Exploring prokaryotic taxonomy

Timothy G. Lilburn and George M. Garrity

1Bioinformatics Group, American Type Culture Collection, Manassas, VA 20110, USA
2Bergey’s Manual Trust and Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA

Techniques drawn from exploratory data analysis, using tools found in the S-Plus statistical software package, have been used to inspect and maintain the Bergey’s Taxonomic Outline and to move towards an automated and community-based means of working on the outline. These techniques can be used to classify sequences from unnamed and uncultured organisms, to visualize errors in the taxonomy or in the curation of the sequences, to suggest emendations to the taxonomy or to the classification of extant species and to complement the visualization of phylogenies based on treeing methods. A dataset of more than 9200 aligned small-subunit rRNA sequences was analysed in the context of the current taxonomic outline. The use of the algorithm in exploring and modifying the taxonomy is illustrated with an example drawn from the family Comamonadaceae.

INTRODUCTION

Readers of this journal are well aware that, prior to the era of rapid and inexpensive sequencing, the outlook for developing a comprehensive and usable taxonomy of the prokaryotes was bleak. While workable schemes for restricted groups such as clinically relevant species had been developed, a universally applicable method was not available until the introduction of sequence analysis of small-subunit (SSU) rRNA (Fox et al., 1980). By the 1990s, this approach had become the principal method of establishing phylogenetic relationships among the prokaryotes; today, it is more likely that a 16S rRNA sequence will be the first piece of data collected for an unknown organism, rather than a Gram strain. From both a deterministic and an informatics perspective, this molecule has become the de facto primary key in prokaryotic systematics. Recently, there have been proposals to use a similar wide-scale, sequence-based approach (using a different sequence) to the taxonomy of the Eukaryotes as well (Hebert et al., 2003).

The legacy of 15 years of SSU rRNA sequencing is tens of thousands of sequences, some nearly full length, some short, most of high quality, and some of more dubious value. This collection is an invaluable resource for the establishment of a comprehensive prokaryotic taxonomy. There is a consensus that SSU rRNA-based phylogenies are largely consistent with the evolutionary history of the organisms, since the groups formed using this approach are often confirmed by other data, i.e. by phenotypic properties. Furthermore, while attempts to build a universal tree have revealed that different molecules produce different trees, it would appear that such trees are generally, though not absolutely, consistent with the SSU rRNA tree (Brown & Koretke, 2000). Whole-proteome comparisons have also produced trees similar to the SSU rRNA tree (Tekaia et al., 1999). Therefore, we expect that a taxonomy based on the wealth of available SSU rRNA sequences should have predictive as well as organizational value.

Until recently, very few attempts have been made to produce a comprehensive taxonomy of the prokaryotes that (i) takes advantage of the large numbers of SSU rRNA sequences available, (ii) is reconcilable with our knowledge of other genotypic and phenotypic information, and (iii) provides a link between the phylogenetic models and the nomenclatural record. We recently published our first attempts in this direction (Garrity & Lilburn, 2002), which drew heavily on techniques from exploratory data analysis (Tukey, 1977). A principal components analysis (PCA) of a matrix of evolutionary distances of >9000 full-length sequences to a set of 223 benchmark sequences revealed the structure of the higher-level relationships amongst the prokaryotes. Despite the fact that the approach reliably reduced the dimensionality of the data from 223 dimensions to 3 and allowed the generation of meaningful two-dimensional plots, the underlying cause for the placement of a given species was not immediately apparent, as PCA involves a transformation of the original data. Moreover, some distortion of visual perception was likely to occur.

In this paper, we extend our exploratory data analysis approach to include techniques that have also found useful
application in the field of microarray analysis (Eisen et al., 1998), i.e. supervised clustering and visualization via heat maps. Our results demonstrate that these graphs (which are a recent adaptation of shaded distance matrices; see, for example, Sneath & Sokal, 1973) provide an informative overview of the current taxonomy and make errors in classification obvious. The heat maps will provide a simple way of placing sequences from novel organisms in the taxonomy, thus allowing simultaneous identification and classification of organisms. We also introduce a comprehensive prokaryotic taxonomy.

