Protein-coding genes as molecular markers for ecologically distinct populations: the case of two Bacillus species

T. Palys
E. Berger, Wesleyan University
I. Mitrica
L. K. Nakamura
Frederick M Cohan, Wesleyan University
Protein-coding genes as molecular markers for ecologically distinct populations: the case of two *Bacillus* species

Thomas Palys, Evelyn Berger, Ionel Mitrica, L. K. Nakamura and Frederick M. Cohan

Author for correspondence: Frederick M. Cohan. Tel: +1 860 685 3482. Fax: +1 860 685 3279. e-mail: FCohan@Wesleyan.edu

*Bacillus globisporus* and *Bacillus psychrophilus* are one among many pairs of ecologically distinct taxa that are distinguished by very few nucleotide differences in 16S rRNA gene sequence. This study has investigated whether the lack of divergence in 16S rRNA between such species stems from the unusually slow rate of evolution of this molecule, or whether other factors might be preventing neutral sequence divergence at 16S rRNA as well as every other gene. *B. globisporus* and *B. psychrophilus* were each surveyed for restriction-site variation in two protein-coding genes. These species were easily distinguished as separate DNA sequence clusters for each gene. The limited ability of 16S rRNA to distinguish these species is therefore a consequence of the extremely slow rate of 16S rRNA evolution. The present results, and previous results involving two *Mycobacterium* species, demonstrate that there exist closely related species which have diverged long enough to have formed clearly separate sequence clusters for protein-coding genes, but not for 16S rRNA. These results support an earlier argument that sequence clustering in protein-coding genes could be a primary criterion for discovering and identifying ecologically distinct groups, and classifying them as separate species.

**Keywords:** species demarcation, sequence cluster, *Bacillus*, bacteria

**INTRODUCTION**

The theory of evolutionary genetics provides a compelling rationale for using sequence data to characterize bacterial diversity. For every gene in the genome, ecologically distinct populations are predicted to diverge into separate sequence clusters (Cohan, 1994a, 1994b, 1995, 1996; Palys et al., 1997; Majewski & Cohan, 1999). That is, ecologically distinct populations are expected to form separate monophyletic groups, based on sequence data, with much greater sequence divergence between populations than within them. Indeed, Palys et al. (1997) showed that closely related but ecologically distinct bacterial taxa can nearly always be distinguished as separate sequence clusters for protein-coding genes. That sequence clusters correspond to ecologically distinct populations has been useful for bacterial systematics in two ways. Firstly, established sequence differences between taxa can be used diagnostically to identify unknown isolates from the environment. For this purpose, taxa need only be fixed for alternative bases at a single or very few sites in the sequence. For example, clinical isolates can be distinguished as *Mycobacterium avium* or *Mycobacterium intracellulare* on the basis of a fixed difference at only one site within the 16S rRNA molecule (De Beenhouwer et al., 1995).

The correspondence between ecotypes and sequence clusters is also useful for discovering cryptic ecological diversity within a taxon. In several cases, a survey of sequence diversity within a named species has revealed multiple sequence clusters which were later found to be ecologically distinct (Baranton et al., 1992; Balmelli & Piffaretti, 1996; Roberts & Cohan, 1995; Roberts et al., 1994, 1996; Nakamura et al., 1999). For example, sequence and allozyme surveys of *Borrelia burgdorferi* (sensu lato) led to the discovery of several *Borrelia* species with different pathogenic properties (Balmelli & Piffaretti, 1996; Baranton et al., 1992).

While sequence data may be used for both species diagnosis and species discovery, discovery requires...
Table 1. Restriction-digest patterns and haplotypes for gene pyk

