

Introduction

The Ribosomal Database Project (RDP-II) provides data, programs and services related to the ribosome. The RDP is a value-added database available to the research community through the RDP website. The RDP organizes sequence data into alignments, annotates ribosomal RNA (rRNA) sequence data, provides a phylogenetic overview of life and offers a suite of services and tools to assist in the handling and analysis of the data. It is not solely a collection of rRNA sequences extracted from the public nucleotide databases. RDP provides information about where a sequence fits in an overall phylogenetic scheme, along with up-to-date nomenclature and other identifying information about the source organism.

The RDP arose out of research conducted by two University of Illinois at Urbana-Champaign (UIUC) faculty members, Carl R. Woese and Gary J. Olsen. Woese recognized early that rRNA could be used to elicit phylogenetic relationships between organisms due to its conserved sequence. A collection of 473 aligned prokaryotic 16S rRNA sequences, many of which were generated in Woese's laboratory as well as those generated by other researchers, was made available to the public in the first release (Release 1) of the RDP on January 5, 1992. Argonne National Laboratory hosted the RDP ftp and public sites until Release 3.0 in August 1993 when the public sites were moved to UIUC. The RDP subsequently moved to Michigan State University (MSU) in 1998.

Growth of the Database

The RDP database has grown to include 16,277 aligned prokaryotic sequences, 2055 aligned eukaryotic sequences and 1503 aligned small subunit (SSU) rRNA mitochondrial sequences. The phylogenetic tree displayed on the right side of the poster is based on sequences that represent the taxonomic families of prokaryotes as defined in the forthcoming edition of Bergey's *Manual of Systematic Bacteriology* plus sequences amplified from the environment that appear to fall outside these families. The sequences on this

