January 1997

Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data

T. Palys  
US Army

L. K. Nakamura  
USDA

Frederick M. Cohan  
Wesleyan University, fcohan@wesleyan.edu

Follow this and additional works at: https://wesscholar.wesleyan.edu/div3facpubs

Part of the Biodiversity Commons, Ecology and Evolutionary Biology Commons, and the Environmental Microbiology and Microbial Ecology Commons

Recommended Citation

https://wesscholar.wesleyan.edu/div3facpubs/355

This Article is brought to you for free and open access by the Natural Sciences and Mathematics at WesScholar. It has been accepted for inclusion in Division III Faculty Publications by an authorized administrator of WesScholar. For more information, please contact anelson01@wesleyan.edu, jmlozanowski@wesleyan.edu.
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. COHAN1*

Department of Biology, Wesleyan University, Middletown, Connecticut 06459-0170, 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
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 elate. 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, $E(d_{ij})$, and for cells in the same population, $E(d_{ii})$. The fraction of all sites that are divergent within $[E(p_{ii})]$ or between $[E(p_{ij})]$ populations may be approximated from $E(d_{ii})$ or $E(d_{ij})$, by dividing the latter by 3, provided that $E(d)$ is small (i.e., <0.09) (14).

**Materials and Methods**

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, $E(d_{ij})$, to the expected divergence within populations, $E(d_{ii})$; this ratio $E(d_{ij}) / E(d_{ii})$, is approximately equal to $E(p_{ij}) / E(p_{ii})$. We derive the distinctness ratio of populations as follows.

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

Dividing each side by $E(d_{ii})$ yields:

$$k = \frac{E(d_{ij})}{E(d_{ii})} = 1 + \frac{\mu_0}{c_p E(d_{ii})}$$

The expected divergence within populations, $E(d_{ii})$, 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 systematist 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 U04769 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, U14499, U14501, and U14503 to U14509. Analysis of divergence at gap4 in Salmonella and Escherichia was based on GenBank sequences M66853 to M66869, M66870 to M66877, and M66879 to M66882. Analysis of divergence at pir in Salmonella and Escherichia was based on GenBank sequences L01132 and L01133 and L01135 to L01139. Analysis of divergence at glnA in Citrobacter was based on GenBank sequences U14336, U14424 and U14425, U14427 to U14429, U14432, and U14466. Analysis of divergence at flu in Listeria was based on GenBank sequences X70034. Analysis of divergence at gap4 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., $E(d_{ij}) < 0.03$, or $E(p_{ij}) < 0.01$] are distinguishable by sequence data (at the level of $k = 2$) whenever the interpopulation recombination rate falls below $2.3 \times 10^{-6}$ 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 $E(d_{ij}) = 0.09$ would be distinguishable at $k > 2$ only when $c_p < 7.7 \times 10^{-6}$.

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.
is based on equation 1 and assumes a neutral mutation rate considered distinct sequence clusters. (for would disfavor interpopulation recombinants at these loci.)

Borrelia burgdorferi' were known to contain several ecologically distinct subspecies (i.e., housekeeping loci). It is expected that populations would be analyzed as separate groups in Table 1.

sites (i.e., contained two or more clearly distinct sequence clusters (e.g., specific adaptations (i.e., housekeeping genes). Some species species at protein-coding genes not involved in population-
tween a pair of subspecies or closely related species was almost (7, 8). Other ecologically distinct populations discovered as sequence clusters include Bacillus mojavensis, Bacillus walsworthii, 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 corre-
spond to ecological differences, the clusters do correspond with geographic origin (38).

**DISCUSSION**

Previous theoretical studies have shown that periodic selec-
tion 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 distinct-
ness of populations' gene sequences (k) and shows how pop-
ulation distinctness at housekeeping loci is determined by periodic selection and interpopulation recombination [repre-
sented by the parameters \( E(d_w) \) and \( \epsilon_p \), 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 dis-
tinct sequence clusters.

