Relationship of 16S rRNA sequence similarity to DNA hybridization in prokaryotes

Jyoti Keswani† and William B. Whitman

INTRODUCTION

Because of the importance of DNA hybridization and 16S rRNA sequence similarity in prokaryotic systematics, the relationship between them is of great interest. Devereux et al. (1990) proposed that, if the extent of DNA hybridization was considered equivalent to the sequence similarity, then \( \log S = K \log D \), where \( S \) was the sequence similarity of the 16S rRNA,

\[ D \] was the extent of DNA hybridization and \( K \) was a constant. Consistent with this assumption, a very significant correlation was found between \( \log S \) and \( \log D \) (Devereux et al., 1990). However, this study was limited by the small number of values compared and because most of the values for \( S \) were estimated from \( S_{ab} \) values. While subsequent investigations have also found correlations between \( S \) and \( D \) (Stackebrandt & Goebel, 1994; Hauben et al., 1997, 1999), the relationship between \( \log S \) and \( \log D \) has not been further tested. Therefore, the goal of the current study was to examine the relationship between \( S \) and \( D \) more fully.

Although highly correlated to phenotypic similarity (cf. Colwell, 1970; Colwell et al., 1974; Stackebrandt & Goebel, 1994), the precise chemical meaning of the

**Keywords**: DNA hybridization, 16S rRNA, cophenetic correlation

---

**Abbreviations**: \( D \), extent of DNA hybridization; \( K \), ratio of \( \log S/\log D \); \( r \), correlation coefficient; \( S \), 16S rRNA sequence similarity; UPGMA, unweighted pair group method with arithmetic averages.

---

[† Present address: Department of Biochemistry, West Virginia University, Morgantown, WV 26506, USA.]

**Authors for correspondence**: William B. Whitman. Tel: +1 706 542 4219. Fax: +1 706 542 2674. e-mail: whitman@arches.uga.edu
**Table 1** Sources of 16S rRNA sequences and DNA hybridization data

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Genus</th>
<th>rRNA sequence</th>
<th>DNA hybridization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hickman-Brenner et al. (1988)</td>
</tr>
<tr>
<td>Bacillus</td>
<td>Bacillus</td>
<td>Ash et al. (1991)</td>
<td>Fritze et al. (1990)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rössler et al. (1991)</td>
<td>Kaneko et al. (1978)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Seki et al. (1978)</td>
</tr>
<tr>
<td>Bdellovibrio</td>
<td>Bdellovibrio</td>
<td>Donze et al. (1991)</td>
<td>Seidler et al. (1972)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Farrow et al. (1983)</td>
</tr>
<tr>
<td>Fusobacterium</td>
<td>Fusobacterium</td>
<td>Lawson et al. (1991)</td>
<td>Dzink et al. (1990)</td>
</tr>
<tr>
<td>Leuconostoc</td>
<td>Lactobacillus</td>
<td>Martinez-Murcia &amp; Collins (1990)</td>
<td>Farro et al. (1989)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Shaw &amp; Harding (1989)</td>
</tr>
<tr>
<td>Methanobacteriaceae</td>
<td>Methanobacterium</td>
<td>Lechner et al. (1985)</td>
<td>Kotelnikova et al. (1993)</td>
</tr>
<tr>
<td></td>
<td>Methanothermobacter</td>
<td>Nölling et al. (1993)</td>
<td>Patel et al. (1990)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Östergaard et al. (1987)</td>
<td>Touzel et al. (1992)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Zellner et al. (1989)</td>
</tr>
<tr>
<td>Methanococcales</td>
<td>‘Methanocaldococcus’</td>
<td>Burggraf et al. (1990)</td>
<td>Burggraf et al. (1990)</td>
</tr>
<tr>
<td></td>
<td>‘Methanothrix’</td>
<td>Keswani et al. (1996)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methanomicrobiaceae</td>
<td>Methanocorpusculum</td>
<td>Rouviere et al. (1992)</td>
<td>Xun et al. (1989)</td>
</tr>
<tr>
<td></td>
<td>Methanoculleus</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Methanogenium</td>
<td>Zellner et al. (1999)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Methanofollis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methanosarcinaceae</td>
<td>Methanococoides</td>
<td>Rouviere et al. (1992)</td>
<td>Boone et al. (1993)</td>
</tr>
<tr>
<td></td>
<td>Methanolobus</td>
<td></td>
<td>Maestrojuan et al. (1992)</td>
</tr>
<tr>
<td></td>
<td>Methanosarcina</td>
<td></td>
<td>Sowers et al. (1984)</td>
</tr>
<tr>
<td>Methanosarcinaceae  (halophilic)</td>
<td>Methanohalobium</td>
<td>Rouviere et al. (1992)</td>
<td>Boone et al. (1993)</td>
</tr>
<tr>
<td></td>
<td>‘Methanosalsus’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pasteurellaceae</td>
<td>Actinobacillus</td>
<td>Dewhirst et al. (1992)</td>
<td>Borr et al. (1991)</td>
</tr>
<tr>
<td></td>
<td>Haemophilus</td>
<td></td>
<td>Coykendall et al. (1983)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Potts &amp; Berry (1983)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Potts et al. (1986)</td>
</tr>
<tr>
<td>Pseudonocardiacae</td>
<td>Amycolata</td>
<td>Bowen et al. (1989)</td>
<td>Bowen et al. (1989)</td>
</tr>
<tr>
<td></td>
<td>Faenia</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kibdelosporangium</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pseudonocardia</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Saccharomonospora</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Saccharopolyspora</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Saccharothrix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spirochaeta</td>
<td>Borrelia</td>
<td>Marconi &amp; Garon (1992)</td>
<td>Johnson et al. (1987)</td>
</tr>
<tr>
<td></td>
<td>Serpula</td>
<td>Stanton et al. (1991)</td>
<td>Miao et al. (1978)</td>
</tr>
</tbody>
</table>
extent of DNA hybridization is ambiguous. It is the change in the melting temperature of the hybrids formed in a DNA reassociation experiment and not the extent of hybridization that is linearly related to sequence similarity, and a change of 1 °C in the melting temperature of the hybrids is equal to a 1.7% change in the sequence similarity (Caccone et al., 1988). The nature of the DNA which does not form hybrids is also not measured. Additional ambiguity results from the fact that reciprocal measurements of the extent of DNA hybridization differ by 4–8% (Sneath, 1989). Thus, when the hybridization experiment is performed by labelling DNA from one organism and comparing the extent of hybridization between homologous and heterologous DNA, the observed value depends to some extent just upon the choice of which DNA is labelled. Lastly, a number of the commonly employed methods of DNA hybridization may give somewhat different results. For instance, the S1 nuclease method is reported to yield 15–20% lower hybridization values than the hydroxyapatite and membrane filter methods (Grimont et al., 1980). In spite of these difficulties in determination and interpretation of the DNA hybridization, it is one of the three criteria used in defining prokaryotic species (Wayne et al., 1987).

