgarno sequence and the start codon, thereby reducing the usual 6-bp interval to 5 bp.

Among the 16 *Salmonella* strains, there were 48 polymorphic control region sites, with an average pairwise difference of 3.4%; the average nucleotide differences within and between subspecies were similar to those for *putP* (Table 2).

The 12 strains of *E. coli* were much less variable than those of *Salmonella* in both the length and the sequence of the control region. In 10 strains, the length was 422 bp, but strain EC70 was missing a single A in a run of 7 A's, and E851819 was missing both this A and a T from a run of 6 T's. There were 18 polymorphic sites, with an average divergence of less than 1% between pairs of strains (Table 2). However, at 14 of these sites, variant nucleotides were found in single strains, and 12 of these substitutions occurred in the sequence of strain EC40 (Fig. 2).

The average pairwise sequence difference between *Salmonella* and *E. coli* control regions was 17.5%, which is similar to the species difference in *putP* (Table 2).

Distribution of polymorphic amino acids among domains. In the inferred proline permease sequence, 297 (61%) of the 489 amino acid positions analyzed are located in membrane-spanning regions. Of the 21 polymorphic amino acid positions in the *Salmonella* proline permeases, 12 (57%) were located in membrane-spanning regions, 5 were in loops, and 4 occurred in the tail region (Fig. 3). Of the seven polymorphic positions in the *E. coli* permeases, one was located in a membrane region, five were in loops, and one occurred in the tail region. These distributions are not significantly different ($\chi^2_{(1)} = 2.34$, $P = 0.126$).

Of the 20 amino acid positions at which all strains of *Salmonella* differed from all strains of *E. coli*, 6 (30%) were in membrane-spanning regions and 14 (70%) were in the other two regions. On the basis of the proportions of amino acids in these domains, the expected numbers were 12 (61%) and 8 (39%), respectively, but the difference is not statistically significant ($\chi^2_{(1)} = 2.52$, $P = 0.11$).

In general, amino acid substitutions within and between species were conservative, especially those occurring in membrane-spanning regions, where 15 of the 20 polymorphisms involved replacement of one hydrophobic amino acid by another hydrophobic amino acid (Fig. 3).

Rates of synonymous and nonsynonymous substitution. For *putP*, we estimated the numbers of synonymous substitutions per 100 synonymous sites (d_S) and nonsynonymous substitutions per 100 nonsynonymous sites (d_N) (29, 30) (Table 3). Overall, there was evidence of strong selective constraint against amino acid replacement, with d_N being less than 7% of d_S in all comparisons within and between species. In neither species were values of d_S for segments of *putP* corresponding to the membrane-spanning and loop regions significantly different, but d_S for *Salmonella* strains was almost twice as large as d_S for *E. coli*. In *Salmonella* strains, d_N was larger for the membrane-spanning regions (0.78) than for the loop and tail regions (0.35), but in *E. coli*, d_N was an order of magnitude larger for the loop and tail regions (0.34) than for the membrane-spanning regions (0.03).

