INTERNATIONAL JOURNAL OF SYSTEMATIC BACTERIOLOGY, Jan. 1977, p. 44-57 Vol. 27, No. 1
 Printed in U.S.A.

Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Procaryotic Systematics

GEORGE E. FOX, KENNETH R. PECHMAN, AND CARL R. WOESE

Department of Genetics and Development, University of Illinois, Urbana, Illinois 61801

A taxonomy for *Bacillus subtilis*, *B. megaterium*, *B. cereus*, *B. pumilus*, *B. pasteurii*, *B. stearothermophilus*, and *Sporosarcina ureae* has been constructed from comparisons of T_1 ribonuclease digests of their respective 16S ribosomal ribonucleic acids. This molecular approach to systematics is shown to give results in essential agreement with traditional techniques for this group of organisms. In addition, the technique appears well suited for higher order classification, an area which has been difficult to approach with traditional techniques.

A prime objective of bacterial systematics certainly must be the establishment of a classification that spans the procaryote kingdom. However, classification by traditional techniques has been difficult, especially beyond the level of genera, reflecting the relative simplicity and/or antiquity of these organisms. It has been proposed that this problem might be resolved by integrating existing taxonomic studies where "suitable" overlap occurs (6). A direct experimental approach powerful enough to reveal the more distant relationships would certainly be more desirable, however.

Such an approach can be based on a molecular characterization of what Zuckerkandl and Pauling (24) call the "semantides"—informational macromolecules. Two considerations necessitate care. Not only are bacterial origins much more ancient than metazoan origins, but the genomes in the former case have a characteristically higher mutation rate (3). Consequently, one might expect an unmanageable degree of divergence in primary structures of any given semantide across the spectrum of procaryotes, again precluding higher order classification. Furthermore, bacterial genetic transfer mechanisms hold the potential for reticulate evolution (8), which could make for an unmanageable complexity.

It is herein contended that these difficulties can be avoided by primary structural characterization of ribosomal ribonucleic acids (rRNAs). The ribosome is clearly of very ancient origin and is necessarily ubiquitous. At least two of its RNA components (5S and 16S rRNA) apparently have functionally equivalent forms over a wide range of procaryotes, since both of these molecules can form functional chimera ribosomes in vitro (1, 9, 21). The primary structures of these rRNA molecules are sufficiently constrained that on the whole they have not changed rapidly in time. Moreover, they contain regions of both extreme conservation (4, 19) and hypervariablity (16, 19) so that both distant and close relationships can be examined. Since three distinct rRNA molecules exist, one can ultimately examine the extent to which a particular classification is dependent on the choice of RNA molecule. Finally the large number of ribosomal components (18), whose genes are not all necessarily contiguous (7), argues that the ribosomal system would not be readily transferred genetically from one organism to another.

In the present study the 16S rRNA molecule was selected since its ≈1,600 nucleotides provide a more reliable classification than the small 5S rRNA (≈120 nucleotides), and experimentally it is more manageable than the larger 23S rRNA (≈3,300 nucleotides). Determination of 16S rRNA primary structure is sufficiently difficult at present that full sequence information cannot be used as a basis for phylogenetic study. However, a partial sequence characterization, in terms of a "comparative cataloging" approach, is both feasible and useful (11, 12, 20, 23). Herein this approach is applied to a group of bacilli to display the method and examine its potential for phylogenetic and general taxonomic classification.

In essence the method involves digesting the 16S rRNA from each organism under investigation with an endonuclease of restricted specificity and determining the sequence of the resulting oligonucleotides. This produces a unique catalog of digestion products from each organism. These catalogs are then intercompared by two essentially independent approaches. In the first, families of homologous digestion products are identified which allow direct examination

of nucleotide replacements which have been introduced during the course of evolution. In the second, each individual digestion product is treated as a taxonomic character, with a dendrogram being produced by essentially normal taxonomic procedures.

MATERIALS AND METHODS

Strains. The strains included in this study were as follows: *Bacillus subtilis* 168; *Bacillus megaterium* KM; a *Bacillus cereus* strain obtained from P. Starr, University of Illinois; *Bacillus pumilus*, isolated and characterized by B. J. Lewis, University of Illinois, a characterization confirmed by R. Gordon, Rutgers University (personal communication); *Bacillus pasteurii*, ATCC 11859; *Bacillus stearothermophilus* strain 10 (10); and *Sporosarcina ureae*, ATCC 6473.

Growth, labeling, and RNA isolation. *B. subtilis*, *B. megaterium*, *B. pumilus*, and *B. cereus* were generally grown aerobically at 37°C in a yeast extract-peptone medium (pH adjusted to 7.0) and labeled with $^{32}PO_4$ in a dephosphorylated version of that medium (15). *B. pasteurii* and *S. ureae* were grown and labeled in the same medium with 0.3 M urea (final concentration) added and the pH adjusted to 8.5 (11). *B. stearothermophilus* was grown in a peptone medium at 60°C (10).

For labeling RNA, $^{32}PO_4$ (New England Nuclear Corp.) was added to each culture (10 to 30 cm^3) in early log phase at a final concentration of 0.5 to 1.0 mCi/ml. After three to four divisions, cells were collected by centrifugation, washed, and opened by passage through a French pressure cell at 15,000 lb/in^2. The rRNA was then phenol-extracted and the 16S rRNA was separated by polyacrylamide gel electrophoresis, with final purification obtained by passage over a Whatman CF-11 cellulose column (K. J. Pechman, Ph.D. thesis, University of Illinois, Urbana, Ill., 1975).

Determination of oligonucleotide catalog. Each purified 16S rRNA was digested with T_1 ribonuclease, and the resulting oligonucleotides were separated and sequenced by the two-dimensional paper electrophoresis technique of Sanger and co-workers (13) with certain modifications introduced in this laboratory (17, 18a).

In brief, the analysis involves electrophoresis of the digest on cellulose acetate at pH 3.5 (in the presence of 7 M urea), followed by transfer of the resulting oligonucleotide pattern to diethylaminoethyl-cellulose paper, with subsequent orthogonal electrophoresis in one of two buffer systems (17, 18a). The resulting oligonucleotide "fingerprint" comprises a pattern of "isopleths." Each isopleth grouping has a characteristic uridylic acid (U) content per oligonucleotide, i.e., the U = 0 isopleth, the U = 1 isopleth, and so on. Within each isopleth, oligonucleotides are displayed by order of size, and within each size "isocline" they are separated to some extent on the basis of composition. Figure 1 is a representative fingerprint of this type.

