JOURNAL OF BACTERIOLOGY, Nov. 1991, p. 7257–7268
0021-9193/91/227257-12$02.00/0

Vol. 173, No. 22

Recombination in *Escherichia coli* and the Definition of Biological Species†‡

DANIEL E. DYKHUIZEN[1]* AND LOUIS GREEN[2]

Department of Ecology and Evolution, State University of New York at Stony Brook, Stony Brook, New York 11794-5245,[1] and Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado 80309[2]

Received 19 April 1991/Accepted 10 September 1991

The DNA sequence of part of the *gnd* (6-phosphogluconate dehydrogenase) gene was determined for eight wild strains of *Escherichia coli* and for *Salmonella typhimurium*. Since a region of the *trp* (tryptophan) operon and the *phoA* (alkaline phosphatase) gene have been sequenced from the same strains, the gene trees for these three regions were determined and compared. Gene trees are different from species or strain trees in that a gene tree is derived from a particular segment of DNA, whereas a species or strain tree should be derived from many such segments and is the tree that best represents the phylogenetic relationship of the species or strains. If there were no recombination in *E. coli*, the gene trees for different genes would not be statistically different from the strain tree or from each other. But, if the gene trees are significantly different, there must have been recombination. Methods are proposed that show these gene trees to be statistically different. Since the gene trees are different, we conclude that recombination is important in natural populations of *E. coli*. Finally, we suggest that gene trees can be used to create an operational means of defining bacterial species by using the biological species definition.

In bacteria, reproduction and exchange of chromosomal genes are discrete and independent functions, not tied together in one process as in most animals and plants. Thus, all reproduction in bacteria is asexual, and all genetic exchange is due to processes other than reciprocal recombination. In this paper, we shall continue to refer to transfer of genetic material from one strain to another as recombination, even though horizontal gene transfer would be more accurate, since recombination is the term used in population genetics to indicate the process that breaks down linkage disequilibrium. The amount of genetic transfer or sex will vary from species to species, and the size of the exchanged fragment of DNA will depend upon which of the three known mechanisms is involved (18). For example, transduction-mediated recombination is expected to transfer sections in the range of tens of kilobases of DNA or a minute or two of the genetic map, whereas conjugation can transfer regions of hundreds to thousands of kilobases. Transformation is unlikely to be important in *Escherichia coli*.

The extent to which recombination affects the genetic structure of bacteria remains a question. If there is no chromosomal recombination in a bacterial species, then all individuals of a species are related by clonal descent, even though they are phenotypically different because of accumulated mutations. If a species is clonal, the phylogenies of different genes from the same strains will be the same; i.e., genes of one strain would share the same most recent common ancestor with genes of another strain. If, on the other hand, recombination is important, the phylogenies of different genes from the same strains will be different; i.e., the common ancestors of different genes from the same pair of strains will be different.

Figure 1 illustrates this idea. Imagine a cell dividing into two cells, each of which is the ancestor to a lineage. One lineage gives rise to both strains A and B, and the other gives rise to the C strain. Later, there is a cell that is the last common ancestor of strains A and B. Then, gene X is transferred from an ancestor of the C strain to an ancestor of the B strain. To determine the relationship of the three strains, A, B, and C, the same gene is sequenced from each strain. The majority of their DNA, if sequenced, would show that A and B are more similar to each other than either is to C. This would lead to the branching order seen in the open bars of Fig. 1. If, however, gene X is sequenced, the data would indicate a strain relationship such as that given by the dark lines in Fig. 1, i.e., that strain B is more closely related to strain C than to strain A. This difference is a consequence of the fact that recombination mixes the phylogenetic relationships of the strains. Thus, if one can show that different genes from the same strains have statistically different phylogenies, either in the branching order as in the example above or in the time of the last common ancestor as judged by the relative rates of accumulated base pair change, then the result indicates that recombination is an important parameter in creating the observed distribution of genotypes in the species.

Over the last few years much information has accumulated to support the clonal model of the population structure of *E. coli* (1, 30, 35). This model suggests that chromosomal recombination is restricted in nature to such a degree that individual cells lines persist as stable clones over long periods of time, so long that members of the same clone can be found around the world. Certain pathogenic clones, in which the various isolates are indistinguishable by electrophoresis, biotyping, and serotyping, have a world-wide distribution and have been isolated over the last 40 years (1, 2, 28). Evidence of the clonal population structure of *E. coli* as a whole is indicated by the strong linkage disequilibrium among many of the 12 enzyme loci analyzed in 302 electrophoretic types representing 1,690 isolates (29, 46).

Even if there is a significant but low rate of recombination,

* Corresponding author.

† This paper is dedicated to the memory of Ralph V. Evans.

‡ Contribution no. 801 from the Graduate Studies in Ecology and Evolution, State University of New York at Stony Brook.

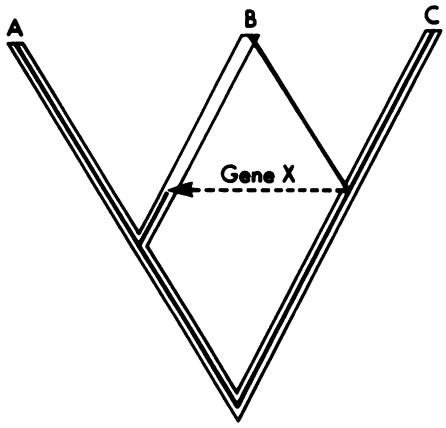

FIG. 1. Effect of recombination on the apparent phylogenetic relationship of three strains. The gene tree of gene X will put strains B and C together, whereas the gene tree for any other gene will place strains A and B together. This difference in the gene trees is caused by the transfer of gene X from an ancestor of strain C into an ancestor of strain B.

E. coli can have a clonal population structure if the species often undergoes purifying or periodic selection (17). If periodic selection is common, then one would expect to find a limited number of isolated clones when the DNA sequences are compared. The sequence results of Milkman and Crawford (22) for the *trpABC* region support this expectation. The sequences can be divided into three groups based on similarity. Nine *E. coli* stains were identical or differed at only a single base from the K-12 sequence. Three strains were identical to each other and different from K-12 at 10 bases. The last strain was different from the K-12 sequence at 44 bases. This result suggests that *E. coli* is composed of a relatively limited number of geographically widespread clones and that recombination plays little if any role in the genetic structure of *E. coli*. The prediction from this model of the population structure of *E. coli* is that the same DNA phylogeny should be found for any segment of the chromosome.

To test this hypothesis, we cloned and sequenced 770 bases in the middle of the *gnd* (6-phosphogluconate dehydrogenase) gene for eight of the *E. coli* strains that were sequenced for *trp* and the homologous region of *gnd* from *Salmonella typhimurium* (9). The data from these two genes, plus additional data on the sequence of the alkaline phosphatase (*phoA*) gene from the same strains, is analyzed by using statistical methods to determine the likelihood of recombination. This analysis has implications for the way species can be defined in bacteria.

MATERIALS AND METHODS

Strains. *E. coli* RM39A, RM191F, RM201C, RM217T, RM45E, RM70B, RM224H, and RM202I were obtained from R. Milkman (22). They are included in the ECOR collection (27) as numbers 4, 16, 45, 67, 69, 70, 68, and 65, respectively. We confirmed that the strains from R. Milkman and the comparable strains from the ECOR collection are the same strains by comparing mobility of 6-phosphogluconate dehydrogenase on cellulose-acetate strips (20) and by comparing the insertion sequence pattern on Southern blots (33). The *S. typhimurium* LT2 used was obtained from J. Roth.

Cloning and sequencing. The DNA sequence of *gnd* for the K-12 strain was obtained from R. E. Wolf (23). The sequence used represents positions 405 through 1172 in the numbering system of Nasoff et al. (23). Southern blots of restriction enzyme digests of genomic DNA from the nine strains were probed with the *gnd* fragment, which was a 2.9-kb *Bgl*II-*Eco*RI fragment covering the 1.4-kb *gnd* gene. This probe was isolated from pMN4, which was provided by R. E. Wolf (24). Restriction enzymes that yielded single hybridizing fragments between 3 and 8 kb in size were used for preparative digests of genomic DNA. For each allele, fragments of about the size of the *gnd* fragment were isolated from agarose gels and cloned into pBR322. Transformants were selected for *gnd* enzyme function (growth on gluconate).

