Pipelining RDP Data to the “Taxomatic”

Timothy G. Lilburn, PI/Co-PI

George M. Garrity, PI/Co-PI (Collaborative)

James R. Cole, Co-PI (Collaborative)

Project ID 0010734

Grant No. DE-FG02-04ER63932

Background

This project was conceived to build on and enhance the results of previously funded research by integrating data and software that were used in building resources for the preparation of Bergey’s Manual of Systematic Bacteriology, 2nd Edition (Volumes 1 & 2A-C) and the Ribosomal Database Project-II (RDP-II). Our objectives were to both enhance the value of the data and create a pipeline approach to keeping the data current. Earlier, we demonstrated the value of using exploratory data analysis (EDA) to visualize the relationships among large sets of SSU rRNA gene sequences that were used to construct a comprehensive phylogeny of prokaryotes. We developed Self-Organizing Self-Correcting Classification (SOSCC) algorithms that were computationally efficient and useful for unraveling problems within the underlying data (e.g., annotation errors, unresolved synonymies, taxonomic and nomenclatural errors). We deployed a web site, referred to as the “*Taxomatic*”, to make the results of our EDA analyses available and to enable comparisons of classifications. However, bottlenecks at the preprocessing stage limited deployment of our applications and data, making the web site essentially static and in need of frequent updates. This limited the usefulness of the web site to end users. To overcome the bottlenecks (which included hand alignment and computation of large matrices of pair-wise evolutionary distances), we proposed building a data pipeline between the *Taxomatic* applications and RDP-II web services.

The main goals of the current project were to accelerate the production of the updated versions of the prokaryotic taxonomy in lock-step with the publication of new taxa and the rearrangement of existing taxa, and to distribute these data via the RDP-II to other stakeholders in the taxonomic community and to the research community at large. A related goal of the current project was to deploy our visualization techniques as part of an interactive web application, enabling users to view, manipulate, and select data sets of particular interest based upon phylogenetic and genomic criteria, and to access sequence data and, ultimately, the scientific literature where the original observations and papers that extend the original observations are found.

Accomplishments vs objectives

As noted previously, we proposed completing this project during 2007, but the unanticipated departure of a postdoc leading the work resulted in delays. This ultimately proved advantageous because it provided an opportunity to revisit some of the underlying assumptions and methods that were in used in prototypes, leading to a more stable and robust implementation of the application.

Early prototypes of the heatmap visualization tool and classifier, based on the SOSCC, were developed in S-Plus and R. While useful for concept testing, these environments proved unsuitable for deploying client applications because of underlying limitations. We re-implemented the SOSCC algorithm as a Java web service and optimized it, addressing a previous limitation that prevented correct placement of some sequences when the algorithm was run in a fully unsupervised, automated version. Statistical evidence for group membership by bootstrapping (currently set to 1000 iterations) within the SOSCC optimized hierarchy was also added, to provide confidence estimates of group membership for each taxon, along with confidence limits of placement in alternative higher taxa. These data are then fed back into the optimization routine to provide a final smoothing of the matrix in which placements with little statistical support are relocated to

Figure 1. The revised SOSCC routine

the position in the matrix that is best supported by the experimental data (Figure 1). These data are then bundled together with links to download the optimized matrix in dnadist format and to view the report and heatmap in the *Taxomatic*. The improvements provide a more satisfactory user experience (e.g. 30 seconds to produce a maximally smoothed matrix of 1000 sequences) and allow the entire application to reside on the RDP server(s), where the interface is now part of the web services offered by RDP-II.

The output of the *Taxomatic* is shown in Figure 2. Distance matrices are visualized as heat maps and options for accessing the underlying matrix, the images and the taxonomic information are offered. The tool accepts raw distance matrices or aligned sequence information as data sources. When sequence information is provided, the

distance matrix is computed using the uncorrected distance model. Users can upload files to the *Taxomatic* website or sequences can be submitted by a SOAP service. This SOAP service is used by RDP to streamline *Taxomatic* use with RDP data. In addition to

supplying source information, users can (i) supply their own taxonomic information by uploading it in XML format, (ii) retrieve taxonomic information from the RDP using either RDP or Genbank identifiers as source data, with or without classification by the RDP Classifier web service, or (iii) completely omit taxonomic data. In the latter case, the input distance matrix can be viewed in the order in which it was loaded.

The SOSCC can now be accessed through the *Taxomatic* either as a preprocessing option or as a SOAP service in which a matrix can be reorganized. SOSCC classification can be done in two ways. A supervised method can be used where an existing taxonomy is fitted to the reorganized matrix or, alternatively, an experimental unsupervised method can be used where boundaries are predicted directly from the resulting matrix. The supervised classification method can be bootstrapped to determine the confidence of the placements.

Figure 2. A screen shot of the output from the *Taxomatic* for the phylum *Tenericutes*. On the left is the heatmap representing the phylogenetic distances among the sequences that represent the members of the phylum. In the center is the taxonomy of the phylum. On the right, the data handling flow for the *Taxomatic* web tool is shown.