Sequences of oligonucleotides on this fingerprint, the primary pattern, are determined (where not obvious from their positions) by various "secondary" and "tertiary" enzymatic digestions which produce a sufficient number and kind of recognizable products that the sequence of the original oligonucleotide can be deduced therefrom (17, 18a).

Data analysis. To identify families of related oligomers, a systematic search was conducted with the aid of an IBM 360/75 electronic computer. In this procedure all nonidentical sequences in any two catalogs, A and B, were examined. Every sequence in A was individually aligned so as to maximize similarity with each sequence in B. A matrix resulted wherein each row represented a particular sequence in catalog A and each column represented one in catalog B. Every entry in the matrix thus represented the number of residues in common between two specific sequences when they were aligned in the most favorable way. Next, each row in the matrix was scanned for an absolute maximum, and if, when identified, that entry was also an absolute maximum in its respective column, the two sequences were presumed to be corresponding related sequences. Finally, the column and row corresponding to the two newly identified homologous sequences were deleted and the process was continued.

This procedure was most powerful when applied to a group of related organisms, since the various binary comparisons tended to bear upon one another. For example, corresponding oligonucleotides from two distantly related species were often undetectable because of "noise" (usually tie scores in a column or row). However, if any one of the alternative possibilities was unequivocally identified in a comparison involving a third organsim, more closely related to one of the original organisms, then it was justifiable to define this relationship to be the correct one in the more distant comparison as well. Such subjective aspects of the analysis required considerable care during application if reliable results were to be obtained.

In a reasonably compact group of organisms, such as the bacilli, this approach identified a significant portion, but not all, of the corresponding related sequences between any two catalogs. In general then, families of related oligomers were identified, many of which had representatives from each organism under study. These families could be used in their entirety to produce individual binary comparisons. However, it alleviated any question as to a proper normalization procedure if only complete families were utilized so that the individual binary comparisons could all be based on counts of the variations at equivalent nucleotide positions. This information could then be used, in principle at least, to calculate a true phylogenetic tree.

Alternatively, an appropriate association coefficient could be defined as follows: $S_{AB} = 2N_{AB}/(N_A + N_B)$, where N_A = total number of residues represented by oligomers of at least length L in catalog A, N_B = total number of residues represented by oligomers of at least length L in catalog B, and N_{AB} = total number of residues represented by all the coincident oligomers between the two catalogs, A and B, of at least length L. For a reasonably compact group

FIG. 1. *Representative 16S rRNA oligonucleotide fingerprint.*

TABLE 1. *Oligonucleotide catalogs*[a]

Oligonucleotide	*B. subtilis*	*B. pumilus*	*B. megaterium*	*B. cereus*	*B. pasteurii*	*S. ureae*	*B. stearothermophilus*	*S. inulinus*
5-mers								
CCCCG	0	0	0	0	0	0	1	−
CCCAG	1	1	1	1	1	1	1	−
CCACG	1	1-2	1	1[b]	1	?[c]	0	+
(C, C, AC)G	0	0	0	0	0	1	0	−
CACCG	0	0	0	1	0	?	0	+
ACCCG	1	0	0	1	1	1	0	−
CAACG	2	1-2	2	2	1	1	1	+
ACACG	1	1	1	1	1	1	1	+
AACCG	0	1	0	0	0	0	0	−
ACAAG	0	0	0	0	0	0	1	+
AACAG	1	1	0	0	0	0	0	−
AAAAG	0	0	1	1	0	0	0	−
UCCCG	1	1	1	1	1	1	1	+
CCCUG	0	0	0	0	1	0	0	−
CUCAG	1	1	1	1	1	1	1	+
CUACG	0	1	0	0	0	0	0	−
ACCUG	1	1	2	1	2-1	2	1[b]	+
AUCCG	0	0	0	1	0	0	0	−
CUAAG	1	1	2	1	1	1	1[b]	−
UACAG	0	0	0	1	1-2	1	0	−
UACCG	1	?	0	1/2	1	1	1	−
UAACG	1	0	1	1	1	0	1	+
CAUAG	0	0	1	0	0	0	0	+
CAAUG	1	1	1	1	0	1	1	−
AUCAG	1	1	1	1	1	1	1	+
ACAUG	0	0	0	0	0	1	0	−
AAUCG	1	1	1	1	1	1	1	−
AACUG	0	0	1	0	0	1	0	−
UAAAG	2	2-1	1	1	2	1-2	1	++
AUAAG	0	0	0	0	0	0	1	−
AAUAG	1	0	0	0	0	0	0	−
AAAUG	1	1	1	1	1	1	1	++
CCUUG	1	1	1	1	0	0	2-1	+
CUCUG	1	1	0	0	1	1	2-1	+
(UC)UCG	1	0	0	0	0	0	0	−
CUUCG	?	0	0	1	0	0	0	−
UCAUG	0	0	0	0	0	1	0	−
CUAUG	0	0	0	0	1	0	0	−
UAUCG	1	1-2	1	1	1	1	0	+
ACUUG	0	0	0	1	0	1	1	−
UUAAG	2	1	1	1-2	1	1	1	+
UAAUG	0	1	0	0	1	1	0	+
UAUAG	0	0	0	0	1	0	0	−
AUUAG	1	1	2	1	1	1	1	+
AUAUG	0	0	0	1	0	0	0	−
AAUUG	1	1	1	1	1	1	1	++
UUUCG	0	1	1[b]	0	0	0	0	+
UCUUG	1	1	1	1	2	2	1	+
CUUUG	0	0	0	0	0	0	1	−
UUUAG	0	1	0	0	0	0	0	−
AUUUG	0	0	0	0	0	0	1	−

