Automating the Quest for Novel Prokaryotic Diversity (Revisited)

[1,6]George M. Garrity, [5]Timothy G. Lilburn, [1]Scott H. Harrison, [2]Yun Bai,
[3]Yuan Zhang, [6]Catherine Lyons and [4,6]James, R. Cole

[1]Department of Microbiology & Molecular Genetics, [2]Department of Mathematics, [3]
Department of Computer Science and Engineering, and [4]Center for Microbial Ecology,
Michigan State University, East Lansing, MI,
[5]Science Information Systems ,
American Type Culture Collection, Manassas, VA
and [6]NamesforLife, LLC, Okemos, MI

Previously, we demonstrated the value of using techniques drawn from the field of exploratory data analysis (EDA) for the analysis and visualization of large sets of sequence data (notably SSU rRNA gene sequences) that are used to construct a comprehensive taxonomy of prokaryotes. While the approach is computationally efficient and quite useful in uncovering a variety of taxonomic and annotation errors, the methods suffered from some practical limitations; notably bottlenecks in the preprocessing of data for our analyses. Work is currently underway to address these limitations that will greatly expedite the preprocessing steps through a pipeline approach. In addition, new methods are under active development that will automatically flag misidentified and potentially novel sequences within a given dataset and automatically place such sequences into close proximity to their nearest neighbors, based on 16S rDNA sequence homology. These methods will also permit linking of EDA plots, derived from such analyses to external data and information resources.

Introduction

Over the past 15 years, a reasonably good picture of the evolutionary relationships among *Bacteria* and *Archaea* (the prokaryotes) has emerged. This is largely the result of widespread use of the 16S rRNA gene as the marker of choice in phylogenetic studies. At present, the microbiological community relies on two large-scale phylogenetic models of the prokaryotes (the ARB and RDP trees (Cole *et al.*, 2003; Ludwig *et al.*, 2004; Maidak *et al.*, 2001)) which exist in different states of completeness. These models also serve as the underpinnings of the comprehensive taxonomy used in the Second Edition of *Bergey's Manual of Systematic Bacteriology* (Cole *et al.*, 2003; Garrity, 2001; Garrity & Holt, 2001; , Garrity, 2005 #16), the RDP-II (Cole *et al.*, 2003), and GenBank. In addition, analyses of 16S rDNA sequences now serve as the principle means of defining community structure, detecting novel evolutionary lineages in virtually any environment, and establishing the identity of unknown isolates. Sequence data and the associated annotation information are often used to suggest how individual strains and entire microbial communities might function. Knowing the identities of community members or their closest relatives tells us something about their likely physiological capabilities, the role they might play in a given environment, or their potential as a source of novel products or processes.

Despite the power of these methods, there are limitations to contemporary phylogenetic models that warrant consideration. At present, it is not possible to incorporate all, or even most of the available sequence data into an all-inclusive taxonomy that can be dynamically maintained and modified as new taxa are described and existing taxa emended. Thus biases in taxon sampling can have profound effects on the interpretation of the data. So too, can sequencing errors, the presence of chimeric sequences in the dataset, and hidden biases such as alignments, masks, rate constants that are passed to phylogenetic algorithms, and differences in positional homology. Further confounding problems include the high number of annotation errors in the sequence data which can lead to the misidentification and mischaracterization of unknown isolates; the lack of clear criteria for defining boundaries of species and higher taxa; and the relative insensitivity of large-scale models to small groups having (one to three members) that may represent novel lineages, either within larger, well-defined regions or that are distantly related to all previously recognized taxa. Such limitations can to have a profound effect on predications of taxonomic, physiological or functional diversity. Finally, the lack of unique identifiers that can persistently link together the strains, data, and the associated literature, backwards and forwards in time, presents a significant obstacle to those who must use this information for decision making and policy formulation.

Previously, we developed methods for using the phylogenetic information in the global 16S rDNA dataset to frame and modify a comprehensive taxonomy of the prokaryotes (Garrity & Holt, 2001; Garrity & Lyons, 2004; Garrity *et al.*, 2004a; Garrity *et al.*, 2004b; Lilburn & Garrity, 2004; Wang *et al.*, 2004). As our research progressed, it became clear that this sequence information could transcend its phylogenetic role and serve as an index to other information pertaining to the organism from which the sequences were derived. Theoretically, 16S rRNA gene sequences could be treated as unique and persistent identifiers because these remain the same no matter how our ideas about taxonomies, classifications, nomenclature, or phylogenies change. Realizing this potential would, however, require overcoming a number of interrelated technical and sociological problems associated with strain and taxon definition, registration, and tracking over time. These concepts were first discussed during the 10th International Congress of Culture Collections in October 2004 in Tsukuba, Japan (Garrity *et al.*, 2004b). We report here on our further progress in solving these problems.