METHODS

Alignment. Alignments are based on the Ribosomal Database Project II release 8.0 prokaryotic alignment (Maidak et al., 2001). The alignment includes 200 sequences from type strains that are not found in release 8.0; 1101 positions from the alignment were used, selected according to (i) the measured variability at each position and (ii) information presented in the conservation maps published by Cannone et al. (2002) and Wuyts et al. (2001).

Sequence data. Only relatively long prokaryotic sequences were used in the analyses in order to maximize the information content and to ensure that the sequences contained as many homologous positions as possible. The 9206 sequences used were more than 1399 bases long and had less than 4 % ambiguity. If sequences contained no data (‘N’s) in more than 10 consecutive alignment positions, they were eliminated from the dataset.

The data were grouped and 223 benchmark sequences incorporated as discussed previously (Garrity & Lilburn, 2002). In the benchmark set of sequences, each sequence represented, where possible, a type species and type genus on which the families are based (Garrity et al., 2002). All 25 phyla in the Bergey’s Taxonomic Outline are represented.

Estimation of evolutionary distances. Prior to estimation of evolutionary distance, subsets of sequences were created, ranging from 750 to 900 sequences in total. Each subset contained the benchmark sequences as the first 223 sequences. Matrices of evolutionary distances were calculated in PAUP* (version 4.08) (Swofford, 2000) using the Jukes–Cantor model (Jukes & Cantor, 1969). Following computation, each matrix was exported as a tab-delimited file, using a short identifier to tag each sequence.

Data structures. Matrices of evolutionary distances were imported into the statistical package S-Plus 6.1 (Insightful), edited and joined in a single data frame and finally linked to a data frame containing taxonomic and physiological information, as described previously (Garrity & Lilburn, 2002). By invoking functions that are part of S-Plus, we were able to arrange the sequence order on the axes of the matrix according to the current version of the taxonomy, based on the hierarchy of names, which were treated as ordered factor variables. Sequences without names were moved to the ends of the lists. The matrix was then colour-coded to allow data patterns to be seen. This allows the identification of any potential misplacements that arise because of incorrect annotation of sequences or because of a failure to identify synonyms.

In the next cycle, the misidentified sequences were extracted and arranged in a new matrix according to their similarities to the benchmark sequences. The unnamed benchmark sequences were also reordered.

Fig. 1. First-pass shaded distance matrix showing the distances from 9206 sequences in our dataset to the 223 benchmark sequences. Sequences have been ordered from left to right (x-axis) and from bottom to top (y-axis) according to their position in the July 2002 version of the Bergey’s taxonomy. The colour scale on the right indicates the evolutionary distance between each pair of sequences, as shown in the corresponding cell of the matrix. Labelled bars on the y-axis indicate the ranges of sequences assigned to the 25 phyla or to the group of sequences from unnamed organisms (‘Clones’) along with the number of benchmarks in each group. Colours denote the same groups on the x-axis, which also indicates the cumulative sequence numbers.
Similar routines were carried out on two subsets of the data: sequences from the class ‘Betaproteobacteria’ and from the family Comamonadaceae.