TABLE 1—*Continued*

Oligonucleotide	*B. subtilis*	*B. pumilus*	*B. megaterium*	*B. cereus*	*B. pasteurii*	*S. ureae*	*B. stearothermophilus*	*S. inulinus*
6-mers								
CCCCCG	1	1	0	0	0	0	0	−
CCCACG	0	0	1	1	1	1	?	−
(C, C, C, AC)G	0	0	0	0	0	0	1	−
AACCCG	0	0	0	0	0	0	1	−
CACAAG	1	1	1	1	1	1	1	+
CAAACG	0	0	0	0	1	0	0	−
AAACCG	1	1	1	1	1	1	1	+
AACAAG	1	1	1	1	1	1	0	−
UCCACG	1	1	1	1	0	1	1	−
UAACCG	0	0	1	0	0	0	0	−
CUAACG	1	1	1	0	0	1	1	−
CCUAAG	0	0	1	1	0	0	0	−
CAACUG	0	0	0	0	0	0	1	−
ACACUG	0	0	0	1	0	1	0	−
UAAACG	1	1	1	1	1	1	1	+
AUACAG	0	0	0	0	0	1	0	−
AAUACG	1	1	1	1	1	1	1	−
AAACUG	1	1	1	1	1	1	1	+
AAUAAG	0	0	0	1	0	0	0	−
UUCCCG	1	1	1	1	1	1	1	+
CCUUCG	0	0	0	0	0	0	1	−
UCCAUG	0	0	0	0	1	0	0	+
UCACUG	0	0	0	0	1	1	0	−
AUUCCG	0	0	0	0	0	1	0	−
AUCCUG	1	1	1	1	1	1	1	+
UUCAAG	0	1	0	0	0	0	0	−
CAUUAG	1	1	1	1	1	1	1	+
UAAUCG	1	1	1	1	1	1	1	+
UAACUG	1	1	1	1	1	1	0	−
AACUUG	1	0	1	0	0	0	0	−
AUUAAG	0	0	0	1	0	0	0	−
UUUCCG	1	1	1	1	1	1	0	−
CUUUCG	0	0	?	1	0	0	0	−
(CU)UUCG	0	0	1	0	0	0	0	−
CCUUUG	0	0	0	0	0	0	1	−
UCAUUG	1	1	1	1	1	1	1	+
UUUUCG	1	1	0	0	1	1	0	+
CUUUUG	0	1[b]	0	0	0	0	1	−
7-mers								
CAAACAG	0	0	1	1[b]	1	1	1	+
AACAAAG	1	1	0	0	0	0	0	−
CACUCCG	1	1	1	1	1	1	1	−
CAACUCG	1	1	1	1	1	1	1	+
CAACCUG	0	0	1	0	1	1	1	+
UAACACG	1	1	1	1	1	1	1	+
UACAAAG	0	0	1	1	0	0	1	+
UAAAAAG	0	0	0	0	0	1[b]	0	−

TABLE 1—*Continued*

Oligonucleotide	*B. subtilis*	*B. pumilus*	*B. megaterium*	*B. cereus*	*B. pasteurii*	*S. ureae*	*B. stearothermophilus*	*S. inulinus*
CUCCCUG	1	0	0	0	0	0	0	−
CCUCUAG	0	0	0	0	0	0	1	−
CUCUCAG	0	0	0	0	0	0	1	−
UUCCCAG	0	0	0	0	1	1	0	−
CACCUUG	0	1	0	1	0	0	0	−
CAUUCAG	1	0	?	0	1	0	1	−
UAACCUG	1	1	0	1	0	0	0	−
ACCUUAG	0	0	0	0	1	0	0	−
AACUCUG	0	0	0	0	1	0	0	−
CAUUAAG	1	1	1	1	1	1	0	+
UAAUACG	1	1	1	1	1	1	1	+
AAACUUG	0	1	0	0	1	0	0	−
CUCUCUG	0	0	0	0	0	0	1	+
UUCUCAG	1	1	1	1	0	0	0	−
UCACUUG	0	0	0	0	0	0	1	−
UUAUCCG	0	0	1	1	0	0	0	−
UACCUUG	1	0	1	0	1	1	0	−
CAUUUAG	0	1	1	0	0	0	0	−
AUCUUAG	1	1	1	1	0	0	0	+
CUUUCUG	0	0	0	0	1	1	0	−
(U_5, C)G	0	0	1	0	0	0	0	−
8-mers								
CCAACCCG	0	0	0	0	1	0	0	−
CAAACCCG	0	0	0	0	0	1	0	−
ACAACCCG	0	0	0	0	0	0	1	−
ACAAACCG	1[b]	1	1[b]	1	1	1[b]	0	+
AACACCAG	1	1	1[b]	1[b]	1	1	1	−
CUCAACCG	1	1	1	1[b]	0	1	1	+
CCACACUG	1	1	1	1	1	1	1	−
ACAUCCCG	0	0	0	0	1	0	0	−
CCCCUUAG	1[b]	1	0	0	1	1	0	−
CCCUUCAG	0	0	0	0	0	1	0	−
CCUACAUG	0	0	1[b]	1	0	0	0	−
CUUAACCG	0	0	0	0	1	0	0	−
CACUCUAG	0	0	0	0	0	0	1	−
AUACCCUG	1	1	1	1	1	1	1	+
CUACAAUG	1	1	1[b]	1	1	1	1	+
CCCUUUAG	0	0	1[b]	1	0	0	0	−
ACUCUCUG	1	1	?	0	0	0	0	−
$A_{1\text{-}0}$CUCUCUG	0	0	1	0	0	0	0	−
UUCCUUCG	0	0	0	0	1	0	0	−
CUCUUAUG	0	0	1	1	0	0	0	−
ACUUUCUG	0	0	0	1	0	0	0	−
AAUUAUUG	1	1	1	1	1	1	1	+
9-mers								
CAACCCUCG	0	0	0	0	0	0	1	−
CUCACCAAG	1[b]	1	1[b]	1	0	1	1	−
CUACACACG	1	1	1	1	1	1	1	+
UACACACCG	1	1	1	1	1	1	1	+
UAACACCCG	1	1	1[b]	1	1	1	0	+