The numbers of restriction-digest patterns for each restriction enzyme correspond to the numbers in the restriction maps of Fig. 1.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Restriction-digest pattern</th>
<th>pyk haplotype*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>AluI</strong></td>
<td><strong>BsrUI</strong></td>
</tr>
<tr>
<td><em>B. globisporus</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B-3396</td>
<td>1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>NRS-1527</td>
<td>1 1 1 1 2 1 1 1</td>
<td>1 1 1 2 2 2 2 2</td>
</tr>
<tr>
<td>NRS-1532</td>
<td>1 1 1 1 2 1 1 1</td>
<td>1 1 1 2 2 2 2 2</td>
</tr>
<tr>
<td>NRS-1533†</td>
<td>1 2 2 3 1 1 1 3</td>
<td>1 2 1 3 3 3 3 3</td>
</tr>
</tbody>
</table>

| *B. psychrophilus* | | | | | | | | |
| NRS-1515       | 2 3 2 4 4 2 1 2 | 1 2 1 4 4 4 4 4 |               |             |             |             |             |             |
| NRS-1524       | 3 4 2 4 2 2 2 2 | 1 2 2 5 5 5 5 5 |               |             |             |             |             |             |
| NRS-1526       | 4 3 2 5 5 2 1 2 | 1 2 2 6 6 6 6 6 |               |             |             |             |             |             |
| NRS-1528       | 3 3 2 4 4 2 1 2 | 1 2 2 7 7 7 7 7 |               |             |             |             |             |             |
| NRS-1530†      | 5 3 3 5 5 2 1 2 | 1 2 2 8 8 8 8 8 |               |             |             |             |             |             |

† Each haplotype number indicates a distinct restriction-digest pattern over all restriction enzymes.

much greater inter-species divergence than diagnosis. This is because diagnosing strains into already characterized species requires only a single nucleotide difference, such that species-specific probes can readily identify new isolates (e.g. te Giffel et al., 1997). Clearly, a single substitution, even if fixed between two species, would not provide an efficient means for discovering the species. If a mixed pool of strains from two such species were surveyed, the single-nucleotide polymorphism would be most parsimoniously interpreted as variation within a single species.

In protein-coding genes, the degree of sequence divergence is sufficient for diagnosis as well as discovery of species: sequence data from protein-coding genes typically yields clear sequence clusters that correspond to species (or subspecies). The mean sequence divergence between species is usually in the range of 1–2%, whereas the mean sequence divergence between closely related species is frequently much greater (Palys et al., 1997). Palys et al. (1997) quantified the sequence distinctness of taxa as the ratio \( k \) of the mean sequence divergence between taxa to the mean sequence divergence within them, and found that protein-coding genes typically yielded a distinctness ratio between 2 and 10 for closely related species.

In contrast, closely related taxa are often extremely similar in their 16S rRNA sequences (Fox et al., 1992; Stackebrandt & Goebel, 1994; Palys et al., 1997). For example, Palys et al. (1997) listed seven pairs of closely related taxa with between zero and two nucleotide sites distinguishing them at the 16S rRNA level. While divergence of the 16S rRNA sequence is frequently sufficient for diagnosis of known species (because only a single fixed difference is required; Barry et al., 1990; Kljun et al., 1991, 1994; Wang et al., 1992; te Giffel et al., 1997), differences in only 0–2 sites typical for 16S rRNA are too few to allow discovery of new species.

Why do 16S rRNA sequences so poorly distinguish ecologically distinct taxa? Several explanations are possible (Palys et al., 1997). Firstly, some taxa may have diverged so recently that neutral sequence divergence has not yet accumulated at any gene locus. Alternatively, some taxa may recombine at a rate that is high enough to prevent neutral sequence divergence at all loci while still allowing the taxa to diverge in ecological characteristics (Cohan, 1994a). Finally, the failure of 16S rRNA sequences to provide multiple diagnostic sites could stem from the unusually slow rate of evolution of this molecule. Rates of evolutionary substitution in 16S rRNA genes are an order of magnitude lower than those for protein-coding genes (Ochman & Wilson, 1987; Yamamoto & Harayama, 1998). Thus, some pairs of ecologically distinct taxa may have had time to accumulate neutral sequence divergence at rapidly evolving loci but not yet at 16S rRNA.

The hypothesis that the low rate of 16S rRNA evolution is responsible for the failure of this molecule to provide multiple diagnostic sites for closely related but ecologically distinct taxa was tested. Our approach is to determine whether pairs of taxa that are very similar in 16S rRNA sequence (i.e. with 1–2 diagnostic sites) may be distinguished as separate clusters for protein-coding genes, with high distinctness ratios \( k \) typically observed for other species pairs (i.e. \( 2 < k < 10 \); Palys et al., 1997).