### METHODS

**Sources of 16S rRNA sequences.** Published 16S rRNA sequences found in the literature before 1994 for which DNA hybridization values could also be identified for the same strains were compiled (Table 1). In addition, sequences determined in part in our laboratory but published later were included. Partial sequences of 500 bp or less were excluded. Sequences were accessed through the GenBank database.

**Determination of S.** Sequences belonging to the same genus or closely related genera were aligned by the program PALIUP contained in Genetics Computer Group (GCG) Sequence Analysis package v. 7.01 (Devereux et al., 1984) on a VAX computer. Any position containing undetermined nucleotides in the aligned sequences were excluded from the subsequent analysis. This alignment was used in the DISTANCE program of GCG for computing the similarity matrices (uncorrected for back mutations). This procedure was used for all the sequences despite the availability of similarity matrices in some of the references to avoid inconsistencies due to the use of different alignment programs.

**Determination of cophenetic correlation coefficients.** The matrices of 16S rRNA sequence similarity values calculated above were used to calculate the cophenetic matrices with an unweighted pair group method with arithmetic averages (UPGMA) program (Sneath & Sokal, 1973). A cophenetic matrix consisted of the estimated similarity values derived from the calculation of the UPGMA tree. The cophenetic correlation coefficient for each taxon was the correlation coefficient (r) calculated from the linear regression between the corresponding values of the similarity matrix and the cophenetic matrix.

**Sources of DNA hybridization values.** DNA hybridization values were taken from published reports (Table 1) and classified according to four general methods. The ‘S1 nuclease’ method refers to those studies based upon the method of Crosa et al. (1973). For the ‘membrane filter’ method, studies using the general technique as described by Johnson (1985) but employing a variety of types of membranes were combined. The ‘hydroxyapatite’ method refers to those studies based upon the method of Brenner et al. (1969). The ‘renaturation’ method refers to those studies based upon the method of De Ley et al. (1970). When applicable, values obtained under optimal conditions, usually at 25 °C below the melting temperature, were used. Data from methods based upon dot-blot hybridizations were not included in this study.