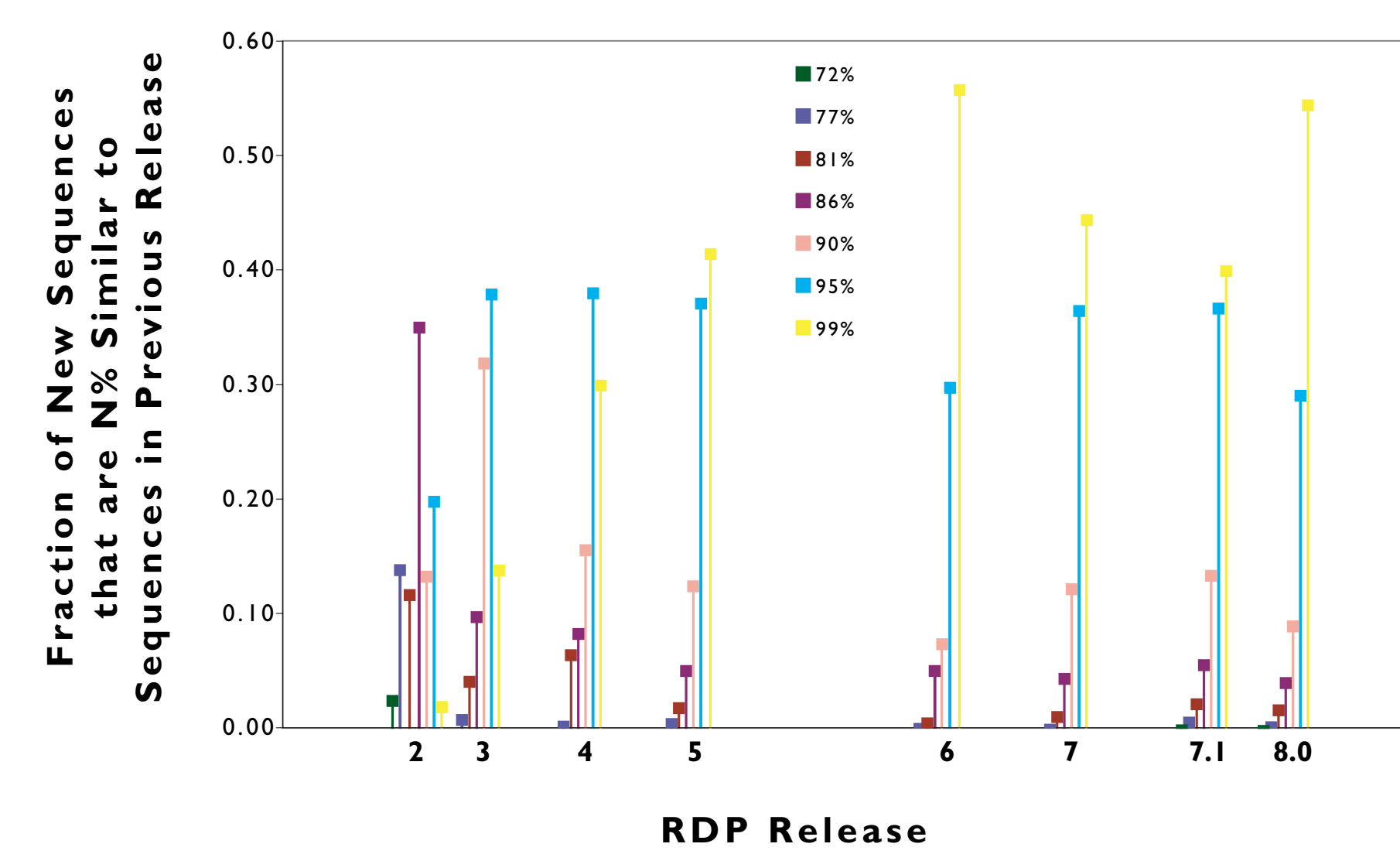


Figure 1. In each RDP release, most of the new sequences are fairly similar (that is, more than 90% similar) to sequences already in the database. Some are less similar and might be thought to represent novel diversity. This figure shows, in pink, the percentage of sequences in each release that are less than 90% similar to any sequence in the previous release. Not surprisingly, the first few releases added considerably to the diversity represented in the database. Since Release 5, about 2 to 3% of the new sequences have represented "diverse" taxa. This proportion shows no sign of decreasing.

tree were included in Release 7.1. Taxa on yellow branches were represented in a paper that was among the first to attempt an outline of prokaryotic phylogeny using 16S SSU rRNA sequences (G. E. Fox et al. "The Phylogeny of Prokaryotes." 1980. *Science* 209: 457-463). Taxa on blue branches plus the yellow branch taxa were represented by sequences in RDP Release 1. Taxa on pink branches have been added to the RDP database since Release 1. It is striking how well the phylogenetic breadth of the prokaryotes was represented in the Fox et al. 1980 paper, compared to the breadth represented by the new backbone tree of 217 taxa calculated for Release 8.0. Figure 1 shows that diversity is still being discovered at a steady rate.

The change in composition of the database is shown in Figure 2. Release 1 contained sequences from 15 of Bergey's 25 phyla and five of these phyla (*Proteobacteria*, *Firmicutes*, *Bacteroidetes*, *Actinobacteria* and *Euryarchaeota*) represented almost 90% of the sequences. Release 8.0 contains sequences from over 20 phyla (the smallest pie wedge in Figure 2 represents about 30 sequences) and the top five phyla represent about 83% of the taxa. In Release 1, five modern-day phyla had no sequences, and at that time, few PCR-generated environmental clones from microbial ecology studies had entered the database. In Figure 2, the percentages included in the parentheses for Release 8.0 are the named genera and environmental clone sequences respectively. Due to the large increase in the number of sequences from *Proteobacteria*, the distribution of sequences amongst the phyla is less equitable in Release 8.0 than in Release 1, although the phylogenetic breadth of Release 8.0 is greater.

Figures 3 and 4 depict the trends for 20 of Bergey's phyla over the time spanned by RDP Releases. Figure 4 depicts more detailed analysis for three widely-studied groups: *Proteobacteria*, *Archaea*, and Gram-positive bacteria.