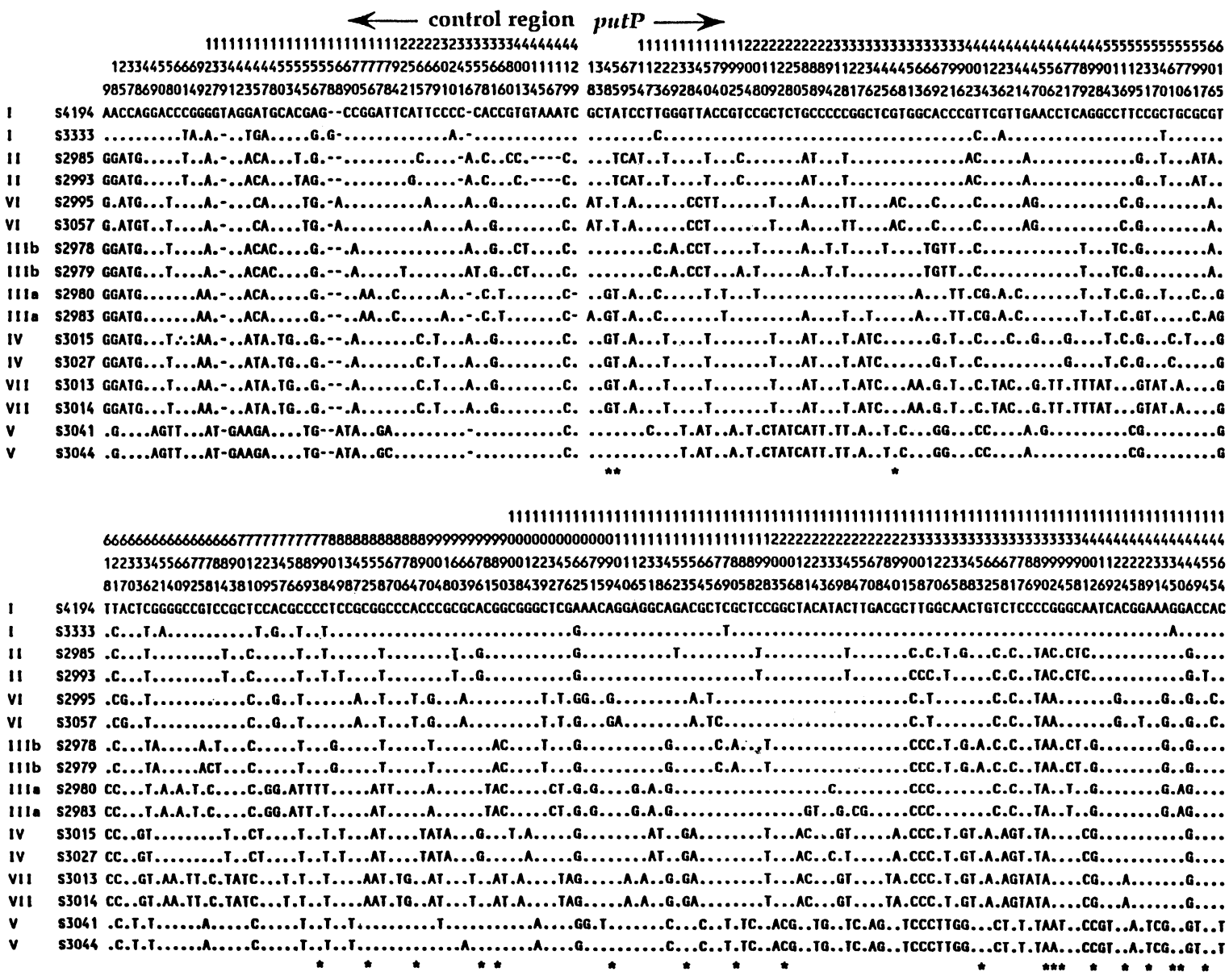

FIG. 1. Nucleotides at polymorphic sites among the *putP* and *put* control region sequences of 16 strains of *Salmonella*. Invariant sites are not shown. Control region sites are numbered (listed vertically) from the alignment of all 28 sequences, with the adenine (A) of the start codon of the *putA* gene numbered 1; *putP* sites are numbered as in the published sequence of *Salmonella* serovar Typhimurium LT2 (26). Asterisks indicate nucleotide substitutions that result in an amino acid replacement.

TABLE 1. Sequence variation in the *putP* gene in 16 strains of *Salmonella* and 12 strains of *E. coli*

Sequence, organism, and sample	No. of polymorphic sites or positions	Mean no. (%) of differences between strains
Nucleotide sites[a]		
Salmonella strains	216	68 (4.6)
Within subspecies		7.2 (0.5)
Between subspecies		75 (5.1)
E. coli	108	35 (2.4)
Salmonella strains vs *E. coli*	370	239 (16.3)
Amino acid positions[b]		
Salmonella strains	21	6.4 (1.3)
Within subspecies		0.5 (0.1)
Between subspecies		6.8 (1.4)
E. coli	7	1.7 (0.3)
Salmonella strains vs *E. coli*	43	26.9 (5.5)

[a] 1,467 nucleotide sites.
[b] 489 amino acid positions.

Distribution of polymorphic nucleotide sites. To test for nonrandom clustering of polymorphic nucleotide sites, a pattern that may be indicative of intragenic recombination, we used Stephens' method (55), which examines the distribution of polymorphic sites relative to the phylogenetic partitions they support. Polymorphic sites supporting a particular phylogenetic partition are expected to be randomly distributed along the sequence if there is no intragenic recombination.

For the 16 *Salmonella* sequences of *putP* and the control region, Stephens' test identified 100 distinctive phylogenetic partitions, but 88 of these partitions were trivial, being supported by five or fewer sites. Analyses of the remaining 12 partitions and of all variable sites tested together are shown in Table 4. The results for all sites indicated that, in general, there were no regions in which consecutive nucleotides were either unusually highly conserved or highly polymorphic. However, there are three partitions for which significant P values for clustered polymorphic sites or segments composed of consecutive unvaried sites were obtained. (In these and other tests of clustering, $P \leq 0.01$ was adopted as the criterion of statistical significance.) Partitions 7 and 10 had significantly long segments of monomorphic sites, and these partitions were further examined for clustered sites. Removal of the long segment of consecutive unvaried sites (between site 38 of the control region and site 858 of *putP* in Fig. 1) from partition 10 indicated that the remaining 4 sites were significantly clustered [$P(d \leq d_o) \leq 3.2 \times 10^{-4}$] ($d_o$ is defined in Table 4, footnote *c*). Further analysis of partition 7 indicated that sites were not clustered [$P(d \leq d_o) \leq 0.61$] and failed to identify other long segments of unvaried sites ($P \leq 0.11$). The only partition for which nonrandom clustering of sites was clearly evident is partition 9. In this case, in the sequences of both strains of subspecies VII (which are identical), 25 unique polymorphic sites were clustered relative to the 5′ region of the *putP* sequence, which had a 785-bp segment of nucleotide sites (from the first site of the control region to site 351 of *putP*) that is identical in sequence to that of subspecies IV (Fig. 1).

For the 12 *E. coli* sequences, Stephens' test identified 46 partitions, of which 39 were trivial. Of the seven partitions with four or more sites, three partitions had significantly clustered sites and two partitions had significantly long segments of unvaried sites (Table 5). Partition 2 corresponds

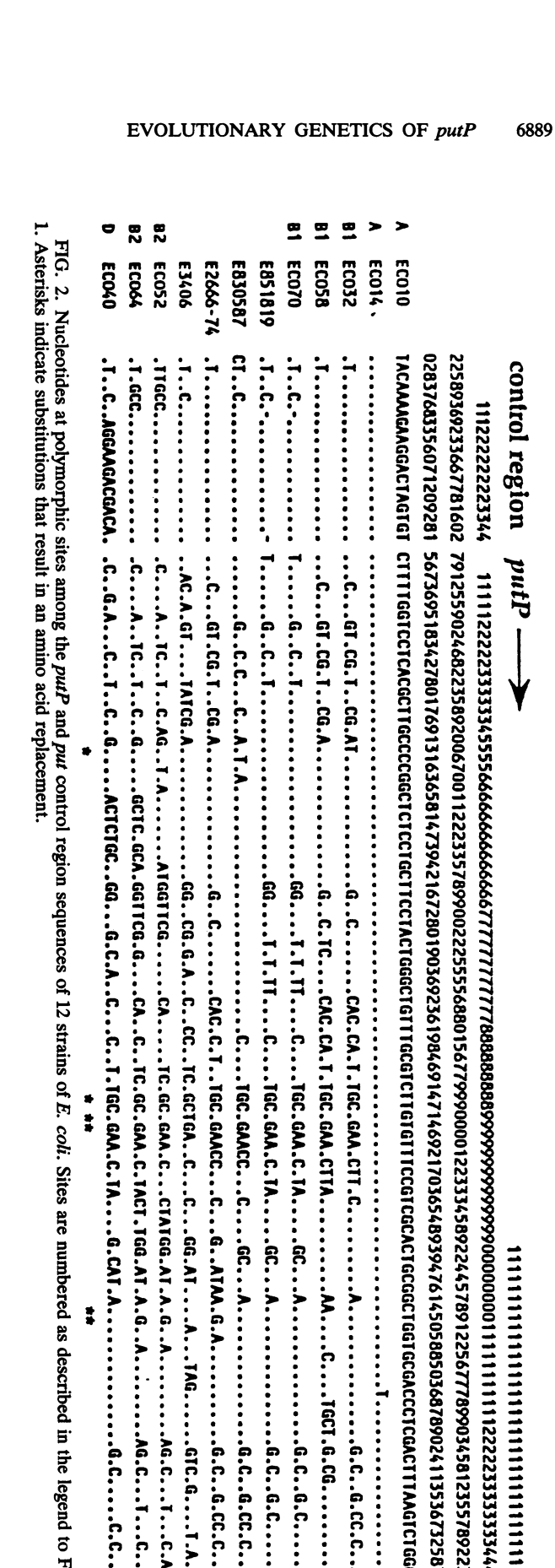

FIG. 2. Nucleotides at polymorphic sites among the *putP* and *put* control region sequences of 12 strains of *E. coli*. Sites are numbered as described in the legend to Fig. 1. Asterisks indicate substitutions that result in an amino acid replacement.