RESULTS

The initial heat map is shown in Fig. 1. Over 2 million distances are represented. Two hundred and twenty-three benchmark sequences, selected to represent the breadth of prokaryotic diversity, are on the y-axis, clustered according to their position in the July 2002 release of the Bergey’s Taxonomic Outline. At values of y above 190 are sequences that (i) are from organisms not in culture at the time of publication and (ii) form deep branches within the Ribosomal Database Project II tree. On the x-axis are 9206 sequences from Ribosomal Database Project II and GenBank, also clustered by name according to their place in the taxonomy; sequences not associated with a named organism are placed at values of x above 7463. Since the benchmarks and the bulk of the sequences are clustered according to the same hierarchy, the bright-green colour indicating the highest level of sequence similarity is seen on the diagonal between positions (0, 0) and (7463, 190). Misplaced sequences are seen as areas (rectangles) of colour that contrast with the background colour located off the diagonal and above the x co-ordinate corresponding to the misplaced sequence(s). For example, a misplaced sequence can be seen at y = 200 as a red line extending across most of the plot. Beyond co-ordinates (7463, 190), the diagonal pattern breaks down. Sequences in this region could not be positioned on the basis of known taxonomic affiliations because we had no name information. To identify and classify these unknown sequences, it was necessary to write a routine that, in essence, moved the unknown sequence along the x-axis until the element representing the highest similarity value for that sequence was placed on the diagonal of the heat-map rectangle. Fig. 2 shows the results of this type of reordering with the subset of ‘unnamed’ sequences. Note that the benchmark sequences have also been re-ordered so that the each of the 33 unnamed benchmark sequences is positioned next to the sequence on the x-axis to which it is most similar. All of the sequences have now been successfully placed in a known phylum, although it is clear from the dark lines that cross the heat map that some of the sequences are from organisms that are only distantly related to the phylum in which they have been placed. In other words, the algorithm has forced them into a taxon to which they may not actually belong. There is also a problem with the visualization of the sequence placement. In phyla with
low sequence diversity, such as the *Actinobacteria*, almost the entire block of sequences is shown in a single shade of green, and any problems of classification at subphylum levels are masked. Restricting our view of the data to a subset of the data and resetting the colour scale readily solved this difficulty. When we used this solution with the ‘*Betaproteobacteria*’, as shown in Fig. 3, we were able to visualize the taxonomy and to correct it as necessary down to family level. Resolution below family level relies, to some extent, on the simple visualization of distances, since our benchmarks were selected to represent the prokaryotic families and are therefore ineffectual below family level. In Fig. 4, the family *Comamonadaceae* is visualized along with sequences that once formed part of this family or that are proposed to form part of the family. Genera within the family can be seen (see below). In this case, the sequences have been rearranged according to their order in a neighbour-joining consensus tree.

**DISCUSSION**

The value of a comprehensive taxonomy for biologists is undisputed. It serves as an organizational framework for our knowledge about life, and reduces the possibility of confusion when discussing or researching taxa. Furthermore, such a taxonomy is predictive: when a novel organism is classified as a member of a given taxon, we can immediately make accurate and testable proposals as to the physiology of the novel organism. Although the taxonomy presented here is based on the evolutionary distance between SSU rRNA

![Fig. 3. Shaded distance matrix showing the distances between the 61 Proteobacteria benchmarks and the 392 sequences in our dataset classified as ‘Betaproteobacteria’. The x-axis shows the families within the ‘Betaproteobacteria’ and the cumulative sequence totals. The y-axis indicates the ranges of the five classes of the Proteobacteria. From top to bottom, the classes are ‘Epsilonproteobacteria’, ‘Deltaproteobacteria’, ‘Gammaproteobacteria’, ‘Betaproteobacteria’ and ‘Alphaproteobacteria’.](image-url)
sequences, it is not, strictly speaking, a phylogenetic taxonomy or classification. Rather, it represents an attempt to establish a comprehensive taxonomy using a simple approach to the sequence data. It is worth noting that the exploratory data analysis approach is not hypothesis-driven. Rather, it seeks to let the data ‘speak for themselves’ and to uncover patterns in the data. Thus, we sought to approach the SSU rRNA sequence data with as few assumptions as possible. This is why, for example, we used the simplest evolutionary model, the Jukes–Cantor model, when estimating the distances between sequences. It also explains why, leaving their graphical shortcomings aside, we did not try to force the results into a tree format. We were not attempting to establish the ‘branching order’ that traces common descent of related taxa. We simply took advantage of the breadth of the sequence data as a phenotypic trait with which to establish a taxonomy. Nevertheless, since a correlation in phenotype can be assumed to arise from a correlation in evolutionary history, evolutionary relationships among the prokaryotes can be inferred from their proximity in the plots. Even if it serves only as a working hypothesis, it will be of value as a starting point for the discussion of taxonomic issues. However, as shown below, the approach has practical utility as well as facilitating the delineation and solution of current problems in prokaryotic classification.

