Discovery and Classification of Ecological Diversity in the Bacterial World: the Role of DNA Sequence Data

THOMAS PALYS,[1] L. K. NAKAMURA,[2] AND FREDERICK M. COHAN[1]*

Department of Biology, Wesleyan University, Middletown, Connecticut 06459-0170,[1] and Microbial Properties Research,
National Center for Agricultural Utilization Research, Agricultural Research Service,
U.S. Department of Agriculture, Peoria, Illinois 61604[2]

All living organisms fall into discrete clusters of closely related individuals on the basis of gene sequence similarity. Evolutionary genetic theory predicts that in the bacterial world, each sequence similarity cluster should correspond to an ecologically distinct population. Indeed, surveys of sequence diversity in protein-coding genes show that sequence clusters correspond to ecological populations. Future population surveys of protein-coding gene sequences can be expected to disclose many previously unknown ecological populations of bacteria. Sequence similarity clustering in protein-coding genes is recommended as a primary criterion for demarcating taxa.

For two decades, systematists have applied whole-genome hybridization as a universal criterion for demarcating species of bacteria: systematists have widely recognized bacterial species as phenotypically distinct groups of strains with 70% or greater annealing of genomic fragments in DNA-DNA hybridization (36, 85). This criterion has been widely used because it can be easily applied to any taxon, and most importantly, the groupings of bacteria based on DNA-DNA hybridization are often the same as those based on phenotypic characters and ecology (36).

However, it is becoming increasingly evident that any particular cut-off value (such as 70%) is arbitrary and not guaranteed to yield groups of bacteria that correspond to real ecological units (82). Also, it is not clear what determines the fraction of genomic segments that anneal in hybridization experiments (82; but see reference 43): is it the fraction of genes that are shared or the sequence similarity at shared gene loci? Accordingly, no evolutionary genetic theory predicts why groups of strains with greater than 70% annealing should correspond to ecologically distinct populations.

There is, however, another molecular approach that may provide a universal criterion for classifying bacterial diversity. This approach relies on the observation that all living organisms, both prokaryotic and eukaryotic, fall into clusters of closely related organisms based on the sequence similarity of shared genes (1, 15, 48). That is, bacteria and other organisms fall into clearly distinct sequence clusters, where the average sequence divergence between strains of different clusters is far greater than the average divergence between strains of the same cluster. Recent theory has suggested that each sequence similarity cluster observed in the bacterial world might correspond to an ecologically distinct population (14, 15, 17, 18). If this conjecture is correct, then a classification system based on sequence clusters would have a theoretical grounding that is lacking in the genomic hybridization approach.

In this study, we will demonstrate that the DNA sequences of protein-coding genes are more effective than DNA-DNA hybridization for classifying the ecological diversity of bacteria. We first extend the theoretical argument of Cohan (14) that

each sequence cluster in the bacterial world should correspond to an ecologically distinct population. We then present empirical evidence that sequences of protein-coding genes have successfully separated populations of bacteria into distinct sequence similarity clusters that correspond to ecological units. Finally, we argue that taxon demarcations should be based on the sequence similarity clusters emerging from sequence surveys of protein-coding genes.

Theoretical background. A model of sequence divergence in bacteria must take into account that the genes typically sequenced by bacterial systematists (e.g., 16S rRNA, nucleic acid polymerases, and ribosomal proteins) are not likely to be involved in ecological differences between bacterial populations. Many of these genes appear to perform the same housekeeping duties in every species, as alleles of these genes are functionally interchangeable across taxa (35, 51, 63, 73, 89). A model of sequence diversity must also take into account the fact that nearly all nucleotide substitutions (and amino acid substitutions) detected in surveys are likely to be neutral in their fitness effects (40). Therefore, the patterns of sequence divergence observed by systematists largely involve substitutions that have no fitness consequence in genes that are not involved in population-specific adaptations.

Cohan (14, 16) has developed a model of neutral sequence divergence within and between ecological populations of bacteria. The model posits that each adaptive mutation within a population confers a competitive advantage upon a mutant cell and its clonal (or nearly clonal) descendants; this advantage allows the mutants to replace all competing cells of the same population. However, because different populations use at least somewhat different resources, an adaptive mutant from one population is not expected to outcompete members of other populations (even if the populations are sympatric). In this model, an ecological population is defined as the domain of competitive superiority of an adaptive mutant.

This definition leads to two predictions about sequence divergence within and between ecological populations (14, 15). First, natural selection favoring each adaptive mutation in a given population is expected to purge that population of its genetic diversity at all loci. This is because recombination is rare in bacteria (50, 69, 75, 86), so that the entire genome originally associated with the adaptive mutation remains intact as it sweeps through the population. Therefore, the population loses all (or nearly all) its genetic diversity at all loci. This

* Corresponding author. Mailing address: Department of Biology, Wesleyan University, 237 Church St., Middletown, CT 06459-0170. Phone: (860) 685-3482. Fax: (860) 685-2141. E-mail: FCohan@Wesleyan.edu.

purging of diversity within a population following each adaptive mutation is called periodic selection (4, 42, 46).

Second, because the adaptive mutant cannot outcompete cells from other ecological populations, each periodic selection event has very little effect on the divergence between populations (14, 15, 18). Because it purges the diversity within but not between populations, periodic selection increases the distinctness of ecological populations at all loci. Each round of periodic selection fosters the divergence of different populations into separate sequence similarity clusters for any gene of interest, whether or not the particular gene is responsible for the ecological differences among populations.

Consider next the diversity-purging effect of periodic selection from a phylogenetic perspective. Let us suppose that a new ecological population is derived clonally from one mutant cell that is adapted to a new ecological niche. The new population is a monophyletic group descending from this original cell, but the population is not yet a separate sequence similarity cluster (Fig. 1A). This is because the average divergence within the new ecological population is not much less than the average divergence between the new population and the most closely related clade. After the first periodic selection event, however, the diversity within the new population is purged (Fig. 1B). Likewise, periodic selection events in the ancestral population will purge diversity within that population as well (Fig. 1C and 1D). Thus, each population will eventually appear as a monophyletic group and as a distinct cluster (Fig. 1E).

While periodic selection tends to make each population more distinct, even rare recombination between populations tends to homogenize different populations at housekeeping gene loci. This is because no fitness penalty accrues for incorporating another species' allele at the housekeeping gene loci studied by systematists (as we have postulated above). Therefore, each recombination of genetic sequence across populations will tend to make the populations more similar. It is a quantitative question whether the neutral sequence divergence between ecological populations is dominated by the diversifying effect of periodic selection or by the homogenizing effect of interpopulation recombination. Cohan's coalescence model (14) shows how the distinctness of ecological populations is determined by the interplay of these factors.

Cohan's coalescence model predicts the expected sequence divergence levels within and between populations, where each such expectation is the average divergence over all pairwise comparisons among cells averaged over all time (14). Thus, the average divergence observed at any one time might be higher or lower than the expected value, depending on how recently periodic selection has purged sequence diversity.

The model assumes that all neutral nucleotide substitutions are synonymous. (The model ignores neutral nucleotide substitutions that cause amino acid changes; in many proteins, especially those that are highly constrained, nearly all neutral substitutions are synonymous [57].) Neutral substitutions are assumed to occur only at the third-base sites of codons, and every third-base substitution is assumed to be neutral. In this model, 33% of mutations are synonymous, which is very close to the actual fraction of mutations that are synonymous (29%) (57).

The model yields the expected number of substitutions per third-base site for cells in different populations, $E(d_b)$, and for cells in the same population, $E(d_w)$. The fraction of all sites that are divergent within [$E(\pi_w)$] or between [$E(\pi_b)$] populations may be approximated from $E(d_w)$ or $E(d_b)$, by dividing the latter by 3, provided that $E(d)$ is small (i.e., <0.09) (14).

Model of distinctness of sequences from different populations. The present study departs from previous work (14–18) by focusing on the distinctness of sequences from different ecological populations. We define the distinctness of populations as the ratio (k) of the expected divergence between populations, $E(d_b)$, to the expected divergence within populations, $E(d_w)$; this ratio $E(d_b)/E(d_w)$ is approximately equal to $E(\pi_b)/E(\pi_w)$. We derive the distinctness ratio of populations as follows.

The expected number of substitutions per third-base site between cells of different populations, $E(d_b)$, is the following (based on equations 2 to 4 of reference 14): $E(d_b) = E(d_w) + \mu_0/c_\beta$, where the value μ_0 is the neutral mutation rate per third-base site (per genome per generation) and the value c_β is the rate of recombination between populations (per gene segment per generation).