Figure 2 shows the restriction maps of these clones. The allele numbers refer to the strains from which the alleles were obtained. The *gnd* alleles are aligned and oriented such that the direction of synthesis is from left to right. All of the alleles contain a *Pst*I site at nucleotide 1172 (numbering as in reference 23). Thus this site was used for subcloning into M13, and all of the clones were oriented such that this site was adjacent to the universal primer. Two alleles contained *Pst*I sites at nucleotide 405, defining a sequence of 770 bases within *gnd* which was used for the analysis. The other alleles were subcloned into M13 by using convenient restriction sites so that the homologous 770 bases could be sequenced in all alleles. The LT2 strain had no convenient sites, and two subclones had to be used.

Sanger dideoxy sequencing was accomplished initially with an M13 universal primer and later with custom-made oligonucleotides. The sequence of the K-12 allele was compared with the new sequence generated, checking each difference. About 20% was sequenced from both directions. No errors were found when the two strands were compared.

Tree construction and testing. Gene phylogenies were constructed as follows. The percent divergence (p) of DNA sequence for each pair of strains was calculated and converted to the Jukes-Cantor distance (d), which corrects for multiple changes at the same site. The formula for this distance measure and its variants were published by Nei (25). The distance gene trees were constructed by the unweighted pair-group method with the arithmetic mean (25). The standard errors of the branch nodes were calculated by the method of Nei et al. (26).

The parsimony gene trees were constructed by using the phylogenetic analysis program PAUP (version 2.3.2; David Swofford, Illinois Natural History Survey, Champaign) and rooted by using the *Salmonella* sequence as the outgroup. The statistical comparison of different gene trees was done by using the method of Templeton (45) as modified by Felsenstein (11). This test, in its simplest form, is a binomial sign test of whether there are more characters supporting one tree over another, with the null hypothesis being that any character is equally likely to support either of the two trees.

A statistical test for the difference in time to the last common ancestor is based on the assumption that branch length is a measure of time and that the rates of change are the same for different genes. In this test which is based on the t test, the percentage change (p) is used. The distance (d) could also have been used. Since the variance of p is a function of p, the variable has to be transformed such that the variance is constant and independent. Therefore the statistic used is

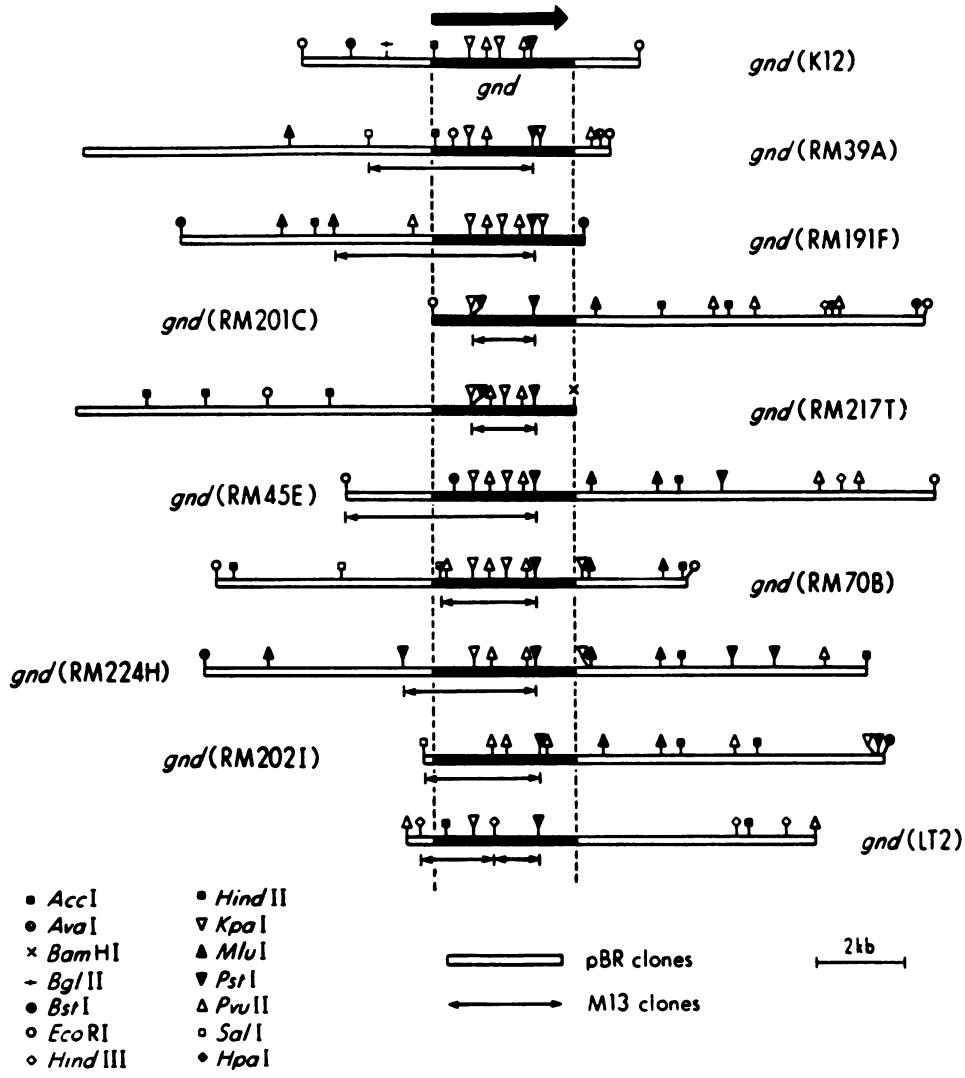

FIG. 2. Restriction map of clones of various alleles of *gnd*. The large thick arrow shows the direction of transcription. The segments cloned into pBR322 are shown as thick bars, with the region containing *gnd* darkened. The regions subcloned into M13 for sequencing are indicated by the double-headed arrows.

$$t = \frac{\arcsin\sqrt{p_{i1}} - \arcsin\sqrt{p_{i2}}}{\sqrt{820.5\,(1/n_1 + 1/n_2)}} \quad (1)$$

where p_{i1} is the proportion of nucleotides different between strain pair i for gene 1, p_{i2} is the proportion for the same pair of strains for gene 2, and n_1 and n_2 are the numbers of nucleotides sampled for genes 1 and 2, respectively (42). Since we do not know whether even the arcsin-transformed percentages will follow a t distribution, given that they were derived from an evolutionary process rather than independent sampling from a constant distribution, we will call these pseudo-t values.

There will be $n(n - 1)/2$ of these pseudo-t tests for n strains. We expect some number of comparisons to be significant by chance. The critical value can be adjusted so that one has confidence at a certain level that comparisons are significant. With multiple tests, the equation of Sidák (38) is used to adjust the significance values:

$$\alpha' = 1 - (1 - \alpha)^{(1/k)} \quad (2)$$

where k is the number of comparisons and α is the desired significance level (usually 0.05). This measure assumes that the tests are independent, which in this case they are not.

Nucleotide sequence accession numbers. The nucleotide sequence data in this paper have been submitted to GenBank. The primary strain numbers are the ECOR numbers with the RM numbers as isolate number. The accession numbers are as follows: M64324 for ECOR4 (RM39A), M64325 for ECOR16 (RM191F), M64326 for ECOR45 (RM201C), M64327 for ECOR67 (RM217T), M64328 for ECOR69 (RM45E), M64329 for ECOR70 (RM70B), M64330 for ECOR68 (RM224H), M64331 for ECOR65 (RM202I), and M64332 for *S. typhimurium* LT2.

RESULTS

Sequences. The sequences are given in Fig. 3. There are about 10^3 chi sites in the genome of *E. coli* or one about every 5,000 bp (39). These sites promote recombination. No chi sites were found in or around the sequenced *gnd* alleles.