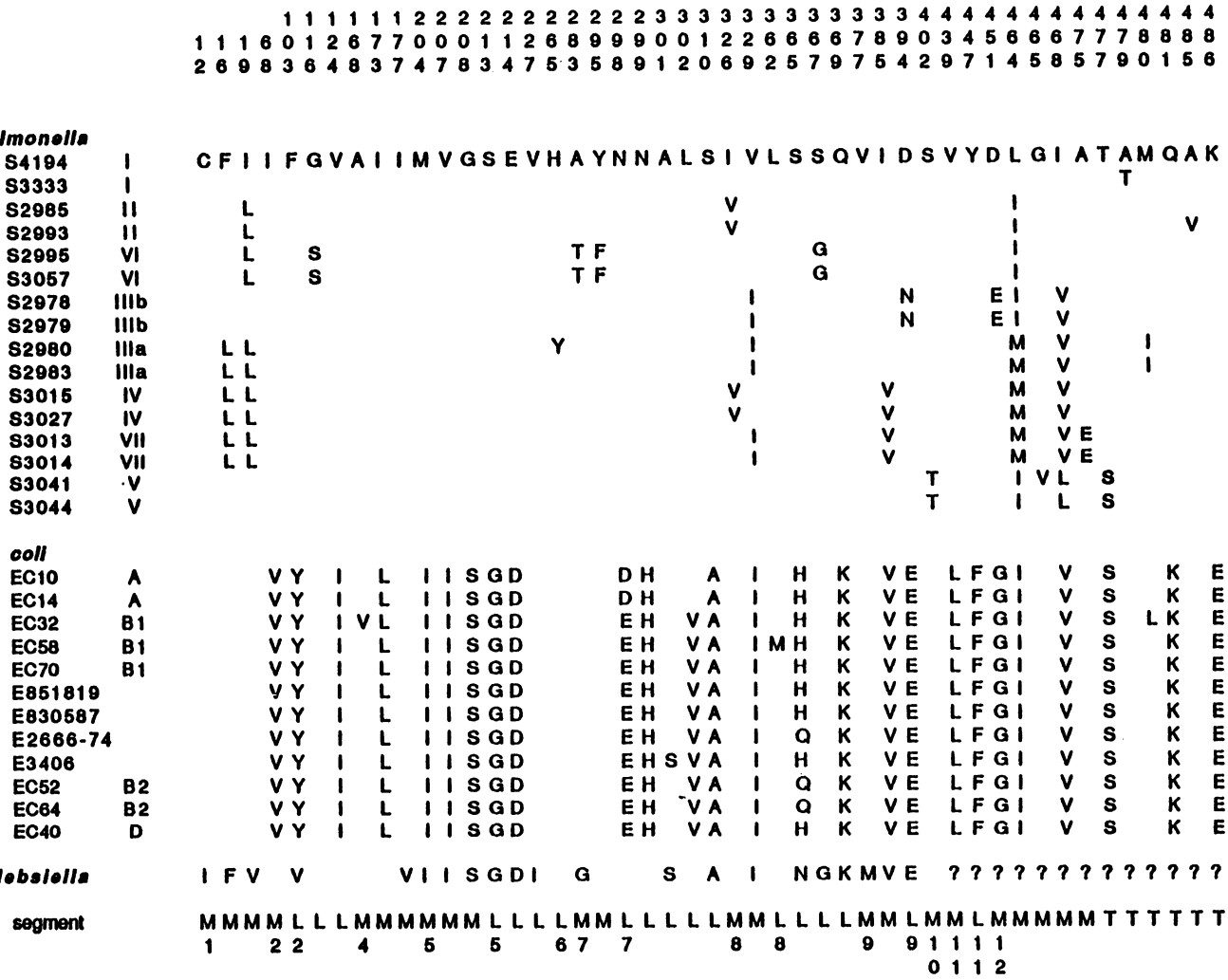

FIG. 3. Distribution of the 43 amino acid polymorphisms among the *putP* sequences of 28 strains of *Salmonella* and *E. coli* and one strain of *K. pneumoniae*. Codon positions (listed vertically) are numbered from the ATG start codon. Standard single-letter amino acid abbreviations are used. Letters at the bottom indicate the structural domains of the protein: M, membrane-spanning region; L, cytoplasmic or periplasmic loop; and T, 3′ tail region.

to a group of seven polymorphic sites in a 227-bp segment at the 3′ end of the *putP* sequence of strain EC58. Partition 4 consists of a cluster of 14 polymorphic sites in the control region and the 5′ end of *putP*, all of which are unique to strain EC40; removal of the significantly long segment of unvaried sites did not affect the clustering of the remaining 13 sites [$P(d \leq d_o) \leq 1.9 \times 10^{-6}$]. Partition 5 includes the sequences of strains of ECOR group A, and removal of two significantly long segments of consecutive unvaried sites indicated that the remaining six polymorphic sites were clustered [$P(d \leq d_o) \leq 3.6 \times 10^{-3}$]. Partition 7 is the most interesting in that the clustering of five polymorphic sites (positions 618, 621, 624, 627, and 633 in Fig. 2) in a 21-bp region in EC40 and EC64 is phylogenetically inconsistent with partition 6.

TABLE 2. Sequence variation in the *put* operon control region in 16 strains of *Salmonella* and 12 strains of *E. coli*

Organism and sample	No. of polymorphic sites	Mean no. (%) of differences between strains[a]
Salmonella strains	48	14.1 (3.4)
Within subspecies		2 (0.5)
Between subspecies		15 (3.6)
E. coli	18	3.8 (0.9)
Salmonella strains vs *E. coli*	104	73.3 (16.3)

[a] The average numbers of sites compared were 420 for *Salmonella* strains, 422 for *E. coli*, and 428 for *Salmonella* strains versus *E. coli*.