Materials and methods

Sequence data – Three experimental datasets were used. The first was a vetted data set of 7673 high quality sequences for which the annotation information has been confirmed and/or updated to reflect the current prokaryotic taxonomy.(Garrity & Lilburn, 2002; Garrity & Lilburn, 2005a; Lilburn & Garrity, 2004) The second dataset was defined by Hugenholtz (Hugenholtz, 2002)as containing representatives of the known and potential phyla of *Bacteria* and *Archaea*. The third consists of 4649 high quality sequences of cultivable *Bacterial* phyla (named and unnamed) that were drawn from our own data set.

Alignments – Manual alignment followed the RDP-II (Release 8.0) alignment, as previously described (Garrity & Lilburn, 2002; Lilburn & Garrity, 2004). Auto-alignment was done using the RDP auto-aligner, which is based on the probabilistic model of Brown et al. (Brown, 2000; Cole *et al.*, 2003).

Estimation of evolutionary distances - Matrices (symmetrical and asymmetrical) of evolutionary distance were calculated using the Jukes and Cantor model (Jukes & Cantor, 1969) and served as the input for EDA models.

Exploratory Data Analyses – All modeling and data visualizations were done in S-Plus Enterprise Developer (Version 7.0, Insightful, Seattle, WA) as previously described (Cole *et al.*, 2003; Garrity & Holt, 2001; Garrity & Lilburn, 2002; Lilburn & Garrity, 2004).

Nomenclature and taxon modeling – We follow the NamesforLife (N4L) data model as described Garrity and Lyons (Garrity & Lyons, 2003; Garrity & Lyons, 2004) in which names, taxa, exemplars, nomoi, practitioners, and references are instantiated as XML information objects. Each information object is uniquely and persistently identified by one or more N4L Digital Object Identifiers (DOIs). Schema development was done using Turbo XML (Tibco Software, Palo Alto, CA) and Epic Architect/Styler (Version 5.0 Arbortext, Ann Arbor, MI). Taxonomic data was derived from an SGML instantiation of the Taxonomic Outline of the Prokaryotes (Garrity *et al.*, 2004a) and supplemented with additional bibliographic information.

Results

Initial studies – In 2001, we demonstrated that EDA techniques such as principal components analysis could be successfully applied to large asymmetric matrices of evolutionary distances to yield highly stable models that were significantly reduced in dimensionality and reconcilable with phylogenetic trees and phenotypic data. While these models proved useful in establishing boundaries of many of the higher taxa, problems of overlapping groups that led to taxon occlusion were evident. Further research into alternative methods of projecting and visualizing data gradually led to the development of phylogenetic heatmaps (Lilburn & Garrity, 2004) These graphics proved useful in revealing higher order relationships (e.g., family, class, order, phylum), uncovering taxonomic and sequence annotation errors, and revealing novel lineages. Heatmaps also proved useful as diagnostic tools for dissecting more complex relationships among taxa that are not readily resolved in PCA plots, phylogenetic trees or comparable models in which data dimensionality is reduced. (becuse of problems attributable to intermediate level relationships, long branch attraction, etc.).

While heatmaps readily revealed annotation and taxonomic errors based on an input taxonomy, automatic correction of those errors was not possible, nor was presumptive identification of unknown isolates. To that end, we devised a self-organizing, self-correcting classifier that both optimized the arrangement of an input taxonomy structure and "corrected" sequences that were most-likely mis-annotated (Garrity & Lilburn, 2005a; Garrity & Lilburn, 2005b; Lilburn & Garrity, 2004). Using this tool, we were able to produce a set of "vetted" sequences that are correctly annotated with regard to their current taxonomic status.

