BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 10 2005, pages 2309–2314
doi:10.1093/bioinformatics/bti346

Phylogenetics

Self-organizing and self-correcting classifications of biological data

George M. Garrity[1,*] and Timothy G. Lilburn[2]

[1]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA and [2]Science Information Systems, American Type Culture Collection, Manassas, VA 20110, USA

Received on December 23, 2004; revised on January 27, 2005; accepted on February 19, 2005
Advance Access publication February 24, 2005

ABSTRACT

Motivation: Rapid, automated means of organizing biological data are required if we hope to keep abreast of the flood of data emanating from sequencing, microarray and similar high-throughput analyses. Faced with the need to validate the annotation of thousands of sequences and to generate biologically meaningful classifications based on the sequence data, we turned to statistical methods in order to automate these processes.

Results: An algorithm for automated classification based on evolutionary distance data was written in S. The algorithm was tested on a dataset of 1436 small subunit ribosomal RNA sequences and was able to classify the sequences according to an extant scheme, use statistical measurements of group membership to detect sequences that were misclassified within this scheme and produce a new classification. In this study, the use of the algorithm to address problems in prokaryotic taxonomy is discussed.

Availability: S-Plus is available from Insightful, Inc. An S-Plus implementation of the algorithm and the associated data are available at http://taxoweb.mmg.msu.edu/datasets

Contact: garrity@msu.edu

INTRODUCTION

In the face of ever-increasing amounts of information in biology, the need for automated ways for organizing this information is paramount. Sequence data and the associated annotation, microarray data, and taxonomic data could all benefit from an automated means of categorization or classification. Furthermore, a framework for assessing any such organizational tools would be of great benefit for comparing results from different approaches to classifying the data. Here, we present an algorithm that uses dynamically reordered heatmaps (Lilburn and Garrity, 2004) to visualize an automatically generated classification. The algorithm allows one to review and modify the classification, identify possible classification errors and facilitate *ad hoc* testing of alternative classifications/hypotheses. Although hierarchical clustering techniques are employed, our approach circumvents two of the major shortcomings of such methods. Our algorithm does not force-fit all the entities into a hierarchy. Rather, it provides a means of identifying and selectively excluding entities that fail to meet minimal criteria for group membership, thereby permitting 'good' clusters to form. The excluded entities are then added back to the global model to allow for a more precise placement wherever possible or recognition of a likely new group in cases where precise placement is not possible. In addition, since hierarchical clustering techniques are used only on a localized level, the effect of evolutionary rate variability across disparate entities (e.g. widely divergent taxa) is minimized, leading to classifications that are consistent with established phylogenetic models. The algorithm we describe is applicable to any classification or taxonomy for which a metric exists or can be derived.

*To whom correspondence should be addressed.

SYSTEMS AND METHODS

Software

The algorithm is implemented in the S programming language (Becker *et al*., 1988) and relies on functions that are part of the S-Plus environment (Version 6.1; Insightful, Seattle, WA). The code we had developed is available as an S-Plus script and can be implemented, with some modification, in R (Venables and Smith, 2002).

Data structure

A matrix of evolutionary distances was imported into the statistical package S-Plus 6.1 (Insightful), edited and joined in a single data frame and finally linked to a second data frame containing taxonomic information, as described previously (Garrity and Lilburn, 2002). The test dataset consists of 1436 SSU rDNA sequences that were initially classified (based on the GenBank annotation) as belonging to the *Gammaproteobacteria*, one of the five classes constituting the phylum *Proteobacteria* within the domain *Bacteria*. Sequences were obtained from the RDP-II database (Cole *et al*., 2003) and from GenBank. Only relatively long 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. All sequences used were more than 1399 bases long and had <4% ambiguities. If sequences contained no data in more than 10 consecutive alignment positions, they were eliminated from the dataset. The sequences were aligned and a sequence difference matrix was prepared as described previously (Lilburn and Garrity, 2004) except that a fully reflected distance matrix was used rather than the rectangular benchmarked matrix. The classification used was drawn from the Taxonomic Outline of the Prokaryotes, Release 3.0 (http://dx.doi.org/10.1007/bergeysoutline).