Evolutionary tree for *putP* and the control region. A neighbor-joining evolutionary tree (41) for the entire segment of the *put* operon sequenced is shown in Fig. 4, together with a comparable tree for the *gapA* gene based on data previously reported by Nelson et al. (32). The root of the tree was placed by the use of the sequence of a strain of *K. pneumoniae* as an outgroup, and the robustness of the branching order was determined by bootstrap analysis of 1,000 computer-generated trees.

At the first node of the *put* tree, all *Salmonella* sequences diverge from those of *E. coli*. Among the salmonellae,

TABLE 3. d_S and d_N for segments of the *putP* gene encoding different functional domains of proline permease in *Salmonella* and *E. coli* strains

Sample	Results for indicated regions					
	Membrane spanning[a]		Loop and tail[b]		All[c]	
	d_S	d_N	d_S	d_N	d_S	d_N
Salmonella strains (n = 16)	15.90 ± 2.30	0.78 ± 0.33	18.10 ± 3.22	0.35 ± 0.25	16.70 ± 1.88	0.60 ± 0.23
E. coli (n = 12)	9.18 ± 1.76	0.03 ± 0.03	9.84 ± 2.52	0.34 ± 0.25	9.04 ± 1.46	0.15 ± 0.11
Total	50.34 ± 3.26	1.94 ± 0.53	67.62 ± 4.07	4.22 ± 0.95	56.73 ± 2.59	2.86 ± 0.50

[a] 297 codons.
[b] 192 codons.
[c] 489 codons.

sequences of strains of the same subspecies are much more similar to one another than to sequences of other subspecies, and the subspecies V sequences are the most divergent, followed by those of subspecies IV and VII.

Among the *E. coli put* sequences, those of EC40, EC52, and EC64 are the most divergent, and those of EC52 and EC64 of the B2 group of ECOR are very similar, as are those of EC10 and EC14 of the A group of ECOR. Sequences of strains of the B1 group of ECOR (EC32, EC58, and EC70) and those of several other strains are weakly associated (51% or more of bootstrapped trees).

DISCUSSION

Selective constraints on proline permease and the *put* control region. Inasmuch as proline permease is not an essential metabolic enzyme in either *Salmonella* strains or *E. coli*, it should be subject to fewer selective constraints on amino acid replacement than is glyceraldehyde-3-phosphate dehydrogenase (encoded by *gapA*), a key glycolytic enzyme. And, as expected, the average difference in amino acid sequence between pairs of strains was larger for *putP* (5.5%) than for *gapA* (1.3%) (32). Nonetheless, there obviously is strong selection against amino acid replacement in *putP*, as evidenced by the fact that the mean ratio of d_S to d_N was 19:1 for pairwise sequence comparisons.

Our analysis demonstrated that the rate of nucleotide substitution in the *put* control region is no greater than that for *putP*, notwithstanding the fact that most of the control region consists of an unusually long untranslated leader sequence. This conservation in sequence has been explained, at least in part, by evidence that regulation of expression of *putP* involves multisite binding of the *putA* protein to DNA of the control region (34).

Rates of nucleotide substitution. For structural genes of both *Salmonella* strains and *E. coli*, frequencies of the use of alternative codons for amino acids have been shown to vary, depending in part on the rate of gene expression (16), and comparisons of rates of synonymous substitution and CAIs for more than 60 genes of *Salmonella* serovar Typhimurium LT2 and *E. coli* K-12 have identified an inverse relationship between these two variables (50, 52).

Although the respective CAIs of *putP* and *gapA* are virtually the same in *Salmonella* strains and *E. coli*, there is a substantial species difference in the relative rates of synonymous substitution in the two genes. Among the 16 strains of *Salmonella*, d_S for *putP* (CAI = 0.33) was 16.70, which is virtually the same as the value of 15.55 for *gapA* (CAI = 0.79). However, d_S for *gapA* is inflated by inclusion of the strains of subspecies V, which carry a highly divergent recombinant segment derived from a source outside the genus *Salmonella* (32; also see below). With the subspecies V sequences omitted, d_S for *gapA* of *Salmonella* strains is reduced to 10.31, which is 38% less than the value for *putP*. In contrast, for strains of *E. coli*, d_S is 11.6 times greater for *putP* (CAI = 0.33; d_S = 9.04) than for *gapA* (CAI = 0.83; d_S

TABLE 4. Phylogenetic partitions of 16 *Salmonella putP* and *put* control region sequences and tests of nonrandom clustering of polymorphic sites (55)

Partition[a]	s[b]	d_o[c]	g_o[d]	$P(d \le d_o)$	P[e]
1. I/others	8	1,367	674	0.30	0.11
2. I/others	7	1,458	594	0.50	0.36
3. II/others	7	1,465	363	0.50	0.81
4. IIIa/others	17	1,697	324	0.46	0.49
5. IIIb/others	11	1,632	376	0.53	0.63
6. IV/others	6	667	336	0.02	0.27
7. V/others	40	1,842	387	0.70	4.0×10^{-3}
8. VI/others	11	1,619	288	0.51	0.85
9. VII/others	25	1,073	159	1.2×10^{-5}	0.44
10. I, V/others	5	1,276	1,247	0.47	3.8×10^{-5}
11. IV, VII/others	10	1,639	520	0.60	0.34
12. IIIa, IV, VII/others	5	1,183	456	0.38	0.65
All sites	264	1,882	42	0.49	0.36

[a] Subspecies are designated by roman numerals (see Materials and Methods). A slash indicates a partition.
[b] s, number of polymorphic sites.
[c] d_o, observed distance between the two terminal polymorphic sites, in base pairs.
[d] g_o, length of the segment of consecutive nonpolymorphic sites, in base pairs.
[e] P, probability that at least one of $s - 1$ random, independently observed segments is as long as or longer than g_o.
[f] Serovar Typhimurium only.