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

For Escherichia coli, the neutral mutation rate is \( 6.93 \times 10^{-10} \) 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, \( \pi_u \) (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., \( \pi_u \approx 0.01 \) [Table 1]). The average divergence be-
tween 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 se-
quence clusters; this was also the case for two of the three subspecies pairs within Bacillus subtilis (23). 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 sub-
specific sequence cluster was usually about 1% or less over all sites (i.e., \( \pi_u \approx 0.01 \) [Table 1]). The average divergence be-
tween 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 se-
quence 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 patho-
genic properties; these clusters were later given species status (7, 8). Other ecologically distinct populations discovered as sequence clusters include Bacillus mojavensis, Bacillus walsworthii, 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 corre-
spond to ecological differences, the clusters do correspond with geographic origin (38).
TABLE 1. Sequence divergence within and between ecologically distinct taxa at protein-coding genes not involved in taxon-specific adaptations

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Gene</th>
<th>Mean divergence within group&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Mean divergence between groups&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Ratio (between/within)&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bacillus</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. burgdorferi vs. B. afzelii</td>
<td>fla</td>
<td>0.003&lt;sup&gt;f&lt;/sup&gt;, 0.009 ± 0.002</td>
<td>0.055 ± 0.001</td>
<td>12.64</td>
<td>27</td>
</tr>
<tr>
<td>B. afzelii vs B. garinii</td>
<td>fla</td>
<td>0.006 ± 0.002, 0.010 ± 0.001</td>
<td>0.052 ± 0.001</td>
<td>6.60</td>
<td>27</td>
</tr>
<tr>
<td>B. garinii vs B. burgdorferi</td>
<td>fla</td>
<td>0.010 ± 0.001, 0.003&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.061 ± 0.000</td>
<td>9.50</td>
<td>27</td>
</tr>
<tr>
<td><strong>Neisseria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N. meningitidis vs N. gonorrhoeae</td>
<td>recA</td>
<td>0.014 ± 0.002&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.016 ± 0.002</td>
<td>1.11</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>argF</td>
<td>0.004 ± 0.001&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.056 ± 0.001</td>
<td>13.91&lt;sup&gt;f&lt;/sup&gt;</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>fbp</td>
<td>0.004 ± 0.000&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.010 ± 0.000</td>
<td>2.38</td>
<td>91</td>
</tr>
<tr>
<td><strong>Salmonella and Escherichia</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. enterica vs S. bongori</td>
<td>mdh</td>
<td>0.007 ± 0.001, 0.003 ± 0.001, 0.002&lt;sup&gt;f&lt;/sup&gt;, 0.020 ± 0.008&lt;sup&gt;f&lt;/sup&gt;, 0.004&lt;sup&gt;f&lt;/sup&gt;, 0.000&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.044 ± 0.004</td>
<td>8.06</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>gnd</td>
<td>0.030 ± 0.004, 0.038 ± 0.005, 0.023&lt;sup&gt;f&lt;/sup&gt;, 0.035 ± 0.006, 0.037 ± 0.006, 0.034 ± 0.006, 0.000&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.057 ± 0.002</td>
<td>2.03</td>
<td>59, 81</td>
</tr>
<tr>
<td></td>
<td>gapA</td>
<td>0.002&lt;sup&gt;f&lt;/sup&gt;, 0.011 ± 0.004, 0.000&lt;sup&gt;f&lt;/sup&gt;, 0.000&lt;sup&gt;f&lt;/sup&gt;, 0.001&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.026 ± 0.003</td>
<td>6.80</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>putP</td>
<td>0.044 ± 0.002, 0.004&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.066 ± 0.001</td>
<td>2.75</td>
<td>58</td>
</tr>
<tr>
<td><strong>E. coli vs S. enterica</strong></td>
<td>mdh</td>
<td>0.011 ± 0.000, 0.036 ± 0.001</td>
<td>0.154 ± 0.001</td>
<td>6.55</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>gnd</td>
<td>0.070 ± 0.006, 0.053 ± 0.001</td>
<td>0.168 ± 0.001</td>
<td>2.73</td>
<td>59, 81</td>
</tr>
<tr>
<td></td>
<td>gapA</td>
<td>0.003 ± 0.000, 0.025 ± 0.001</td>
<td>0.061 ± 0.000</td>
<td>4.36</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>putP</td>
<td>0.022 ± 0.001, 0.044 ± 0.002</td>
<td>0.206 ± 0.000</td>
<td>6.24</td>
<td>58</td>
</tr>
<tr>
<td><strong>E. coli vs S. bongori</strong></td>
<td>mdh</td>
<td>0.011 ± 0.000, 0.004 ± 0.001</td>
<td>0.160 ± 0.000</td>
<td>21.3</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>gnd</td>
<td>0.070 ± 0.006, 0.037 ± 0.007</td>
<td>0.165 ± 0.002</td>
<td>3.08</td>
<td>59, 81</td>
</tr>
<tr>
<td></td>
<td>gapA</td>
<td>0.003 ± 0.000, 0.004&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.086 ± 0.000</td>
<td>24.57</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>putP</td>
<td>0.022 ± 0.001, 0.004&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0.207 ± 0.001</td>
<td>15.92</td>
<td>58</td>
</tr>
</tbody>
</table>