ALGORITHM

Given a matrix of distances among the items to be classified and an initial hierarchical classification, the algorithm first restructures the matrix, so that the ordering of items in the matrix matches the

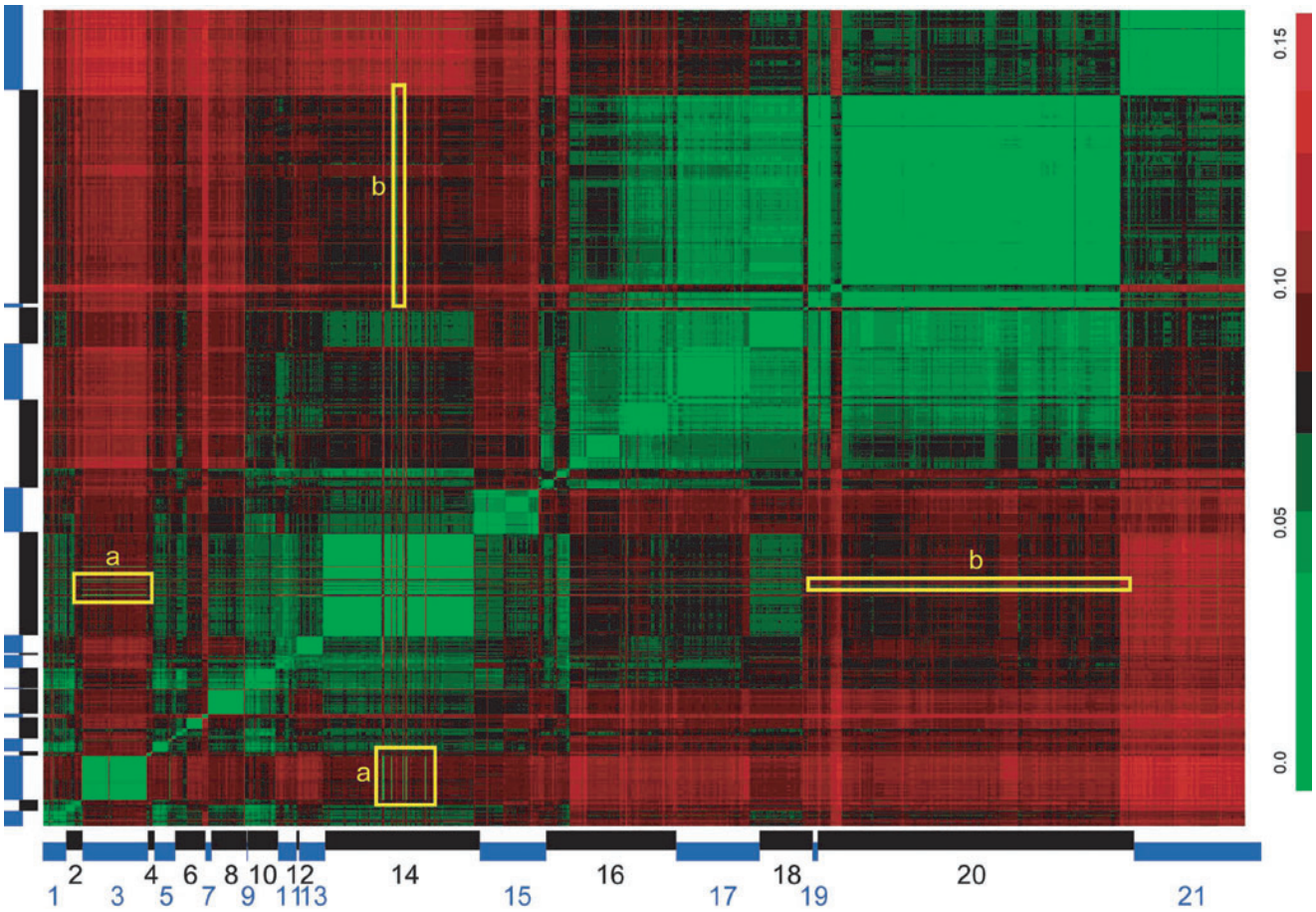

Fig. 1. Heatmap of the *Gammaproteobacteria* based on an evolutionary distance matrix with unnamed and uncorrected sequences removed. The distances, expressed as the differences per position in the sequence alignment, are encoded as indicated on the scale at the right. Ordering of the sequences within the heatmap is based on the sequence of taxa in the Taxonomic Outline of the Prokaryotes, Release 3.0. Solid bars along axis indicate the range encompassed by individual families within the heatmap. Families are as follows: *Chromatiaceae* (1), *Ectothiorhodospiraceae* (2), *Xanthomonadaceae* (3), *Cardiobacteriaceae* (4), *Thiotrichaceae* (5), *Piscirickettsiaceae* (6), *Francisellaceae* (7), *Legionellaceae* (8), *Coxiellaceae* (9), *Methylococcaceae* (10), *Oceanospirillaceae* (11), *Alcanovoraxaceae* (12), *Halomonadaceae* (13), *Pseudomonadaceae* (14), *Moraxellaceae* (15), *Alteromonadaceae* (16), *Vibrionaceae* (17), *Aeromonadaceae* (18), *Succinovibrionaceae* (19), *Enterobacteriaceae* (20) and *Pasteurellaceae* (21). The green lines within the yellow boxes (a and b) are indicative of misplaced taxa.

ordering of items in the classification. Thus, members of a given group appear in the same region of the matrix and in close proximity to other items presumed to be members of the next higher group within the hierarchy. Next, the distance from each item to its second-nearest neighbor is extracted from the matrix and the 90th percentile of this value is estimated. The 90th percentile value serves as a 'goodness-of-fit' (gof) measure and was chosen to provide a reasonable stringency, without being overly restrictive. Using this gof value as a logical test for group membership, a binary transformation of each submatrix, representing the distances among the members of a group, is then created and rearranged by hierarchical clustering along both dimensions of the matrix. These submatrices are then used to guide the global rearrangement of the complete input matrix, which may then be visualized as a colorized distance matrix or, as it is also known, a heatmap (Fig. 1).

In the next iteration, items that fail to meet the gof test (indicative of items that were misclassified in the original classification) are excluded from the analysis and the rearrangement of items is further refined using the above heuristic. On completion, the misidentified items can be added back to the 'cleaned' matrix, and placed adjacent to their nearest neighbors, based on distance rather than the presumed identity from the original classification. The classification table is then revised based on the matches, and the sorting and visualization routine is repeated using the revised classification. The process is then repeated to test for any items that fail to meet the gof criterion, in their new location. Those items that fail to meet the gof criteria are then excluded from further analysis, as they are likely members of more distantly related groups that lie outside the natural boundaries of the group being studied.

To establish an optimal ordering for the grouped items at higher levels, we used a summary statistic, the medioid, for the distance from each group to the other groups. This statistic represents a distance from the hypothetical center of each group to the hypothetical centers of all the other groups. First, we