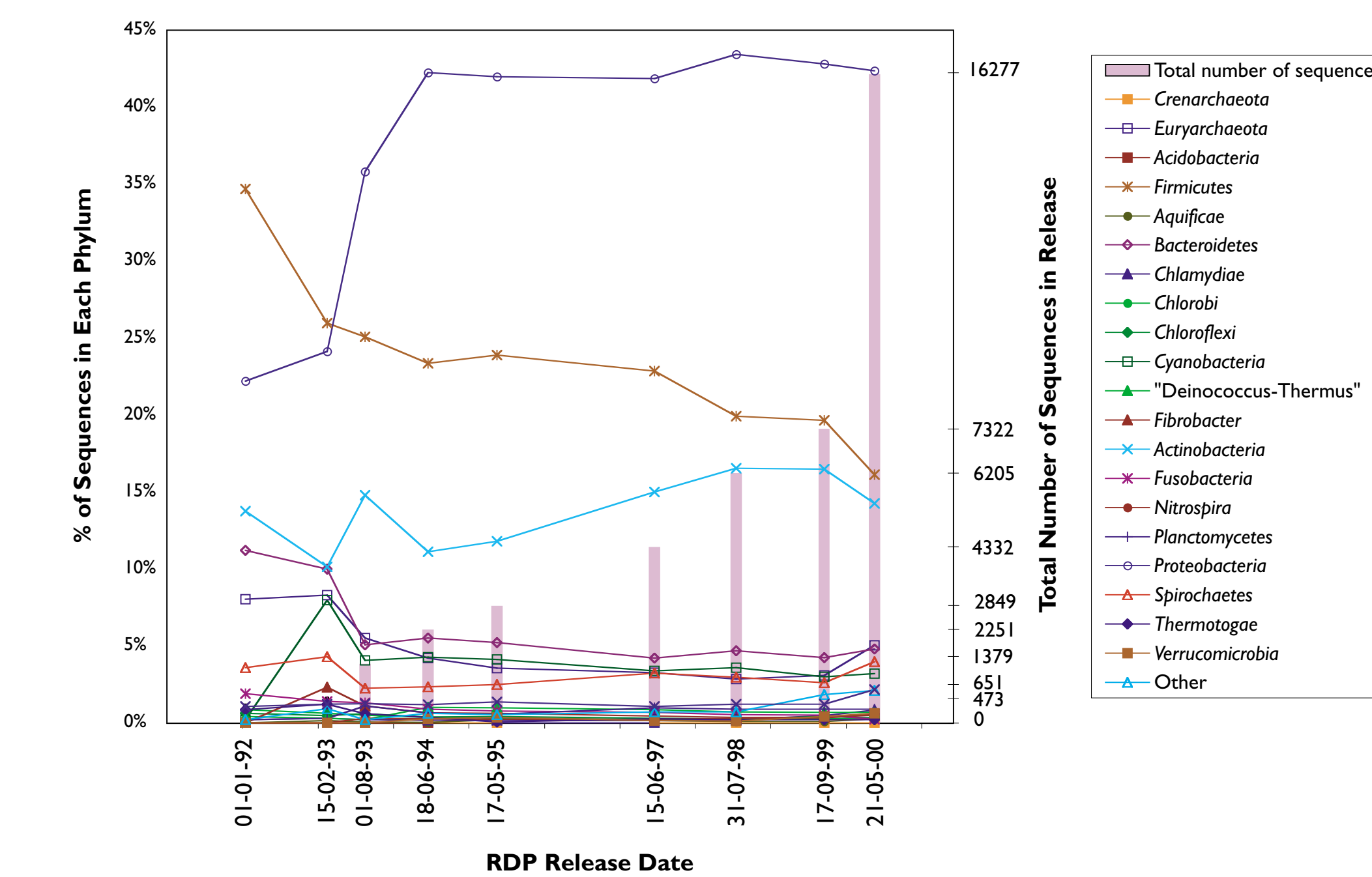


Figure 3. A breakdown of each RDP release by Bergey's phylum. The bars show the total number of prokaryotic sequences in the release (right axis) and the points show the percentage (left axis) of sequences in a given release that fall into the phyla.

A tree designed to show the phylogenetic breadth of the sequences in the RDP-II database is at the right of this poster. The role of environmental clones in the expansion of prokaryotic diversity is clearly seen; four clusters in the tree are mainly composed of environmental clones. We are also seeking new ways of visualizing and exploring our data. Figure 5 shows a principal component analysis (PCA) of the evolutionary distances between ~4,500 full-length sequences in Release 7.1. Each dot represents a single sequence and taxonomic or phylogenetic groups of sequences can be found by overlaying the prepared transparencies. When more than a few hundred sequences are involved, this way of viewing the data is more tractable than the traditional tree view.