TABLE 1—*Continued*

Oligonucleotide	*B. subtilis*	*B. pumilus*	*B. megaterium*	*B. cereus*	*B. pasteurii*	*S. ureae*	*B. stearothermophilus*	*S. inulinus*
ACAUCCCAG	0	0	0	0	0	1	0	−
ACCAAAUCG	0	0	0	0	0	0	1	−
AUAACACCG	0	0	0	0	0	0	1	−
CAACCCUUG	1	1	1[b]	1	1[b]	0	0	+
UACCUCACG	0	0	0	0	0	0	1	−
ACUCCUACG	1	1	1	1[b]	1	1	1	−
CUUACCAAG	0	0	0	0	1	0	0	−
CUAACUACG	1	1	1	1	1	1	1	+
CUAAUACCG	1	1	1	1	1	1	1	+
CACUCUAAG	1[b]	1	1	1	1	0	0	+
AUAACUCCG	1	1	0	1	1	1	1	−
AAUUCCACG	1	1	1	1	1	1	1	+
AAUAAUCAG	0	0	0	0	1[b]	0	0	−
UCCCCUUCG	1	0	0	0	0	0	0	−
C(C_{1-0}CU)CUUAG	0	0	1	0	0	0	0	−
CCCCUUAUG	1	1	1	1	1	1	1	+
CAUCCUCUG	0	0	1[b]	0	0	0	0	−
CACUCUAUG	0	0	0	0	0	1[b]	0	−
UCCCUUAAG	0	0	0	0	0	0	1	−
AAUCUUCCG	1[b]	1	1[b]	1	0	0	1	−
AUAACUUCG	0	0	1[b]	0	0	0	0	+
AAUCUCAUG	0	0	0	1	0	0	0	−
UUCCCUUCG	0	0	0	0	0	0	1[b]	+
CCUUUUAAG	0	0	0	0	1[b]	1[b]	0	−
UCACUUAUG	0	0	0	1	0	0	0	−
UUUCUUAAG	1	1	1[b]	1	0	0	0	−
UUUAAUUCG	1	1	1	1	1	1	1	+
≥10-mers								
ACAACCCAAG	0	0	0	0	0	0	1	−
ACAACCCUAG	0	1	0	1	0	0	0	−
AAACUCAAAG	1	1	1	1	1	1	1	+
ACAUCCCCUG	0	0	0	0	0	0	1	−
A(C,A)ACUCUAG	0	0	1	0	0	0	0	−
ACAAUCCUAG	1	0	0	0	0	0	0	−
UAAAACUCUG	0	0	1	1	0	0	0	−
AAAUUCAAAG	0	0	0	1	0	0	0	−
ACAUCCUCUG	1[b]	1	1[b]	1	0	0	0	+
UCACUUACAG	0	1	0	0	0	0	0	−
CUUCCCUUCG	0	0	1[b]	1[b]	0	0	0	−
UUU(CU)CUUUG	0	0	0	0	1	0	0	−
UACCUAACCAG	1[b]	1	1	1	0	0	0	−
AACCUUACCAG	1	1	1	1	1	1	1	+
CCUAAUACAUG	1	1	1[b]	1	1	1	1	+
CCAUCAUUCAG	0	0	0	0	0	1	0	−
UACCUUACCAG	0	0	0	0	0	1[b]	0	−
CCAUCAUUAAG	0	0	0	1	0	0	0	−
($C_{\sim 5}$, U_4)AG	0	0	0	0	0	1	0	−
UACCUCAUUAG	0	0	0	0	1	0	0	−
CUUUCCCUUCG	0	1	0	0	0	0	0	−
UAACCCUUUUG	0	0	0	0	1	0	0	−

TABLE 1—*Continued*

Oligonucleotide	*B. subtilis*	*B. pumilus*	*B. megaterium*	*B. cereus*	*B. pasteurii*	*S. ureae*	*B. stearothermophilus*	*S. inulinus*
UAACCCUUAG	0	0	0	0	0	1	0	
UAACCUUUAUG	0	1	0	0	0	0	0	−
$UAAC_{1-2}CUUUUAG$	1	0	0	0	0	0	0	−
AUAACAUUUUG	0	0	0	1[b]	0	0	0	−
UAAC(CU)UUUUG	0	0	0	1	0	0	0	−
AAUCCCAAAAAG	0	0	0	0	0	0	1	−
CAACCCUUAAUG	0	0	0	0	0	1[b]	0	−
AAUCUUCCACAAUG	0	0	0	0	1	1	0	−
UCACACCCUUUAG	0	0	0	0	0	0	1[b]	−
UCAAAUCAUCAUG	1	1	1	1	*[e]	1	1	+
(AU, C_x, U_y, CUUA, CUA)AG	0	0	1	0	0	0	0	−
(U_3, $C_{\sim 4}$)UCUAUG	0	0	1	0	0	0	0	−
$AUUUC_{0-1}UCCCUUCG$	0	0	0	0	0	1	0	−
CCAAUCCCACAAAAUCG	0	0	0	0	0	1	0	−
$UUCAAAC_{0-1}CAUAAAAG$	1[b]	0	0	0	0	0	0	−
(AUAA, UCCA)U(CU, CUU)CG	0	0	0	0	0	1	0	−
(C, C, AAU, C, C, C, AC, AAAU, C, U)G	1	0	0	0	0	0	0	−
CUAAUCCCAUAAAACCG	0	0	0	0	1	0	0	−
CCAAUCCCAUAAAUCUG	0	1	0	0	0	0	0	−
CUAAUCUCAUAAAACCG	0	0	0	1[b]	0	0	0	−
AUAUACCUUUCCCUUCG	0	0	0	0	1	0	0	−
Modified oligomers[d]								
AAĠ	1	1	1	1	1	1	1	+
ȦÄG	1	1	1	1	1	1	1	+
ĠCCG	1	1	1	1	1	1	1	+
CĊCCG	1	1	1	1	1	1	1	+
ĊAACG	1	1	1	1	?	1	1	+
ȦAACAAG	1	1	1	1	1	1	1	+
UCAĊACCACG	1	1	1[b]	1	1	1	1	+
UĊAAAUCAUCAUG	0	0	0	0	1	0	0	−
Terminii								
$AUCACCUCCUUUU_{OH}$	1[b]	1[b]	1[b]	1[b]	1[b]	1[b]	1[b]	+
pUCUUAUG	0	0	0	0	0	1[b]	0	−
pUUUAUCG	1	0	0	0	0	0	0	−
pUAUUAUG	0	0	0	0	1	0	0	−
pCUUUUUG	0	0	0	0	0	0	0	+
pUUCUUUG	0	0	0	0	0	0	1	−
pUUUUUCG	0	1	1	0	0	0	0	−
pUUUAUUG	0	0	0	1	0	0	0	−

[a] Symbols: 1/0, the mole fraction of each oligonucleotide is indicated for each organism except *S. inulinus*; +/−, the presence or absence of each oligonucleotide in *S. inulinus* is indicated (the complete catalog for this organism will be presented elsewhere); 2-1/1-2, the oligonucleotide in question may be present in multiple copies (the first number indicates the more likely quantitation).