This approach requires estimates of sequence divergence both within and between populations, so multiple strains of each population must be sampled. Unfortunately, to our knowledge, only two eco-
logically distinct taxa that are barely distinguishable by 16S rRNA data (*M. avium* and *M. intracellulare*) have previously been surveyed for sequence variation within and between taxa at protein-coding genes. While these groups are 99.8% identical in terms of 16S rRNA sequence (Boddinghaus et al., 1990), they do fall into distinct sequence clusters for two protein-coding genes, with a distinctness ratio of 1-7 (Soini et al., 1994; Kapur et al., 1995; Ros & Belak, 1996; Palys et al., 1997).

This study has tested whether the particularly low rate of 16S rRNA evolution is responsible for this molecule’s limited ability to distinguish another pair of species, *Bacillus globisporus* and *Bacillus psychrophilus*. These are two very closely related species differing in their optimum temperatures and other conditions for growth (Larkin & Stokes, 1967; Gordon et al., 1973; Nakamura, 1984). The 16S rRNA sequences of two strains of *B. psychrophilus* and one strain of *B. globisporus* were previously determined (Ash et al., 1991; Fox et al., 1992; Ludwig et al., 1992; Suzuki & Yamasato, 1994); these species were found to have only two consistent nucleotide differences. Nevertheless, these species are easily distinguished by DNA–DNA reassociation criteria and several phenotypic characters (Nakamura, 1984). Here, evidence is presented that *B. globisporus* and *B. psychrophilus* can be clearly distinguished into separate clusters by the sequences of two protein-coding genes, pyk (encoding pyruvate kinase) and ald (encoding alanine dehydrogenase).

**METHODS**

**Bacterial strains and culture conditions.** Four *B. globisporus* and five *B. psychrophilus* strains, listed in Table 1, were used in this study. The strains were provided by the Agricultural Research Service Culture Collection at the National Center for Agricultural Utilization Research. Strains were maintained on tryptone-glucose-yeast extract (TGY) agar plates.

**DNA isolation.** Genomic DNA was extracted and purified as described previously (Cohan et al., 1991), except that strains were cultured on TGY medium.

**PCR amplification.** PCR was used to amplify pyk and ald fragments from purified genomic DNA preparations. Four oligonucleotide primers (PK1, PK2, PK3 and PK4) for pyk were designed from the *B. psychrophilus* sequence (Tanaka et al., 1995). Primers extended from: bp 340–359 (PK1; 5'-ATT GAG GCT GGT ATG GT-3'); bp 2001–1981 (PK2; 5'-AAC GCC AGA TTC AGC ATC CAT-3'); bp 487–506 (PK3; 5'-CGG ACG CAT TCT ATG ATG AA-3'); and bp 1869–1849 (PK4; 5'-TTC CGT AAT GAG TCC TGC GCA-3'). An initial PCR of pyk with primers PK1 and PK2 was performed in 100 µl reaction mix with the following final concentrations of reagents: PCR buffer (Promega, 10 mM Tris/HCl, pH 9.0 and 50 mM KCl), 5 mM MgCl₂, 200 mM each dNTP, 1 µM each primer, 2.5 units Taq DNA polymerase (Promega) and ~0.5 µg genomic DNA. The thermocycler profile consisted of 30 cycles: denaturation at 94 °C for 30 s, annealing at 50 °C for 45 s and extension at 72 °C for 2 min. PCR products were run on 1% TAE (Tris-acetate-EDTA)-agarose gels and fragments corresponding to the expected size of pyk (1662 bp) were excised and purified using GeneClean II (Bio 101). A second, nested PCR was performed with primers PK3 and PK4, using ~0.1 µg purified pyk PCR product as template. Reagent concentrations, reaction volume and thermocycler profile were the same as for the initial pyk PCR reaction.