TABLE 5. Phylogenetic partitions of 12 *E. coli putP* and *put* control region sequences and tests of nonrandom clustering of polymorphic sites (55)

Partition[a]	s[b]	d_o[b]	g_o[b]	$P(d \le d_o)$	P[b]
1. EC52/others	7	1,796	797	0.96	0.28
2. EC58/others	7	227	102	1.8×10^{-5}	0.26
3. E3406/others	14	1,308	450	0.04	0.08
4. EC40/others	14	866	471	3.0×10^{-4}	9.4×10^{-4}
5. EC10, EC14/others	8	1,683	1,296	0.78	9.9×10^{-4}
6. EC52, EC64/others	11	1,696	533	0.69	0.29
7. EC40, EC64/others	5	21	9	5.6×10^{-8}	0.46
All sites	126	1,842	99	0.15	0.10

[a] A slash indicates a partition.
[b] See Table 4, footnotes *b* through *e*, for explanation.

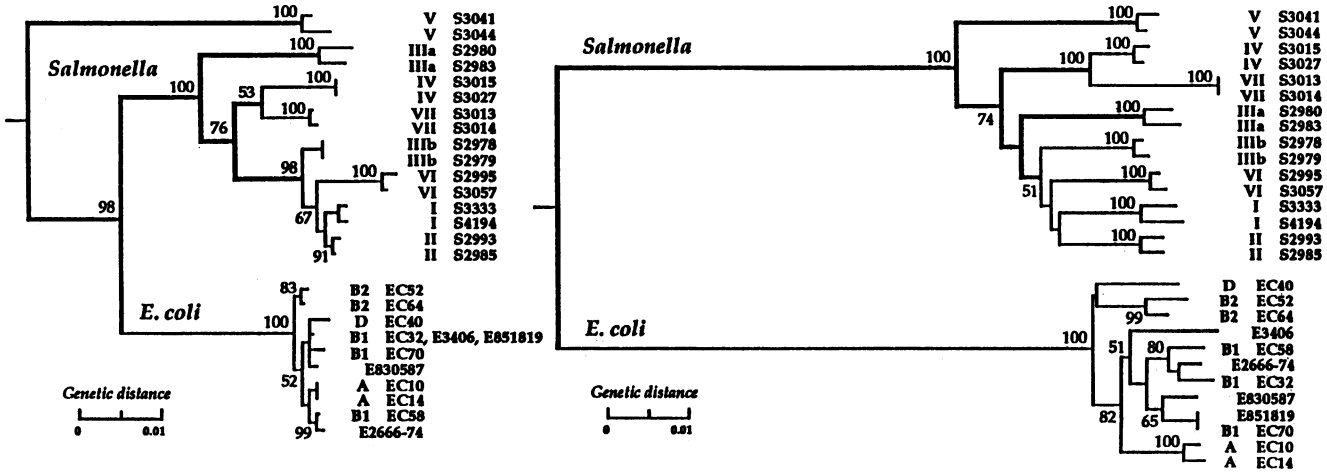

FIG. 4. Evolutionary trees for the *put* operon and *gapA* gene sequences of 16 strains of *Salmonella* and 12 strains of *E. coli*, constructed by the neighbor-joining method (41) from matrices of pairwise distances. A number adjacent to a node indicates the percentage of bootstrap trees that contained that node.

= 0.78). Thus, within species, the CAI seems to be a poor predictor of the rate of synonymous substitution.

Evolutionary relationships among strains. If the rate of substitutive recombination, whether intragenic or assortative (involving entire genes) (58), is low, cell lineages may be expected to evolve more or less independently and phylogenetic trees for different genes will be generally congruent. Hence, comparisons of gene trees, undertaken in conjunction with statistical analyses of the distribution of polymorphic sites in sequences, may permit identification of recombination events.

Several lines of evidence indicate that the genus *Salmonella* and *E. coli* are very distinct groups, between which there is little if any genetic exchange in natural populations (37, 39). Our studies of the *put* operon and *gapA* have failed to identify any recombination events involving the transfer of genetic material between these two bacteria. For the salmonellae, the eight subspecies that have been distinguished on the basis of biochemical characteristics, DNA hybridization experiments, and multilocus enzyme electrophoresis are similarly identified by the nucleotide sequences of both *put* and *gapA*. Moreover, analyses of both genes have indicated that V is the most divergent subspecies and have identified subspecies IV and VII as relatively closely related groups, thus confirming evidence from DNA hybridization experiments (21, 35), multilocus enzyme electrophoresis (38, 45), and biotyping (20).

A comparison of the *put* and *gapA* trees (Fig. 4) shows several differences in topology, some or all of which may be attributed to recombination of gene segments. In the *put* tree, *Salmonella* subspecies V clusters with the other seven subspecies of *Salmonella*, but in the *gapA* tree, it forms a branch apart from both the other salmonellae and *E. coli*, which is inconsistent with all other lines of evidence relating to evolutionary genetic relationships, including DNA hybridization experiments and multilocus enzyme electrophoresis. Our suggestion (32) that the unusual degree of divergence of *gapA* in subspecies V is a consequence of the acquisition, by horizontal transfer, of a segment of the gene from a source outside both the genus *Salmonella* and *E. coli* has since been supported by the discovery of a region of almost identical sequence in *K. pneumoniae* (18, 31).

Apart from the position of subspecies V, the topologies of the *put* and *gapA* trees for *Salmonella* subspecies are generally similar, with subspecies I, II, IIIb, and VI showing the same relationships. However, the positions of the branch leading to subspecies IV and VII and that leading to subspecies IIIa are reversed in the two trees; in *gapA*, subspecies IIIa is separated from subspecies I, II, IIIb, VI, IV, and VII at the second node of the *Salmonella* cluster, whereas the subspecies IV and VII branch occupies a comparable position in the *put* tree. This difference in branching order is attributable to the occurrence of a cluster of 25 unique polymorphic sites in the central part of the *putP* sequence in strains of subspecies VII. (The association of subspecies IV with subspecies VII remains because the *put* sequences of strains of these two subspecies are otherwise quite similar.) This part of *putP* of subspecies VII apparently was acquired by horizontal transfer, but we have yet to identify the source. It is clear, however, that the donor must have been a fairly close relative of the known types of *Salmonella*, because interspersed among the 25 unique polymorphic sites in the sequence of subspecies VII are 17 other polymorphic sites that are shared with one of more the *Salmonella* subspecies, particularly subspecies IV. We suggest that the donor was an as-yet-unrecognized form of *Salmonella*, and it is relevant to note that a survey of *Salmonella* strains recovered from cold-blooded vertebrate hosts has recently identified several strains that are strongly differentiated in multilocus enzyme genotype from all eight of the currently recognized subspecies (31).