[b] The sequence given is probable but not certain.

[c] Question mark indicates that presence or absence of the oligonucleotide in question is not determined.

[d] Dot indicates base which is post-transcriptionally modified.

[e] Asterisk indicates that the identical oligonucleotide, but containing a post-transcriptionally modified base, is found in *B. pasteurii*.

such as the bacilli examined here, $L = 5$. The similarity coefficient, S_{AB}, was calculated for each individual binary comparison and the resulting similarity matrix was used to produce a dendrogram in one of the usual ways, in this study single-linkage clustering.

RESULTS

Table 1 lists the oligonucleotide catalogs for the various species of the genus *Bacillus* (11; Pechman, thesis; 18a). Table 1 also records whether any given sequence was present or absent in the catalog of *Sporolactobacillus inulinus*, a closely related but presumably "non-*Bacillus*" organism. In Table 2 the 52 sequences of universal occurrence within the *Bacillus* species examined here are shown. Table 3 summarizes the 39 families of corresponding related oligomers that were identified.

Single-linkage clustering utilizing the association coefficient, S_{AB}, defined the dendrogram shown in Fig. 2. *Escherichia coli* (17) is here included to provide perspective. The individual binary comparisons upon which this dendrogram was based are indicated in Table 4.

The closest relationship found was between *B. subtilis* and *B. pumilus*. *B. megaterium* and *B. cereus* were also found to cluster as were *B. pasteurii* and *S. ureae*. Assuming this tree can be attributed phylogenetic significance, the *B. megaterium-B. cereus* divergence from the common *Bacillus* ancestor appeared to have been considerably more recent than that of the *B. pasteurii-S. ureae* couple. Hence, the seven organisms screened could be crudely thought of as defining three groups of species of *Bacillus*—one group containing *B. subtilis*, *B. pumilus*, *B. megaterium*, and *B. cereus*; a second here represented only by *B. stearothermophilus*; and a third containing *B. pasteurii* and *S. ureae*. That all of these groups are indeed best considered divisions within the genus *Bacillus* rather than individual genera was indicated by the fact that they are each more closely related to the others than any one is to the outside "reference" organism—*Sporolactobacillus inulinus*.

Binary comparison of the individual organisms over the entire set of 27 complete families allowed a direct determination of the nucleotide replacements which were introduced during the evolution of the portions of the 16S rRNA molecule represented by these families. These numbers are displayed in Table 4. In the present case the complete families represented 239 nucleotide positions, whereas the 52 universal sequences in Table 1 represented another 364 positions. The combined total of 603 nucleotide positions is the vast majority of the total number of positions accessible by the method used herein (i.e., the average *Bacillus* catalog given here contains 746 nucleotides in digestion products of size five and larger) and is a significant portion, $\approx$ 37%, of the entire 16S rRNA sequence. Since ancestral sequences can be postulated for each family, it is in principle possible to calculate a phylogenetic tree from these families (Fox and Woese, unpublished data). Such a tree is topologically very similar to that obtained with the association coefficient employed here.

In any event, many of the families reaffirm the branching points suggested in the tree generated by the association coefficeint S_{AB}. In particular, families 1, 5, 14, 15, 22, and 32 indicate *Sporolactobacillus* to be outside the main group; families 3, 7, 11, 24, and 34 indicate the close relationship between *B. subtilis* and *B. pumilus*; families 8, 19, 20, 21, 23, and 35 emphasize the interrelatedness among *B. subtilis*, *B. pumulis*, *B. megaterium* and *B. cereus*; whereas numbers 8, 11, 19, 21, 24, 30, and 38 reflect the group comprising *B. pasteurii* and *S. ureae*. In addition the families allow some insight into the evolutionary process and how it has affected the RNAs. The incorporated mutations appear in some sense nonrandom; transition mutations appear far more readily than do transversions. Likewise, insertions or deletions of single nucleotides appear to be rare.

TABLE 2. *Universal oligomers*

AAĜ	AAACCG	ACUCCUACG
ÂÂG	UAAACG	CUAACUACG
ĜCCG	AAUACG	CUAAUACCG
CĈCCG	AAACUG	AAUUCCACG
ĈAACG[a]	UUCCCG	CCCCUUAUG
ÛAACAAG	AUCCUG	UUUAAUUCG
UCAĈACCACG	CAUUAG	
	UAAUCG	AAACUCAAAG
CCCÂG	UCAUUG	
CAACG		AACCUUACCAG
ACACG	CACUCCG	CCUAAUACAUG
UCCCG	CAACUCG	
CUCAG	UAACACG	UCAAAUCAUCAUG[b]
ACCUG	UAAUACG	AUCACCUCCUUUU$_{OH}$
CUAAG		
AUCAG	AACACCAG	
AAUCG	CCACACUG	
UAAAG	AUACCCUG	
AAAUG	CUACAAUG	
UUAAG	AAUUAUUG	
AUUAG		
AAUUG	CUACACACG	
UCUUG	UACACACCG	
CACAAG		

[a] Presence or absence not determined in *B. pasteurii*.

[b] This oligomer is post-transcriptionally modified in *B. pasteurii*—e.g., UĈAAAUCAUCAUG.