Current studies – While our initial efforts proved quite successful, several technical issues limited routine use of these methods. Some of these limitations were operational and centered on requirements for manual intervention in the process (e.g. hand alignment of sequences, movement of large data sets across multiple systems, lack of persistent identifiers for strains and sequences). Other limitations included the need for refinement of our global model, which until recently has been based on sequence similarity to a set of 223 internal reference points. This approach may provide adequate resolution of most groups, but not all (see below).

We are currently pursuing independent lines of research to address both operational and theoretical limitations with our earlier approaches. Two significant barriers have existed that restricted our ability to automate our methods and deploy them as a network service to the community: the requirement of hand alignment of 16S sequences and data curation/annotation of the underlying sequence and taxonomic records.

With regard to the sequence alignment problem, we have confirmed that that evolutionary distance matrixes produced from hand-aligned and auto-aligned *Bacterial* sequences are highly correlated (r=0.964, t=429.1, df=2, n=10655). We have also developed a data pipeline and supporting application programs that permit retrieval of aligned sequences from the RDP server and preprocessing of the data (e.g. calculation of evolutionary distance matrices) that serve as input to our EDA routines. Work is currently in progress to modify the RDP auto-aligner to accommodate 16S sequences from *Archaea*.

We are continuing to refine and optimizing our EDA models. In addition to providing a superior means of visualizing large data sets, the use of asymmetric matrices has provided significant improvements in computational efficiency over conventional treeing methods. In theory, this approach permits us to accurately place any given sequence into a precise position within a taxonomy that incorporates the entire 16S rDNA sequence dataset, which now totals > 180,000 sequences. However, the accuracy of a given location and the assignment of most probable identity by our self-organizing self-correcting classifier (SOSCC) algorithm are dependent on the benchmarks used in building the model. While our initial version was quite robust and compared favorably with common nearest-neighbor classification methods, such as SEQ-MATCH (Cole *et al.*, 2003) and BLAST, there are areas within the taxonomy that would benefit from the incorporation of additional benchmarks. Likewise, there are areas within the taxonomy that are over-defined. To that end, we are actively exploring strategies for selecting additional sequences to serve as internal reference points. The ability to draw on autoaligned sequences and to compute full evolutionary distance matrices "on the fly" is proving to be quite useful in resolving areas of uncertainty within the Bergey's taxonomy and ascertaining truly novel lineages for which only a small number of sequences currently exist. This also provides us with an opportunity to examine alternative taxonomic views of the same or overlapping datasets, to discover optimal compositions of higher taxa, and to test alternative algorithms for forming higher taxa. Ultimately, this should lead to the discovery of an optimized set of benchmarks that can be used for high-throughput identification of unknowns based on a comparison against the full breadth of prokaryotic diversity, whether or not the reference points represent cultivated or yet-to-be cultivated species.

A related line of research involves the use of interactive graphics as a means of linking together taxonomic information with phenotypic, genotypic, and genomic data. The 16S rRNA gene already serves as a unique and persistent identifier for prokaryotic taxa. It might also serve as a handle that could provide a means of linking to other information sources that could be mapped back onto a taxonomy or phylogeny. Thus, the graphics produced from our EDA models could also serve as a GUI for data mining, providing end-users with a novel means of viewing and harvesting data and the literature as it pertains to one or more dynamically evolving taxonomic frameworks. Such an approach would, however, require the use of unique and persistent identifiers that will *always* resolve to the appropriate information resources over time. A data architecture, built on DOI technology is currently under development that we believe will ultimately fulfill these requirements.

Discussion

Rapid advances in sequencing technology, coupled with the dramatic decrease in sequencing costs, have permanently changed the way in which we practice microbiology. No longer is the question when to sequence an unknown strain. Rather, the question now is what does a 16S rDNA analysis reveal about the likely identity of an unknown isolate or the nature of an entire community? Like a Gram-stain in the past, the 16S sequence has become one of the first pieces of information collected when characterizing an unknown isolate. However, a 16S sequence is infinitely more powerful when the quality of the underlying databases that are interrogated and the and the power of the algorithms used in the analysis are sufficiently high. If any of these have hidden methodological weaknesses or are affected by annotation errors, the response to a query may be wrong. So too, may be the decisions and actions that are taken. Whether or not an end user is capable of recognizing such problems is difficult to ascertain. Methods that return long lists of names and numbers or graphics that are subject to widely differing expert interpretations leave much to be desired.