Two oligonucleotide primers (AD1 and AD2) for ald were designed from the ald sequence of *Bacillus sphaericus* (Kuroda et al., 1990). Primer AD1 extended from bp 491–510 (5'-ATG AAG ATT GGT ATT CCA AA-3') and primer AD2 extended from bp 1480–1461 (5'-AGC TGC CTT ATT GGC AAT TT-3'). An initial PCR of ald was performed using primers AD1 and AD2 in 100 µl reaction mix, with the following final concentrations of reagents: PCR buffer (Promega, 10 mM Tris/HCl, pH 9.0 and 50 mM KCl), 5 mM MgCl₂, 200 mM each dNTP, 2 µM each primer, 2.5 units Taq DNA polymerase (Promega) and ~0.5 µg genomic DNA. The thermocycler profile consisted of 30 cycles: denaturation at 94 °C for 30 s, annealing at 35 °C for 45 s and extension at 72 °C for 2 min. As with pyk, PCR products of ald were electrophoresed, and fragments corresponding to the expected size of ald (990 bp) were excised and purified. A second reaction of ald was performed with AD1 and AD2 primers, using ~0.1 µg purified PCR product as template, under the same reaction conditions used for pyk. The final products of the pyk and ald PCR reactions were run on 1% TBE (Tris-borate-EDTA) gels to confirm that the sizes of products were as expected.

**Fig. 1.** Restriction site map of the pyk gene segment (bp 487–1869 of Tanaka et al., 1995) for each restriction-digest pattern observed. Presence and absence of restriction sites were determined from restriction fragment patterns, using the computer algorithm of Cohan et al. (1991).
Two primers for 16S rRNA were designed from the sequence of strain DSM 4 of *B. globisporus* (Ludwig et al., 1992; GenBank accession no. X68415). Primer GLF1 (5’-GCT GAA ACT CAA AGG AAT TGA-3’) extended from bp 910–930 and primer GLR2 (5’-ATT GCT GGC AAC TAA GAT CAA-3’) extended from bp 1144–1124. PCR was performed for *B. psychrophilus* strains NRS-1515, NRS-1524 and NRS-1526, and for *B. globisporus* strains NRS-1527 and NRS-1532, in 100 µl reaction mixes with the following concentrations of reagents: 200 mM each dNTP, 2 µM each primer, 1.25 units Taq DNA polymerase (Qiagen) and 8 mM MgCl₂. The thermocycler profile consisted of 3 min at 94 °C, then 35 cycles of 45 s at 94 °C, 45 s at 42 °C and 1 min at 72 °C, followed by 3 min at 72 °C. The PCR products were purified by 8% agarose gel electrophoresis and extracted from gels using the Qiagel gel purification kit (Qiagen).


**Sequence analysis.** The 16S rRNA genes were sequenced directly from the purified PCR products at the University of Pennsylvania DNA Sequencing Facility.

**Phylogenetic analysis.** Phylogenies for pyk and *ald* and for a composite of the two genes were constructed on the basis of restriction site variation, using the PAUP 3.1 algorithm for

---

**Table 2. Restriction-digest patterns and haplotypes for gene *ald***

The numbers of restriction-digest patterns for each restriction enzyme correspond to the numbers in the restriction maps of Fig. 1.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Restriction-digest pattern</th>
<th><em>Haplotypes</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Alu</em></td>
<td><em>Bst</em>UI</td>
</tr>
<tr>
<td><em>B. globisporus</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B-3396</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NRS-1527</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NRS-1532</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><em>B. psychrophilus</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NRS-1515</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>NRS-1524</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>NRS-1526</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>NRS-1528</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>NRS-1530*</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

*Each haplotype number indicates a distinct restriction-digest pattern over all restriction enzymes.*

† Gene *ald* of strain NRS-1533* could not be amplified under the PCR conditions used.
branch and bound with bootstrap (Swofford, 1993). The pyk phylogeny was rooted by the sequence of the Bacillus licheniformis type strain ATCC 14580T (Tanaka et al., 1995) and the ald phylogeny was rooted by the sequence of B. sphaericus strain IFO 3525 (Kuroda et al., 1990).