Intragenic recombination. The procedure used to deter-

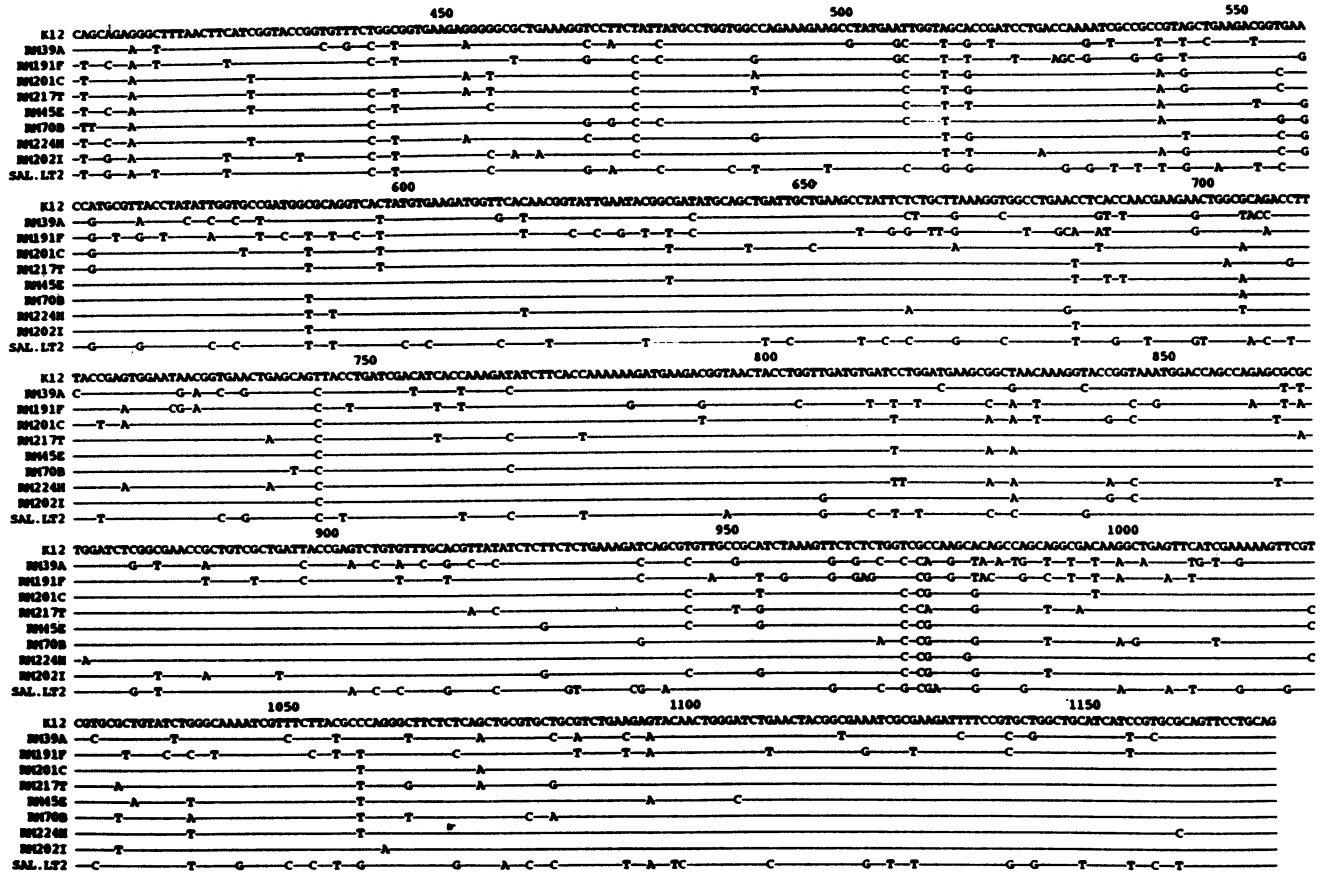

FIG. 3. DNA sequence for a 770-bp internal region of *gnd*. Only nucleotides that are different from the one found in K-12 at the site are listed. Thus sites such as 745, where K-12 contains a T and all other strains contain a C, are sites where the mutation presumably happened in the lineage to the K-12 strain, and the T is a derived character state. The predicted amino acid sequences are different for every strain. The amino acid changes were listed by Sawyer et al. (34). The characteristics of the synonymous site polymorphisms were presented by Sawyer (32).

mine the importance of recombination in *E. coli* is to compare the trees created from different genes in the same strains. This procedure is based on the assumption that the sequences have evolved independently, i.e., that there is no intragenic recombination. Intragenic recombination decreases the resolution of these tests, since it blends sequences together until each is about the same distance from all others. Thus the relative rates of intragenic versus intergenic recombination are important. If recombination typically introduces pieces considerably larger than the region of each gene sequenced, then the frequency of intragenic recombination will be less than the frequency of recombination between two well-spaced genes.

The presence of intragenic recombination has been shown for both *gnd* (4, 32) and *phoA* (7) sequences, where short sequences of 200 to 500 bp have been inserted, and for the region near *trp* (44), where a unique deletion and a unique rearrangement are found in all possible combinations. In addition, although the variation at the *trp* locus is too low to analyze the data statistically, the variation seems to be clustered as if there are either hot spots of mutation (22) or intragenic recombination. It is clear that intragenic recombination occurs, but is it an amount sufficient to invalidate the use of tree-building algorithms?

To answer this question, pairwise G tests for goodness of fit were performed on the *gnd* data. The sequence was broken into 12-bp blocks, giving 64 blocks. The numbers of blocks for each pair of sequences with zero, one, two, three, four, or five and greater differences were determined. The expected number of blocks in each class was estimated by determining the total number of differences and distributing these into blocks by the Poisson process. The determined G values are given in Table 1. For individual values, the chi-square value for five degrees of freedom at 5% significance is 11.07. Only 2 of the 45 values are above this value, which is the expected frequency of false-positives. Since there are multiple nonindependent tests, the critical value for significance was adjusted by using equation 2 to 20.3 for $k = n(n-1)/2$ and 16.3 for $k = n$. No values are significant, and this implies that the changes along the sequence are homogeneous enough to treat the sequences as independent lineages for the purpose of estimating gene trees.

Evidence of intergenic recombination. Table 2 gives the percent differences between pairs of strains for *gnd* and for *trp*. The percent divergence between the *Salmonella* sequence and the *E. coli* sequences for *trp* is between 16.2 and 16.7% with an average of 16.5%; for *gnd* it is between 14.7 and 18.3% with an average of 15.7%. This shows that these genes are evolving at about the same rate. If anything, *trp* seems to be changing a little faster than *gnd*. Thus the generally smaller differences between strains of *E. coli* for

TABLE 1. G values testing for intragenic recombination in *gnd*

Strain	G value for intragenic recombination in *gnd*								
	RM39A	RM191F	RM201C	RM217T	RM45E	RM70B	RM224H	RM202I	LT2[a]
K-12	1.17	6.33	2.76	5.82	1.45	12.83	6.23	3.58	6.38
RM39A		2.48	1.50	1.70	1.70	5.73	4.71	1.83	10.12
RM191F			8.62	11.14	10.41	10.71	9.71	3.60	6.20
RM201C				3.10	9.52	9.50	5.00	1.39	11.03
RM217T					4.83	2.49	2.32	0.97	7.19
RM45E						1.51	4.37	5.18	7.92
RM70B							3.60	2.53	6.30
RM224H								1.81	8.39
RM202I									9.43

[a] *S. typhimurium* strain.

trp than *gnd* cannot be explained by the rate of evolution of *gnd* being faster than that of *trp*.