EDA techniques provide an independent and theory neutral way of examining various transformations of extremely large sequence datasets. In addition to being computationally efficient, graphical methods such as heatmaps can provide unambiguous answers to some of the more vexing problems facing contemporary microbiologists. The ability to use such graphics as a means of linking together the underlying sequence data with virtually any other type of data or information about a particular strain, species, higher taxon, or community could be highly advantageous. It would also provide a key step in automating the process of identifying truly novel taxa at different ranks in a more objective manner. The subject of this presentation will be how we might arrive at this point.

Acknowledgements
This research was supported by the Office of Science (BER), U.S. Department of Energy, Grant Nos. DE-FG02-04ER63933 and DE-FG02-99ER62848, National Science Foundation Award No. DBI-0328255 and a Michigan University Commercialization Initiative Fast (MUCI) Track Grant.

References cited

Brown, M. P. S. (2000). Small subunit ribosomal RNA modeling using stochastic context-free grammar. In *Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*, pp. 57–66. San Diego, California, USA.

Cole, J. R., Chai, B., Marsh, T. L., Farris, R. J., Wang, Q., Kulam, S. A., Chandra, S., McGarrell, D. M., Schmidt, T. M., Garrity, G. M. & Tiedje, J. M. (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. (2001). Bergey's Manual of Systematic Bacteriology. In *Bergey's Manual of Systematic Bacteriology*, pp. 742 pp. Edited by G. M. Garrity. New York: Springer-Verlag.

Garrity, G. M. & Holt, J. G. (2001). The Road Map to the *Manual*. In *Bergey's Manual of Systematic Bacteriology*, pp. 119-166. Edited by D. R. Boone, R. W. Castenholz & G. M. Garrity. New York: Springer-Verlag.

Garrity, G. M. & 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 Newsletter* **35**, 5-15.

Garrity, G. M. & Lyons, C. (2003). Future-proofing biological nomenclature. *OMICS* **7**, 31-33.

Garrity, G. M. & Lyons, C. (2004). Systems and Methods for Resolving Ambiguity Between Names and Entities. *U.S. Patent Application 10/759,817* **Filing date January 16, 2004**.

Garrity, G. M. & Lilburn, T. G. (2005a). Self-organizing and self-correcting classifications of biological data. *Bioinformatics* **21**, 2309-2314.

Garrity, G. M. & Lilburn, T. G. (2005b). Methods for Data Classification. *U.S. Patent Application EV618124130US* **Filing date June 16, 2005**.

Garrity, G. M., Bell, J. & Lilburn, T. G. (2004a). Taxonomic Outline of the Procaryotes, Bergey's Manual of Systematic Bacteriology, Second Edition, Release 5.0.: Springer-Verlag.

Garrity, G. M., Bell, J. A. & Lilburn, T. G. (2005). The Revised Road Map to the Manual. In *Bergey's Manual of Systematic Bacteriology, Volume 2 The Proteobacteria*, pp. 2791 pp. Edited by G. M. E. Garrity. New York: Springer.

Garrity, G. M., Zhang, Y., Lilburn, T. G. & Cole, J. R. (2004b). Automating the Quest for Novel Prokaryotic. In *Innovative Roles of Biological Resource Centers*. Edited by M. Watanabe, K. Suzuki & T. Seki. Tsukuba, Japan: Japan Society for Culture Collections & World Federation of Culture Collections.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3**, REVIEWS0003.

Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, pp. 21-132. Edited by Munzo. New York: Academic Press.

Lilburn, T. G. & Garrity, G. M. (2004). Exploring prokaryotic taxonomy. *Int. J. System. Evol. Micro.* **54**, 7-13.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Forster, W., Brettske, I., Gerber, S., Ginhart, A., Gross, O., Grumann, S., Hermann, S., Jost, R., Konig, A., Liss, T., Lussmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A. & Schleifer, K. (2004). ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363-1371.

Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T., Jr., Saxman, P. R., Farris, R. J., Garrity, G. M., Olsen, G. J., Schmidt, T. M. & Tiedje, J. M. (2001). The RDP-II (Ribosomal Database Project). *Nucleic Acids Res* **29**, 173-174.

Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. (2004). A naïve Bayesian classifier for rapid assignment of rRNA sequences into the new Bacterial taxonomy. In *RECOMB 2004, Abstract H25*, pp. 292-293. San Diego.