The total extent of diversity in DNA sequence is much less among strains of *E. coli* than among strains of *Salmonella*, and, consequently, relationships are harder to define; but analyses of both *put* and *gapA* have substantiated the distinctiveness of the A and B2 subgroups of ECOR, as originally defined by multilocus enzyme electrophoresis (13, 47).

Among the *putP* sequences of *E. coli*, those of EC40 and EC64 share a cluster of seven unique polymorphic sites in a

21-bp region. This cluster of sites is phylogenetically inconsistent with evidence from *gapA* sequence analysis (32) and multilocus enzyme electrophoresis (13, 47) that EC64 and EC52 are in total genomic character more similar to one another than either is to EC40. Even the *putP* sequence of EC64 is more similar to that of EC52, except for the 21-bp segment, in which it resembles EC40. The simplest explanation for these shared unique polymorphic sites is an intragenic recombination event between the EC40 and EC64 lineages, and it is interesting that both EC40 and EC64 were recovered in Sweden from women with urinary tract infections, whereas EC52 was isolated from an orangutan housed in a zoo in Seattle, Wash.

A second probable case of recombination in *E. coli* involves the occurrence of a cluster of 14 unique polymorphic sites in a small segment of the control region of strain EC40, but the donor remains to be identified. Two other cases of clustering of polymorphic sites were identified by our analysis, but because each involved a small number of sites in a single strain, the evidence for recombination is equivocal at best.

In sum, our comparative studies of nucleotide sequence variation in the *put* operon and in *gapA* have identified four cases for which the most plausible explanation for nonrandom distribution of polymorphic sites is horizontal gene transfer and intragenic recombination.

Frequency and evolutionary significance of horizontal gene transfer and substitutive recombination. Population genetic studies of human pathogenic and other bacteria, based largely on the determination of multilocus chromosomal genotypes by enzyme electrophoresis (46), have demonstrated that natural populations of most species have a basically clonal structure and that for many pathogenic species, including *Salmonella* spp., a small proportion of existing genotypes is predominant and widely if not globally distributed (1, 4, 47–49). At the same time—and, at first sight, somewhat paradoxically—studies of nucleotide sequence variation have indicated that chromosomes may have a mosaic phylogenetic structure (25) as a result of the exchange of genetic material among strains of the same or even different species (24, 54).

For the enterobacteria, major concerns of research in which evidence of substitutive recombination has been detected are several genes of *E. coli* (3, 5, 10, 11, 25, 44, 56), the phase 1 flagellin-encoding *fliC* gene of *Salmonella* strains (53), and genes of the *rfb* cluster of *Salmonella* strains, which mediate synthesis of the antigenic O subunit of the cell surface lipopolysaccharide (19, 57). However, numerous examples are available for other bacteria as well (9, 17, 22, 36, 54). In the case of *E. coli*, several studies have indicated that intragenic recombination has had a major part in the generation of allelic diversity at the *gnd* locus, which encodes the metabolic enzyme 6-phosphogluconate dehydrogenase (3, 5, 11, 43). Our own analysis of *gnd* in 25 strains of *E. coli* has extended and confirmed previously reported evidence of a relatively high rate of transfer of segments among strains and has demonstrated that parts of the gene have even been recruited from other species of bacteria, including *K. pneumoniae* (31). In an analysis of variation at synonymous sites in sequence data reported for several genes in strains of *E. coli*, Whittam and Ake (58) found that the value of Hudson's (15) estimator of the neutral-recombination parameter was 2 1/2 to 8 times greater for *gnd* than for the *phoA* alkaline phosphatase locus (10), two open reading frames in the *trp* operon region (56), and *gapA* (32). Whittam and Ake's (58) finding that the level of allozyme variation in 6-phosphogluconate dehydrogenase is nearly three times that expected on the basis of the size of the protein further supports the hypothesis that intragenic recombination is a major factor generating allelic variation at the *gnd* locus. In contrast to the situation in *E. coli*, however, we have identified only a few unequivocal cases of recombination in *gnd* sequences among strains of *Salmonella* (31). Another example of variation in recombination rates among related phylogenetic lineages of bacteria is provided by *Neisseria meningitidis*, in which genes encoding immunoglobulin A1 proteases have recombined with considerably different frequencies in various lineages marked by serogroup (22).

Concluding comments. The picture emerging from comparative studies of sequence variation in genes of the salmonellae and *E. coli* is that intragenic recombination is an important mechanism promoting allelic diversity. However, for most genes, recombination apparently does not occur at rates sufficiently high or over regions sufficiently large to completely obscure phylogenetic relationships dependent upon mutational divergence of lineages or to prevent particular multilocus genotypes from persisting for periods on the order of 100 years, at least, and, in many cases, achieving global distribution. This is true for *gapA* and for *putP* and the *put* control region, and the general concordance of individual neighbor-joining trees for these genes with a tree for the same strains based on electrophoretically demonstrable allelic variation at multiple enzyme loci (32) suggests that it is true for metabolic enzyme genes in general. Nonetheless, it is clearly not the case for *gnd* in *E. coli* or for the phase 1 flagellin-encoding gene (*fliC*) in *Salmonella* spp. (53), in which recombination is a primary proximate source of allelic diversity within populations.

From an adaptive evolutionary standpoint, it seems reasonable to expect that a mosaic structure will most often be evident for genes encoding highly antigenic cell surface proteins, such as flagellins, and those mediating the synthesis of polysaccharides, since a recombination event may be followed by an increase in frequency of the recombinant strain more often than in the case of most other types of genes because of the selective advantage to a cell of presenting altered cell surface structures to the environment (host defense mechanisms and phages) (46). In addition, of course, recombination events that increase resistance to antibiotics may confer a tremendous selective advantage, as in the case of chromosomal genes encoding penicillin-binding proteins in penicillin-resistant strains of *N. meningitidis* and *Neisseria gonorrhoeae* (54). However, for many genes, such as *putP* or *gapA*, that encode polypeptides for which there may be no adaptive premium on diversity in amino acid sequence per se, it seems unlikely that either intragenic or assortative recombination would confer an adaptive advantage to the recipient cell. If a recombinant has no advantage, its likely fate is to be lost from the population through drift or to remain in low frequency. The hypothesis that the recombination rate varies among genes in relation to the functional type of gene product will be tested as additional sequence data become available. For *gnd*, which is a conspicuous exception to this generalization, it has been suggested that the proximity of the *rfb* gene region, which presumably is subject to strong selection for antigenic diversity in the cell surface lipopolysaccharide, diminishes the chance of loss of recombinant *gnd* alleles by genetic drift (5; see also reference 58). If this is so, however, it remains to be determined why the postulated effect apparently has been less severe in *Salmonella* spp. than in *E. coli*.