**Data analysis.** Statistical analysis were performed with the statistical analysis package SAS.

### RESULTS

**Evaluation of the relationship of logD versus logS**

Using data obtained from 387 pairs of organisms belonging to 20 different taxa, a highly significant relationship was found between the logarithmic transformations of 16S rRNA sequence similarity (S) and the extent of DNA hybridization (D). The correlation coefficient (r) between logS and logD was 0.571, which was significant at $P < 0.0001$ (Fig. 1). The DNA hybridization values were obtained from studies using four different general methods. Because different methods have been reported to yield different
hybridization values (Grimont et al., 1980), the relationship of log D and log S was also tested for each method separately. For D values obtained by the S1 nuclease, membrane filter, hydroxyapatite and renaturation methods, the correlation coefficients of log D with log S were 0.623, 0.636, 0.834 and 0.550 for sample sizes of 188, 153, 23 and 23, respectively. These correlation coefficients were all significant at P < 0.0001.

Although the relationship between log S and log D was highly significant, the scatter of values in Fig. 1 was also high, and the amount of the variance accounted for (r²) was low. To determine the sources of variability, the ratio log S/log D (K) was examined. K was expected to be a constant. Pairs of organisms for which S = 1.0 or D = 1.0 were removed from the analysis because the logarithmic transformations of these values were zero, resulting in uninformative or nonsensical K values. The total number of such pairs was only five. In addition, the five remaining values from the genus Bdellovibrio were also removed because they were based upon 16S rRNA sequences of less than 1000 positions. Reducing the data set to 377 values had little consequence on the correlation coefficient (data not shown), and K ± 1 sd was 0.030 ± 0.25.

**Ultrametric properties of S**

If the ultrametric properties of the 16S rRNA from a specific taxon were poor, S might no longer reflect the phylogeny of the taxon (Sneath, 1993). In this case, K would no longer measure the biological relationships between D and S but would be artefactual. If this reasoning was correct, the inclusion of these taxa would be expected to increase the variability of K. The cophenetic correlation, which is a correlation between the S value calculated during tree-building and the observed S (see above), tests the ultrametric properties of the sequences (Sneath & Sokal, 1973). For the taxa Aeromonas, Fusobacterium, Pasteurellaceae, Serratia, Methanothermobacter and halophilic Methanosarcinaceae, the cophenetic correlation was below 0.85 and lower than found for the other taxa (Table 2 and data not shown). To test whether taxa with low cophenetic correlations were a source of variability of K, these taxa, excepting the halophilic Methanosarcinaceae, were removed from the data set. For the halophilic Methanosarcinaceae, visual inspection of the UPGMA matrix of the observed and estimated S indicated that ‘Methanosalsus zilliniæ’ was the source of the low cophenetic correlation, and removal of the 16S rRNA sequence of this species increased the cophenetic correlation to 0.90. Therefore, only the two K values that involved ‘Methanosalsus zilliniæ’ were removed from the data set. This analysis indicated that many of the extreme K values were derived from taxa with low cophenetic correlations. For instance, the sd of K decreased when the nonultrametric taxa were removed, and K ± sd was 0.029 ± 0.018, n = 262 (where n is the number of values tested). Thus, the non-ultrametric properties of the 16S rRNA gene for some taxa were a major source of variability of K.

Inspection of the K values suggested that taxa with low cophenetic correlations contributed greatest to the variability when D was also high. There are a number of biological reasons why S might prove a poor indicator of evolutionary relationships for closely related taxa (Sneath, 1993; Keswani et al., 1996). To examine this effect quantitatively, the set of values where D was determined by the S1 nuclease method was divided into two groups with either D ≤ 0.40 or D > 0.40. D = 0.40 was chosen to divide the groups in order to obtain a sufficient number of values with a high D. The values of K ± sd were 0.028 ± 0.017 (n = 142) and 0.040 ± 0.054 (n = 45) when D was ≤ 0.40 and > 0.40, respectively. When taxa with low cophenetic correlations were removed, the K ± sd were 0.028 ± 0.019 (n = 81) and 0.012 ± 0.009 (n = 18), respectively. Thus, when D was ≤ 0.40, the presence of taxa with a low cophenetic correlation had little effect on the mean and sd of K. Removal of taxa with a low cophenetic correlation strongly affected the mean and sd of K only when D was also > 0.40. However, if the cophenetic correlation was inversely related with the mean S or sd of K, removal of taxa with low cophenetic correlations might also decrease the sd of K or bias the mean value of K. However, this did not appear to be
Correlation of RNA similarity with DNA relatedness

Table 2 Variation of the relationship between \( S \) and \( D \) for different taxa