**Diversity, divergence and distinctness.** The nucleotide diversity at pyk and ald within B. globisporus and within B. psychrophilus \((\pi_{a1} \text{ and } \pi_{a2})\) respectively and the nucleotide divergence between species \((\pi_s)\) were calculated from restriction-site divergence, for all pairwise comparisons of divergence between strains, using Equation 5.50 of Nei (1987). The distinctness ratio \((k)\) between species was calculated as follows:

\[
k = \pi_s/[0.5(\pi_{a1} + \pi_{a2})]
\]  

(Equation 1),

where \(\pi_s\) is the mean sequence divergence between strains of different populations, and \(\pi_{a1}\) and \(\pi_{a2}\) are the mean sequence divergence levels within population 1 and within population 2, respectively.

**RESULTS**

**Gene pyk**

For gene *pyk*, each restriction enzyme revealed from two to five distinct restriction-digest patterns among the nine strains (Table 1 and Fig. 1). There were 40 restriction sites among all nine strains, and 24 of these sites were variable across strains. Eight different haplotypes were found among the nine strains assayed.

There were four most-parsimonious phylogenies based on *pyk* restriction site variation (with tree size of 41 restriction site changes; Fig. 3). Monophyly of the four *B. globisporus* strains and monophyly of the five *B. psychrophilus* strains were strongly supported by bootstrap values of 100% and 91%, respectively.

Restriction site variation clearly distinguishes the two species as separate sequence clusters. Twelve restriction sites appear to be diagnostic characters, in being fixed in one species’ sample but absent in the other’s (Table 1 and Fig. 1). There is little nucleotide sequence variation within each species (0-9% and 1-5% within *B. globisporus* and within *B. psychrophilus*, respectively). Much greater sequence divergence occurs between species (9-8%), yielding a distinctness ratio of 8:17 (Table 3).

**Gene ald**

Analysis of the variation in *ald* gene does not include strain NRS-1533\(^T\) because the *ald* gene of this strain could not be amplified by PCR. Each restriction enzyme revealed from one to three distinct restriction digest patterns for gene *ald* (Table 2 and Fig. 2). There were 27 restriction sites among the eight strains surveyed, and 13 of these were variable among strains. Five haplotypes were found among the eight strains.

There was one most parsimonious phylogeny based on *ald* restriction site variation (with tree size of 27 restriction site changes; Fig. 4). Monophyly of *B. globisporus* is strongly supported (with 99% bootstrap support). While *B. psychrophilus* appears as a monophyletic group in the phylogeny, monophyly of this group is less well supported (62% bootstrap support).

As with *pyk*, restriction site variation at *ald* clearly distinguishes *B. globisporus* and *B. psychrophilus* as separate sequence clusters. There is little variation within each species (0-0% within *B. globisporus* and 1-5% within *B. psychrophilus*) and much greater variation between species (6-4%) (Table 3), yielding a distinctness ratio of 8:53. Moreover, five restriction sites appear to be diagnostic, in being fixed in one species but absent in the other (Table 2, Fig. 2).

---

**Table 3.** Nucleotide divergence within and between *B. globisporus* and *B. psychrophilus*, and the distinctness ratio \((k)\), for genes *pyk* and *ald*

<table>
<thead>
<tr>
<th>Gene</th>
<th>Species</th>
<th>Within-species divergence</th>
<th>Between-species divergence</th>
<th>Distinctness ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>pyk</em></td>
<td><em>B. globisporus</em></td>
<td>0.009 ± 0.002</td>
<td>0.098 ± 0.002</td>
<td>8.17</td>
</tr>
<tr>
<td></td>
<td><em>B. psychrophilus</em></td>
<td>0.015 ± 0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>ald</em></td>
<td><em>B. globisporus</em></td>
<td>0.000 ± 0.000</td>
<td>0.064 ± 0.001</td>
<td>8.53</td>
</tr>
<tr>
<td></td>
<td><em>B. psychrophilus</em></td>
<td>0.015 ± 0.003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DISCUSSION

B. globisporus and B. psychrophilus are among many pairs of ecologically distinct taxa that are extremely similar in their 16S rRNA sequences (Fox et al., 1992; Stackebrandt & Goebel, 1994; Palys et al., 1997). The present sequence data indicate that whereas the sequence differences between these species are few (at only bp 1028 and 1029), these differences appear to be fixed between species and could serve as diagnostic characters. On the basis of protein-coding genes, these species are much more distinct. Our restriction data show that B. globisporus and B. psychrophilus can be clearly distinguished as separate sequence clusters for two protein-coding genes, with at least 17 diagnostic nucleotide sites between species. These observations indicate that the limited ability of the 16S rRNA gene sequence to distinguish B. globisporus and B. psychrophilus stems from the unusually slow evolutionary rate of this gene. Ruled out are the following two alternative explanations: 1) that the time of divergence has been too short for neutral sequence divergence to accumulate at any locus; or 2) that recombination between species has been too frequent to allow neutral sequence divergence at any locus.