TABLE 3. *Oligonucleotide families*[a]

Families	*B. subtilis*	*B. pumilus*	*B. megaterium*	*B. cereus*	*B. pasteurii*	*S. ureae*	*B. stearothermophilus*	*S. inulinus*
Complete families								
1. AAUACG	1	1	1	1	1	1	1	−
AAUCCG	0	0	0	0	0	0	0	+
2. UAACUG	1	1	1	1	1	1	0	−
UUACUG	0	0	0	0	0	0	0	+
CAACUG	0	0	0	0	0	0	1	−
3. CAAACAG	0	0	1	1	1	1	1	+
AACAG[b]	1	1	0	0	0	0	0	−
4. AACAAG	1	1	1	1	1	1	0	−
ACAAG[b]	0	0	0	0	0	0	1	+
5. CACUCCG	1	1	1	1	1	1	1	−
CAUUCCG	0	0	0	0	0	0	0	+
6. UCCACG	1	1	1	1	0	1	1	−
UCCAUG	0	0	0	0	1	0	0	+
7. CCCCCG	1	1	0	0	0	0	0	−
CCCACG	0	0	1	1	1	1?[c]	1	−
8. UUCUCAG	1	1	1	1	0	0	0	−
UUCCCAG	0	0	0	0	1	1	0	−
CUCUCAG	0	0	0	0	0	0	1	−
9. UAACCUG	1	1	0	1	0	0	0	−
CAACCUG	0	0	1	0	1	1	1	+
10. pUAUUAUG	0	0	0	0	1	0	0	−
pUCUUAUG	0	0	0	0	0	1	0	−
pUUUAUUG	0	0	0	1	0	0	0	−
pUUUAUCG	1	0	0	0	0	0	0	−
pUUUUUCG	0	1	1	0	0	0	0	−
pUUCUUUG	0	0	0	0	0	0	1	−
pCUUUUUG	0	0	0	0	0	0	0	+
11. ACUUUCUG	0	0	0	1	0	0	0	−
ACUCUCUG	1	1	1?	0	0	0	0	−
CUCUCUG	0	0	0	0	0	0	1	+
CUUUCUG	0	0	0	0	1	1	0	−
12. ACAAACCG	1	1	1	1	1	1	0	+
ACAACCCG	0	0	0	0	0	0	1	−
13. CUCAACCG	1	1	1	1	0	1	1	+
CUUAACCG	0	0	0	0	1	0	0	−
14. AACACCAG	1	1	1	1	1	1	1	−
AAUACCAG	0	0	0	0	0	0	0	+
15. CCACACUG	1	1	1	1	1	1	1	−
CCACAUUG	0	0	0	0	0	0	0	+

TABLE 3—*Continued*

	Families	*B. subtilis*	*B. pumilus*	*B. megaterium*	*B. cereus*	*B. pasteurii*	*S. ureae*	*B. stearothermophilus*	*S. inulinus*
16.	AUAACUCCG	1	1	0	1	1	1	1	−
	AUAACUUCG	0	0	1?	0	0	0	0	+
17.	CUCACCAAG	1	1	1	1	0	1	1	−
	CUUACCAAG	0	0	0	0	1	0	0	−
	CCCACCAAG	0	0	0	0	0	0	0	+
18.	CACUCUAAG	1	1	1	1	1	0	0	+
	CACUCUAUG	0	0	0	0	0	1	0	−
	CACUCUAG	0	0	0	0	0	0	1	−
19.	UUUCUUAAG	1	1	1	1	0	0	0	−
	CUUCUUAAG	0	0	0	0	0	0	0	+
	UCCCUUAAG	0	0	0	0	0	0	1	−
	CCUUUUAAG	0	0	0	0	1	1	0	−
20.	UACCUAACCAG	1	1	1	1	0	0	0	−
	UACCUCAUUAG	0	0	0	0	1	0	0	−
	UACCUUACCAG	0	0	0	0	0	1	0	−
	UACCUCACG	0	0	0	0	0	0	1	−
21.	AAUCUUCCG	1	1	1	1	0	0	1	−
	AAUCUUCCACAAUG	0	0	0	0	1	1	0	−
22.	CCCAG/ACUCCUACG	1	1	1	1	1	1	1	−
	CCCAAACUCCUACG	0	0	0	0	0	0	0	+
23.	ACAUCCUCUG	1	1	1	1	0	0	0	+
	ACAUCCCCUG	0	0	0	0	0	0	1	−
	ACAUCCOG	0	0	0	0	1	0	0	−
	ACAUCCCAG	0	0	0	0	0	1	0	−
24.	CCCCUUAG	1	1	0	0	1	1	0	−
	CCCUUUAG	0	0	1	1	0	0	0	−
	UCCAACCCCUUAG[b]	0	0	0	0	0	0	0	+
	UCACACCCUUUAG[b]	0	0	0	0	0	0	1	−
25.	UCCCCUUCG	1	0	0	0	0	0	0	−
	UUCCCUUCG	0	0	0	0	0	0	1	+
	CUUCCCUUCG	0	0	1	1	0	0	0	−
	AUUUCUCCCUUCG	0	0	0	0	0	1	0	−
	0–1								
	CUUUCCCUUCG	0	1	0	0	0	0	0	−
	AUAUACCUUUCCCUUCG	0	0	0	0	1	0	0	−
26.	CAACCCUUG	1	1	1	1	1	0	0	+
	CAACCCUCG	0	0	0	0	0	0	1	−
	CAACCCUUAAUG	0	0	0	0	0	1	0	−
27.	UAAAACUCUG	0	0	1	1	0	0	0	−
	UAAAG/CUCUG	1	1	0	0	1	1	1	+
Incomplete families									
28.[d]	CAUUCAG	1	0	0	0	1	0	1	−
	CAUUUAG	0	1	1	0	0	0	0	−