Dividing each side by $E(d_w)$ yields:

$$k = \frac{E(d_b)}{E(d_w)} = 1 + \frac{\mu_0}{c_\beta E(d_w)} \tag{1}$$

The expected divergence within populations, $E(d_w)$, is determined largely by the diversity-purging effect of periodic selection, which is in turn determined by the rate of recombination within populations and the intensity of periodic selection (14).

Survey of sequence divergence within and between bacterial taxa. We tested the hypothesis that sequence similarity clusters based on housekeeping protein genes should correspond to ecologically distinct groups of bacteria. Since species and subspecies are generally considered to be ecologically distinct groups (82), our approach was to test whether housekeeping genes could distinguish pairs of species or subspecies as separate sequence clusters. Sequence data were obtained from the systematics literature and from GenBank. We aimed to include in the survey every pair of closely related taxa (species or subspecies) for which a protein-coding gene had been sequenced in more than one strain per taxon. We aligned sequences from GenBank by using the Clustal algorithm of DNASTAR's MegAlign program. The mean sequence divergence level within each taxon and between each pair of closely related taxa was calculated as the mean of all pairwise comparisons; distinctness ratios were calculated as the ratio of the mean between-taxon divergence to the mean within-taxon divergence (the latter calculated as the average of the mean within-taxon divergence levels for the two taxa).

Analysis of divergence at *mdh* in *Salmonella* and *Escherichia* was based on GenBank sequences U04742 to U04768 and U04770 to U04784. Analysis of divergence at *gnd* in *Salmonella* and *Escherichia* was based on GenBank sequences U14337, U14344, U14346 and U14347, U14351 to U14353, U14356, U14361, U14365, U14367, U14435, U14437, U14441 to U14444, U14447, U14450, U14452 and U14453, U14455 and U14456, U14459 to U14461, U14479, U14481 and U14482, U14484, U14487, U14489, U14494 and U14495, U14497 to U14499, U14501, and U14503 to U14509. Analysis of divergence at *gapA* in *Salmonella* and *Escherichia* was based on GenBank sequences M66853 to M66868, M66870 to M66877, and M66879 to M66882. Analysis of divergence at *putP* in *Salmonella* and *Escherichia* was based on GenBank sequences L01132 and L01133 and L01135 to L01159. Analysis of divergence at *gnd* in *Citrobacter* was based on GenBank sequences U14336, U14424 and U14425, U14427 to U14429, U14432, and U14466. Analysis of divergence at *fla* in *Listeria* was based on GenBank sequences X86979 to X87005. Analysis of divergence at *hsp* in *Mycobacterium* was based on GenBank sequences U17922 and U17923, U17940 to U17944, and U55826 and U55827. Analysis of divergence at *32kDa* in *Mycobacterium* was based on GenBank sequences Z33657 to Z33662.

RESULTS

Theoretical prediction of divergence. Figure 2 shows how the distinctness ratio of populations is affected by the rate of interpopulation recombination and the expected level of divergence within populations (based on equation 1). For any level of within-population diversity, there exists a level of interpopulation recombination that allows populations to be distinguished by sequence data. For example, populations containing little diversity [e.g., $E(d_w) < 0.03$, or $E(\pi_w) < 0.01$] are distinguishable by sequence data (at the level of $k \geq 2$) whenever the interpopulation recombination rate falls below 2.3×10^{-8} per gene segment per genome per generation. Populations containing greater sequence diversity are distinguishable only under lower rates of interpopulation recombination. For example, a population with a diversity of $E(d_w) = 0.09$ would be distinguishable at $k > 2$ only when $c_\beta < 7.7 \times 10^{-9}$.

It should be emphasized that equation 1 and Fig. 2 predict the distinctness of populations only at loci that are functionally interchangeable across populations (which we will refer to as
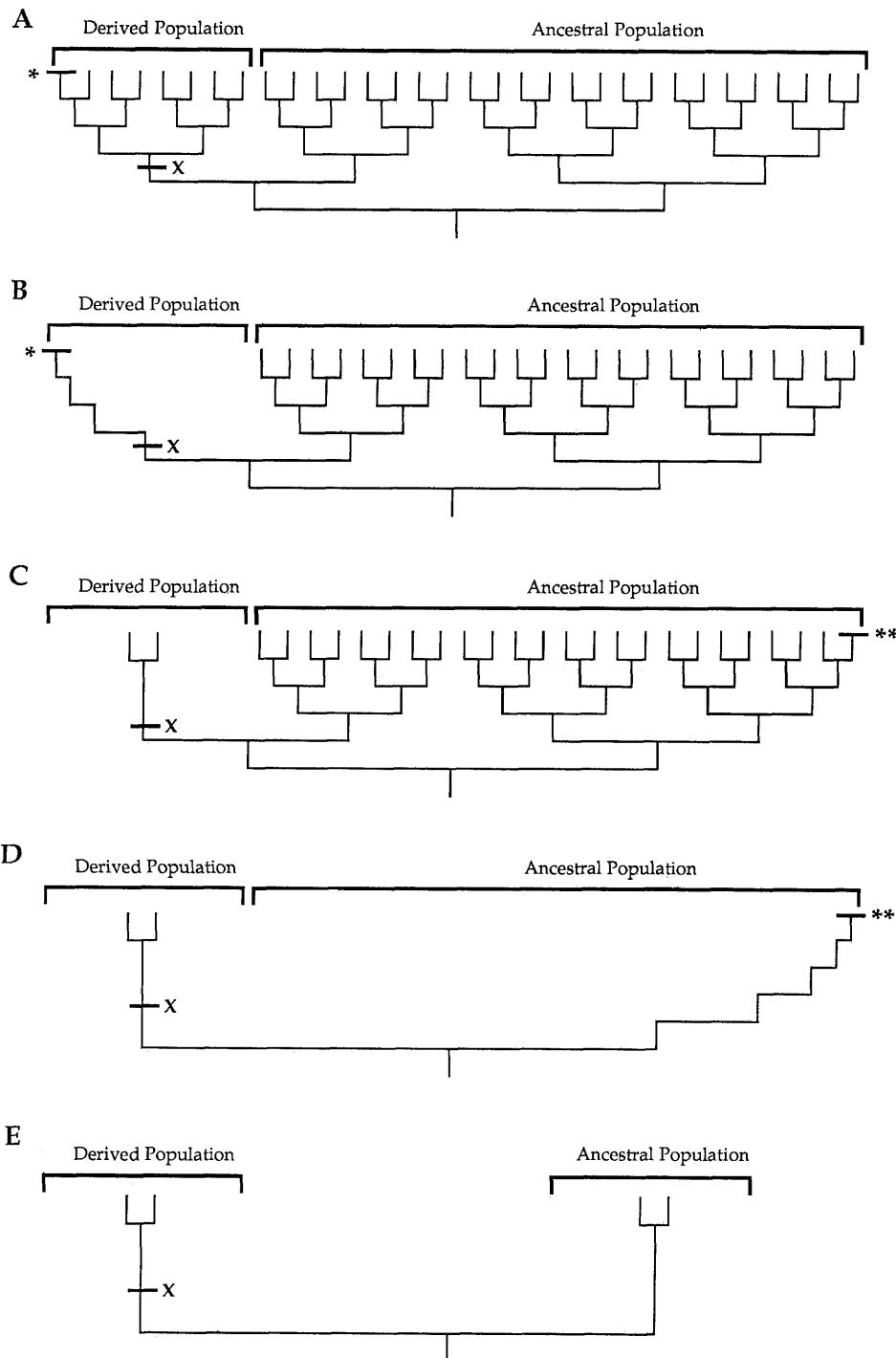
FIG. 1. A phylogenetic perspective on periodic selection. The figure demonstrates that two populations will become distinct sequence similarity clusters, where the between-population divergence is much greater than the within-population divergence. (A) The derived population consists of the descendants of a mutant (X) capable of utilizing a new ecological niche. The adaptive mutant in the derived population (*) is capable of outcompeting all other members of the derived population. Note that at this point the two populations do not appear as distinct clusters. Moreover, the ancestral population is not even a monophyletic group. (B) The adaptive mutant (*) has driven all the other lineages within the derived population to extinction. (C) With time, the derived population becomes more genetically diverse. One cell in the ancestral population (**) has developed a mutation that allows it to outcompete other members of the population. (D) The adaptive mutant (**) has outcompeted other members of the ancestral population. (E) The ancestral population becomes more genetically diverse. At this point, each population is a distinct sequence cluster as well as a monophyletic group.
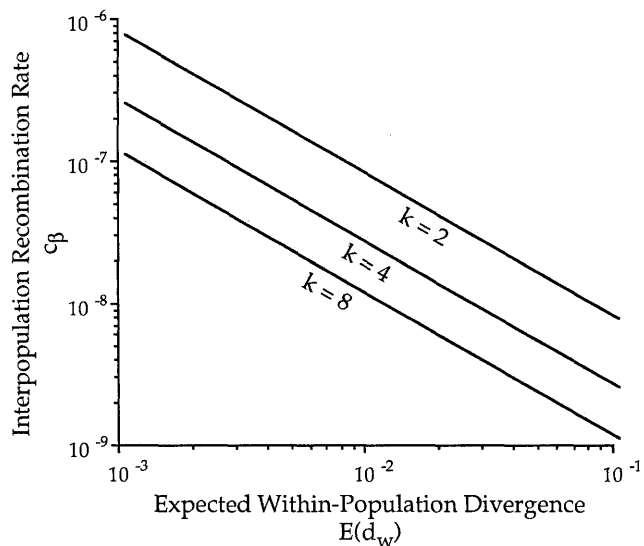
FIG. 2. Effects of the interpopulation recombination rate c_β and the expected within-population divergence $E(d_w)$ on the distinctness ratio k. The graph is based on equation 1 and assumes a neutral mutation rate μ_0 of 6.93×10^{-10} (for *E. coli*) (23). Populations to the left and below the $k = 2$ line may be considered distinct sequence clusters.