**Citrobacter spp.**

C. diversus vs C. freundii

**Erwinia carotovora**

E. carotovora subsp. carotovora vs E. carotovora subsp. odorifera

E. carotovora subsp. carotovora vs E. carotovora subsp. atroseptica

E. carotovora subsp. atroseptica vs E. carotovora subsp. odorifera

**Vibrio**

V. cholerae cluster A vs V. cholerae cluster B

V. cholerae cluster A vs V. cholerae cluster C

V. cholerae cluster A vs V. cholerae cluster D

V. cholerae cluster B vs V. cholerae cluster C

V. cholerae cluster B vs V. cholerae cluster D

V. cholerae cluster C vs V. cholerae cluster D

V. cholerae cluster A vs V. mnicicus

V. cholerae cluster B vs V. mnicicus

V. cholerae cluster C vs V. mnicicus

V. cholerae cluster D vs V. mnicicus

**Bacillus**

B. subtilis 168 cluster vs B. subtilis W23 cluster

B. subtilis 168 cluster vs B. vallismortis

B. subtilis W23 cluster vs B. vallismortis

Continued on following page
populations \((c_{\text{EW}})\)? We can imagine certain circumstances under which the rates of recombination within and between populations would be equal. Equality of \(c_{\text{EW}}\) and \(c_{\text{EW}}\) 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 \(c_{\text{EW}}\) and \(c_{\text{EW}}\) 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, \(E(d_{\text{EW}})\), is \(\sim 0.09\) \((16)\) and the between-population recombination rate is very low (assuming \(c_{\text{EW}} < c_{\text{EW}} = 10^{-5}\) \((69)\), yielding a distinctiveness ratio \(k\) greater than 2 (equation 1) (Fig. 2). Likewise, in the case of *E. coli*, \(E(d_{\text{EW}})\) and \(c_{\text{EW}}\) 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 \([E(d_{\text{EW}})]\) 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 \(c_{\text{EW}}\) 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-

**TABLE 1—Continued**

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Gene</th>
<th>Mean divergence within group*</th>
<th>Mean divergence between groups*</th>
<th>Ratio (between/within)*</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. subtilis</em> 168 cluster vs. <em>B. meiowais</em></td>
<td>gyrA</td>
<td>0.012 ± 0.002, 0.031 ± 0.004</td>
<td>0.158 ± 0.002</td>
<td>7.35</td>
<td>69, 70</td>
</tr>
<tr>
<td></td>
<td>polC</td>
<td>0.014 ± 0.003, 0.010 ± 0.002</td>
<td>0.131 ± 0.002</td>
<td>10.92</td>
<td>69, 70</td>
</tr>
<tr>
<td></td>
<td>rpoB</td>
<td>0.003 ± 0.001, 0.003 ± 0.001</td>
<td>0.031 ± 0.001</td>
<td>10.33</td>
<td>69, 70</td>
</tr>
<tr>
<td><em>B. subtilis</em> W23 cluster vs. <em>B. meiowais</em></td>
<td>gyrA</td>
<td>0.045 ± 0.004, 0.031 ± 0.004</td>
<td>0.156 ± 0.002</td>
<td>4.11</td>
<td>69, 70</td>
</tr>
<tr>
<td></td>
<td>polC</td>
<td>0.050 ± 0.008, 0.010 ± 0.002</td>
<td>0.121 ± 0.002</td>
<td>4.03</td>
<td>69, 70</td>
</tr>
<tr>
<td></td>
<td>rpoB</td>
<td>0.012 ± 0.002, 0.003 ± 0.001</td>
<td>0.026 ± 0.001</td>
<td>3.47</td>
<td>69, 70</td>
</tr>
<tr>
<td><em>B. globosphorus</em> vs. <em>B. psychrophilus</em></td>
<td>pyt</td>
<td>0.004 ± 0.002, 0.018 ± 0.004</td>
<td>0.096 ± 0.002</td>
<td>8.73</td>
<td>64</td>
</tr>
<tr>
<td><em>Listeria</em></td>
<td><em>L. monocytogenes</em> type 1 vs. <em>L. monocytogenes</em> type 2</td>
<td>fla</td>
<td>0.000 ± 0.000, 0.001 ± 0.000</td>
<td>0.018 ± 0.000</td>
<td>36.00</td>
</tr>
<tr>
<td></td>
<td><em>L. monocytogenes</em> type 1 vs. <em>L. monocytogenes</em> type 3</td>
<td>fla</td>
<td>0.000 ± 0.000, 0.005*</td>
<td>0.008 ± 0.001</td>
<td>1.60</td>
</tr>
<tr>
<td></td>
<td><em>L. monocytogenes</em> type 2 vs. <em>L. monocytogenes</em> type 3</td>
<td>fla</td>
<td>0.001 ± 0.000, 0.005*</td>
<td>0.025 ± 0.001</td>
<td>8.33</td>
</tr>
<tr>
<td><em>Mycobacterium</em></td>
<td><em>M. avium</em> vs. <em>M. intracellulare</em></td>
<td>hsp</td>
<td>0.005 ± 0.001, 0.036 ± 0.007</td>
<td>0.036 ± 0.003</td>
<td>1.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32kDa</td>
<td>0.047 ± 0.021, 0.012 ± 0.003</td>
<td>0.046 ± 0.010</td>
<td>1.56</td>
</tr>
</tbody>
</table>