The consistency of the distances between the *Salmonella* allele and the various alleles in *E. coli* suggest that the rates of evolution within *E. coli* have been the same for all strains (with the exception of the *gnd* of strain RM191F; see below). This means that the branch lengths will be about equal and that the unweighted pair-group method with the arithmetic mean can be used to approximate the true tree. The percent differences were converted to distances, and the trees for *gnd* and *trp* were derived with standard errors of the node position. It thus seems clear from comparing these trees (Fig. 4) that multiple recombinations have taken place. Not only is the order of branching different between the trees, but also the distances are different. Thus different genes within the same pair of strains must have had different cells as the last common ancestors. They could only have been brought into the same strains by recombination.

Statistical analysis of trees to determine intergenic recombination. Although in the case above recombination seems the only explanation for the differences in gene trees, this will not always be the case, since the genetic divergence of DNA depends upon chance events. The gene trees for two genes can be quite different because of chance convergences and chance variations in rates. Thus, statistical analysis must be done to determine when the differences between gene trees are too large to explain by chance and must have been caused by recombination. We present three ways of analyzing gene trees to show that intergenic recombination must be an explanation for the differences between trees.

(i) Sets of strains for each gene with statistically different ancestors. When there is no information about similarity in the rate of evolution for the different genes, a method that depends upon determining which nodes are significantly different from each other and which are not can be used. For example, in Fig. 4 the error bars on nodes 4 through 9 for the *gnd* tree overlap. All of these strains can be considered as having the same common ancestor, and the tree can be represented as seven lines radiating from a common point.

In this manner, strains can be divided into groups that could have had the same common ancestor. The groups are defined by nodes that are not significantly different from each other. By this criterion, the *gnd* sequences form two groups. One group contains the genes from strains RM39A, LT2, and RM191F, whereas the other group contains the genes from all other strains. The *trp* sequences form four groups. The first group includes the *trp* sequences from RM191F, K-12, RM201C, RM39A, and RM217T. The second group contains the *trp* sequences from strains RM70B, RM45E, and RM224H. The third group contains only the *trp* sequence from strain RM202I, and the last group contains only the *trp* sequence from *S. typhimurium* LT2. If there is no recombination, the *trp* and *gnd* groups should match. However, the *trp* groups divide and combine the *gnd* groups (Fig. 5).

The distance tree for *phoA* (Fig. 6) is derived from the data of DuBose et al. (7). Note that the scale is much different for this tree. This tree is more complex, so a systematic method of group creation has to be formalized. The *phoA* alleles from strains K-12, RM191F, and RM201C form a group. This branch is then eliminated, and consequently node 4 is eliminated. The error bars for nodes 5 and 6 overlap, but

TABLE 2. Divergence between pairs of strains[a]

Strain	% Differences between pairs									
	K-12	RM39A	RM191F	RM201C	RM217T	RM45E	RM70B	RM224H	RM202I	LT2[b]
K-12		**14.20**	**15.70**	**5.60**	**5.70**	**4.50**	**4.20**	**5.20**	**4.90**	**15.20**
RM39A	0.10		**17.30**	**15.30**	**14.20**	**14.70**	**13.80**	**15.60**	**15.50**	**15.30**
RM191F	0.00	0.10		**14.90**	**16.40**	**14.70**	**15.60**	**14.70**	**15.60**	**18.30**
RM201C	0.05	0.15	0.05		**5.20**	**5.10**	**6.20**	**5.10**	**5.70**	**16.10**
RM217T	0.10	0.30	0.10	0.15		**5.50**	**5.60**	**6.10**	**5.50**	**15.30**
RM45E	0.90	1.00	0.90	0.95	1.00		**5.20**	**4.50**	**4.40**	**15.20**
RM70B	0.90	1.00	0.90	0.95	1.00	0.00		**6.00**	**5.30**	**14.70**
RM224H	0.90	1.00	0.90	0.95	1.00	0.00	0.00		**5.70**	**15.70**
RM202I	3.10	3.20	3.10	3.15	3.10	3.70	3.80	3.70		**15.20**
LT2	16.30	16.50	16.30	16.30	16.20	16.70	16.70	16.70	16.60	

[a] Values in boldface type are for *gnd*, and those in lightface type are for *trp*.
[b] *S. typhimurium* strain.

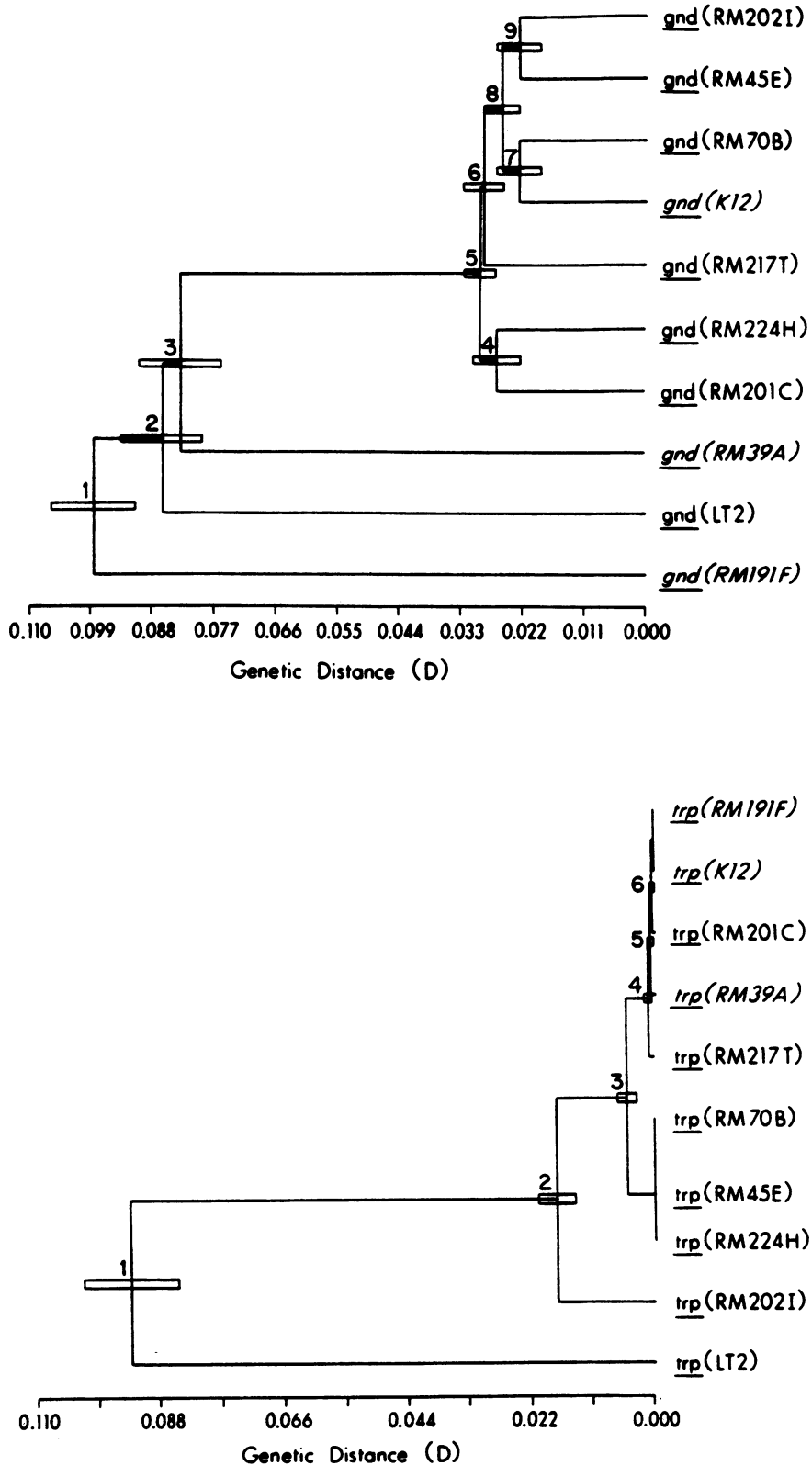

FIG. 4. Distance trees for alleles of *gnd* and *trp* from the same set of strains. All alleles are from *E. coli* except the one from *S. typhimurium* LT2. The bars on the nodes represent the standard errors of the branch points. There is no error bar for the node of *trp* from RM191F and K-12 because the sequences are identical. Likewise for the *trp* sequences from strains RM70B, RM45E, and RM224H. The three stains in italics are the only group A strains.