<table>
<thead>
<tr>
<th>Representative genera</th>
<th>Slope ± SE*</th>
<th>Intercept ± SE*</th>
<th>( n )</th>
<th>Cophenetic correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.53 ± 0.02</td>
<td>2.20 ± 0.08</td>
<td>327</td>
<td>NA</td>
</tr>
<tr>
<td>Aeromonas</td>
<td>0.01 ± 0.23</td>
<td>0.19 ± 0.87</td>
<td>9</td>
<td>0.65</td>
</tr>
<tr>
<td>Bacillus</td>
<td>0.68 ± 0.20</td>
<td>2.93 ± 0.65</td>
<td>40</td>
<td>0.89</td>
</tr>
<tr>
<td>Enterococcus</td>
<td>0.18 ± 0.21</td>
<td>0.82 ± 0.73</td>
<td>21</td>
<td>0.95</td>
</tr>
<tr>
<td>Fusobacterium</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>0.72</td>
</tr>
<tr>
<td>Leuconostoc</td>
<td>0.28 ± 0.19</td>
<td>1.53 ± 0.64</td>
<td>43</td>
<td>0.99</td>
</tr>
<tr>
<td>Listeria</td>
<td>0.68 ± 0.22</td>
<td>3.17 ± 0.81</td>
<td>11</td>
<td>0.99</td>
</tr>
<tr>
<td>Methanobacteriaceae</td>
<td>1.08 ± 0.21</td>
<td>4.89 ± 0.80</td>
<td>15</td>
<td>1.00</td>
</tr>
<tr>
<td>Methanococcales</td>
<td>0.55 ± 0.19</td>
<td>2.16 ± 0.65</td>
<td>24</td>
<td>0.94</td>
</tr>
<tr>
<td>Methanomicrobacteriace</td>
<td>0.30 ± 0.19</td>
<td>1.55 ± 0.64</td>
<td>17</td>
<td>0.99</td>
</tr>
<tr>
<td>Methanosarcinaceae</td>
<td>0.35 ± 0.33</td>
<td>1.59 ± 0.99</td>
<td>6</td>
<td>0.80</td>
</tr>
<tr>
<td>Methanosarcinaceae</td>
<td>0.73 ± 0.21</td>
<td>3.19 ± 0.69</td>
<td>15</td>
<td>0.99</td>
</tr>
<tr>
<td>Pasteurellaceae</td>
<td>−0.47 ± 0.22</td>
<td>−2.72 ± 0.72</td>
<td>38</td>
<td>0.80</td>
</tr>
<tr>
<td>Propionibacterium</td>
<td>−0.04 ± 0.35</td>
<td>3.85 ± 1.11</td>
<td>6</td>
<td>0.87</td>
</tr>
<tr>
<td>Pseudomonocardiaae</td>
<td>−0.68 ± 1.37</td>
<td>−0.70 ± 3.19</td>
<td>15</td>
<td>0.93</td>
</tr>
<tr>
<td>Serratia</td>
<td>0.49 ± 0.26</td>
<td>1.75 ± 0.89</td>
<td>29</td>
<td>0.65</td>
</tr>
<tr>
<td>Spirochaeta</td>
<td>0.55 ± 0.20</td>
<td>2.13 ± 0.68</td>
<td>9</td>
<td>1.00</td>
</tr>
<tr>
<td>Streptococcus</td>
<td>0.93 ± 0.19</td>
<td>3.52 ± 0.61</td>
<td>28</td>
<td>0.91</td>
</tr>
</tbody>
</table>

*From the analysis of covariance with genera. Slopes and intercepts are for the linear regression equation \( \ln(-\ln D) = m[\ln(-\ln S)] + b \), where \( m \) is the slope and \( b \) is the intercept.

The correlations between the mean \( S \) or the \( SD \) of \( K \) of taxa with their cophenetic correlations were poor, 0.226 (\( n = 20 \)) and 0.260, respectively. Therefore, the variability of \( K \) associated with taxa with low cophenetic correlations was only associated with those members with a high \( D \). Visual inspection of \( K \) from taxa with a low cophenetic correlation further indicated that most of the extreme values were associated with taxa where \( D \) was greater than 0.70. Other methods to identify specific sequences contributing to the low cophenetic correlations were unsuccessful. For instance, exclusion of taxa represented by sequences with high Cook’s distance or where the difference between the calculated and actual similarity values were greater than 2 SDs above the mean did not increase the cophenetic correlations.

Nonlinearity of the relationship between \( \log D \) and \( \log S \)

Upon the removal of taxa with a low cophenetic correlation, the presence of a systematic variation in \( K \) with \( \log S \) (or \( S \)) was detected. Using the nearly complete data set described above (\( n = 377 \)), the correlation coefficient of \( K \) with \( \log S \) was 0.421. Although highly significant, this value was less than the correlation coefficient between \( \log D \) and \( \log S \) and appeared to support a simple linear relationship. However, comparison with three other sets of data suggested that the relationship between \( \log D \) and \( \log S \) could be better described by more complex functions.