housekeeping loci). It is expected that populations would be more distinct than predicted by equation 1 at loci responsible for population-specific adaptations, since natural selection would disfavor interpopulation recombinants at these loci.

Survey of sequence divergence levels within and between taxa. Table 1 shows the mean sequence divergence levels within and between pairs of closely related species and subspecies at protein-coding genes not involved in population-specific adaptations (i.e., housekeeping genes). Some species contained two or more clearly distinct sequence clusters (e.g., the 168 and W23 sequence clusters within *Bacillus subtilis*). Whether or not such clusters had subspecies status, they were analyzed as separate groups in Table 1.

The average divergence observed within a specific or subspecific sequence cluster was usually about 1% or less over all sites (i.e., $\pi_w \approx 0.01$ [Table 1]). The average divergence between a pair of subspecies or closely related species was almost always 2 or more times greater than the divergence within taxa (i.e., $k > 2$).

Two species, *Salmonella enterica* and *Erwinia carotovora*, were known to contain several ecologically distinct subspecies at the time they were surveyed for sequence diversity. The seven subspecies of *S. enterica* fell clearly into separate sequence clusters; this was also the case for two of the three subspecies pairs within *E. carotovora* (Table 1). However, one pair of *E. carotovora* subspecies was not well distinguished as separate clusters, showing a distinctness ratio of only 1.28. We are aware of no other case in which ecologically distinct taxa have failed to be distinguished as separate sequence clusters.

Most interesting are the species that were not known to contain multiple ecologically distinct populations at the time they were surveyed but were nevertheless found to contain multiple sequence clusters. In several cases, these clusters were later shown to represent different ecological populations. For example, several sequence similarity clusters discovered within *Borrelia burgdorferi* (sensu lato) were subsequently found to represent different ecological populations with different pathogenic properties; these clusters were later given species status

(7, 8). Other ecologically distinct populations discovered as sequence clusters include *Bacillus mojavensis*, *Bacillus vallismortis*, and the *B. subtilis* 168 and W23 sequence clusters (69–71). Several sequence clusters have also been discovered for *Vibrio cholerae*. While these clusters are not known to correspond to ecological differences, the clusters do correspond with geographic origin (38).

DISCUSSION

Previous theoretical studies have shown that periodic selection fosters the distinctness of different bacterial populations at all gene loci (Fig. 1), while recombination tends to homogenize the gene sequences of different populations (14–18). The present study introduces a quantitative measure of the distinctness of populations' gene sequences (k) and shows how population distinctness at housekeeping loci is determined by periodic selection and interpopulation recombination [represented by the parameters $E(d_w)$ and c_β, respectively]. Figure 2 shows a wide range of parameter values under which each ecologically distinct population should diverge into its own sequence cluster. By estimating the parameters of equation 1 for natural bacterial populations, we can then determine whether actual bacterial populations should diverge into distinct sequence clusters.

Estimates of parameters. Here we consider the actual values of the parameters of equation 1: the neutral mutation rate, μ_0; the expected sequence diversity within populations, $E(d_w)$; and the rates of interpopulation recombination, c_β, occurring in nature.

For *Escherichia coli*, the neutral mutation rate is 6.93×10^{-10} per third-base site per genome per generation (24). The rate is expected to be nearly the same for any bacterium with a genome of about the same size as *E. coli*'s (16, 24).

We may estimate $E(d_w)$ from the mean level of sequence divergence observed within sequence similarity clusters, π_w (16). Table 1 shows that the average divergence observed within clusters is usually (though not always) about 1% or less over all sites (i.e., $\pi_w \approx 0.01$). Assuming that most substitutions occur at third-base sites, for many taxa, the expected number of substitutions per third-base site, $E(d_w)$, is then about 0.03 (i.e., $3\pi_w$).

The rate of recombination between populations (c_β) has not been directly measured, but we may obtain an upper bound for c_β from estimated rates of recombination within sequence clusters. The rate of recombination within bacterial taxa has been estimated from DNA sequence survey data by the method of Hudson (34). For *B. subtilis* and *B. mojavensis*, the within-cluster recombination rate has been estimated at 10^{-7} (69), and the rate within *E. coli* has been estimated at 10^{-8} (86). Allozyme surveys of many other bacterial taxa have yielded high levels of linkage disequilibrium consistent with rates of within-taxon recombination no greater than the rate of mutation (50, 75).

If each taxon studied represents a single ecological population, then the above estimates of recombination rates each represent the within-population recombination rate. If instead each taxon represents a pool of multiple ecological populations, then these estimates represent a weighted average of within- and between-population recombination rates. Because the rate of recombination between populations should be no greater than the within-population rate, the estimated rate of recombination within clusters represents an upper bound for the rate of between-population recombination (14).

How should we expect the rates of recombination between populations (c_β) to compare to the recombination rates within

TABLE 1. Sequence divergence within and between ecologically distinct taxa at protein-coding genes
not involved in taxon-specific adaptations[a]