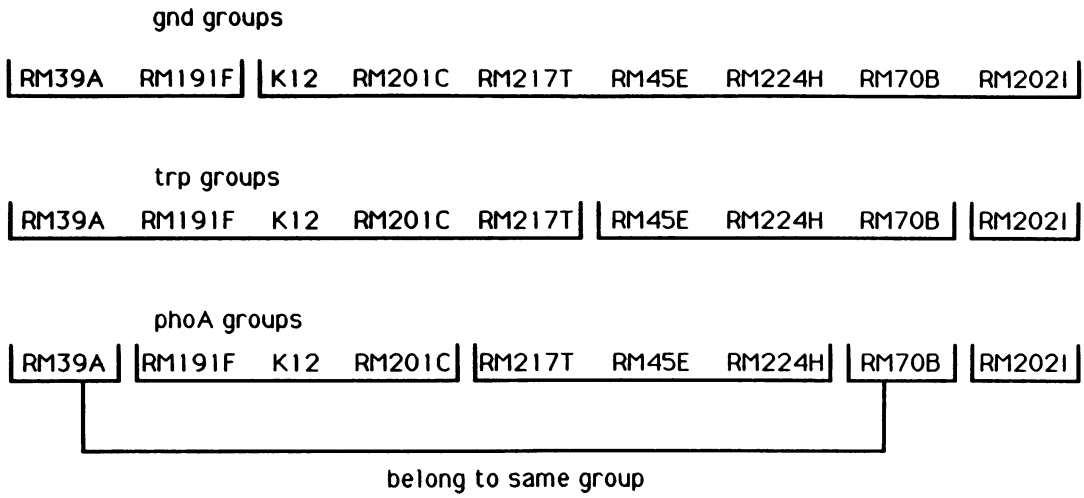

FIG. 5. Diagram of grouping of strains that must have had different ancestors. Without recombination, these groupings should be the same for different genes in the same set of strains.

they do not overlap the error bar for node 2. Thus the alleles from RM45E, RM217T, and RM224H form a group. Then this branch is eliminated, which removes node 2. This leaves only alleles from strains RM70B, RM39A, and RM202I and nodes 1 and 3. Since the error bars from these nodes are nonoverlapping, these three strains are divided into two groups, with one group containing the first two alleles and the other containing the last allele. When this is added to the other grouping from the other genes, *phoA* breaks up previous groups and joins strains that previously had been separated. Thus groups are further divided, showing multiple recombinations. This graphic method shows that recombination is important without requiring the assumption that all genes must evolve at the same rate. If they do, then other tests that test for distance can also be done (see below).

(ii) Test for significant differences in branch order with parsimony trees. The *gnd* sequence data provided 119 informative sites that gave three minimum-length parsimony trees with a length of 277 and a consistency of 0.577. The trees obtained differed in their placement of the *gnd* gene from strains K-12, RM45E, RM70B, and RM224H. The tree most unlike the distance tree was chosen and is shown in Fig. 7. Neither this tree nor either the other two was significantly different from the distance tree. There are a large number of most parsimony *trp* trees. For each one of these, the minimum length for the *gnd* data was calculated. The one that best fit the *gnd* data (that gave the smallest number of steps) was chosen. This *trp* tree is significantly different from both the parsimony and the distance *gnd* trees (Fig. 7). Thus, the branching orders are significantly different for these gene trees, implying recombination.

It is of some interest that distance and the parsimony trees

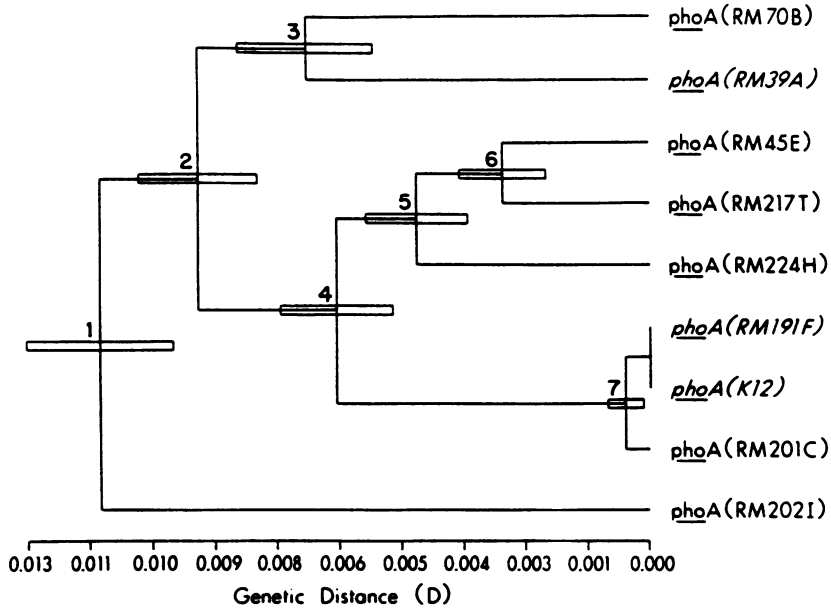

FIG. 6. Distance tree for *phoA* alleles. These are the same strains as shown in Fig. 4, with the exception of LT2. Note that the scale is different.

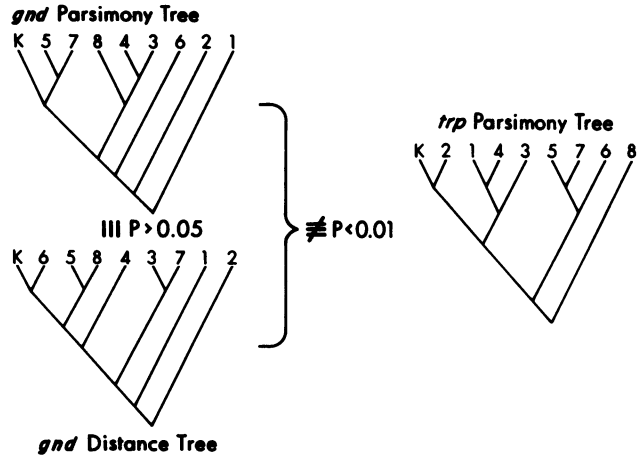

FIG. 7. Branching relationships: parsimony and distance trees of *gnd* and parsimony tree of *trp*. The numbers 1 through 8 indicate the alleles of the following *E. coli* strains: 1, RM39A; 2, RM191F; 3, RM201C; 4, RM217T; 5, RM45E; 6, RM70B; 7, RM224H; 8, RM202I. K, *E. coli* K-12. The distance and parsimony trees of *gnd* are not significantly different from each other at the 5% level; the *trp* tree is significantly different from both *gnd* trees at the 1% level.

place the *gnd* allele from strain RM191F in very different places. The distance tree places it outside the *Salmonella* allele, whereas the parsimony tree places it (strain 2 in Fig. 7) within both the *Salmonella* outgroup (data not shown) and the allele from RM39A (strain 1 in Fig. 7). The neighbor joining method (31) is more likely to give the correct tree than is either of the other methods (15, 43). This method placed the gnd^+ from RM191F within the alleles from *S. typhimurium* and RM39A, as did the parsimony tree (Fig. 8). The longer branch length of the gnd^+ from RM191F suggests that the rate of evolution of the gnd^+ from RM191F was faster than expected. Consequently, the equal rate assumption of the unweighted pair-group method with the arithmetic mean is invalidated for this strain, and the correct tree would place the common ancestor of the gnd^+ from RM191F and the other *gnd* alleles from *E. coli* after the species split with *S. typhimurium*.

(iii) Test for significant of differences in distances to common ancestors. The differences in branch lengths are obvious for the *gnd* and *trp* trees (Fig. 4). This could be because of differences in the rate of evolution of the genes or because of recombination. The statistic used (equation 1) to test for a significant difference in time to the common ancestor for different genes in the same strains is distributed as the Student *t* with infinite degrees of freedom. The variance used is the theoretical variance that assumes that all the changes happen independently.