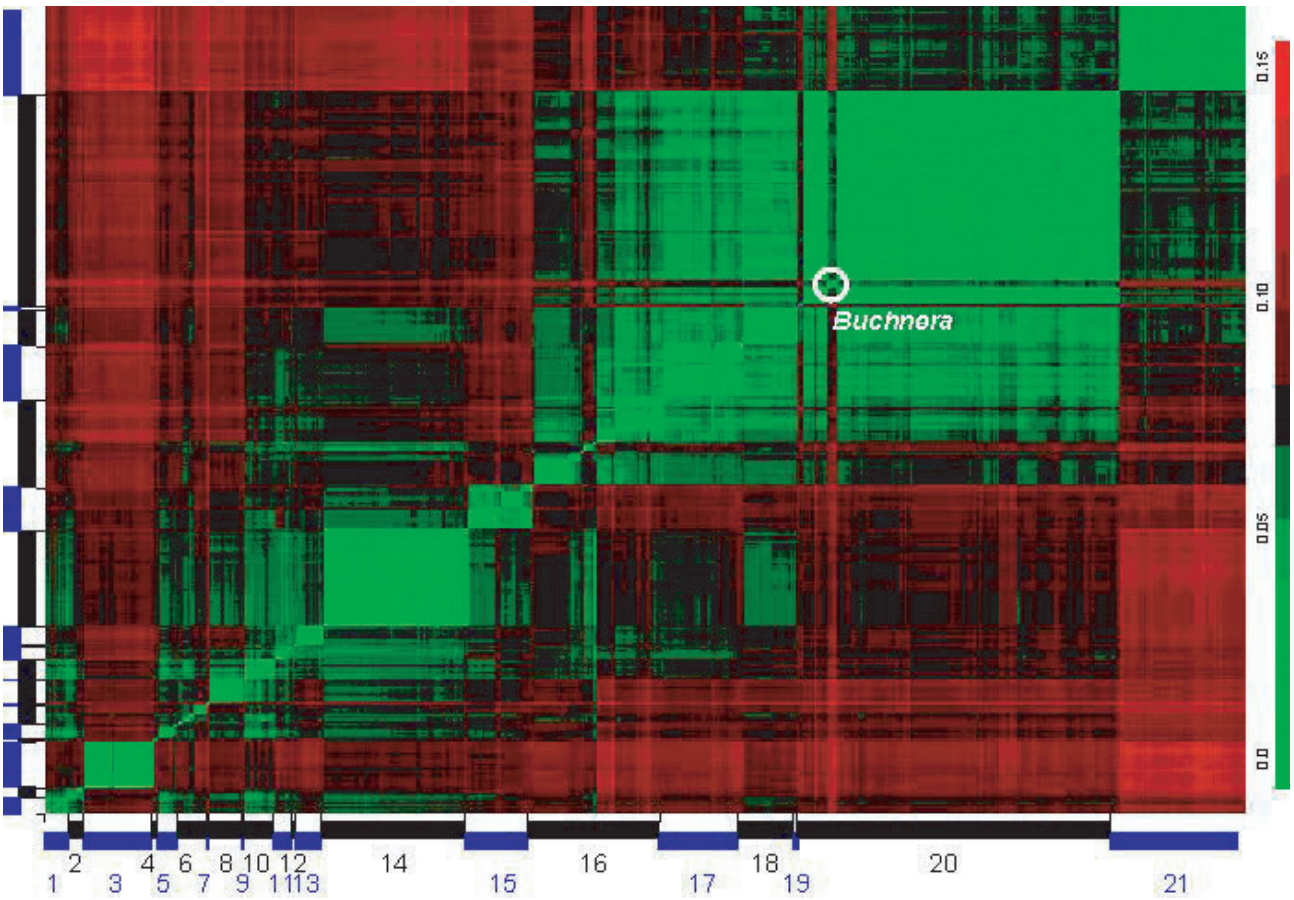

Fig. 2. Heatmap of the *Gammaproteobacteria* with unnamed and uncorrected sequences added back following supervised clustering. The ordering of the families is according to the Taxonomic Outline of the Prokaryotes and they are numbered as in Figure 1. The genus *Buchnera* is enclosed within the white circle.

estimated the column means for each group-level submatrix, effectively reducing the initial matrix from a square matrix to a rectangular matrix with the medioids on one axis and the items on the other. The matrix was then transposed and column means computed in a second iteration to yield a reduced matrix in which each group was represented by a single medioid in a vector of distances.

The matrix of medioids was subjected to a round of supervised sorting, as was performed at the group level, resulting in a rearrangement of the medioid matrix. In this pass, a smoothing routine was also introduced, which involved an iterative column sort [analogous to the process of seriation (Sneath and Sokal, 1973)] where the dimensions of medioid matrix were successively decreased by a single row and the column on each successive pass. The final ordering of medioids was then used to create a new index (using the order of appearance of each group) into the full distance matrix.