Comparison	Gene	Mean divergence within group[b]	Mean divergence between groups[c]	Ratio (between/within)[d]	Reference(s)
Borrelia					
B. burgdorferi vs. *B. afzelii*	*fla*	0.003[e], 0.006 ± 0.002	0.055 ± 0.001	12.64	27
B. afzelii vs *B. garinii*	*fla*	0.006 ± 0.002, 0.010 ± 0.001	0.052 ± 0.001	6.60	27
B. garinii vs *B. burgdorferi*	*fla*	0.010 ± 0.001, 0.003[e]	0.061 ± 0.000	9.50	27
Neisseria					
N. meningitidis vs *N. gonorrhoeae*	*recA*	0.014 ± 0.002[f]	0.016 ± 0.002	1.11	91
	argF	0.004 ± 0.001[f]	0.056 ± 0.001	13.91[g]	91
	fbp	0.004 ± 0.000[f]	0.010 ± 0.000	2.38	91
Salmonella and *Escherichia*					
S. enterica subspecies I, II, IIIa, IIIb, IV, VI, and VII	*mdh*	0.007 ± 0.001, 0.003 ± 0.001, 0.002[e], 0.002[e], 0.020 ± 0.008, 0.004[e], 0.000[e]	0.044 ± 0.004	8.06	11
	gnd	0.030 ± 0.004, 0.038 ± 0.005, 0.023[e], 0.035 ± 0.006, 0.037 ± 0.006, 0.034 ± 0.006, 0.000[e]	0.057 ± 0.002	2.03	59, 81
	gapA[h]	0.002[e], 0.011 ± 0.004, 0.009[e], 0.000[e], 0.000[e], 0.001[e]	0.026 ± 0.003	6.80	60
S. enterica vs *S. bongori*	*mdh*	0.036 ± 0.001, 0.004 ± 0.001	0.086 ± 0.001	4.64	11
	gnd	0.053 ± 0.001, 0.037 ± 0.007	0.075 ± 0.001	1.67	59, 81
	gapA	0.025 ± 0.001, 0.004[e]	0.087 ± 0.001	6.00	60
	putP	0.044 ± 0.002, 0.004[e]	0.066 ± 0.001	2.75	58
E. coli vs *S. enterica*	*mdh*	0.011 ± 0.000, 0.036 ± 0.001	0.154 ± 0.001	6.55	11
	gnd	0.070 ± 0.006, 0.053 ± 0.001	0.168 ± 0.001	2.73	59, 81
	gapA	0.003 ± 0.000, 0.025 ± 0.001	0.061 ± 0.000	4.36	60
	putP	0.022 ± 0.001, 0.044 ± 0.002	0.206 ± 0.000	6.24	58
E. coli vs *S. bongori*	*mdh*	0.011 ± 0.000, 0.004 ± 0.001	0.160 ± 0.000	21.3	11
	gnd	0.070 ± 0.006, 0.037 ± 0.007	0.165 ± 0.002	3.08	59, 81
	gapA	0.003 ± 0.000, 0.004[e]	0.086 ± 0.000	24.57	60
	putP	0.022 ± 0.001, 0.004[e]	0.207 ± 0.001	15.92	58
Citrobacter spp.					
C. diversus vs *C. freundii*	*gnd*	0.009 ± 0.001, 0.115[e]	0.210 ± 0.054	3.39	12, 59, 81
Erwinia carotovora[i]					
E. carotovora subsp. *carotovora* vs *E. carotovora* subsp. *odorifera*	*pel*	0.063 ± 0.001, 0.031 ± 0.002	0.060 ± 0.001	1.28	21
E. carotovora subsp. *carotovora* vs *E. carotovora* subsp. *atroseptica*	*pel*	0.063 ± 0.001, 0.019 ± 0.001	0.087 ± 0.001	2.12	21
E. carotovora subsp. *atroseptica* vs *E. carotovora* subsp. *odorifera*	*pel*	0.019 ± 0.001, 0.031 ± 0.002	0.078 ± 0.001	3.12	21
Vibrio[j]					
V. cholerae cluster A vs *V. cholerae* cluster B	*asd*	0.006 ± 0.001, 0.013 ± 0.001	0.020 ± 0.001	2.11	38
V. cholerae cluster A vs *V. cholerae* cluster C	*asd*	0.006 ± 0.001, 0.009 ± 0.001	0.010 ± 0.001	1.33	38
V. cholerae cluster A vs *V. cholerae* cluster D	*asd*	0.006 ± 0.001, 0.001[e]	0.041 ± 0.001	11.71	38
V. cholerae cluster B vs *V. cholerae* cluster C	*asd*	0.013 ± 0.001, 0.009 ± 0.001	0.020 ± 0.001	1.82	38
V. cholerae cluster B vs *V. cholerae* cluster D	*asd*	0.013 ± 0.001, 0.001[e]	0.053 ± 0.001	7.57	38
V. cholerae cluster C vs *V. cholerae* cluster D	*asd*	0.009 ± 0.001, 0.001[e]	0.042 ± 0.001	8.40	38
V. cholerae cluster A vs *V. minicus*	*asd*	0.006 ± 0.001[k]	0.100 ± 0.000	16.67	38
V. cholerae cluster B vs *V. minicus*	*asd*	0.013 ± 0.001[k]	0.104 ± 0.001	8.00	38
V. cholerae cluster C vs *V. minicus*	*asd*	0.009 ± 0.001[k]	0.102 ± 0.001	11.33	38
V. cholerae cluster D vs *V. minicus*	*asd*	0.001[e,k]	0.090 ± 0.001	90.00	38
Bacillus					
B. subtilis 168 cluster vs *B. subtilis* W23 cluster	*gyrA*	0.012 ± 0.002, 0.045 ± 0.004	0.066 ± 0.001	2.32	69
	polC	0.014 ± 0.003, 0.050 ± 0.008	0.076 ± 0.003	2.38	19, 69
	rplX	0.001 ± 0.000, 0.003[e]	0.015 ± 0.001	7.69	76
	rpoB	0.003 ± 0.001, 0.012 ± 0.002	0.014 ± 0.001	1.87	69
B. subtilis 168 cluster vs *B. vallismortis*	*gyrA*	0.012 ± 0.002, 0.004 ± 0.001	0.105 ± 0.001	13.13	69, 70
	polC	0.014 ± 0.003, 0.000 ± 0.000	0.055 ± 0.001	7.86	19, 69, 70
	rpoB	0.003 ± 0.001, 0.000 ± 0.000	0.017 ± 0.001	11.33	69, 70
B. subtilis W23 cluster vs *B. vallismortis*	*gyrA*	0.045 ± 0.004, 0.004 ± 0.001	0.100 ± 0.001	4.08	69, 70
	polC	0.050 ± 0.008, 0.000 ± 0.000	0.069 ± 0.002	2.76	19, 69, 70
	rpoB	0.012 ± 0.002, 0.000 ± 0.000	0.018 ± 0.001	3.00	69, 70

TABLE 1—*Continued*

Comparison	Gene	Mean divergence within group[b]	Mean divergence between groups[c]	Ratio (between/within)[d]	Reference(s)
B. subtilis 168 cluster vs *B. mojavensis*	*gyrA*	0.012 ± 0.002, 0.031 ± 0.004	0.158 ± 0.002	7.35	69, 70
	polC	0.014 ± 0.003, 0.010 ± 0.002	0.131 ± 0.002	10.92	19, 69, 70
	rpoB	0.003 ± 0.001, 0.003 ± 0.001	0.031 ± 0.001	10.33	69, 70
B. subtilis W23 cluster vs *B. mojavensis*	*gyrA*	0.045 ± 0.004, 0.031 ± 0.004	0.156 ± 0.002	4.11	69, 70
	polC	0.050 ± 0.008, 0.010 ± 0.002	0.121 ± 0.002	4.03	19, 69, 70
	rpoB	0.012 ± 0.002, 0.003 ± 0.001	0.026 ± 0.001	3.47	69, 70
B. globisporus vs *B. psychrophilus*	*pyk*	0.004 ± 0.002, 0.018 ± 0.004	0.096 ± 0.002	8.73	64
Listeria[l]					
L. monocytogenes type 1 vs *L. monocytogenes* type 2	*fla*	0.000 ± 0.000, 0.001 ± 0.000	0.018 ± 0.000	36.00	67
L. monocytogenes type 1 vs *L. monocytogenes* type 3	*fla*	0.000 ± 0.000, 0.005[e]	0.008 ± 0.001	1.60	67
L. monocytogenes type 2 vs *L. monocytogenes* type 3	*fla*	0.001 ± 0.000, 0.005[e]	0.025 ± 0.001	8.33	67
Mycobacterium					
M. avium vs *M. intracellulare*	*hsp*	0.005 ± 0.001, 0.036 ± 0.007	0.036 ± 0.003	1.76	37, 72
	32kDa	0.047 ± 0.021, 0.012 ± 0.003	0.046 ± 0.010	1.56	78

[a] All divergence estimates are based on sequences of segments ranging from 787 to 1895 bp, except for the following: *L. monocytogenes* comparisons are based on a 186-bp segment; the *rplX* comparisons for *B. subtilis* are based on a 309-bp segment; and the *hsp* and *32kDa* segments compared in *Mycobacterium* are 360 and 407 bp, respectively. The following comparisons are based on restriction site analyses: the *pel* gene of *Erwinia carotovora* and the *gyrA*, *polC*, *pyk* and *rpoB* genes of *Bacillus*.

[b] Mean ± standard error for pairwise divergence within each of the groups (π_w) shown in order of appearance in the Comparison column.

[c] Mean ± standard error for the divergence between groups (π_b), based on all pairwise comparisons of strains from different groups. When only two groups are compared, the standard error is based on the variance among pairs of strains. In cases where more than two groups are compared, the mean intergroup divergence over all pairs of groups is presented; the standard error in these cases is based on the variance among group pairs.

[d] Ratio of the between-group divergence to the average of the within-group divergence levels (π_b/π_w). With few exceptions, each group listed can be distinguished as a separate sequence similarity cluster on the basis of protein-coding genes, with the ratio of between- to within-population divergence greater than 2. Note that some pairs of groups that clearly fall into separate sequence similarity clusters on the basis of protein-coding genes are nearly identical in their 16S rRNA sequences (i.e., less than 0.5% divergence; see Table 2).

[e] Standard error is not available because only two strains were sampled.

[f] Only one strain of *N. gonorrhoeae* was sampled, so an estimate of within-taxon divergence is not available.

[g] The ratio is high because the gene in *N. meningitidis* was transferred from *N. cinerea* (91).

[h] Only one strain from subspecies VI has been sequenced for *gapA*, so within species divergence is unavailable for that taxon.