Dynamic links to NamesforLife information objects, which provide additional information about individual source organisms, their current taxonomic position, and bibliographic information, have been implemented and await a final clean-up of that data by NamesforLife, LLC. Once that task is completed (estimated 3Q 2009), the complete taxonomic hierarchy based on 16S will be rebuilt and published as a new release of the Taxonomic Outline of Bacteria and Archeae (TOBA). This task was originally scheduled

for the latter part of 2008, but is on hold pending resolution of a number of taxonomic and nomenclatural anomalies that have accumulated in the over time.

Students associated with this project:

Scott Harrison, Microbiology and Molecular Genetics, Michigan State University.

Paul Saxman, Medical Informatics Program, University of Michigan State University

Jordan Fish, Computer Science, Michigan State University

Sheena Tapo, Microbiology and Molecular Genetics, Michigan State University

Nicole Osier, Microbiology and Molecular Genetics, Michigan State University.

Publications in chronological order

Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 37 (Database issue): D141-D145; doi: 10.1093/nar/gkn879. [Oxford University Press: <http://nar.oxfordjournals.org/cgi/content/full/gkn879>]

Lilburn, T.G., S.H. Harrison, J.R. Cole, and G.M. Garrity. 2006. Computational aspects of systematic biology. *Briefings in Bioinformatics* 7: 186-195

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

Published Abstracts in chronological order

Fish, J., Q. Wang, S. H. Harrison, T. G. Lilburn, P. R. Saxman, J. R. Cole, and G. M. Garrity. 2009. Release of the Taxomatic and Refinement of the SOSCC Algorithm, February 8-11, 2009, GTL (Genomes to Life) Awardee Workshop VII, Bethesda, Maryland.

Cole, J. R. 2008. Thirty Years of Ribosomal RNA Sequencing, September, 20th, SCOPE (Scientific Committee on Problems of the Environment) Workshop presentation, Changsha, China.

Cole, J. R. 2008. The Ribosomal Database Project. Max Planck Institute for Marine Microbiology "International Workshop on Molecular Markers: Ribosomal RNA", April 7-9, Max Planck Institute Workshop presentation Bremen, Germany.

Chai, B., Q. Wang, R. Farris, J. Fish, E. Cardenas, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, G. M. Garrity, J. M. Tiedje, J. R. Cole. 2008. Ribosomal Database Project - II: Tools and Sequences for rRNA Analysis. Session 292/R Bioinformatics and Databases; Poster R-122. ASM 108th General Meeting, June 1-5, Boston, Massachusetts.

Wang, Q., B. Chai, W. Sul, D. M. Tourlousse, R. C. Penton, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, J. M. Tiedje, J. R. Cole. 2008. A Protocol for Rapid and Efficient Bacterial Community Analysis Using Pyrosequencing. Session 175/N Molecular Microbial Ecology Communities - III; Poster N-203. ASM 108th General Meeting, June 1-5, Boston, Massachusetts.

Chai, B., Q. Wang, R. Farris, J. Fish, E. Cardenas, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, G. M. Garrity, J. M. Tiedje, J. R. Cole. 2008. Ribosomal Database Project - II: Tools and Sequences for rRNA Analysis. ISME-12 Symposium "Sustaining the Blue Planet", August 17-22, Cairns, Australia.

S.H. Harrison, T.G. Lilburn, J.R. Cole, P.R. Saxman, and G.M. Garrity. 2007. Recognizing and Dealing with Taxonomic Distortions Caused By the Wealth of Sequence Data. ASM 107th General Meeting, May 21-25, Toronto, Canada.

J. Fish, Q. Wang, S.H. Harrison, T. G. Lilburn, P. R. Saxman, J. R. Cole, and G. M. Garrity. 2007. Further refinement and deployment of the SOSCC algorithm as a web service for automated classification and identification of Bacteria and Archaea. DOE Genomes to Life Contractor and Grantee Workshop, Bethesda, MD

Harrison, S.H., P. Saxman, T.G. Lilburn, J.R. Cole, and G.M. Garrity. 2006. Pipelining RDP Data to the Taxomatic and linking to external data. DOE Genomes to Life Contractor and Grantee Workshop, Bethesda, MD

Garrity, G.M., C.M. Lyons, J.R. Cole 2006 Knowledge bleed, NamesforLife, and Rumsfeld's axiom. FEMS2006, 2nd Annual Meeting Federation of European Microbiology Societies. Symposium on Biodiversity, Madrid, Spain

Lilburn, T. G., Y. Bai, Y. Zhang, J. R. Cole and G. M. Garrity. 2005. Projections, trees and evolutionary space. For the XI th International Congress of Bacteriology and Applied Microbiology, San Francisco, CA.

Lilburn, T. G., Y. Bai, Y. Zhang, J. R. Cole and G. M. Garrity. 2005. Exploring evolutionary space. For the DOE Genomes to Life Contractors and Grantees Workshop III, Washington, DC.

Electronic Publications

Garrity, G. M., Lilburn, T. G., Cole, J. R., Harrison, S. H. , Euzeby, J. , and Tindall, B. J.. *The Taxonomic Outline of Bacteria and Archaea* [Online], Volume 7 Number 7 (3 April 2007) <http://www.taxonomicoutline.org>