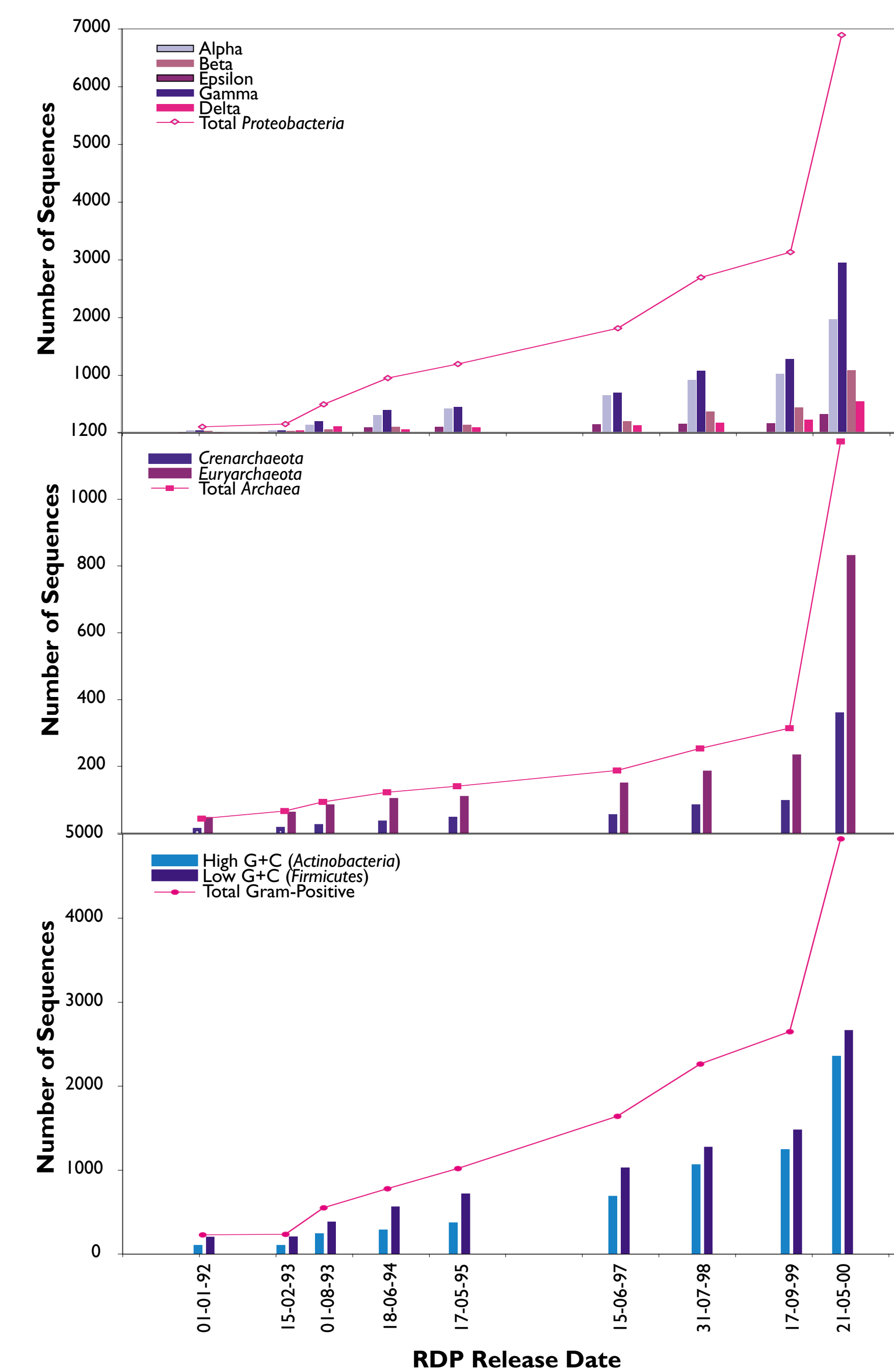


Figure 4. Breakdown of three widely-studied groups according to the representation of their subgroups in each release. Top panel: *Proteobacteria*, center panel: *Archaea*, and bottom panel: Gram-positive bacteria.