[i] By using restriction site data from Darrasse et al. (21), sequence divergence was calculated for each pair of strains with equations 5.50 and 10.5 of Nei (57).

[j] *V. cholerae* clusters A, B, C, and D have not yet been shown to represent different ecological populations. Because these clusters are allopatric, they may simply represent geographically separate subpopulations of the same ecological population (38).

[k] Only one strain of *V. minicus* was sampled, so an estimate of within-group divergence is not available.

[l] *L. monocytogenes* clusters 1 and 3 have not yet been shown to represent different ecological populations.

populations (c_w)? We can imagine certain circumstances under which the rates of recombination within and between populations would be equal. Equality of c_w and c_β would first require that a cell encounters members of other populations as often as it encounters members of its own. This might be the case, for example, for two planktonic bacterial populations using different soluble resources that are homogeneously distributed in one pond. Also, equality of c_w and c_β requires that all vectors of recombination infect both populations equally well. Finally, the populations must not be divergent in DNA sequences because sequence divergence directly causes sexual isolation (68, 90).

It is far more likely that the rate of recombination between populations will be much lower than that within populations, almost certainly by many orders of magnitude. In the case of pathogens, for example, ecological populations frequently differ in the host species or the tissues they infect. In these cases, the microgeographic differences in habitat will ensure that between-population recombination is reduced far below the within-population rate, if only because cells of one population would rarely encounter cells from the other population. We conservatively assume that recombination between populations occurs at a rate at least 10 times lower than that within populations, at least for nonplanktonic bacteria.

Prediction that ecological populations should diverge into separate sequence clusters. Consider next whether ecological

populations should be expected to be distinguishable as separate sequence clusters. In the case of *B. subtilis*, the expected within-population divergence, $E(d_w)$, is ~0.09 (16) and the between-population recombination rate is very low (assuming $c_\beta \ll c_w \approx 10^{-7}$) (69), yielding a distinctness ratio k greater than 2 (equation 1) (Fig. 2). Likewise, in the case of *E. coli*, $E(d_w)$ and c_β are also sufficiently low (52, 86) so that $k > 2$. We may conclude that each ecological population within these taxa should fall into its own distinct sequence similarity cluster. If the low rates of recombination in *E. coli* and *B. subtilis* are typical for the bacterial world and if the levels of sequence diversity within populations [$E(\pi_w)$] are typically around 1% or less (as seen in Table 1), then each ecological population of bacteria should fall into its own sequence similarity cluster (equation 1) (Fig. 2).

There are, however, several circumstances that may prevent ecological populations from distinguishing themselves as separate sequence clusters. First, it is possible that in some taxa, interpopulation recombination rates (c_β) are too high to allow sequence divergence between populations (especially for clusters with high levels of sequence diversity) (Fig. 2).

Second, the differences in adaptation between some populations may be determined entirely by the presence or absence of a plasmid or prophage. For example, symbiotic populations of *Rhizobium* species require a symbiosis plasmid (49), while the nearby rhizosphere populations of the same bacterial spe-

cies are best adapted when they lack the plasmid (74). In such cases, a cell lineage can repeatedly adapt to one population's environment and then to another's by gaining or losing a plasmid. We should not expect ecological populations whose difference is determined by the presence or absence of a highly mobile plasmid to diverge at genomic loci.

Third, the coalescence model predicts the expected divergence between populations over all time. Therefore, populations that are newly divergent may have failed to diverge into distinct sequence clusters simply because they have not yet had sufficient time.

Finally, for very highly conserved genes (such as 16S rRNA), where the rates of neutral mutation are extremely low (5), there may not be any discernible variation either within or between ecological populations. This last caveat is not a problem for protein-coding genes, however, since even the most highly conserved protein-coding genes usually have at least 0.5% sequence divergence (at all sites) within clusters (Table 1).

In summary, we conclude that each ecological population in the bacterial world should diverge eventually into its own sequence similarity cluster for protein-coding genes, with some exceptions. The exceptions are taxa with recombination rates that are significantly higher than those found in *Bacillus* and *Escherichia* and populations whose ecological differences are coded on plasmids.

Why an ecological population cannot be split into multiple sequence clusters. While each sequence cluster is generally expected to contain only one ecological population, we conversely expect no ecological population to be split into two or more sequence clusters (Fig. 3) (18). This is because divergence within a population into two or more sequence clusters would be unstable with respect to periodic selection. Each adaptive mutant within the population would drive to extinction cells from all the clusters of the population, and the cluster bearing the adaptive mutant would be all that survives this purge of diversity (Fig. 3). It appears, then, that two long-standing, highly divergent clusters cannot each contain cells from the same population.

There is one exception to this conclusion: geographically isolated members of the same ecological population could diverge into separate sequence clusters. This is because geographically isolated subpopulations would have no opportunity to exchange genes, and adaptive mutants from one geographically isolated subpopulation would have no opportunity to compete with other subpopulations. Thus, neutral sequence divergence between subpopulations would be allowed to proceed without bound. Divergence among geographically isolated members of the same ecological population would be especially likely for bacteria with low mobility but would not be possible for highly vagile organisms like *Bacillus*, where intercontinental migration of spores occurs extremely frequently (69). Nevertheless, we can be sure that two highly divergent sequence clusters from the same geographical area (i.e., within migration range) must represent different ecological populations.

In summary, the theory predicts that if the low interpopulation recombination rates of *Escherichia* and *Bacillus* are typical, then each long-standing ecological population of bacteria should eventually diverge into its own sequence cluster for any gene of interest. We will be better able to assess the generality of this conclusion when further data on recombination rates in other taxa become available. In any case, the theory predicts that classifying the ecological diversity of bacteria according to sequence clusters can only underestimate the ecological diversity in nature, since sympatric members of the same population should not be found in two separate sequence clusters.
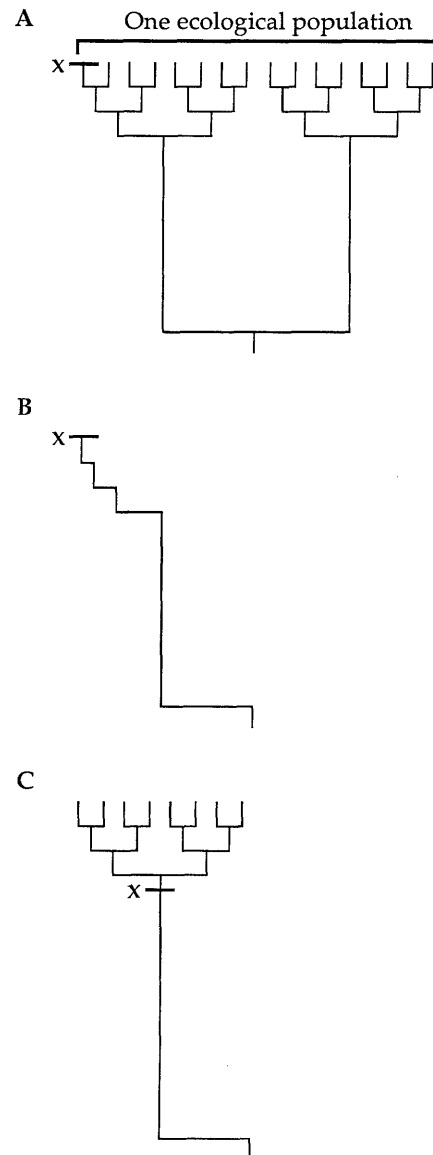

FIG. 3. Sympatric members of a single ecological population cannot be split among multiple sequence clusters. (A) The population initially contains two distinct sequence clusters. Then an adaptive mutation occurs in lineage X. (B) Because the adaptive mutant can purge diversity from the entire population, only one cluster survives periodic selection. (C) After periodic selection, variation within the population is reestablished, but the population now forms a single sequence cluster.

We next consider how well ecological populations empirically correspond to the DNA sequence clusters described by systematists.

Sequence similarity clusters based on 16S rRNA and protein-coding genes. The 16S rRNA gene is ubiquitous among cellular organisms and is highly conserved (88) and so may be amplified from any bacterial taxon by universal primers (77). Consequently, hundreds of bacterial species have been sequenced at this locus, and these sequence data have been extremely useful in determining the evolutionary tree of all cellular and multicellular life, as well as the trees of bacterial families and genera (82). The closest relatives of any newly discovered bacterial species may be readily determined by 16S rRNA sequence data.