The application of exploratory data analysis techniques to the problem of a comprehensive prokaryotic taxonomy has proved to be quite fruitful. The two techniques adopted, PCA and shaded distance matrices, have their own advantages for constructing the taxonomy. PCA provides a three-dimensional map of the sequence space or evolutionary space defined by the dataset (Garrity & Lilburn, 2002). This structure was consistent with extant large-scale phylogenetic trees, and the ability to visualize the position of individual sequences within this three-dimensional higher-level structure proved invaluable. The map-like qualities of the PCA plots enabled us to spot misclassified organisms, poorly curated sequences and other problems or anomalies. The disadvantages of the three-dimensional view include...
occlusion of points (one can’t see the entire dataset at once) and a lack of resolution between taxonomic groups below phylum level unless further processing is done. Nevertheless, a three-dimensional point-cloud view of sequence space is quite provocative, encouraging questions about the evolutionary forces that drove the sequences to adopt the positions we see.

Heat maps can present the data as they are clustered in both extant and proposed taxonomies. Since the heat maps give a two-dimensional view, there is, in principle, no occlusion of data, and one can review the positions of all the sequences in the taxonomic hierarchy. Thus, in our matrix showing all of the sequence distances (Fig. 1), we can first see that our initial taxonomy is generally consistent. Problems are visible, such as the misplaced sequence in the ‘Clones’ benchmarks, but solutions immediately offer themselves. We can see where this misplaced sequence might belong (in the *Archaea*), and the correct placement of many of the ‘unnamed’ sequences is suggested by the colour and position of the matrix elements. The heat maps make assignment of new sequences to the appropriate taxon simple: the sequence is aligned, distances to the benchmark sequences are obtained, and the new sequence is placed in the complete distance matrix next to the sequence that has the most similar set of distances with respect to the benchmarks. We note that not all the information in the heat map is immediately visible because screen and printer resolutions lead to the presentation of multiple distances from the x-axis sequences to a single benchmark sequence as a single block of colour. If we zoom in on the matrix or create a subset of cells from the matrix, as in Fig. 2, this problem disappears.

Fig. 2 shows the distribution of the 1743 ‘unnamed’ sequences amongst the 25 phyla. This is an example of how heat maps can be used to generate speculative classifications – none of the organisms associated with the ‘unnamed’ sequences have validly published names. Fig. 2 makes the broad diversity of the benchmarks evident: dark lines representing one taxon in the *Gammaaproteobacteria* and four taxa from the *Actinobacteria* stand out. These dark lines indicate a relatively low level of similarity to all the other taxa represented in the matrix. This could mean that there is a problem with these sequences, or it may imply that these sequences are from rare, extremely under-represented lineages, or from lineages yet to be described.

Fig. 3 relates to the exploration of the taxonomy of a second subset of the sequences, those representing the ‘Beta-proteobacteria’. It is easy to see in this figure that eight of the 392 sequences from organisms classified as members of the ‘Betaproteobacteria’ are probably misclassified. Two of the first three sequences along the x-axis are probably from the ‘Alphaproteobacteria’, whereas the last five sequences appear to be most closely related to the ‘Gammaaproteobacteria’, but it is possible that they are not members of the *Proteobacteria* at all. The correct placement of these last five sequences can only be achieved if the set of benchmarks includes sequences that are related to the unknowns. This result illustrates the importance of a comprehensive selection of benchmark sequences.