Implications for systematics

In principle, DNA sequences for either rRNA or protein-coding genes should provide a means to discover ecologically distinct groups and to classify them as separate species. This is because evolutionary genetic theory predicts that each ecologically distinct group will eventually diverge into its own sequence cluster for any gene in the genome (Cohan, 1994a, 1994b, 1995, 1996; Palys et al., 1997; Majewski & Cohan, 1999). One caveat is that the theory’s expectation of eventual divergence does not necessarily predict the state of divergence now at any particular gene locus. The divergence level depends on the age of the taxa as well as the rate of evolution of the gene in question.

Apparently the rate of evolution for protein-coding genes is usually high enough to distinguish ecologically distinct populations even when they are very closely related. Palys et al. (1997) found a nearly complete correspondence between sequence clusters for protein-

16S rRNA

On the basis of sequencing two B. psychrophilus strains and one B. globisporus strain, the study of Fox et al. (1992) showed these species to be distinguished by two diagnostic nucleotide sites (bp 1028 and 1029). Our sequencing of bp 910–1144 in three additional B. psychrophilus strains and two additional B. globisporus strains confirmed that sites 1028 and 1029 are diagnostic: at sites 1028 and 1029, each B. psychrophilus strain has bases A and T, respectively, and each B. globisporus strain has bases G and C.

Composite of both genes

A composite-gene phylogeny based on the 40 restriction sites of pyk and ald is shown in Fig. 5. There were six most-parsimonious trees, with a tree size of 69 restriction site changes. The monophyly of each species is demonstrated by strong bootstrap support (100% support for B. globisporus and 95% support for B. psychrophilus).

Fig. 4. Phylogeny of B. globisporus and B. psychrophilus, based on restriction site variation in ald. The level of bootstrap support, indicated by percentage values, is based on 100 replicate bootstrap runs.

Fig. 5. Phylogeny of B. globisporus and B. psychrophilus, based on restriction site variation in both pyk and ald. The level of bootstrap support, indicated by percentage values, is based on 100 replicate bootstrap runs.

16S rRNA

On the basis of sequencing two B. psychrophilus strains and one B. globisporus strain, the study of Fox et al. (1992) showed these species to be distinguished by two diagnostic nucleotide sites (bp 1028 and 1029). Our sequencing of bp 910–1144 in three additional B. psychrophilus strains and two additional B. globisporus strains confirmed that sites 1028 and 1029 are diagnostic: at sites 1028 and 1029, each B. psychrophilus strain has bases A and T, respectively, and each B. globisporus strain has bases G and C.

DISCUSSION

B. globisporus and B. psychrophilus are among many pairs of ecologically distinct taxa that are extremely similar in their 16S rRNA sequences (Fox et al., 1992; Stackebrandt & Goebel, 1994; Palys et al., 1997). The present sequence data indicate that whereas the sequence differences between these species are few (at only bp 1028 and 1029), these differences appear to be fixed between species and could serve as diagnostic characters. On the basis of protein-coding genes, these species are much more distinct. Our restriction data show that B. globisporus and B. psychrophilus can be clearly distinguished as separate sequence clusters for two protein-coding genes, with at least 17 diagnostic nucleotide sites between species. These observations indicate that the limited ability of the 16S rRNA gene sequence to distinguish B. globisporus and B. psychrophilus stems from the unusually slow evolutionary rate of this gene. Ruled out are the following two alternative explanations: 1) that the time of divergence has been too short for neutral sequence divergence to accumulate at any locus; or 2) that recombination between species has been too frequent to allow neutral sequence divergence at any locus.