TABLE 2. Ecologically distinct bacterial taxa that cannot be distinguished by 16S rRNA sequence variation (i.e., between-taxon sequence divergence levels average 0.5% or less over all pairs of strains)

Taxon	Mean 16S rRNA divergence[a]		Reference
	Within taxon	Between taxa	
Deleya aquamarina *Halomonas meridiana*		0.000	22
Francisella tularensis biovar A *F. tularensis* biovar B		0.001	25
Bacillus cereus *B. anthracis*	0.001	0.000 ± 0.000	2
Bacillus psychrophilus *B. globisporus*		0.002	26
Bacillus anthracis *B. thuringiensis*		0.003	2
Bacillus cereus *B. thuringiensis*		0.003	2
Bacillus cereus *B. mycoides*		0.004	2
Bacillus anthracis *B. mycoides*		0.005	2
Bacillus thuringiensis *B. mycoides*		0.006	2
Bacillus subtilis *B. atrophaeus*		0.005	3
Clostridium acetobutylicum *C. beijerinckii*	0.000 ± 0.000 0.002 ± 0.002	0.001 ± 0.002	39
Enterococcus casseliflavus *E. gallinarium*		0.002	87
Enterococcus avium *E. raffinosus*		0.003	87
Enterococcus durans *E. faecium*		0.003	87
Aerococcus viridans *Pediococcus urinae-equi*		0.000	20
Mycobacterium intracellulare serovars 4, 5, 6, and 8 *M. intracellulare* serovar 9[b]	0.000 ± 0.000	0.000 ± 0.000	9
Mycobacterium intracellulare serovars 4, 5, 6, 8, 10, and 11 *M. avium*	0.000 ± 0.000	0.002 ± 0.000	9
Cellulomonas biazotea *C. fimi*		0.003	66
Fusobacterium nucleatum *F. periodonticum*		0.000	61

[a] Mean ± standard error. When there is data for only one pair of strains available, the standard error is not given.

[b] *Mycobacterium intracellulare* serovars 4, 5, 6, and 8 are in one DNA-DNA hybridization group, and serovar 9 is in another.

Nevertheless, the 16S rRNA gene has failed to distinguish many closely related but ecological distinct groups of bacteria. In these cases (listed in Table 2), there is little or no variation within or between ecological populations. One possible explanation is that these populations may have only recently diverged, so that neutral divergence has not yet had time to accumulate at any gene locus. Alternatively, these populations may have had time to accumulate neutral divergence at rapidly evolving loci, but not yet at the highly conserved 16S rRNA loci (5). Finally, the groups in question may undergo recombination at such a high rate that between-population divergence is always reduced to nearly the level of divergence within populations at ay housekeeping gene (i.e., so that $k \approx 1$ [equation 1]).

We may test these hypotheses with data for sequence diversity at protein-coding loci, since these loci evolve at a much higher rate than 16S rRNA (5). We first consider the case of *Bacillus globisporus* and *Bacillus psychrophilus*, two very closely related species differing in their preferred temperatures and other conditions for growth (28, 44, 54). They are nearly identical in their 16S rRNA sequences (Table 2) (26), yet they clearly fall into different sequence similarity clusters on the basis of their sequences for the pyruvate kinase gene (Table 1) (64): the average sequence divergence between groups exceeds that within groups by a factor of 8.

Unfortunately, very few ecologically distinct populations that are indistinguishable by 16S rRNA data (listed in Table 2) have been surveyed for variation at protein-coding gene loci. Nevertheless, those taxa that have been so surveyed corroborate the pattern seen in *B. globisporus* and *B. psychrophilus*: ecologically distinct populations that fail to be distinguished by 16S rRNA data fall into separate sequence similarity clusters for protein-coding genes. Examples include *Bacillus atrophaeus* and the 168 group of *B. subtilis* (3, 69) and *Mycobacterium intracellulare* and *Mycobacterium avium* (9, 37, 72, 78) (Tables 1 and 2).

We conclude that the inability of 16S rRNA sequences to distinguish some taxa is neither a result of frequent recombination between taxa nor a result of insufficient time to accumulate neutral divergence at any gene. Rather, it is a consequence of the low evolutionary rate of 16S rRNA genes. While 16S rRNA sequence data are useful for distinguishing moderately divergent populations into separate sequence clusters, protein-coding genes provide a better opportunity for distinguishing very closely related ecological populations.

Population surveys of housekeeping protein genes have almost unanimously shown closely related species and subspecies to fall into separate sequence clusters (Table 1). Indeed, we are aware of only one case in which a sequence survey of protein-coding genes has failed to distinguish ecologically distinct populations with taxonomic status (the case of *Erwinia carotovora* subsp. *odorifera* and *E. carotovora* subsp. *carotovora* [Table 1]) (21). Housekeeping protein gene sequences clearly distinguish ecologically distinct populations that have already been given the status of species or subspecies (Table 1). Moreover, these gene sequences have the power to discover ecologically distinct populations not distinguished by other molecular techniques. In several cases, ecologically distinct populations were discovered only because they formed a sequence similarity cluster separate from those of known taxa (e.g., sequence clusters within *Borrelia burgdorferi* [sensu lato] [7, 8], within *Bacillus subtilis* [sensu lato] [69–71], and in *Frankia* [62]).

We are aware of no case where sympatric strains from distinct sequence similarity clusters (i.e., $k > 2$) have failed to show ecological differences. Conversely, we are aware of only one instance (above) where ecologically distinct populations have failed to be distinguished into separate sequence similar-

TABLE 3. Ecologically distinct groups that are distinguishable as separate sequence similarity clusters but are not distinguishable as separate species on the basis of the DNA-DNA hybridization criterion

Taxon	Ecologically distinct groups distinguishable by sequence data	Ecological difference	Reference(s)
Salmonella enterica	Subspecies I, II, IIIa, IIIb, IV, VI, and VII	Different biotypes	58
Neisseria	*N. gonorrhoeae*, *N. meningitidis*, and *N. lactamica*	Virulence properties	84, 91
Bacillus subtilis	168 and W23 sequence clusters	Found in different deserts	69
Listeria monocytogenes	Cluster 1 and cluster 2	Human and environmental isolates	67
Mycobacterium	*M. intracellulare* and *M. avium*	Host range, virulence properties	6, 9

ity clusters on the basis of protein-coding genes. We conclude that the correspondence between ecological populations and DNA sequence similarity clusters is grounded not only in evolutionary genetic theory but also in the practice of bacterial systematics. Surveying sequence diversity should continue to be a useful technique for discovering the ecological diversity within the bacterial world, among culturable and nonculturable organisms alike (13, 41, 53).

Rationale for using DNA sequence data to classify nonculturable bacteria. Increasingly often, populations of nonculturable bacteria are being described on the basis of 16S rRNA sequence data and a minimum of phenotypic information (53). This practice has recently been recognized by the International Committee on Systematic Bacteriology; provisional taxa whose uniqueness is evidenced solely by DNA sequence data are now given the status *Candidatus* (53). The rationale for classifying nonculturable organisms on the basis of DNA sequence data has been that among culturable bacteria, sequence clusters

have empirically corresponded to ecological populations (82). The present study gives further rationale for the *Candidatus* status by providing an evolutionary genetic basis for the correspondence between sequence clusters and ecology.

DNA-DNA hybridization compared to sequence data as tools for identifying ecological diversity. DNA-DNA hybridization has been a useful molecular technique for assigning bacterial strains into ecologically coherent clusters, even when the ecological properties of the strains are unknown. However, there are several cases in which sequence data have been more effective in distinguishing ecologically different groups (Table 3). For example, *Neisseria meningitidis* and *Neisseria gonorrhoeae* fall into the same group by DNA-DNA hybridization (29) but clearly fall into separate clusters on the basis of sequences at protein-coding genes (Table 1) (91). On the other hand, we are aware of no set of ecologically distinct populations that are distinguishable by DNA-DNA hybridization but not by protein-coding sequence data. We conclude that pro-

TABLE 4. Closely related, ecologically distinct taxa that are not distinguished by DNA-DNA hybridization and have not yet been surveyed for variation in protein-coding sequences

Ecologically distinct groups	Ecological difference	Reference(s)
Yersinia enterocolitica biotype 1 and biotypes 2, 3, and 4	Habitat (human vs water)	23
Francisella tularensis biovars A, B, and novidica	Virulence properties	25
Pseudomonas syringae pathovars	Host specificity	45
Xanthomonas campestris pathovars	Host specificity	83
Ralstonia solanacearum biovars Blood disease bacterium of banana	Host specificity, virulence properties	80
Bacillus thuringiensis subspecies B. cereus B. anthracis	Virulence properties	32, 56
Bacillus thiaminolyticus groups 1 and 2	Habitat (human feces vs honeybee larvae)	55
Paenibacillus larvae subsp. *larvae* P. larvae subsp. *pulvifaciens*	Virulence properties	33
Phytoplasma 16S rRNA groups	Host specificity	30
Mycoplasma F38 group Mycoplasma capricolum	Virulence properties	10
Mycoplasma buteonis M. falconis M. gypis	Host specificity	65
Porphyromonas macacae cat and monkey biovars	Host specificity	47
Fusobacterium necrophorum biovars A, B, and AB	Virulence properties	61

description of *Tetragenococcus* gen. nov. FEMS Microbiol. Lett. **70:**255–262.