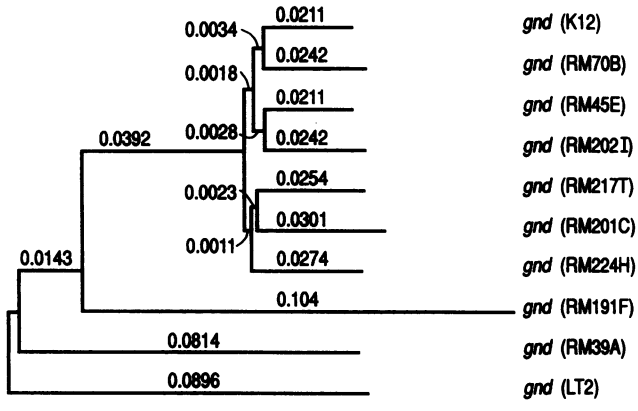

FIG. 8. *gnd* tree generated by the neighbor joining method. Branch lengths reflect distances.

Table 3 shows the values from the comparison between *gnd* and *trp*. When the value is positive the *gnd* distance between the two strains is longer than the *trp* distance and when the value is negative, it is reversed. The comparisons between the *Salmonella* strain and the various *E. coli* strains are not significant, showing that the rates of evolution for *gnd* and *trp* alleles are not significantly different. Since most of the values for the comparison between *S. typhimurium* and *E. coli* are negative, the *trp* locus may be evolving faster than *gnd*. The only positive value is the comparison between *Salmonella* alleles and the alleles from strain RM191F, confirming the previous observation that the *gnd* allele from this strain is evolving a little faster than expected. These effects are very small compared with the differences in distance between genes within the *E. coli* strains. Consequently the differences in the distances in *trp* and *gnd* cannot be explained by differences in the rate of evolution of the genes. Any significant differences in distance between pairs of strains for these genes must be because of different common ancestors, i.e., recombination. For example, the ancestral gene of the K-12 *trp* and the RM191F *trp* is recent, since these sequences are identical; but the ancestors of the K-12 *gnd* and the RM191F *gnd* must have diverged very long ago, since these sequences are so different.

All the values comparing *E. coli* strains are positive, and most are significant if tested individually (32 out of 36). Since there are multiple comparisons, the significance level for 5% for the group is 3.21. Even with this value, 30 out of 36 values are significant.

An outgroup sequence was not determined to confirm that the rate of evolution is the same in *phoA* as in the other two genes, since *S. typhimurium* does not contain a gene that is homologous to *phoA* (8). Sharp and Li (36) have shown there is a strong negative correlation between the rate of divergence of homologous *E. coli* and *Salmonella* genes and the codon adaptation index (37). The codon adaptation index is a measure of the degree of codon usage bias toward codons that are favored by highly expressed genes. The codon adaptation index for *phoA* is 0.35 compared with 0.38 for the *trp* region and 0.55 for *gnd*. This is about what is expected for a gene that codes for a protein synthesized intermittently or in moderate amounts. Thus we will assume that the rates of evolution of all three genes are equivalent and that significant tests imply recombination.

Table 4 shows the values from the comparison between *gnd* and *phoA* and between *phoA* and *trp*. As expected, most of the values for the comparison between *gnd* and *phoA* are significant. The importance of recombination may be only for *gnd*. Thus the test between *trp* and *phoA* is important. Even after correction for multiple comparisons so the critical value is 3.21, 12 values would be considered significant, implying multiple recombinations.

DISCUSSION

Recombination in *E. coli*. The results from this study clearly show that there is a tremendous amount of recombination and that this recombination is important in structuring the genetics of *E. coli*. Similar results were also seen when a gene tree for *gnd* for other wild strains of the ECOR

TABLE 3. Comparison of the distance between *gnd* alleles and *trp* alleles[a]

Strain	Pseudo-*t* value								
	RM39A	RM191F	RM201C	RM217T	RM45E	RM70B	RM224H	RM202I	LT2[b]
K-12	13.96	16.03	8.52	8.24	4.67	4.38	5.31	1.81	−0.59
RM39A		15.63	14.29	13.05	11.54	11.04	12.03	8.85	−0.66
RM191F			12.81	15.16	11.76	12.24	11.75	9.01	1.04
RM201C				7.53	5.12	6.06	5.12	2.46	−0.11
RM217T					5.37	5.49	5.88	2.35	−0.49
RM45E						9.05	8.41	0.70	−0.81
RM70B							9.74	1.52	−1.08
RM224H								1.87	−0.53
RM202I									−0.75

[a] The significance levels for two-tailed *t* tests are as follows: at 0.05, 1.963; at 0.01, 2.583; at 0.005, 2.816; at 0.001, 3.304.
[b] *S. typhimurium* strain.

collection was compared with the phenogram generated from multilocus enzyme electrophoresis (4).

Figure 5 shows that all strains except the pair K-12 and RM201C and the pair RM45E and RM224H have to have had a recombination event that replaced all or most of at least one of the genes. Even these exceptional pairs of stains show evidence of recombination when distances are considered. Tables 3 and 4 show that the percent differences are significantly different for the different genes in these pairs of strains. For the pair K-12 and RM201C the *gnd* alleles are clearly more different from each other than are either of the alleles from the other two genes, which are not significantly different from each other. This strongly suggests that at least one of the *gnd* alleles was recombined into a strain, whereas the other genes could represent clonal frame genes. Clonal frame refers to those genes in a particular strain that have not been replaced by recombination since some designated time (21). For the pair RM45E and RM224H, the percent difference for every pair of alleles is significant. Thus there have to have been at least two recombination events. The large percentage of significant values in Tables 3 and 4 shows that most pairs of strains are separated by at least two recombination events. This is a minimum estimate.

Even those species pairs in which there is no evidence for different ancestors for the two genes (i.e., the percentages are not significantly different for the two genes) may have undergone recombination. The test simply examines whether the ancestral genes could have arisen at the same time. If the ancestral genes arose at about the same time but in different strains, recombination would be required but would not be identified. In addition, the record of earlier recombinations is eliminated by later ones. Thus, rates of recombination will be hard to estimate.

One way of estimating the rate would be to compare strains with recent common ancestors so that the clonal frame can be identified. This allows the number of segments inserted by recombination to be determined and rates of recombination to be estimated relative to the mutation rate. There is evidence of recombination within a clone as defined by all the strains having the same electrotype (16). By using extensive data around the *trp* locus, a tentative recombination rate has been estimated at 7×10^{-12} replacements per bp per generation (21). In this collection of strains, there are three strains where the pairwise percentages for *phoA* and *trp* are not significantly different from each other or from zero. One should be able to use these strains to estimate rates of recombination. However, there is something very odd about these three strains (K-12, RM191F, and RM201C), which makes us suppose that the recombination rate will be higher than expected from a simple application of this technique. Analysis of protein gel electrophoretic data and biotype data has consistently separated *E. coli* group A strains from all others (22a, 28, 35). Group A contains K-12 and ECOR strains 1 through 25. The presumption is that this is a recently arisen clone in which the clonal frame has not been made unrecognizable by recombination. Since K-12, RM191F, and RM201C are very similar in two of three genes, it could be surmised that these genes are part of the clonal frame and thereby that all three of these strains are group A strains. RM191F is the same as ECOR16, which is a group A strain, but RM201C is the same as ECOR45, which is not a group A strain. Thus, either both genes have

TABLE 4. Comparison of the distance between *phoA* alleles and *trp* and *gnd* alleles[a]

Strain	Pseudo-*t* value								
	K-12	RM39A	RM191F	RM201C	RM217T	RM45E	RM70B	RM224H	RM202I
K-12		**11.06**	**18.19**	**6.04**	**5.07**	**3.85**	**1.78**	**4.96**	**3.44**
RM39A	4.80		**12.97**	**11.64**	**11.68**	**12.00**	**10.92**	**12.69**	**10.63**
RM191F	0.00	4.80		**16.52**	**12.92**	**11.87**	**10.69**	**12.26**	**11.61**
RM201C	0.18	4.59	0.18		**4.45**	**4.35**	**3.70**	**4.73**	**4.13**
RM217F	4.30	3.14	4.30	4.11		**6.49**	**3.13**	**5.96**	**3.33**
RM45E	1.46	1.10	1.46	1.47	−0.35		**2.93**	**5.97**	**2.39**
RM70B	3.21	1.61	3.21	3.19	3.07	7.37		**4.42**	**2.93**
RM224H	1.08	0.98	1.08	1.09	0.71	3.59	6.66		**4.46**
RM202I	−1.38	−0.59	−1.38	−1.28	−0.66	−1.60	−1.21	−2.35	

[a] The significance levels for two-tailed *t* tests are as follows: at 0.05, 1.963; at 0.01, 2.583; at 0.005, 2.816; at 0.001, 3.304. Values in boldface type are for *phoA* versus *gnd*; those in lightface type are for *phoA* versus *trp*.

been moved out of the K-12 type clonal frame into an ancestor of RM201C or so many other genes have been moved into RM201C that it is no longer recognizable as a group A strain. Consequently, the use of these strains to estimate the recombination rate would underestimate the rate.