Set 1 consisted of data where \( D \) was determined by either the SI nuclease or membrane-binding methods (\( n = 338 \); see below). Set 2 consisted of the same data with the values from taxa with low cophenetic correlations removed (\( n = 246 \)). Set 3 consisted of the same data as set 1 but with only those values from taxa with low cophenetic correlations where \( D \) was also greater than 0.70 removed (\( n = 327 \)). For these three sets, the correlation coefficients for \( K \) versus \( \log S \) were 0.408, 0.769 and 0.711, respectively (Fig. 2 and data not shown). These correlation coefficients were all highly significant with \( P < 0.0001 \). Very similar results were also obtained when \( S \) was replaced by evolutionary distance to correct for back mutations (data not shown). In contrast, the correlation coefficients of \( K \) with \( \log D \) (or \( D \)) were less than 0.15 within each of the three data sets, indicating that no significant relationship existed between these parameters (data not shown). Presumably, the systematic variation of \( K \) with \( S \) existed because, in contrast to the initial assumptions, \( D \) was a poor surrogate for sequence information.

In the absence of a theoretical relationship between \( D \) and \( S \), it was still possible to look for an empirical relationship that might be useful to predict \( D \) from \( S \). To further explore this possibility, alternative functions were examined for data set 3. Of the functions tested, only a complementary log-log transformation \( [\ln(-\ln D) \text{ versus } \ln(-\ln S)] \) produced a higher correlation coefficient, 0.789 with data set 3, than that found with the \( K \) versus \( \log S \) plot (Fig. 2). Although the correlation coefficients of the two plots...
were comparable, the residuals for the $K$ versus log$S$ plot were significantly skewed ($P < 0.001$). In contrast, the residuals of the complementary log log plot were not skewed and appeared normally distributed (data not shown). For these reasons, further evaluation of the relationship between $D$ and $S$ was performed with the complementary log log transformation.

**Effect of method of measuring $D$**

To determine if the method of measurement of $D$ affected the observed relationship between $D$ and $S$, values obtained by each of the four methods (with the values removed where the taxon had a low cophenetic correlation and $D$ was also $> 0.70$) were examined. For values obtained by the S1 nuclease, membrane filter, hydroxyapatite and renaturation methods, the correlation coefficients of $\ln(-\ln D)$ with $\ln(-\ln S)$ were $0.770$, $0.734$, $0.890$ and $0.629$ for sample sizes of $178$, $149$, $23$ and $16$, respectively. These correlation coefficients were all significant at $P < 0.0001$ except for the renaturation method, which was significant at $P < 0.01$. The slopes for the four methods were $0.514$, $0.454$, $0.844$ and $0.316$, respectively, but only the slope from the hydroxyapatite method was significantly different ($P < 0.05$). However, when the residuals were examined by one-way analysis of variance, the values from both the renaturation and hydroxyapatite methods differed significantly ($P < 0.001$) from values of the S1 nuclease and membrane methods. In an analysis of covariance, these latter methods also appeared significantly different ($P < 0.05$). However, the contribution of the method to the variance was very small, and the correlation coefficient only increased from $0.789$ to $0.793$ when method was included as a variable in data set 3. These analyses indicated that combining the values from the S1 nuclease and membrane methods to form data set 3 was justified. However, it was not possible to conclude that the renaturation and hydroxyapatite methods necessarily affected the observed relationship between $D$ and $S$. Because of their small sample sizes, the values obtained by these methods represented only a few taxa. Because the relationship between $D$ and $S$ varied between taxa (see below), it was not possible to assign the observed differences solely to differences in methodology.

**Taxa-specific effects on the relationship between $D$ and $S$**

In an analysis of covariance, the correlation coefficient for data set 3 increased to $0.884$ when taxon was included as a variable, and the effect of taxon was highly significant ($P < 0.0001$). Similar results were also obtained when the values in data set 3 obtained by either the S1 nuclease or membrane methods were analysed separately. These results suggested that the relationship between $S$ and $D$ was not constant across prokaryotic groups. Moreover, even for taxa within the same subphylum or other higher-order group, the relationship varied significantly (Table 2). For instance, within the low G+C group of the Gram-positive bacteria, the slopes of the linear regression ($\pm$se) for the taxa represented by *Bacillus*, *Enterococcus*, *Leuconostoc*, *Listeria* and *Streptococcus* did not shown. For these reasons, further evaluation of

![Fig. 2. Nonlinearity of the relationship between log$S$ and log$D$.](image)