Our studies demonstrate that, just as one must be cautious in inferring phylogenetic relationships among organisms on the basis of sequence variation in single genes, one should avoid generalizing about recombination rates and other evolutionary processes for entire genomes on the basis of data for single loci, such as *gnd* in *E. coli* or *fliC* in *Salmonella* spp. Moreover, our results also indicate that one cannot safely make generalizations regarding recombination rates in individual genes, let alone genomes, from one bacterial species to another, even if they are phylogenetically closely related.

ACKNOWLEDGMENTS

We thank S. Plock and S. Spigelmyer for technical assistance and S. W. Schaeffer and T. S. Whittam for providing computer programs for data analysis and for helpful discussion and critical reading of various versions of the manuscript. T. S. Whittam also provided strains of *E. coli*. D. Dykhuizen and an anonymous reviewer provided suggestions for improving the manuscript.

This research was supported by grant AI-22144 from the National Institutes of Health.

REFERENCES

1. **Achtman, M.** 1990. Molecular epidemiology of epidemic bacterial meningitis. Rev. Med. Microbiol. **1**:29–38.
2. **Bachmann, B. J.** 1990. Linkage map of *Escherichia coli* K-12, edition 8. Microbiol. Rev. **54**:130–197.
3. **Barcak, G. J., and R. E. Wolf, Jr.** 1988. Comparative nucleotide sequence analysis of growth-rate-regulated *gnd* alleles from natural isolates of *Escherichia coli* and from *Salmonella typhimurium* LT-2. J. Bacteriol. **170**:372–379.
4. **Beltran, P., J. M. Musser, R. Helmuth, J. J. Farmer III, W. M. Frerichs, I. K. Wachsmuth, K. Ferris, J. G. Wells, A. Cravioto, and R. K. Selander.** 1988. Toward a population genetic analysis of *Salmonella*: genetic diversity and relationships among strains of the serotypes *S. choleraesuis*, *S. derby*, *S. dublin*, *S. enteritidis*, *S. heidelberg*, *S. infantis*, *S. newport*, and *S. typhimurium*. Proc. Natl. Acad. Sci. USA **85**:7753–7757.
5. **Biserčić, M., J. Y. Feutrier, and P. R. Reeves.** 1991. Nucleotide sequences of the *gnd* genes from nine natural isolates of *Escherichia coli*: evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. J. Bacteriol. **173**:3894–3900.
6. **Boyd, E. F., K. Nelson, and R. K. Selander.** Unpublished data.
7. **Chen, L.-M., and S. Maloy.** 1991. Regulation of proline utilization in enteric bacteria: cloning and characterization of the *Klebsiella put* control region. J. Bacteriol. **173**:783–790.
8. **Doolittle, R. F., D. F. Feng, K. L. Anderson, and M. R. Alberro.** 1990. A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. J. Mol. Evol. **31**:383–388.
9. **Dowson, C. G., A. Hutchison, N. Woodford, A. P. Johnson, R. C. George, and B. G. Spratt.** 1990. Penicillin-resistant viridans streptococci have obtained altered penicillin-binding protein genes from penicillin-resistant strains of *Streptococcus pneumoniae*. Proc. Natl. Acad. Sci. USA **87**:5858–5862.
10. **DuBose, R. F., D. E. Dykhuizen, and D. L. Hartl.** 1988. Genetic exchange among natural isolates of bacteria: recombination within the *phoA* gene of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **85**:7036–7040.
11. **Dykhuizen, D. E., and L. Green.** 1991. Recombination in *Escherichia coli* and the definition of biological species. J. Bacteriol. **173**:7257–7268.
12. **Hahn, D. R., R. S. Myers, C. R. Kent, and S. R. Maloy.** 1988. Regulation of proline utilization in *Salmonella typhimurium*: molecular characterization of the *put* operon and DNA sequence of the *put* control region. Mol. Gen. Genet. **213**:125–133.
13. **Herzer, P. J., S. Inouye, M. Inouye, and T. S. Whittam.** 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. J. Bacteriol. **172**:6175–6181.
14. **Higuchi, R. G., and H. Ochman.** 1989. Production of single-stranded DNA templates by exonuclease digestion following the polymerase chain reaction. Nucleic Acids Res. **17**:5865.
15. **Hudson, R. R.** 1987. Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50**:245–250.
16. **Ikemura, T.** 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13–34.
17. **Kroll, J. S., and E. R. Moxon.** 1990. Capsulation in distantly related strains of *Haemophilus influenzae* type b: genetic drift and gene transfer at the capsulation locus. J. Bacteriol. **172**: 1374–1379.
18. **Lawrence, J. G., H. Ochman, and D. L. Hartl.** 1991. Molecular and evolutionary relationships among enteric bacteria. J. Gen. Microbiol. **137**:1911–1921.
19. **Lee, S. J., L. K. Romana, and P. R. Reeves.** 1992. Cloning and structure of group C1 O antigen (*rfb* gene cluster) from *Salmonella enterica* serovar *montevideo*. J. Gen. Microbiol. **138**:305–312.
20. **Le Minor, L.** Personal communication.
21. **Le Minor, L., M. Y. Popoff, B. Laurent, and D. Hermant.** 1986. Individualisation d'une septième sous-espèce de *Salmonella*: *S. choleraesuis* subsp. *indica* subsp. nov. Ann. Inst. Pasteur/ Microbiol. **137B**:211–217.
22. **Lomholt, H., K. Poulsen, D. A. Caugant, and M. Kilian.** 1992. Molecular polymorphism and epidemiology of *Neisseria meningitidis* immunoglobulin A1 proteases. Proc. Natl. Acad. Sci. USA **89**:2120–2124.
23. **Maloy, S. R.** 1987. The proline utilization operon, p. 1513–1519. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, vol. 2. American Society of Microbiology, Washington, D.C.
24. **Mazodier, P., and J. Davies.** 1991. Gene transfer between distantly related bacteria. Annu. Rev. Genet. **25**:147–171.
25. **Milkman, R., and M. M. Bridges.** 1990. Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics **126**:505–517.
26. **Miller, K., and S. Maloy.** 1990. DNA sequence of the *putP* gene from *Salmonella typhimurium* and predicted structure of proline permease. Nucleic Acids Res. **18**:3057.
27. **Nakao, T., I. Yamato, and Y. Anraku.** 1987. Nucleotide sequence of *putP*, the proline carrier gene of *Escherichia coli* K12. Mol. Gen. Genet. **208**:70–75.
28. **Nakao, T., I. Yamato, and Y. Anraku.** 1987. Nucleotide sequence of *putC*, the regulatory region for the *put* regulon of *Escherichia coli* K12. Mol. Gen. Genet. **210**:364–368.
29. **Nei, M., and T. Gojobori.** 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.
30. **Nei, M., and L. Jin.** 1989. Variances of the average numbers of nucleotide substitutions within and between populations. Mol. Biol. Evol. **6**:290–300.
31. **Nelson, K., and R. K. Selander.** Unpublished data.
32. **Nelson, K., T. S. Whittam, and R. K. Selander.** 1991. Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli*. Proc. Natl. Acad. Sci. USA **88**:6667–6671.
33. **Ochman, H., and R. K. Selander.** 1984. Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. **157**:690–693.
34. **Ostrovsky de Spicer, P., K. O'Brien, and S. Maloy.** 1991. Regulation of proline utilization in *Salmonella typhimurium*: a membrane-associated dehydrogenase binds DNA in vitro. J. Bacteriol. **173**:211–219.
35. **Popoff, M. Y.** Personal communication.
36. **Poulsen, K., J. Reinholdt, and M. Kilian.** 1992. A comparative genetic study of serologically distinct *Haemophilus influenzae* type 1 immunoglobulin A1 proteases. J. Bacteriol. **174**:2913–2921.
37. **Rayssiguier, C., D. S. Thaler, and M. Radman.** 1989. The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. Nature