*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 mpx 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, pyt, and rpoB genes of *Bacillus*. *Mean ± standard error for pairwise divergence within each of the groups (\(s_{j}\)) shown in order of appearance in the Comparison column.*

*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-

---

**RAW TEXT END**
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 agile 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.
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 any 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. odontofera 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]1), 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 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)**

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Mean 16S rRNA divergence</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within taxon</td>
<td>Between taxa</td>
</tr>
<tr>
<td>Deleya aquamarina</td>
<td>0.000</td>
<td>22</td>
</tr>
<tr>
<td>Halomonas meridiana</td>
<td>0.001</td>
<td>25</td>
</tr>
<tr>
<td>Franciscella tularensis biovar A</td>
<td>0.002</td>
<td>26</td>
</tr>
<tr>
<td>F. tularensis biovar B</td>
<td>0.003</td>
<td>2</td>
</tr>
<tr>
<td>Bacillus cereus</td>
<td>0.004</td>
<td>2</td>
</tr>
<tr>
<td>B. mycoides</td>
<td>0.005</td>
<td>2</td>
</tr>
<tr>
<td>Bacillus anthracis</td>
<td>0.006</td>
<td>2</td>
</tr>
<tr>
<td>B. thuringiensis</td>
<td>0.005</td>
<td>3</td>
</tr>
<tr>
<td>Bacillus subtilis</td>
<td>0.000 ± 0.000</td>
<td>20</td>
</tr>
<tr>
<td>Clostridium acetobutylicum C. beijerinckii</td>
<td>0.003 ± 0.002</td>
<td>270</td>
</tr>
<tr>
<td>Enterococcus casseliflavus E. gallinarum</td>
<td>0.007</td>
<td>87</td>
</tr>
<tr>
<td>Enterococcus avium E. raffinosus</td>
<td>0.003</td>
<td>87</td>
</tr>
<tr>
<td>Enterococcus durans E. faecium</td>
<td>0.003</td>
<td>87</td>
</tr>
<tr>
<td>Aerococcus viridans Pediococcus urinary-equina</td>
<td>0.000</td>
<td>86</td>
</tr>
<tr>
<td>Mycobacterium intracellulare serovars 4, 5, 6, and 8</td>
<td>0.000 ± 0.000</td>
<td>9</td>
</tr>
<tr>
<td>M. intracellulare serovar A</td>
<td>0.000 ± 0.000</td>
<td>9</td>
</tr>
<tr>
<td>Mycobacterium intracellulare serovars 4, 5, 6, 8, 10, and 11</td>
<td>0.000 ± 0.000</td>
<td>9</td>
</tr>
<tr>
<td>M. avium</td>
<td>0.003</td>
<td>66</td>
</tr>
<tr>
<td>Cellulomonas biazotea C. fim *</td>
<td>0.000</td>
<td>61</td>
</tr>
<tr>
<td>Fusobacterium nucleatum F. periodicum</td>
<td>0.000 ± 0.000</td>
<td>9</td>
</tr>
</tbody>
</table>