TABLE 3—*Continued*

	Families	*B. subtilis*	*B. pumilus*	*B. megaterium*	*B. cereus*	*B. pasteurii*	*S. ureae*	*B. stearothermophilus*	*S. inulinus*
29.[d]	CCAUCAUUAAG	0	0	0	1	0	0	0	–
	CCAUCAUUCAG	0	0	0	0	0	1	0	–
	CCAUCACUUACAG	0	0	0	0	0	0	0	+
30.	UACCUUG	1	0	1	0	1	1	0	–
	CACCUUG	0	1	0	1	0	0	0	–
31.	CUAACG	1	1	1	0	0	1	1	–
	CAAACG	0	0	0	0	1	0	0	–
32.	UUUCCG	1	1	1	1	1	1	0	–
	UAUCCG	0	0	0	0	0	0	0	+
33.	CUAAUCUCAUAAAACCG	0	0	1	0	0	0	0	–
	CCAAUCCCAUAAAUCUG	0	0	0	1	0	0	0	–
	CCAAUCCCACAAAUCUG	1?	0	0	0	0	0	0	–
	CUAAUCCCAUAAAACCG	0	0	0	0	1	0	0	–
	CCAAUCCCACAAAAUCG	0	0	0	0	0	1	0	–
	CCAAUCCCAUAAAG	0	0	0	0	0	0	0	+
	AAUCCCAAAAG	0	0	0	0	0	0	1	–
34.	AACAAAG	1	1	0	0	0	0	0	–
	UACAAAG	0	0	1	1	0	0	1	+
	UAAAAAG[b]	0	0	0	0	0	1	0	–
35.	AUCUUAG	1	1	1	1	0	0	0	+
	ACCUUAG	0	0	0	0	1	0	0	
36.	ACAAUCCUAG	1	0	0	0	0	0	0	–
	ACAACUCUAG	0	0	1?	0	0	0	0	–
	ACAACCCUAG	0	1	0	1	0	0	0	–
	ACAACCCAAG	0	0	0	0	0	0	1	–
37.	UAACCUUUUAG	1	0	0	0	0	0	0	–
	UAACCUUUUUG	0	0	0	1?	0	0	0	–
	UAACCUUUAUG	0	1	0	0	0	0	0	–
	AACCUUUAUG	0	0	0	0	0	0	0	+
	UAACCCUUAG	0	0	0	0	0	1	0	–
	UAACCCUUUUG	0	0	0	0	1	0	0	–
38.	CCAACCCG	0	0	0	0	1	0	0	–
	CAAACCCG	0	0	0	0	0	1	0	–
39.	UCACUUAUG	0	0	0	1	0	0	0	–
	UCACUUACAG	0	1	0	0	0	0	0	–

[a] The notation here is similar to that in Table 1, except that tentative sequences are not indicated.
[b] Membership in family is tentative.
[c] 1?, The underlined portion of the sequence indicated is not well determined in this organism.
[d] These families may be tentatively combined to form one complete family.

DISCUSSION

The limited group of species of the genus *Bacillus* examined here serves to illustrate the potential of the comparative cataloging approach in the study of bacterial systematics. To be genuinely useful, the method must not only give reliable results in general, but it must be effective where traditional techniques have not proven adequate.

Since many of the organisms examined here have been well characterized by more conventional methods, relationships herein defined can be compared to existing classifications. The arrangement of strains originally suggested by Smith et al. (14) and recently updated by Gordon et al. (5) was based on the most extensive study available. They classify the genus *Bacillus* into three major groups, based primarily on the shape of the spores and the swelling of the sporangium by the spore. Of the organisms examined here, *B. megaterium*, *B. cereus*, *B. subtilis*, and *B. pumilus* occur in their group 1, *B. stearothermophilus* occurs in group 2, and *B. pasteurii* occurs in group 3. Among the group 1 organisms, Gordon et al. (5) further suggested that *B. subtilis* and *B. pumilus* were very closely related, whereas *B. megaterium* and *B. cereus* were closer to each other than either was to any other species. These same conclusions also emerged from the present approach. In addition, our results provided strong evidence that *S. ureae*, whose classification has traditionally been difficult, should not only be classified as a member of the genus *Bacillus*, but should be placed in group 3 along with *B. pasteurii* (11).

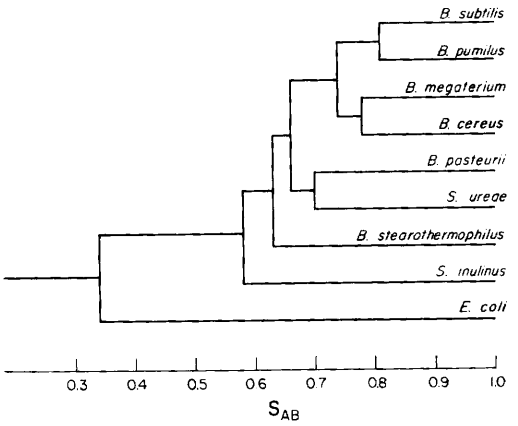

FIG. 2. *Dendrogram for strains indicated derived by single linkage clustering using a Dice type similarity coefficient, S_{AB}.*

Comparative cataloging of 16S rRNA offers several advantages over traditional methods. (i) The identical procedure can be applied to any genus, resulting in similarity coefficients which are directly comparable. Hence, data from a diversity of organisms can be readily and objectively integrated, ultimately permitting quantitative definition of terms such as "species," "genus," and "family." (ii) Exceedingly ancient divergences should be detectable, allowing identification of relationships throughout the procaryotic kingdom. That this is indeed the case has already been demonstrated by the succesful application of the comparative cataloging approach to the elucidation

TABLE 4. *Number of oligonucleotide differences—complete families*

Strain	Similarity coefficient (S_{AB}) with $L = 6$							
	B. subtilis	*B. pumilus*	*B. megaterium*	*B. cereus*	*B. pasteurii*	*S. ureae*	*B. stearothermophilus*	*S. inulinus*
B. subtilis		0.81	0.73	0.73	0.65	0.66	0.63	0.56
B. pumilus	3		0.74	0.73	0.64	0.65	0.63	0.57
B. megaterium	9	7		0.78	0.62	0.63	0.61	0.58
B. cereus	8	8	5		0.61	0.62	0.59	0.54
B. pasteurii	24	21	22	21		0.70	0.56	0.54
S. ureae	22	18	21	20	12		0.62	0.50
B. stearothermophilus	21	20	18	19	24	23		0.53

of the procaryotic nature of the chloroplast (2, 22).

Comparative cataloging is not, however, a panacea. Compared to traditional techniques, it is relatively expensive and time consuming, and it requires considerable specialized expertise. Thus it is appropriate to view comparative cataloging of 16S rRNA as an adjunct to and not a replacement for the more usual approaches.

For phylogenetic purposes it would be highly desirable to define a similarity coefficient that relates in a linear fashion the actual number of base changes between two 16S rRNA molecules. S_{AB} as defined herein is not such a coefficient. A similarity coefficient, P_{AB}, which has this property, was previously proposed (18a, 20). However, the derivation of this coefficient assumes that the probability of a mutation being accepted is the same at all positions in the 16S rRNA. This has been shown to be a nonvalid assumption (19), and therefore P_{AB} is only likely to be a useful approximation of true homology for closely related organisms. Thus, until these matters receive further attention, S_{AB}, which is readily calculated and cannot be misconstrued, seems to be the association coefficient of choice.