(London) **342**:396–401.

38. **Reeves, M. W., G. M. Evins, A. A. Heiba, B. D. Plikaytis, and J. J. Farmer III.** 1989. Clonal nature of *Salmonella typhi* and its genetic relatedness to other salmonellae as shown by multilocus enzyme electrophoresis, and proposal of *Salmonella bongori* comb. nov. J. Clin. Microbiol. **27**:313–320.
39. **Riley, M., and K. E. Sanderson.** 1990. Comparative genetics of *Escherichia coli* and *Salmonella typhimurium*, p. 85–95. *In* K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society of Microbiology, Washington, D.C.
40. **Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich.** 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239**:487–491.
41. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.
42. **Sanderson, K. E., and J. R. Roth.** 1988. Linkage map of *Salmonella typhimurium*, edition VII. Microbiol. Rev. **52**:485–532.
43. **Sawyer, S.** 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6**:526–538.
44. **Sawyer, S. A., D. E. Dykhuizen, and D. L. Hartl.** 1987. Confidence interval for the number of selectively neutral amino acid polymorphisms. Proc. Natl. Acad. Sci. USA **84**:6225–6228.
45. **Selander, R. K., P. Beltran, and N. H. Smith.** 1991. Evolutionary genetics of *Salmonella*, p. 25–57. *In* R. K. Selander, A. G. Clark, and T. S. Whittam (ed.), Evolution at the molecular level. Sinauer Associates, Sunderland, Mass.
46. **Selander, R. K., D. A. Caugant, H. Ochman, J. M. Musser, M. N. Gilmour, and T. S. Whittam.** 1986. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl. Environ. Microbiol. **51**:873–884.
47. **Selander, R. K., D. A. Caugant, and T. S. Whittam.** 1987. Genetic structure and variation in natural populations of *Escherichia coli*, p. 1625–1648. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, vol. 2. American Society of Microbiology, Washington, D.C.
48. **Selander, R. K., and J. M. Musser.** 1990. Population genetics of bacterial pathogenesis, p. 11–36. *In* B. H. Iglewski and V. L. Clark (ed.), Molecular basis of bacterial pathogenesis. Academic Press, Inc., San Diego, Calif.
49. **Selander, R. K., N. H. Smith, J. Li, P. Beltran, K. E. Ferris, D. J. Kopecko, and F. A. Rubin.** 1992. Molecular evolutionary genetics of the cattle-adapted serovar *Salmonella dublin*. J. Bacteriol. **174**:3587–3592.
50. **Sharp, P. M.** 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J. Mol. Evol. **33**:23–33.
51. **Sharp, P. M., and W.-H. Li.** 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**:1281–1295.
52. **Sharp, P. M., and W.-H. Li.** 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4**:222–230.
53. **Smith, N. H., P. Beltran, and R. K. Selander.** 1990. Recombination of *Salmonella* phase 1 flagellin genes generates new serovars. J. Bacteriol. **172**:2209–2216.
54. **Spratt, B. G., L. D. Bowler, Q.-Y. Zhang, J. Zhou, and J. Maynard Smith.** 1992. Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal *Neisseria* species. J. Mol. Evol. **34**:115–125.
55. **Stephens, J. C.** 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2**:539–556.
56. **Stoltzfus, A., J. F. Leslie, and R. Milkman.** 1988. Molecular evolution of the *Escherichia coli* chromosome. I. Analysis of structure and natural variation in a previously uncharacterized region between *trp* and *tonB*. Genetics **120**:345–358.
57. **Wang, L., L. K. Raomana, and P. R. Reeves.** 1992. Molecular analysis of a *Salmonella enterica* group E1 *rfb* gene cluster: O antigen and the genetic basis of the major polymorphism. Genetics **130**:429–443.
58. **Whittam, T. S., and S. E. Ake.** Genetic polymorphisms and recombination in natural populations of *Escherichia coli*. *In* N. Takahata and A. G. Clark (ed.), Molecular paleo-population biology, in press. Japan Scientific Society Press, Tokyo.
59. **Wilson, K.** 1990. Preparation of genomic DNA from bacteria, p. 2.4.1–2.4.5. *In* F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl (ed.), Current protocols in molecular biology, vol. 1. John Wiley & Sons, Inc., New York.