The data presented may give an impression of a higher recombination rate than actually exists. The *gnd* locus is tightly linked to the *rfb* locus, which codes for the O antigen (14, 35). There are over 160 O-antigen types (12), which implies selection for diversity (i.e., rare types have a selective advantage). Thus, recombinants that brought in a rare O antigen and the linked *gnd* would be selected, giving a higher sampling of recombinants. Obviously, more extensive study with more loci and more strains is required to distinguish genes influenced by either diversifying selection or purifying selection from those which are influenced only by mutation, drift, and recombination and to determine which genes are likely to be indicative of clonal frames so that accurate estimates of the recombination rate can be made.

Statistical methods. In this paper we propose various methods for the analysis of recombination. The first two are formulated to statistically analyze trees, with a null hypothesis that trees for different genes in the same strains are not statistically different from each other. Most proposed statistical tests are designed to indicate whether the derived tree is the true tree. The purpose of these tests is different. The last test, which is a pseudo-*t* test, attempts to determine which sequence pairs for the same strains but different genes show evidence of having different ancestors.

These tests rely on assumptions that may not be valid. It is assumed that the changes between strains are selectively neutral and equally probable at all sites. The effects of these assumptions on the statistics of these tests are currently being investigated. These methods and others to be developed will have to be studied to determine their level of significance and their power. However, the pseudo-*t* test does correctly provide nonsignificant values for the differences in the *gnd-trp* distances between *S. typhimurium* and all of the *E. coli* strains, given the assumption that *S. typhimurium* and *E. coli* are different species and consequently do not recombine genes. This result implies that all of the significant values for the distances between *E. coli* strains are real and not just a result of the departure from the assumptions of the *t*-test.

An operational definition for biological species. The two principal definitions of species are the phenetic definition (41) and the biological species definition (19). The phenetic or phenotypic definition distinguishes species as groups of organisms that form a compact unit in character space, well separated from other such groups. Traditionally, a phenetic species is defined as a group of morphologically similar organisms located in a particular geographical region and morphologically distinct from other groups of organisms. In microbiology, this definition has been extended to include biochemical and physiological traits as well as morphological ones. The species definition being employed by Brenner and Falkow (5, 6), that species are groups of strains in which (i) over 70% of their DNA reassociates under moderately restrictive conditions and (ii) the thermal stability of this reassociated DNA is within 4°C of that of homologous reassociated DNA, is a phenetic definition.

Biological species are interbreeding groups of organisms, with each species separated from others through reproductive barriers. This definition implies that the phylogenes of different genes from individuals of the same species should

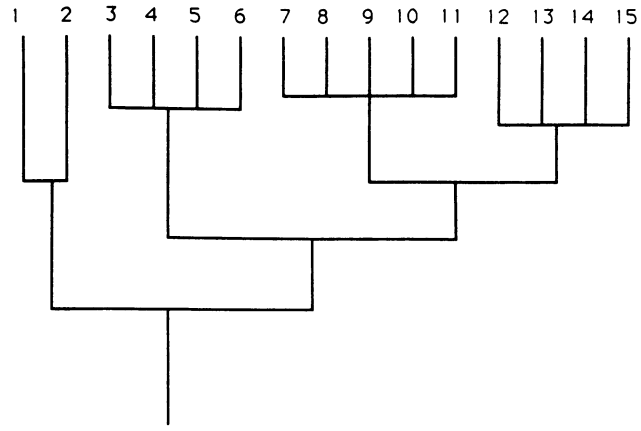

FIG. 9. Best-fit tree for 15 hypothetical strains from a group of gene trees, which defines at least four species. See the text for a discussion.

be significantly different, whereas the phylogeny of genes from individuals of different species should not be significantly different. Thus we have an operational criterion for the defining of bacterial species. Consider 15 hypothetical strains of bacteria isolated from nature from which one is interested in determining the number of species. Portions of a few genes are sequenced in each strain, and gene trees are derived. These trees are tested, and it is found that the gene trees are inconsistent only within certain groups. For example, the branching pattern of strains 3 through 6 (Fig. 9) is different for different genes, but these strains are always grouped together for all genes. Thus they would be judged as members of the same species and as a different species from the other strains. Likewise strains 7 through 11 would be judged as members of a second species, and strains 12 through 15 would be judged as members of a third species. Since at least three strains are required to obtain inconsistencies in the gene trees, this method can not determine if stains 1 and 2 are members of the same species or represent different species. However, a decision can be made by using the genetic distances between strains 1 and 2 for the various genes. One of the other strains can be chosen as the out group, and the rates of evolution can be determined to find out whether they are the same for the various genes. If they are, the distances for pairs of genes for strains 1 and 2 can be tested with the pseudo-*t* test. If the distances are significantly different, they would be judged to be members of the same species, whereas if the distances are not significantly different, they would be members of different species.

In this paper we have shown that, within *E. coli*, the gene trees are significantly different. It could be argued that horizontal gene transfer between what we call different species is so common that the method proposed in this paper will never resolve differences between species. Therefore, the other part of the requirement is to show that different genes from individuals of different species provide an estimation of the same tree. This was not tested directly in this study, but there is evidence (36) that *S. typhimurium* and *E. coli* are separate species by this definition. When homologous genes from *E. coli* K-12 and *S. typhimurium* LT2 are compared, none are so similar that they could represent recent horizontal transfer and the range of percent divergence can be explained by different constraints on codon usage (36). Horizontal gene transfer would destroy this

relationship. Therefore, these data imply that there is little gene transfer between these species.

In this paper we are not concerned with the mechanisms of the transition from a lattice of individual ancestors to a tree of ancestoral species, but we recognize that during the transition any definition of species will be difficult to apply unambiguously. Some of the possible difficulties and complications of this approach are discussed in Avise and Ball (3). An example of a possible difficulty is seen in the various *Neisseria* species. These species, which differ in sequence by up to 23% and therefore are less similar to each other than are *E. coli* and *S. typhimurium*, have transferred pieces of the PBP2B gene among each other (40). The pieces transferred provide resistance to penicillin, and thus any transfer would be strongly selected. This might represent the rare case of horizontal gene transfer between species. Investigations with other genes may show that there is little gene transfer, and thus these would still be considered separate species.

The methodologies for an operational definition of bacterial species in terms of the biological species definition can now be developed. Sequencing of DNA segments amplified by the polymerase chain reaction (10, 13) will permit DNA sequences for a number of genes from a large number of strains to be acquired quickly. These can then be used to define species by the methodology outlined above after the proper statistical procedures have been worked out. Genes chosen for this purpose should be chromosomal and found in most organisms. Examples would be ribosomal genes and genes for enzymes in central metabolism, like *gnd* and *trp*. Genes like *phoA*, which are not found in many enteric species (8), should be avoided, as should genes that provide resistance to antibiotics, since they are more likely to have transgressed species boundaries (40). Horizontal gene transfer across species should be rare enough that it will not be a problem. However, because of this possibility gene trees for more than two genes should be used. The genes chosen should be similar to characters chosen for taxonomic purposes in plants and animals—those that provide traits not important to the peculiar adaptive strategies of particular species.