Implications for systematics

In principle, DNA sequences for either rRNA or protein-coding genes should provide a means to discover ecologically distinct groups and to classify them as separate species. This is because evolutionary genetic theory predicts that each ecologically distinct group will eventually diverge into its own sequence cluster for any gene in the genome (Cohan, 1994a, 1994b, 1995, 1996; Palys et al., 1997; Majewski & Cohan, 1999). One caveat is that the theory’s expectation of eventual divergence does not necessarily predict the state of divergence now at any particular gene locus. The divergence level depends on the age of the taxa as well as the rate of evolution of the gene in question.

Apparently the rate of evolution for protein-coding genes is usually high enough to distinguish ecologically distinct populations even when they are very closely related. Palys et al. (1997) found a nearly complete correspondence between sequence clusters for protein-

16S rRNA

On the basis of sequencing two B. psychrophilus strains and one B. globisporus strain, the study of Fox et al. (1992) showed these species to be distinguished by two diagnostic nucleotide sites (bp 1028 and 1029). Our sequencing of bp 910–1144 in three additional B. psychrophilus strains and two additional B. globisporus strains confirmed that sites 1028 and 1029 are diagnostic: at sites 1028 and 1029, each B. psychrophilus strain has bases A and T, respectively, and each B. globisporus strain has bases G and C.

DISCUSSION

B. globisporus and B. psychrophilus are among many pairs of ecologically distinct taxa that are extremely similar in their 16S rRNA sequences (Fox et al., 1992; Stackebrandt & Goebel, 1994; Palys et al., 1997). The present sequence data indicate that whereas the sequence differences between these species are few (at only bp 1028 and 1029), these differences appear to be fixed between species and could serve as diagnostic characters. On the basis of protein-coding genes, these species are much more distinct. Our restriction data show that B. globisporus and B. psychrophilus can be clearly distinguished as separate sequence clusters for two protein-coding genes, with at least 17 diagnostic nucleotide sites between species. These observations indicate that the limited ability of the 16S rRNA gene sequence to distinguish B. globisporus and B. psychrophilus stems from the unusually slow evolutionary rate of this gene. Ruled out are the following two alternative explanations: 1) that the time of divergence has been too short for neutral sequence divergence to accumulate at any locus; or 2) that recombination between species has been too frequent to allow neutral sequence divergence at any locus.

Implications for systematics

In principle, DNA sequences for either rRNA or protein-coding genes should provide a means to discover ecologically distinct groups and to classify them as separate species. This is because evolutionary genetic theory predicts that each ecologically distinct group will eventually diverge into its own sequence cluster for any gene in the genome (Cohan, 1994a, 1994b, 1995, 1996; Palys et al., 1997; Majewski & Cohan, 1999). One caveat is that the theory’s expectation of eventual divergence does not necessarily predict the state of divergence now at any particular gene locus. The divergence level depends on the age of the taxa as well as the rate of evolution of the gene in question.

Apparently the rate of evolution for protein-coding genes is usually high enough to distinguish ecologically distinct populations even when they are very closely related. Palys et al. (1997) found a nearly complete correspondence between sequence clusters for protein-

16S rRNA

On the basis of sequencing two B. psychrophilus strains and one B. globisporus strain, the study of Fox et al. (1992) showed these species to be distinguished by two diagnostic nucleotide sites (bp 1028 and 1029). Our sequencing of bp 910–1144 in three additional B. psychrophilus strains and two additional B. globisporus strains confirmed that sites 1028 and 1029 are diagnostic: at sites 1028 and 1029, each B. psychrophilus strain has bases A and T, respectively, and each B. globisporus strain has bases G and C.

DISCUSSION

B. globisporus and B. psychrophilus are among many pairs of ecologically distinct taxa that are extremely similar in their 16S rRNA sequences (Fox et al., 1992; Stackebrandt & Goebel, 1994; Palys et al., 1997). The present sequence data indicate that whereas the sequence differences between these species are few (at only bp 1028 and 1029), these differences appear to be fixed between species and could serve as diagnostic characters. On the basis of protein-coding genes, these species are much more distinct. Our restriction data show that B. globisporus and B. psychrophilus can be clearly distinguished as separate sequence clusters for two protein-coding genes, with at least 17 diagnostic nucleotide sites between species. These observations indicate that the limited ability of the 16S rRNA gene sequence to distinguish B. globisporus and B. psychrophilus stems from the unusually slow evolutionary rate of this gene. Ruled out are the following two alternative explanations: 1) that the time of divergence has been too short for neutral sequence divergence to accumulate at any locus; or 2) that recombination between species has been too frequent to allow neutral sequence divergence at any locus.