(a) Correlation between log$S$/log$D$ ($K$) and log$S$. The line was determined by linear regression of the 327 values in data set 3. (b) Complementary log log plot of the same data. (c) The complementary log log plots for representative taxa. For clarity, the $\ln(-\ln D)$ values for *Bacillus* (□) and *Leuconostoc* (△) were offset by $-1$ and $+1$, respectively. The values for *Methanococcales* (●) were unchanged. The expected se estimated from the error in reciprocal determinations of $D$ and from the use of $S$ to estimate the evolutionary distance are indicated at 0.99 $S$ [ln$(-\ln S) = -46$] and 0.93 $S$ [ln$(-\ln S) = -2.6$] by the rectangles with broken outlines. One se is shown.
were $0.68 \pm 0.20$, $0.18 \pm 0.21$, $0.28 \pm 0.19$, $0.68 \pm 0.22$ and $0.93 \pm 0.19$, respectively, and these values were significantly different ($P < 0.05$). Similar variation was observed within the representatives of the $\gamma$-subclass of the Proteobacteria (Aeromonas, Pasteurellaceae and Serratia) and the Methanosarcinaceae (Table 2).

In contrast, analysis of covariance indicated no significant difference between members of the bacterial and archaeal domains. For representatives of the bacterial and archaeal groups, the slopes $\pm$SE were $0.50 \pm 0.05$ ($n = 250$) and $0.56 \pm 0.04$ ($n = 77$), respectively, which were not significantly different at $P > 0.1$. Even though the archaea were represented mostly by mesophilic methanogens, these results suggested that the rates of relative change of $S$ and $D$ were about the same for the representatives of both domains.

**Prediction of $D$ given $S$**

The $r^2$ for the analysis of covariance between genera was 0.78, which suggested that this analysis had accounted for most of the variance. The magnitude of the expected error was estimated for comparison. The inherent statistical error for estimating evolutionary distance from $S$ was taken from Kimura (1983): $\sigma = \sqrt{[1 - S]/S.N}$, where $\sigma$ is the standard error and $N$ is the number of positions determined (assumed to be 1200). This standard error does not include error expected from inaccuracies in the actual sequences or due to differences between multiple copies of the rRNA gene in the same genome and would be expected to underestimate the total error. Similarly, the SD of reciprocal measurements of $D$ is about 5% (W. B. Whitman, unpublished). This factor also probably underestimates the total error because it neglects the error from repeated measures of $D$. Nevertheless, the expected ranges of values produced by these effects at $S = 0.99$ and 0.93 in complementary log log plots were consistent with observed values (Fig. 2c) and the conclusion that these plots predicted most of the variability in the relationship between $D$ and $S$ when differences between taxa were also taken into consideration.

The data set was also examined for additional biases. For most of the taxa, $D$ values were reported prior to measuring $S$. If $D$ was high, it seemed likely that many investigators would not sequence the 16S rRNA, and $S$ would be under reported for high values of $D$. At high $S$, this bias might lower the observed distribution of $D$. Tests were performed for this potential bias. In the first test, the distribution of residuals was compared at $S \geq 0.985$ and $S < 0.985$. At high $S$, about two-thirds of the residuals were positive, as might be expected if values with a high $D$ were under represented. In the second test, log log complementarity plots were examined for high values of $S$. For the data with $0.999 > S > 0.970$ ($n = 105$), the plot was not significantly different ($P > 0.05$) from the plot of the entire data set 3 or of those values where $0.99 > S > 0.90$ ($n = 250$). In conclusion, while the data set appeared biased at high $S$, this bias did not appear strong enough to affect the usefulness of the observed relationship.

Because the residuals of the complementary log log plots were normally distributed, it was possible to estimate the distribution of $D$ given $S$ (Fig. 3). Important characteristics of this distribution were the precipitous decline in $D$ when $S$ was near 1.0 and a much slower decline in $D$ at low values of $S$. Thus, it would also be expected that, given an $S$ of 0.998, $D$ would be $< 0.70$ about 50% of the time. Given an $S$ of 0.992 or 0.986, $D$ would be $< 0.70$ about 95% or 99% of the time, respectively. Moreover, at $S = 0.963$, $D$ would be expected to be $\leq 0.20$ about 50% of the time. At $S = 0.929$, $D$ would be expected to be $\leq 0.10$. To test whether or not the observed distribution was dependent on the type of curve chosen to represent the relationship between $D$ and $S$, the distribution was also estimated from the plot of $K$ vs log$S$. Because the residuals of this plot were not normally distributed (see above), the expectations of 50, 95 and 99% were taken...
Because the relationship of \( D \) to \( S \) varied significantly between genera, a more precise estimate of \( D \) might be made for individual taxa if the complementary log log plots were known for each taxon. To test this possibility, the distribution of \( D \) was estimated for three taxa, *Bacillus*, *Leuconostoc* and *Methanococcales*. These taxa were chosen because they each were represented by data from a wide range of \( S \). For *Bacillus* and *Leuconostoc*, the slopes were significantly greater and smaller, respectively, than the slopes for the entire data set used in Fig. 3. The slope for *Methanococcales* was similar to that of the entire data set. As judged from the breadth of the distribution, use of the taxon-specific regression coefficients produced more precise estimates of \( D \) (Fig. 4). For instance, for *Bacillus* and *Methanococcales*, given an \( S \) of 0·997, 0·993 or 0·991, \( D \) would be expected to be < 0·70 about 50, 95 or 99% of the time, respectively. For *Leuconostoc*, given an \( S \) of 0·999, \( D \) would be expected to < 0·50, < 0·58 or < 0·61 about 50, 95 or 99% of the time.