IMPLEMENTATION

To demonstrate the use of the algorithm, we focus on a problem of biological classification, specifically a fine-grained classification of the Class *Gammaproteobacteria*, the largest phylogenetically coherent group of prokaryotes (Garrity, 2001). The input in our example consists of a matrix of evolutionary distance data and a table containing the complete taxonomic rank assigned to each sequence, based on the source organism annotation. The evolutionary distance data can be exported from PAUP* (Swofford, 2000) or created using the ape library in R (Paradis *et al*., 2004). Names in the taxonomy table occur in the specific order of appearance found in the Taxonomic Outline. Taxonomic names (e.g. genus, family, order, etc.) are treated as factor variables. Starting with an extant classification, we were able to move from a set of unordered sequences to an ordered and revised hierarchical taxonomy that is consistent with the current large-scale phylogenetic models in ∼20 min.

Initially, a heatmap of the distance matrix showed very little structure other than the diagonal line representing self-identity (data not shown) because the order of the sequences was based on their GenBank accession numbers. When unnamed sequences and sequences known to be misidentified were removed and the matrix was reordered according to the Taxonomic Outline (Fig. 1), misplaced sequences, indicative of taxonomic errors, were easily identified, as their color values tended to contrast significantly with the background level of relatedness. In Figure 1, we have highlighted some of the evident problems with this classification.

Figure 2 shows the heatmap after reorganization of all the genera and re-insertion of unnamed/misidentified sequences that meet the criterion for inclusion in the *Gammaproteobacteria*. Note that the