Implications for systematics

In principle, DNA sequences for either rRNA or protein-coding genes should provide a means to discover ecologically distinct groups and to classify them as separate species. This is because evolutionary genetic theory predicts that each ecologically distinct group will eventually diverge into its own sequence cluster for any gene in the genome (Cohan, 1994a, 1994b, 1995, 1996; Palys et al., 1997; Majewski & Cohan, 1999). One caveat is that the theory’s expectation of eventual divergence does not necessarily predict the state of divergence now at any particular gene locus. The divergence level depends on the age of the taxa as well as the rate of evolution of the gene in question.

Apparently the rate of evolution for protein-coding genes is usually high enough to distinguish ecologically distinct populations even when they are very closely related. Palys et al. (1997) found a nearly complete correspondence between sequence clusters for protein-
coding genes and ecologically distinct taxa: out of dozens of pairs of closely related but ecologically distinct taxa surveyed (including species and subspecies comparisons), only one pair of subspecies failed to be distinguished as a separate sequence cluster for protein-coding genes. Moreover, several ecologically distinct populations have been discovered only because they formed a sequence cluster separate from those of known taxa, based on protein-coding genes (Baranton et al., 1992; Palys et al., 1997).

In contrast, 16S rRNA sequence data have frequently shown very few differences between closely related species (Palys et al., 1997), such that 16S rRNA sequence data alone would not suggest the presence of multiple species. The possibility that the unusually slow rate of 16S evolution might be responsible has been addressed directly only for the present Bacillus case and for the case of M. avium and M. intracellulare. Like the Bacillus case, these Mycobacterium species have extremely similar 16S rRNA sequences, but they are easily distinguished as separate sequence clusters for two protein-coding genes (Soini et al., 1994; Kapur et al., 1995; Ros & Belak, 1996; Palys et al., 1997). The Mycobacterium and Bacillus examples thus demonstrate the existence of taxa for which enough time has transpired for clear divergence at protein-coding genes but not for the 16S rRNA gene. More generally, it is suggested that the limited ability of 16S rRNA to distinguish other species is a consequence of the extremely slow rate of 16S rRNA evolution. While theory predicts ecologically distinct groups will diverge eventually into separate clusters for any gene, in practice only protein-coding genes (and perhaps intertranscribed sequences) appear to evolve rapidly enough to reliably distinguish and allow discovery of closely related taxa.

Protein-coding sequences complement the value of DNA reassociation data for demarcation. Whereas theory predicts a correspondence between ecologically distinct populations and sequence clusters, there is no clear theoretical rationale for using the DNA reassociation criterion (Palys et al., 1997). Moreover, protein-coding sequences have distinguished taxa that could not be distinguished by the DNA reassociation criterion (Palys et al., 1997). Sequence clustering of protein-coding genes thus provides a criterion that is both sensitive and theoretically sound for discovering ecologically distinct groups, and classifying them into separate species.

The International Committee on Systematic Bacteriology Subcommittee on the Taxonomy of the Genus Bacillus (International Committee on Systematic Bacteriology, 1993) recently established a collaborative polyphasic taxonomic study of the genus Bacillus, with a view toward establishing minimum standards for the description of species and revising the nomenclature of the group. Methods used in the polyphasic studies have included SDS-PAGE of whole-cell protein, whole-cell fatty acid analysis, whole-cell pyrolysis mass spectrometry, DNA relatedness evaluation, DNA base composition analysis, amplified rDNA-restriction analysis (ARDRA), 16S rRNA gene sequencing and phenotypic (biochemical, morphological and physiological) characterization studies. In view of its demonstrated effectiveness, sequence analysis of protein-coding genes could be added to the arsenal of techniques used for characterizing and establishing Bacillus species.

ACKNOWLEDGEMENTS

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