21. **Darrasse, A., S. Priou, A. Kotoujansky, and Y. Bertheau.** 1994. PCR and restriction fragment length polymorphism of a *pel* gene as a tool to identify *Erwinia carotovora* in relation to potato diseases. Appl. Environ. Microbiol. **60:**1437–1443.

22. **Dobson, S. J., T. A. McMeekin, and P. D. Franzmann.** 1993. Phylogenetic relationships between some members of the genera *Deleya, Halomonas,* and *Halovibrio.* Int. J. Syst. Bacteriol. **43:**665–673.

23. **Dolina, M., and R. Peduzzi.** 1993. Population genetics of human, animal and environmental *Yersinia* strains. Appl. Environ. Microbiol. **59:**442–450.

24. **Drake, J. W.** 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proc. Natl. Acad. Sci. USA **88:**7160–7164.

25. **Forsman, M., G. Sandstrom, and A. Sjostedt.** 1994. Analysis of 16S ribosomal DNA sequences of *Francisella* strains and utilization for determination of the phylogeny of the genus and for identification of strains by PCR. Int. J. Syst. Bacteriol. **44:**38–46.

26. **Fox, G. E., J. D. Wisotzkey, and P. Jurtshuk, Jr.** 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int. J. Syst. Bacteriol. **42:**166–170.

27. **Fukunaga, M., and Y. Koreki.** 1996. A phylogenetic analysis of *Borrelia burgdorferi* sensu lato isolates associated with Lyme disease in Japan by flagellin gene sequence determination. Int. J. Syst. Bacteriol. **46:**416–421.

28. **Gordon, R. E., W. C. Haynes, and C. H.-N. Pang.** 1973. The genus *Bacillus.* Agricultural handbook no. 427. U.S. Department of Agriculture, Washington, D.C.

29. **Guibourdenche, M., M. Y. Popoff, and J. Y. Riou.** 1986. Deoxyribonucleic acid relatedness among *Neisseria gonorrhoeae, N. meningitidis, N. lactamica, N. cinerea* and "*Neisseria polysaccharea.*" Ann. Inst. Pasteur Microbiol. **137B**(2):177–185.

30. **Gundersen, D. E., I.-M. Lee, D. A. Schaff, N. A. Harrison, C. J. Chang, R. E. Davis, and D. T. Kingsbury.** 1996. Genomic diversity and differentiation among phytoplasma strains in 16S rRNA groups I (aster yellows and related phytoplasmas) and III (X-disease and related phytoplasmas). Int. J. Syst. Bacteriol. **46:**64–75.

31. **Guttman, D. S., and D. E. Dykhuizen.** 1994. Clonal divergence in *Escherichia coli* is a result of recombination, not mutation. Science **266:**1380–1383.

32. **Henderson, I., C. J. Duggleby, and P. C. Turnbull.** 1994. Differentiation of *Bacillus anthracis* from other *Bacillus cereus* group bacteria with the PCR. Int. J. Syst. Bacteriol. **44:**99–105.

33. **Heyndrickx, M., K. Vandemeulebroecke, B. Hoste, P. Janssen, K. Kersters, P. De Vos, N. A. Logan, N. Ali, and R. C. W. Berkeley.** 1996. Reclassification of *Paenibacillus* (formerly *Bacillus*) *pulvifaciens* (Nakamura 1984) Ash et al. 1994, a later subjective synonym of *Paenibacillus* (formerly *Bacillus*) *larvae* (White 1906) Ash et al. 1994, as a subspecies of *P. larvae,* with emended descriptions of *P. larvae* as *P. larvae* subsp. *larvae* and *P. larvae* subsp. *pulvifaciens.* Int. J. Syst. Bacteriol. **46:**270–279.

34. **Hudson, R. R.** 1987. Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:**242–250.

35. **Ichige, A., and G. C. Walker.** 1997. Genetic analysis of the *Rhizobium meliloti bacA* gene: functional interchangeability with the *Escherichia coli sbmA* gene and phenotypes of mutants. J. Bacteriol. **179:**209–216.

36. **Johnson, J. L.** 1986. Nucleic acids in bacterial classification, p. 972–975. *In* P. H. A. Sneath, N. S. Mair, N. E. Sharpe, and J. G. Holt (ed.), Bergey's manual of systematic bacteriology, vol. 2. Williams & Wilkins, Baltimore, Md.

37. **Kapur, V., L. L. Li, M. R. Hamrick, B. B. Plikaytis, T. M. Shinnick, A. Telenti, W. R. Jacobs, Jr., A. Banerjee, S. Cole, and K. Y. Yuen.** 1995. Rapid *Mycobacterium* species assignment and unambiguous identification of mutations associated with antimicrobial resistance in *Mycobacterium tuberculosis* by automated DNA sequencing. Arch. Pathol. Lab. Med. **119:**131–138.

38. **Karaolis, D. K. R., R. Lan, and P. K. Reeves.** 1995. The sixth and seventh cholera pandemics are due to independent clones separately derived from environmental, nontoxigenic, non-O1 *Vibrio cholerae.* J. Bacteriol. **177:**3191–3198.

39. **Keis, S., C. F. Bennet, V. K. Ward, and D. T. Jones.** 1995. Taxonomy and phylogeny of industrial solvent-producing clostridia. Int. J. Syst. Bacteriol. **45:**693–705.

40. **Kimura, M.** 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

41. **Knight, I. T., W. E. Holben, J. M. Tiedje, and R. R. Colwell.** 1992. Nucleic acid hybridization techniques for detection, identification, and enumeration of microorganisms in the environment, p. 65–91. *In* M. A. Levin, R. J. Seidler, and M. Rogul (ed.), Microbial ecology: principles, methods, and applications. McGraw-Hill, New York, N.Y.

42. **Koch, A. L.** 1974. The pertinence of the periodic selection phenomenon to prokaryotic evolution. Genetics **77:**127–142.

43. **Lan, R., and P. R. Reeves.** 1996. Gene transfer is a major factor in bacterial evolution. Mol. Biol. Evol. **13:**47–55.

44. **Larkin, J. M., and J. L. Stokes.** 1967. Taxonomy of psychrophilic strains of *Bacillus.* J. Bacteriol. **94:**889–895.

45. **Legard, D. E., C. F. Aquadro, and J. E. Hunter.** 1993. DNA sequence variation and phylogenetic relationships among strains of *Pseudomonas sy-*

ringae pv. syringae inferred from restriction site maps and restriction fragment length polymorphism. Appl. Environ. Microbiol. **59:**4180–4188.

46. **Levin, B. R.** 1981. Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. Genetics **99:**1–23.

47. **Love, D. N.** 1995. *Porphyromonas macacae* comb. nov., a consequence of *Bacteroides macacae* being a senior synonym of *Porphyromonas salivosa.* Int. J. Syst. Bacteriol. **45:**90–92.

48. **Mallet, J.** 1995. A species definition for the modern synthesis. Trends Ecol. Evol. **10:**294–299.

49. **Martinez, E., D. Romero, and P. Palacios.** 1990. The *Rhizobium* genome. Crit. Rev. Plant Sci. **9:**17–19.

50. **Maynard Smith, J., N. H. Smith, M. O'Rourke, and B. G. Spratt.** 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. USA **90:**4384–4388.

51. **Mikulskis, A. V., and G. R. Cornelis.** 1994. A new class of proteins regulating gene expression in enterobacteria. Mol. Microbiol. **11:**77–86.

52. **Milkman, R., and M. M. Bridges.** 1993. Molecular evolution of the *Escherichia coli* chromosome. Genetics **133:**455–468.

53. **Murray, R. G. E., and E. Stackebrandt.** 1995. Taxonomic note: implementation of the provisional status *Candidatus* for incompletely described procaryotes. Int. J. Syst. Bacteriol. **45:**186–187.