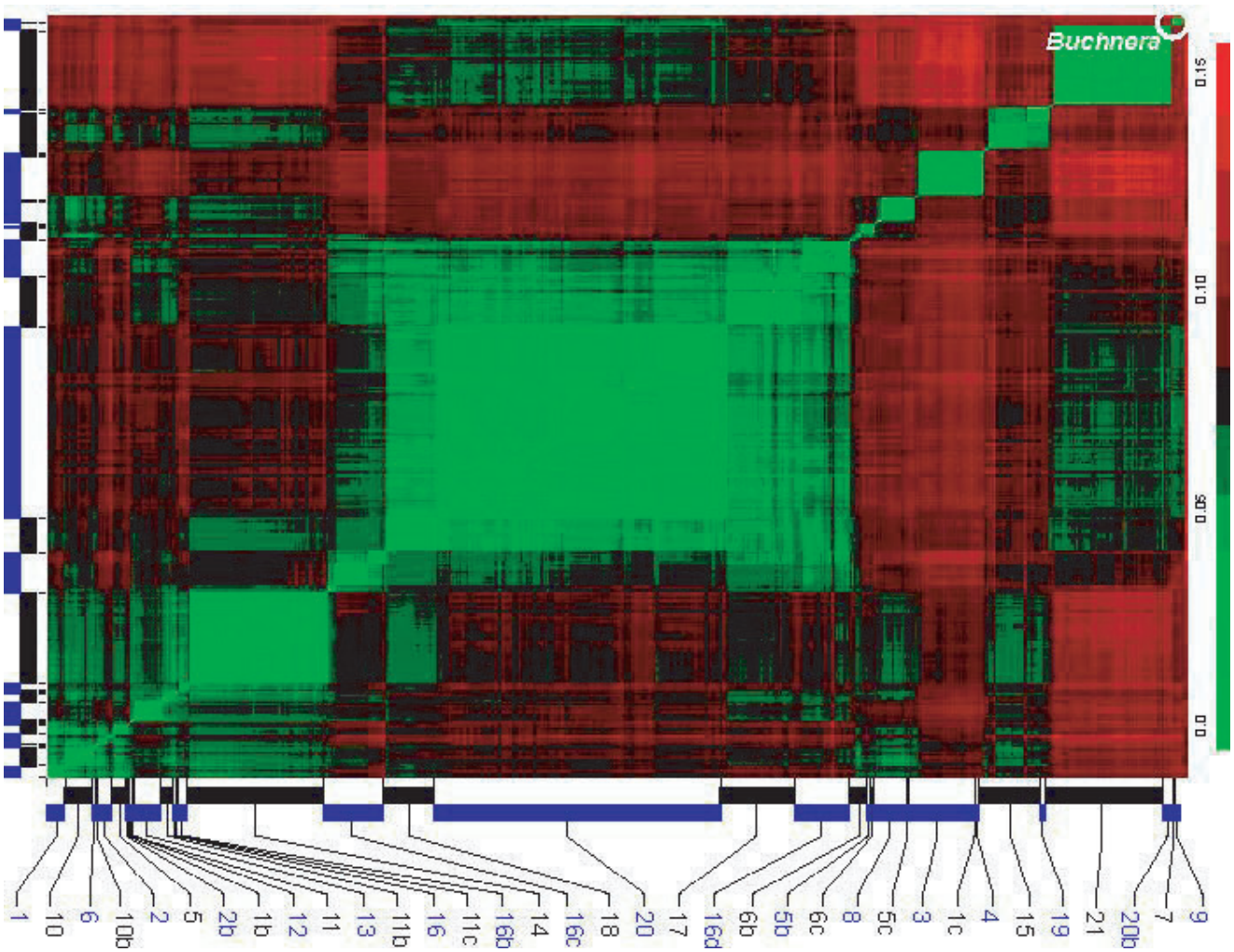

Fig. 3. Optimized heatmap of the *Gammaproteobacteria* based on seriated medioids. The ordering of medioids was used to create a new index into the full distance matrix using the order of appearance of each genus. Note that some taxonomic groups have changed location from the original map, as the algorithm seeks to order taxa according to overall levels of similarity (as defined by 16S sequence analysis). In addition, some families have been spit into multiple subgroups, indicating a high likelihood that those groups are paraphyletic as currently formed. The genus *Buchnera* is enclosed within the white circle. Families are numbered as in Figure 1 with lower case letters designating newly created subgroups of families.

exclusion of a small number of sequences derived from species not belonging to this taxon had essentially no impact on the overall range of evolutionary distances within the matrix. Yet, we do find an improvement in contrast between closely related and distantly related species that accentuates further classification problems. Inspection of this heatmap shows that, while the obvious errors in sequence placement have been corrected, ordering of the genera within the higher taxa (i.e. within the families and orders of the *Gammaproteobacteria*) remains suboptimal. Particularly noteworthy are three sharp breaks: two transect the *Alteromonadaceae* (16) and one transects the *Enterobacteriaceae* (20). Likewise, one can see that the *Xanthomonadaceae* (3), *Piscirickettsiaceae* (6) and *Franciscellaceae* (7) seem to be misplaced, interrupting families that would otherwise be contiguous.

Application of the medioid reordering algorithm resolved most of these discrepancies (Fig. 3). The initial classification, in which the 125 genera were grouped into 21 families, has been subdivided into 35 groups, suggesting that some of the families are paraphyletic. While 13 out of 21 families remained unchanged in composition, the remaining 8 families were subdivided into two [*Methylococcaceae* (10, 10b), *Ectothiorhodospiraceae* (2, 2b), *Enterobacteriaceae* (20, 20b), three [*Chromatiaceae* (1, 1b, c), *Oceanospirillaceae* (11, 11b, c), *Piscirickettsiaceae* (6, 6b, c), *Thiotrichaceae* (5, 5b, c)] or four [*Alteromonadaceae* (16, 16b-d)] subgroups. Also evident is that taxonomic groups [*Enterobacteriaceae* (20), *Xanthomonadaceae* (3) and *Moraxellaceae* (15)] have changed location from the original map, as the algorithm seeks to order taxa according to overall levels of similarity (as defined by 16S sequence analysis).

DISCUSSION

The algorithm described here is capable of generating new classifications and of visualizing and correcting extant classifications.