ACKNOWLEDGMENTS

This study was supported by NASA grant NSG-7044 and Public Health Service grant AI-6457 from the National Institute of Allergy and Infectious Diseases to C. R. W. We thank L. B. Zablen and C. D. Pribula for technical assistance in obtaining data on *Sporolactobacillus inulinus* 16S rRNA.

REPRINT REQUESTS

Address reprint requests to: Dr. C. R. Woese, Department of Genetics and Development, University of Illinois, Urbana, Ill. 61801.

LITERATURE CITED

1. **Bellemare, G., R. Vigne, and B. Jordan.** 1973. Interaction between *Escherichia coli* ribosomal proteins and 5S RNA molecules: recognition of prokaryotic 5S RNAs and rejection of eukaryotic 5S RNA. Biochimie 55:29-35.
2. **Bonen, L., and W. F. Doolittle.** 1975. On the prokaryotic nature of red algal chloroplasts. Proc. Natl. Acad. Sci. U.S.A. 72:2310-2314.
3. **Drake, J. W.** 1974. The role of mutation in microbial evolution, p. 41-58. *In* M. J. Carlile and J. J. Skehel (ed.), 24th Symp. Soc. Gen. Microbiol. Cambridge University Press, London.
4. **Fox, G. E., and C. R. Woese.** 1975. The architecture of 5S rRNA and its relation to function. J. Mol. Evol. 6:61-76.
5. **Gordon, R. E., W. C. Haynes, and C. Hor-Nay Pang.** 1973. The genus *Bacillus*. Agriculture Handbook #427, U.S. Department of Agriculture, Washington, D.C.
6. **Hill, L. R.** 1975. Interlocking numerical taxonomies. Int. J. Syst. Bacteriol. 25:245-251.
7. **Jaskunas, S. R., M. Nomura, and J. Davies.** 1974. Genetics of bacterial ribosomes, p. 333-368. *In* M. Nomura, A. Tissières, and P. Lengyel (ed.), Ribosomes. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
8. **Jones, D., and P. H. A. Sneath.** 1970. Genetic transfer and bacterial taxonomy. Bacteriol. Rev. 34:40-81.
9. **Nomura, M., P. Traub, and H. Bechmann.** 1968. Hybrid 30S ribosomal particles reconstituted from components of different bacterial origins. Nature (London) 219:793-799.
10. **Pace, B., and L. L. Campbell.** 1971. Homology of ribosomal ribonucleic acid of diverse bacterial species with *Escherichia coli* and *Bacillus stearothermophilus*. J. Bacteriol. 107:543-547.
11. **Pechman, K. J., B. J. Lewis, and C. R. Woese.** 1976. Phylogenetic status of *Sporosarcina ureae*. Int. J. Syst. Bacteriol. 26:305-310.
12. **Pechmann, K. J., and C. Woese.** 1972. Characterization of the primary structural homology between the 16S ribosomal RNAs of *Escherichia coli* and *Bacillus megaterium* by oligomer cataloging. J. Mol. Evol. 1:230-240.
13. **Sanger, F., G. G. Brownlee, and B. G. Barrell.** 1965. A two dimensional fractionation procedure for radioactive nucleotides. J. Mol. Biol. 13:373-398.
14. **Smith, N. R., R. E. Gordon, and F. E. Clark.** 1952. Aerobic spore forming bacteria. Agriculture Monograph No. 16. U.S. Department of Agriculture, Washington, D.C.
15. **Sogin, M. L., K. J. Pechman, L. Zablen, B. J. Lewis, and C. R. Woese.** 1972. Observations on the post-transcriptionally modified nucleotides in the 16S ribosomal ribonucleic acid. J. Bacteriol. 112:13-16.
16. **Sogin, S. J., M. L. Sogin, and C. R. Woese.** 1972. Phylogenetic measurement in procaryotes by primary structural characterization. J. Mol. Evol. 1:173-184.
17. **Uchida, T., L. Bonen, H. W. Schaup, B. J. Lewis, L. Zablen, and C. Woese.** 1974. The use of ribonuclease U_2 in RNA sequnce determination: some corrections in the catalog of oligomers produced by ribonuclease T_1 digestion of *Escherichia coli* 16S ribosomal RNA. J. Mol. Evol. 3:63-77.
18. **Wittman, H. G. 1976.** Structure, function, and evolution of ribosomes. Eur. J. Biochem. 61:1-13.
18a. **Woese, C., M. Sogin, D. Stahl, B. J. Lewis, and L. Bonen.** 1976. A comparison of the 16S ribosomal RNAs from mesophilic and thermophilic *Bacilli*: some modifications in the Sanger method for RNA sequencing. J. Mol. Evol. 7:197-213.
19. **Woese, C. R., G. E. Fox, L. Zablen, T. Uchida, L. Bonen, K. Pechman, B. J. Lewis, and D. Stahl.** 1975. Conservation of primary structure in 16S ribosomal RNA. Nature (London) 254:83-86.
20. **Woese, C. R., M. L. Sogin, and L. A. Sutton.** 1974. Procaryote phylogeny I: concerning the relatedness of *Aerobacter aerogenes* to *Escherichia coli*. J. Mol. Evol. 3:293-299.
21. **Wrede, P., and V. A. Erdmann.** 1973. Activities of *B. stearothermophilus* 50S ribosomes reconstituted with prokaryotic and eucaryotic 5S RNA. FEBS Lett. 33:315-319.
22. **Zablen, L. B., M. S. Kissil, C. R. Woese, and D. E. Buetow.** 1975. Phylogenetic origin of the chloroplast and prokaryotic nature of its ribosomal RNA. Proc. Natl. Acad. Sci. U.S.A. 72:2418-2422.
23. **Zablen, L., L. Bonen, R. Meyer, and C. R. Woese.** 1975. The phylogenetic status of *Pasteurella pestis*. J. Mol. Evol. 4:347-358.
24. **Zuckerkandl, E., and L. Pauling.** 1965. Molecules as documents of evolutionary history. J. Theoret. Biol. 8:357-366.