To evaluate the effect of sample size on the taxon-specific complementary log log plots, slopes were calculated for random selections of data. For *Leuconostoc* and *Methanococcales*, which were characterized by an even distribution of measurements over a range of \( D \), the relative sd of the observed slopes for random selections of 10 and 20 measurements were 16 and 8%, respectively (data not shown). In contrast, for *Bacillus*, where most of the data were obtained at low values of \( D \), the relative sd for 10 and 20 randomly selected measurements were 81 and 4%, respectively. Very similar results were also obtained when the intercepts were examined. Therefore, both sample size and the distribution of measurements played important roles in determining the complementary log log plots with a high precision.

**DISCUSSION**

Although this rather statistical analysis of the relationship between \( D \) and \( S \) might appear ‘dry’, it allows for some interesting conclusions to be formed. First, there exists a very significant relationship between \( D \) and \( S \). However, this relationship is obscured by the presence of taxa with rRNA sequences with nonultrametric properties, by variation in the relationship between taxa, by differences in the methods for determining \( D \), and by deviation from the ideal model for this relationship. When these factors are controlled for, it is possible to account for 78% of the variability of \( D \) given \( S \). Moreover, the experimental error associated with \( D \) and the inherent statistical error in using \( S \) to estimate evolutionary distance are of sufficient magnitude to account for much of the remaining variability. These results imply that it is possible to predict \( D \) from \( S \) with a known precision. Given the relative ease in determining \( S \) by automated sequencing, the ability to estimate \( D \) will be of great utility for systematic studies. For organisms that have never been isolated but have been detected in natural

---

**Fig. 4.** Probability of \( D \) when \( S \) is known for the specific taxa *Leuconostoc* (a), *Methanococcales* (b) and *Bacillus* (c). The numbers next to the lines indicate the likelihood that \( D \) is equal to or below the indicated value. The distributions of \( D \) were calculated from the equation \( \ln(-\ln D) = m\ln(-\ln S) + b \) and the sd of the residuals in \( D \), where \( m \) is the slope and \( b \) is the y intercept were as reported in Table 2. The sd of the residuals were, 0·1402, 0·2516 and 0·3383 for *Leuconostoc*, *Methanococcales* and *Bacillus*, respectively.

---

as the median, the 95% and the 99% highest residuals, respectively (data not shown). From this analysis, given an \( S \) of 0·996, 0·986 or 0·982, \( D \) would be expected to be < 0·70 about 50, 95 or 99% of the time, respectively. These estimates are in good agreement with those obtained from the complementary log log plots.
samples by rRNA sequence alone, the ability to estimate \( D \) will provide a clearer understanding of their genetic and phenotypic diversity.

About one-third of the taxa included in this study possess rRNA with nonultrametric properties that confound the estimation of \( D \) from \( S \). These taxa all share high values of \( D \) and much lower values of \( S \) than expected. Sneath (1993) proposed that the nonultrametric properties of the rRNA of Aeromonas spp. were due to intergeneric homologous recombination. Homologous recombination appears to be widespread among certain prokaryotic taxa (Dykhuizen & Green, 1991; Smith et al., 1993), so this explanation may be more general. Alternatively, Keswani et al. (1996) proposed that the allelic variability expected in the large populations typical of prokaryotes might also contribute to the nonultrametric properties of rRNA. Lastly, sequences which are wrong might also prove nonultrametric.

For taxa where rRNA possess nonultrametric properties, alternative phylogenetic methods may be more appropriate. For the sequences examined in this study, the nonultrametric properties arose from comparisons of very closely related strains where \( D > 0.70 \). At these high levels of relatedness, classical taxonomic methods may be the methods of choice for determining relatedness. Even when nonultrametric rRNA sequences are excluded, \( S \) is expected to be very high, about 0.998, when \( D \) is \( > 0.70 \). This result implies that many taxa defined on the basis of \( S > 0.97 \) may be underspeciated (Stackebrandt & Goebel, 1994). Moreover, at these high levels of \( S \), sequencing errors and gene heterogeneity may have a major effect (Clayton et al., 1995). Together, these factors imply that rRNA sequencing is not a robust method for determining evolutionary relationships at the species level.