54. **Nakamura, L. K.** 1984. *Bacillus psychrophilus* sp. nov., nom. rev. Int. J. Syst. Bacteriol. **34:**121–123.

55. **Nakamura, L. K.** 1990. *Bacillus thiaminolyticus* sp. nov., nom. rev. Int. J. Syst. Bacteriol. **40:**242–246.

56. **Nakamura, L. K.** 1994. DNA relatedness among *Bacillus thuringiensis* serovars. Int. J. Syst. Bacteriol. **44:**125–129.

57. **Nei, M.** 1987. Molecular evolutionary genetics. Columbia University Press, New York, N.Y.

58. **Nelson, K., and R. K. Selander.** 1992. Evolutionary genetics of the proline permease gene (*putP*) and the control region of the proline utilization operon in populations of *Salmonella* and *Escherichia coli.* J. Bacteriol. **174:**6886–6895.

59. **Nelson, K., and R. K. Selander.** 1994. Intergenic transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. Proc. Natl. Acad. Sci. USA **91:**10227–10231.

60. **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.

61. **Nicholson, L. A., C. J. Morrow, L. A. Corner, and A. L. M. Hodgson.** 1994. Phylogenetic relationship of *Fusobacterium necrophorum* A, AB, and B biotypes based upon 16S rRNA gene sequence analysis. Int. J. Syst. Bacteriol. **44:**315–319.

62. **Normand, P., S. Orso, B. Cournoyer, P. Jeannin, C. Chapelon, J. Dawson, L. Evtushenko, and A. K. Misra.** 1996. Molecular phylogeny of the genus *Frankia* and related genera and emendation of the family *Frankiaceae.* Int. J. Syst. Bacteriol. **46:**1–9.

63. **Osuna, R., S. A. Boylan, and R. A. Bender.** 1991. In vitro transcription of the histidine utilization (*hutUH*) operon from *Klebsiella aerogenes.* J. Bacteriol. **173:**116–123.

64. **Palys, T., E. Berger, L. K. Nakamura, and F. M. Cohan.** Unpublished data.

65. **Poveda, J. B., J. Giebel, J. Flossdorf, J. Meier, and H. Kirchhoff.** 1994. *Mycoplasma buteonis* sp. nov., *Mycoplasma falconis* sp. nov., and *Mycoplasma gypis* sp. nov., three species from birds of prey. Int. J. Syst. Bacteriol. **44:**94–98.

66. **Rainey, F. A., N. Weiss, and E. Stackebrandt.** 1995. Phylogenetic analysis of the genera *Cellulomonas, Promicromonospora,* and *Jonesia* and the proposal to exclude the genus *Jonesia* from the family *Cellulomonadaceae.* Int. J. Syst. Bacteriol. **45:**649–652.

67. **Rasmussen, F. O., P. Skouboe, L. Dons, L. Rossen, and J. E. Olson.** 1995. *Listeria monocytogenes* exists in at least three evolutionary lines: evidence from flagellin, invasive associated protein and listeriolysin O genes. Microbiology **141:**2053–2061.

68. **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 **342:**396–401.

69. **Roberts, M. S., and F. M. Cohan.** 1995. Recombination and migration rates in natural populations of *Bacillus subtilis* and *Bacillus mojavensis.* Evolution **49:**1081–1094.

70. **Roberts, M. S., L. K. Nakamura, and F. M. Cohan.** 1994. *Bacillus mojavensis* sp. nov., distinguishable from *Bacillus subtilis* by sexual isolation, divergence in DNA sequence, and differences in fatty acid composition. Int. J. Syst. Bacteriol. **44:**256–264.

71. **Roberts, M. S., L. K. Nakamura, and F. M. Cohan.** 1996. *Bacillus vallismortis* sp. nov., a close relative of *Bacillus subtilis,* isolated from soil in Death Valley, California. Int. J. Syst. Bacteriol. **46:**470–475.

72. **Ros, C., and K. Belak.** 1996. Detection and identification of mycobacteria in formalin-fixed, paraffin-embedded tissues by nested PCR and restriction enzyme analysis. J. Clin. Microbiol. **34:**2351–2355.

73. **Sedgwick, S. G., D. Lodwick, N. Doyle, H. Crowne, and P. Strike.** 1991. Functional complementation between chromosomal and plasmid mutagenic DNA repair genes in bacteria. Mol. Gen. Genet. **229:**428–436.

74. **Segovia, L., D. Piñero, R. Palacios, and E. Martinez-Romero.** 1991. Genetic structure of a soil population of nonsymbiotic *Rhizobium leguminosarum*. Appl. Environ. Microbiol. **57:**426–433.

75. **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, San Diego, Calif.

76. **Sharp, P. M., N. C. Nolan, N. N. Cholmain, and K. M. Devine.** 1992. DNA sequence variability at the *rplX* locus of *Bacillus subtilis*. J. Gen. Microbiol. **138:**39–45.

77. **Smart, C. D., B. Schneider, C. L. Blomquist, L. J. Guerra, N. A. Harrison, U. Ahrens, K. H. Lorenz, E. Seemuller, and B. C. Kirkpatrick.** 1996. *Phytoplasma*-specific PCR primers based on sequences of the 16S-23S rRNA spacer region. Appl. Environ. Microbiol. **62:**2988–2993.

78. **Soini, H., E. C. Bottger, and M. K. Viljanen.** 1994. Identification of mycobacteria by PCR-based sequence determination of the 32-kilodalton protein gene. J. Clin. Microbiol. **32:**2944–2947.

79. **Soria, G., J. Barbe, and I. Gibert.** 1994. Molecular fingerprinting of *Salmonella typhimurium* by IS200-typing as a tool for epidemiological and evolutionary studies. Microbiologia **10:**57–68.

80. **Taghavi, M., C. Hayward, L. I. Sly, and M. Fegan.** 1996. Analysis of the phylogenetic relationships of the strains of *Burkholderia solanacearum*, *Pseudomonas syzygii*, the blood disease bacterium of banana based on 16S rRNA gene sequences. Int. J. Syst. Bacteriol. **46:**10–15.

81. **Thampapillai, R. Lan, and P. R. Reeves.** 1994. Molecular evolution in the *gnd* locus of *Salmonella enterica*. Mol. Biol. Evol. **11:**813–828.

82. **Vandamme, P., B. Pot, M. Gillis, P. De Vos, K. Kersters, and J. Swings.** 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol. Rev. **60:**407–438.

83. **Vauterin, L., B. Hoste, K. Kersters, and J. Swings.** 1995. Reclassification of *Xanthomonas*. Int. J. Syst. Bacteriol. **45:**472–489.

84. **Vazquez, J. A., L. de la Fuente, S. Berron, M. O'Rourke, N. H. Smith, J. Zhou, and B. G. Spratt.** 1993. Ecological separation and genetic isolation of *Neisseria gonorrhoeae* and *Neisseria meningitidis*. Curr. Biol. **3:**567–572.

85. **Wayne, L. G., D. J. Brenner, R. R. Colwell, P. A. D. Grimont, O. Kandler, M. I. Krichevsky, L. H. Moore, W. E. C. Moore, R. G. E. Murray, E. Stackebrandt, M. P. Starr, and H. G. Trüper.** 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int. J. Syst. Bacteriol. **37:**463–464.

86. **Whittam, T. S., and S. E. Ake.** 1993. Genetic polymorphisms and recombination in natural populations of *Escherichia coli*, p. 223–245. *In* N. Takahata and A. G. Clark (ed.), Molecular paleopopulation biology. Japan Scientific Society Press, Tokyo, Japan.

87. **Williams, A. M., U. M. Rodriguez, and M. D. Collins.** 1991. Intrageneric relationships of Enterococci as determined by reverse transcriptase sequencing of small-subunit rRNA. Res. Microbiol. **142:**67–74.

88. **Woese, C. R.** 1987. Bacterial evolution. Microbiol. Rev. **51:**221–271.

89. **Yu, H., M. J. Schurr, and V. Deretic.** 1995. Functional equivalence of *Escherichia coli* σ^E and *Pseudomonas aeruginosa* AlgU: *E. coli rpoE* restores mucoidy and reduces sensitivity to reactive oxygen intermediates in *algU* mutants of *P. aeruginosa*. J. Bacteriol. **177:**3259–3268.

90. **Zawadzki, P., M. S. Roberts, and F. M. Cohan.** 1995. The log-linear relationship between sexual isolation and sequence divergence in Bacillus transformation is robust. Genetics **140:**917–932.

91. **Zhou, J., and B. G. Spratt.** 1992. Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. Mol. Microbiol. **6:**2135–2146.