ACKNOWLEDGMENTS

This work was supported by Public Health Service grant GM30201 from the National Institutes of Health.

We thank R. E. Wolf, Jr., R. Milkman, and J. Roth for strains, D. L. Hartl, R. F. DuBose, and J. Kim for discussions, and Margaret Riley, Judith Mongold, and Roger Milkman for their careful reading and suggestions to improve the manuscript.

REFERENCES

1. **Achtman, M., M. Heuzenroeder, B. Kusecek, H. Ochman, D. Caugant, R. K. Selander, V. Valsanen-Rhen, T. K. Korhonen, S. Stuart, F. Orskov, and I. Orskov.** 1983. Six widespread bacterial clones among *Escherichia coli* K1 isolates. Infect. Immun. **39**:315–335.
2. **Achtman, M., and G. Pluschke.** 1986. Clonal analysis of descent and virulence among selected *Escherichia coli*. Annu. Rev. Microbiol. **40**:185–210.
3. **Avise, J. C., and R. M. Ball, Jr.** 1990. Principles of genealogical concordance in species concepts and biological taxonomy. Oxf. Surv. Evol. Biol. **7**:45–67.
4. **Bisercic, M., J. Y. Feutrier, and P. R. Reeves.** 1991. Nucleotide sequence of the *gnd* gene from nine natural isolates of *Escherichia coli*: evidence of intragenic recombination as contributing factor in the evolution of the polymorphic *gnd* locus. J. Bacteriol. **173**:3894–3900.
5. **Brenner, D. J.** 1981. Introduction to the family *Enterobacteriaceae*, p. 1105–1127. *In* M. P. Starr, H. Stolp, H. G. Truper, A. Balows, and H. G. Schlegel (ed.), The prokaryotes: a handbook on habitats, isolation and identification of bacteria, vol. 2. Springer-Verlag, Berlin.
6. **Brenner, D. J., and S. Falkow.** 1971. Molecular relationships among members of the *Enterobacteriaceae*. Adv. Genet. **16**:81–118.
7. **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.
8. **DuBose, R. F., and D. L. Hartl.** 1990. The molecular evolution of bacterial alkaline phosphatase: correlating variation among enteric bacteria to experimental manipulations of the protein. Mol. Biol. Evol. **7**:547–577.
9. **Dykhuizen, D. E., and L. Green.** 1986. DNA sequence variation, DNA phylogeny and recombination. Genetics **113**:s71.
10. **Erlich, H. A.** 1989. PCR technology: principles and applications for DNA amplification. Stockton Press, New York.
11. **Felsenstein, J.** 1985. Confidence limits on phylogenies with a molecular clock. Syst. Zool. **34**:152–161.
12. **Hartl, D. L., and D. E. Dykhuizen.** 1984. The population genetics of *Escherichia coli*. Annu. Rev. Genet. **18**:31–68.
13. **Innis, M. A., D. H. Gelfand, J. J. Sninsky, and T. J. White.** 1990. PCR protocols: a guide to methods and applications. Academic Press, Inc., San Diego.
14. **Jiang, X. M., B. Neal, F. Santiago, S. J. Lee, L. K. Romana, and P. R. Reeves.** 1991. Structure and sequence of the *rfb* (O antigen) of *Salmonella* serovar *typhimurium* (strain LT2). Mol. Microbiol. **5**:695–713.
15. **Jin, L., and M. Nei.** 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**:82–102.
16. **Lawrence, J. G., D. E. Dykhuizen, R. F. DuBose, and D. L. Hartl.** 1988. Phylogenetic analysis using insertion sequence fingerprinting in *Escherichia coli*. Mol. Biol. Evol. **6**:1–14.
17. **Levin, B. R.** 1981. Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. Genetics **99**:1–23.
18. **Levy, S. B., and R. V. Miller.** 1989. Gene transfer in the environment. McGraw-Hill Book Co., New York.
19. **Mayr, E.** 1963. Animal species and evolution. Harvard University Press, Cambridge, Mass.
20. **Milkman, R.** 1973. Electrophoretic variation in *E. coli* from natural sources. Science **182**:1024–1026.
21. **Milkman, R., and M. M. Bridges.** 1990. Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics **126**:505–517.
22. **Milkman, R., and I. P. Crawford.** 1983. Clustered third-base substitutions among wild strains of *Escherichia coli*. Science **321**:378–380.

22a. **Miller, R. D., and D. E. Dykhuizen.** Unpublished data.

23. **Nasoff, M. S., H. V. Baker II, and R. E. Wolf, Jr.** 1984. DNA sequence of the *Escherichia coli* gene, *gnd*, for 6-phosphogluconate dehydrogenase. Gene **27**:253–264.
24. **Nasoff, M. S., and R. E. Wolf, Jr.** 1980. Molecular cloning, correlation of genetic and restriction maps, and determination of the direction of transcription of *gnd* of *Escherichia coli*. J. Bacteriol. **143**:731–741.
25. **Nei, M.** 1987. Molecular evolutionary genetics. Columbia University Press, New York.
26. **Nei, M., J. C. Stephens, and N. Saitou.** 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. **2**:66–85.
27. **Ochman, H., and R. K. Selander.** 1984. Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. **157**:690–693.
28. **Ochman, H., and R. K. Selander.** 1984. Evidence for clonal population structure in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **81**:198–201.
29. **Ochman, H., T. S. Whittam, D. A. Caugant, and R. K. Selander.** 1983. Enzyme polymorphism and genetic population structure

in *Escherichia coli* and *Shigella*. J. Gen. Microbiol. **129**:2715–2726.
30. **Orskov, F., and I. Orskov.** 1983. Summary of a workshop on the clone concept in the epidemiology, taxonomy, and evolution of the *Enterobacteriaceae* and other bacteria. J. Infect. Dis. **148**: 346–357.
31. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.
32. **Sawyer, S. A.** 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6**:526–538.
33. **Sawyer, S. A., D. E. Dykhuizen, R. F. DuBose, L. Green, T. Mutangadura-Mhlanga, D. F. Wolczyk, and D. L. Hartl.** 1987. Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. Genetics **115**:51–63.
34. **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.
35. **Selander, R. K., D. A. Caugant, and T. S. Whittam.** 1987. Genetic structure and variation in natural population 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. American Society for Microbiology, Washington, D.C.
36. **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.
37. **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.
38. **Sidák, Z.** 1967. Rectangular confidence regions for the means of multivariate normal distributions. J. Am. Stat. Assoc. **62**:626–633.
39. **Smith, G. R.** 1983. General recombination, p. 175–209. *In* R. W. Hedrix, J. W. Roberts, F. W. Stahl, and R. A. Weisberg (ed.), Lambda II. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
40. **Smith, J. M., C. G. Dowson, and B. G. Spratt.** 1991. Localized sex in bacteria. Nature (London) **349**:29–31.
41. **Sneath, P. H., and R. R. Sokal.** 1973. Numerical taxonomy. W. H. Freeman & Co., San Francisco, Calif.
42. **Sokal, R. R., and F. J. Rohlf.** 1969. Biometry, 1st ed., p. 607–610. W. H. Freeman & Co., San Francisco, Calif.
43. **Sourdis, J., and M. Nei.** 1988. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. Mol. Biol. Evol. **5**:298–311.
44. **Stoltzfus, A., J. F. Leslie, and R. Milkman.** 1988. Molecular evolution of the *Escherichia coli* chromosome. I. Analysis of the structure and natural variation in a previously uncharacterized region between *trp* and *tonB*. Genetics **120**:345–358.
45. **Templeton, A. R.** 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and apes. Evolution **37**:221–244.
46. **Whittam, T. S., H. Ochman, and R. K. Selander.** 1983. Multilocus genetic structure in natural populations of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **80**:1751–1755.