These considerations lead to some practical suggestions for systematic studies. First, the ultrametric properties of rRNA should be routinely tested. Second, in taxa where the rRNA sequences do not possess nonultrametric properties and the taxon-specific complementary log log plot is not known, \( S \) of \( \leq 0.986 \) provides evidence for different genospecies with a high level of confidence. If the taxon-specific complementary log log plot is known, a higher boundary for \( S \) will probably pertain. Third, \( S > 0.986 \) does not provide evidence for a single genospecies, and \( S \) cannot be used to confidently place organisms in the same genospecies under any circumstances. Fourth, in taxa where the rRNA sequences possess nonultrametric properties, \( S \) should not be used to make systematic decisions regarding the presence or absence of genospecies.

The relationship between \( D \) and \( S \) is complex, and \( D \) is a poor surrogate for sequence similarity. For closely related organisms, \( D \) changes much more rapidly than for distantly related organisms. For instance, within the range \( 1.0 > S > 0.95 \), \( D \) decreases from 1.0 to 0.15. Within the range \( 0.95 > S > 0.90 \), \( D \) decreases only from 0.15 to 0.06. Because \( D \) is such a poor indicator of relationships among distantly related organisms, ranges of \( D \) assigned for intergeneric relationships should be interpreted cautiously (Johnson, 1984). In addition, \( S \) is expected to be strongly related to the DNA sequence similarity of genomes, at least in so far as genomic similarity can be represented by a single gene. This implies that the relationship between \( D \) and the genomic sequence similarity is also likely to be complex and that \( D \) is unlikely to provide reliable, quantitative information about the genomic sequence similarity. Given that the genomic DNA sequence similarity is the reference standard to determine phylogeny and taxonomy in prokaryotes (Wayne et al., 1987), the usefulness of \( D \) in taxonomy would appear to be limited.

The relationship between \( S \) and \( D \) also varied significantly between taxa. While variability in the rates of evolution among rRNA genes from diverse groups of organisms has been noted previously (Woese, 1987), there is no fundamental reason to assume that the rate of change of \( D \) is a constant. Thus, differences between taxa cannot, a priori, be attributed to changes in the rates of evolution of either \( D \) or \( S \). In complementary log log plots, taxa with a high rate of change of \( S \) relative to \( D \) would be detected by a low slope (Table 2). Examples of such taxa include Aeromonas, Enterococcus, Leuconostoc and Methanomicrobiales. Taxa where \( S \) changes slowly relative to \( D \) include Methanobacteriaceae, Propionibacterium, Streptococcus, Pasteurellaceae and Methanosarcinaceae. However, attribution of these differences to changes in \( D \), \( S \) or both \( D \) and \( S \) must await further study. Whatever the cause, these observations argue that \( S \) and \( D \) should be interpreted flexibly when making taxonomic decisions.

In contrast, the relative rates of evolution of \( S \) and \( D \) were the same in the domains Bacteria and Archaea. These results are consistent with changes observed in the rates at the genus or family level provided that the variability within each domain is high. Thus, \( S \) is a ‘quasi-molecular evolutionary clock’ that is approximately constant when averaging effects and stochastic factors are taken into account (Kimura et al., 1989). Similarly, Ochman & Wilson (1987) argued that the mean substitution rate for 16S rRNA was nearly the same for prokaryotes, plants and mammals based on calibrations of branching events in the 16S rRNA phylogenetic tree with specific ecological events. Importantly, the similarity of the relative rates of change of \( D \) and \( S \) suggests that similar evolutionary processes are occurring in the modern populations of both domains and provides strong support for the construction of a universal phylogenetic tree within the prokaryotes based upon this gene.

**ACKNOWLEDGEMENTS**

The authors are grateful to Wendy Smith and Stephen Rathbun of the Statistics Department of the University of Georgia for assistance with the statistical analyses and
suggesting the complementary log log plots, and to Brian Tindall for helpful discussions. This work was supported in part by NSF grant MCB-0084164.

REFERENCES


Gharbia, S. E. & Shah, H. N. (1989). Glutamate dehydrogenase and 2-oxoglutarate reductase electrophoretic patterns and deoxyribonucleic acid-deoxyribonucleic acid hybridization


Correlation of RNA similarity with DNA relatedness.
Int J Syst Bacteriol