The heat maps are more flexible than the PCA plots in that they can be used to visualize relationships down to the taxonomic level at which tree displays are useful, as when we seek resolution at or below the family level (which is the level at which the benchmarks were set). For example, in the visualization of the family *Comamonadaceae* shown in Fig. 4, we can see the genera within the family quite clearly, especially genera that are represented by several sequences, like *Hydrogenophaga*. It is also readily apparent that the family contains two major groups and, indeed, we note that, in the latest edition of the *Bergey’s Manual of Systematic Bacteriology* (Willems & Gillis, 2004), the genera in the smaller group (*Rubrivivax*, *Sphaerotilus*, *Ideonella* and *Aquabacterium*), which were once included in the *Comamonadaceae*, have been removed from this family. At the same time, *Leptothrix* was moved into the *Comamonadaceae*, but it is apparent from this analysis that it should be grouped with the removed genera. It would also appear that *Polaromonas* and *Brachymonas* are probably not members of the *Comamonadaceae*, but this family contains the closest sequenced relatives to these two genera. Fig. 4 also contains five species of the genus *Aquaspirillum* and *Alcaligenes latus*, which are currently not classified as members of the *Comamonadaceae*. These species more closely resemble members of the *Comamonadaceae* than the type strains of the type species of their respective genera, and the placements shown in Fig. 4 illustrate the use of heat maps for the revision of an extant taxonomy. Note also that, at the taxonomic level presented in Fig. 4, the heat maps can be used to enhance (or complement) a tree: the distances from a given taxon to all the other taxa in a tree are given by the colour of the matrix elements in that row.

The overviews provided by heat-map visualizations also invite observations on prokaryotic diversity. For example, by comparing the x-axes in Figs 1 and 2, we can compare the distribution of the ‘named’ and ‘unnamed’ sequences within the 25 phyla. The proportions of *Proteobacteria* and *Firmicutes* are the same in each distribution. In the ‘unnamed’ sequence set, the proportions of *Spirochaetes* and *Actinobacteria* are two to four times lower than in the ‘named’ sequence set, while the proportion of *Bacteroidetes* and *Cyanobacteria* are twofold greater. As mentioned above, rows or columns that are globally dark may be indicative of unusually diverse sequences and, hence, organisms. The dark lines seen within the *Proteobacteria*, *Actinobacteria* and *Dictyoglosmus* in Fig. 2 are associated with sequences that are less similar to all the other sequences in the dataset than normal, and hint at the existence of novel phyla. Similar dark lines are seen within the *Firmicutes* in Fig. 1, and are concentrated in the region of several families that historically have been hard to place. These include *Haemobartonella* and *Eperythrozoon*, two genera that appear well separated from all others in the PCA maps.

A description of the supervised clustering algorithm used to
build and refine the heat maps is outside the scope of this journal and has been submitted for publication elsewhere. The software will be available as part of a web site that will allow researchers to explore prokaryotic taxonomy and classify their own organisms on the basis of the SSU rRNA sequences. A prototype of this site will be available before the end of 2003 (http://www.msu.edu/~garrity/taxoweb/index.html).

The taxonomy developed using, in part, the approaches discussed in this paper is available from the Bergey’s Manual Trust via the World Wide Web (Garrity et al., 2002; http://dx.doi.org/10.1007/bergeysoutline). It is revised twice a year as new data and analyses become available and includes information regarding emendations of the classification along with commentary on taxa in dispute. The outline also includes information about when an organism was first described, strain designation, culture deposit information, synonymies, SSU rRNA sequence deposit information and the Ribosomal Database Project II short identifier. The latter two items are not always available, as high-quality, full-length SSU rRNA sequences are not yet available for all of the type strains. This taxonomy is perhaps the first comprehensive taxonomy of the prokaryotes, and is presented as a work in progress. Comments and suggestions are welcomed.

ACKNOWLEDGEMENTS

We thank Drs B. J. Tindall, Phillip Hugenholtz and Peter Sneath for insightful comments on the manuscript. This research was supported by the Biological and Environmental Research Program, United States Department of Energy (grant no. DE-FG02-02ER63315).

REFERENCES