It was developed to automate a process for examining classifications that has been used by us to establish a comprehensive taxonomy of the prokaryotes (Lilburn and Garrity, 2004). In this example of an application of the algorithm to a problem drawn from prokaryotic classification, we first examined the current classification as it is given in the Taxonomic Outline of the Prokaryotes. Errors in the classification of the organisms represented by the sequences can arise from both known and unknown (i.e. unrecognized) problems in classification and are exemplified in Figure 1 by the bright green lines enclosed within the yellow boxes. Such misplaced sequences can be from organisms that have more than one valid name (taxonomic synonyms, which are both abundant and notoriously problematic) or bear incorrect names. If undetected, these nomenclatural errors can result in the placement of an organism in a taxon that does not contain its closest evolutionary relatives. In our example, we see sequences that are classified as *Pseudomonas* sp. (boxes 'a'), but that are actually much more closely related to *Xanthomonas* sp. A number of validly named species of *Pseudomonas*, which were described prior to the widespread use of 16S rRNA sequence analysis in systematic studies of prokaryotes, are now known to belong to the genus *Xanthomonas*, but not all nomenclatural changes have been formalized in the literature, nor have available sequences in GenBank or EMBL been updated to reflect these changes. A less clear-cut example is bounded by the box labeled 'b'. The enclosed line arises from a sequence that is classified as a *Pseudomonas* sp., but the data suggest that it belongs to one of the many genera within the *Enterobacteriaceae*. As our algorithm is intended to suggest corrections to an extant classification graphically, it should be able to move all the misplaced sequences, as well as those sequences that are misnamed or unnamed, to their correct positions in the taxonomy as visualized in the heatmap.

The functions initially written to re-position the misplaced sequences and correct the errors seen in Figure 1 worked quite well, as evidenced in Figure 2. Two problems did arise, however. The first problem was impaired visualization of the patterns and it arose from the presence in our dataset of outliers, that is, four sequences that were not members of the phylum *Proteobacteria*.

When distant outliers are included in an analysis, the scale of the heatmap changes in our current implementation. In the present example, the inclusion of four outlying sequences meant that the bulk of the distances were encoded by the first two colors in the distance scale, which made visual resolution of groups impossible. This problem was circumvented by screening the matrix against a set of sequences encompassing the entire range of possible distances and then removing those that lie outside the highest level of the classification in question prior to the analysis. Here, this meant removing the sequences originating outside of the *Gammaproteobacteria*. The second problem was that our algorithm only dealt with the data at the lowest level of our classification, in this case, at the genus level. Above that level, discontinuities in the ordering of taxa were visible. For example, in Figure 2, the genus *Buchnera* is clearly not closely related to the flanking genera of the Family *Enterobacteriaceae* and its presence resulted in a sharp fragmentation of an otherwise closely related group of genera. To deal with this second type of problem, we developed the medioids ordering approach, a method of data reduction that speeds the reordering of higher levels of a classification. This resulted in the placement of the genus *Buchnera* (20b) in the upper right corner of the heatmap in Figure 3, which is in accordance with the underlying data that suggest that this genus might be among the most distantly related genera within the *Gammaproteobacteria*.

The appearance of the final heatmap in Figure 3 suggests that some additional refinement on a gross scale is still in order. However, this arrangement is markedly improved as compared with the input and required less than 20 min of compute time to complete. It also reveals that some of the families, as currently defined, may require further refinement. Of the 21 families, 13 remained intact, suggesting that constituent intergeneric relationships were well founded, at least based on the single measure of 16S rDNA sequence similarity. On the other hand, eight of the families were split (see Implementation section), indicating that one or more genera in each of these families might be misclassified, a fact that has been borne out by us using independent, albeit much slower, means of establishing this fact (i.e. BLAST searches and searches of the RDP-II database). Moreover, the approach presented here provides some indication as to which higher taxa those misplaced genera might belong. This is not always possible using BLAST, as the underlying database is not designed for classification purposes.