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

\* Mycobacterium intracellulare serovars 4, 5, 6, and 8 are in one DNA-DNA hybridization group, and serovar 9 is in another.
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 gonorhoeae* 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-
tein-coding sequence data provide a more effective tool for describing and discovering the ecological diversity of bacteria.

There are many ecologically distinct populations of bacteria that fail to be distinguished by DNA-DNA hybridization (Tables 3 and 4). Perhaps future surveys of protein-coding sequences will succeed in distinguishing the population groups listed in Table 4 into separate sequence similarity clusters.

**How distinct must a cluster be?** In nearly all cases for which we have sequence diversity data, groups known to be ecologically divergent have shown a distinctness ratio \( k \) of 2 or higher (Table 1). This suggests that future sequence similarity clusters with this level of distinctness should be ecologically distinct as well.

**Importance of surveying multiple genes.** We suggest that classification of bacteria by sequence similarity clusters should be based on variation at two or more unlinked loci because rare homologous recombination with other taxa can substantially alter the extent of sequence divergence at a single locus (31, 59, 91). If two or more unlinked loci are analyzed, the pattern of sequence divergence is less likely to be affected by recombination.

**Challenge of newly divergent populations.** As discussed earlier, the theory predicts that populations will be distinguishable as sequence similarity clusters only when the gene being studied has had sufficient time to diverge since the populations split from their common ancestor. How, then, might we expect to use molecular data to distinguish pathogenic populations that represent very recent innovations?

It is possible that even very new populations may be distinguished as separate sequence similarity clusters. Let us suppose, for example, that the O157:H7 pathogen of *E. coli* were a new evolutionary innovation of the last decade. Because all O157:H7 strains would be expected to descend from an original innovative mutant, the O157:H7 strains should form a clade for any gene sequence (provided that recombination between populations is sufficiently rare) (Fig. 1A). If the O157:H7 population were to frequently fine tune its new adaptations, the population would undergo many periodic selection events that would not be shared with non-O157:H7 *E. coli* (Fig. 1B). Each such periodic selection event would purge diversity within the O157:H7 population, but it would not purge the divergence between O157:H7 and other ecological populations. Therefore, one would expect that the divergence within O157:H7 (or any newly formed population) would be much lower than that between O157:H7 and its nearest relatives (Fig. 1C).

However, to see that the divergence between new populations is much larger than that within them, one would need a marker with a rapid evolutionary rate. Insertion sequence fingerprinting may provide such a set of markers that would be able to detect O157:H7 and similar populations as separate clusters (79).

**Conclusions.** An understanding of the evolutionary genetics of bacteria makes an improved criterion for species classification possible. The theory of periodic selection predicts that ecologically distinct populations should eventually diverge into distinct DNA sequence clusters at nearly every gene locus. A review of the systematics literature corroborates this prediction: ecologically distinct populations of bacteria nearly always fall into separate clusters based on the DNA sequences of protein-coding genes. Thus, sequence clustering based on protein-coding genes is a useful criterion for distinguishing ecological populations of bacteria. Moreover, the sequences of protein-coding genes are more effective in distinguishing ecologically distinct populations than either DNA-DNA hybridization or 16S rRNA gene sequences.

We therefore recommend that molecular criteria for species demarcation should include protein-coding gene sequences, 16S rRNA gene sequences, and DNA hybridization. Specifically, we recommend that each phenotypically distinguished DNA sequence cluster should be recognized as a separate species or subspecies.

Many systematists have already used sequence similarity clusters as a criterion for species demarcation in the eukaryotic world (48). If bacteriologists were also to accept this criterion, then species of all categories, both prokaryotic and eukaryotic, could be distinguished by the same criterion (15, 48).

**ACKNOWLEDGMENTS**

This work was supported by Environmental Protection Agency grants R82-1388-01-0 and R82-5348-01-0 and by research funds from Wesleyan University.

**REFERENCES**