CONCLUSION

The algorithm we have developed provides an intuitive approach to making and viewing classifications; conceivably, persons with no training could generate classifications and, by looking at the heatmaps, see how a classification might be improved. Our algorithm formalizes and automates the means used to achieve such improvements. Errors in data curation, classification and identification (of both sequences and source organisms) can be easily spotted and their effects corrected. Also, the classification itself can be modified so that the information content of the taxonomy is enhanced. The chief drawback of the approach is that groups formed from entities that are sparsely represented in the dataset may not be as robust or stable as groups formed from more richly represented entities. This is especially true in instances where such entities are equidistant to two or more otherwise unrelated entities. In the present example of a phylogenetically based classification, the tree graphs conventionally used to represent relationships among taxa also suffer from this problem. A possible advantage of heatmaps over tree graphs is that the former reveal such situations more clearly than trees. Even in cases involving relatively few taxa, which are handled well by tree graphs, heatmaps can provide a supplementary tool to aid in the interpretation of treeing problems, as we have shown earlier (Lilburn and Garrity, 2004). Our algorithm can be used to develop and improve classifications of all types. For example, functional assignment of new sequence benefits from a reliable protein classification. Data from gene expression microarrays might also be usefully classified using these techniques. We also note that while we employed hierarchical clustering as a means of optimizing internal groups within the larger matrix, other non-supervised techniques, such as binary recursive partitioning, are equally applicable. The only requirements are that the alternative technique return an ordered list of entities for subsequent manipulation and that the alternative method be suitable for use with the type of data under investigation. In our example application of the algorithm, we demonstrated that significant improvements to a prokaryotic taxonomy can be readily obtained using these statistical approaches to the evaluation of sequence-based evolutionary distances. The approach has been independently validated by comparing the list of misidentified taxa and incorrectly

annotated sequences to the taxonomic record, to placement in large-scale phylogenetic trees, and by BLAST searches. Improvements to the taxonomy of the prokaryotes are of interest not only to microbiologists, but also to the medical, legal and biodefense communities, for example, where it is essential that a microbe should be correctly identified and named. Furthermore, phylogenomic approaches to annotation rely on accurate identification of related species and, in at least one case (Sicheritz-Ponten and Andersson, 2001), such approaches require an accurate and comprehensive taxonomy. We are developing tools based on this algorithm that will allow us to maintain and expand a comprehensive prokaryotic taxonomy (Garrity *et al*., 2004).

ACKNOWLEDGEMENTS

We would like to express our thanks to the three anonymous reviewers for their helpful criticism of the original draft of this manuscript. This research was supported by the Biological and Environmental Research Program (BER), United States Department of Energy (Grant nos DE-FG02-02ER63315, DE-FG02-04ER63933 and DE-FG02-04ER63632).

REFERENCES

Becker,R.A. *et al*. (1988) *The New S language: A Programming Environment for Data Analysis and Graphics*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.

Cole,J.R. *et al*. (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res.*, **31**, 442–443.

Garrity,G.M. (ed.) (2001) *Bergey's Manual of Systematic Bacteriology*, 2nd edn. Springer-Verlag, NY.

Garrity,G.M. and Lilburn,T.G. (2002) Mapping taxonomic space: an overview of the road map to the second edition of Bergey's Manual of Systematic Bacteriology. *WFCC Newslett.*, **35**, 5–15.

Garrity,G.M., Bell,J. and Lilburn,T.G. (2004) *Taxonomic outline of the prokaryotes*. In *Bergey's Manual of Systematic Bacteriology*, 2nd edn, Release 5.0. Springer-Verlag, NY, DOI: 10.1007/bergeysoutline.

Lilburn,T.G. and Garrity,G.M. (2004) Exploring prokaryotic taxonomy. *Int. J. Syst. Evol. Microbiol.*, **54**, 7–13.

Paradis,E. *et al*. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Sicheritz-Ponten,T. and Andersson,S.G.E. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res.*, **29**, 545–552.

Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. W. H. Freeman, San Francisco, CA.

Venables,W.N. and Smith,D.M. (2002) *An Introduction to R*. Network Theory Ltd, Bristol, UK.