The RDP Today

RDP-II is widely used in microbial ecology, molecular phylogeny and evolutionary biology, organism identification, characterization of microbial populations, and studies of biodiversity (Figure 6). The single largest group of RDP users is from the USA, but the majority of users are from other countries (Figure 7). To better serve the international community, the RDP will also be hosted at a mirror site in Japan in the near future.

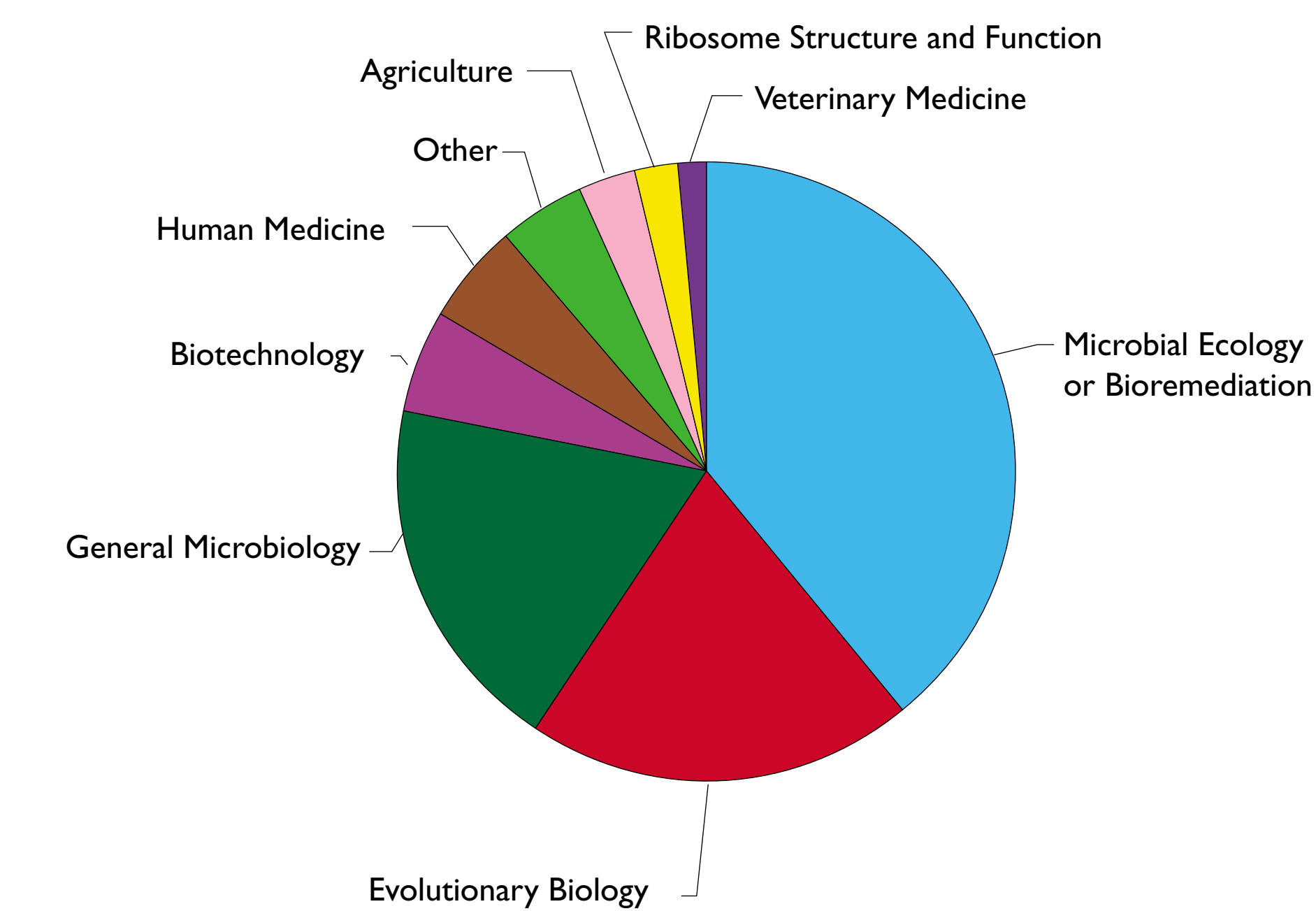


Figure 6. Breakdown of RDP users by primary field of study.

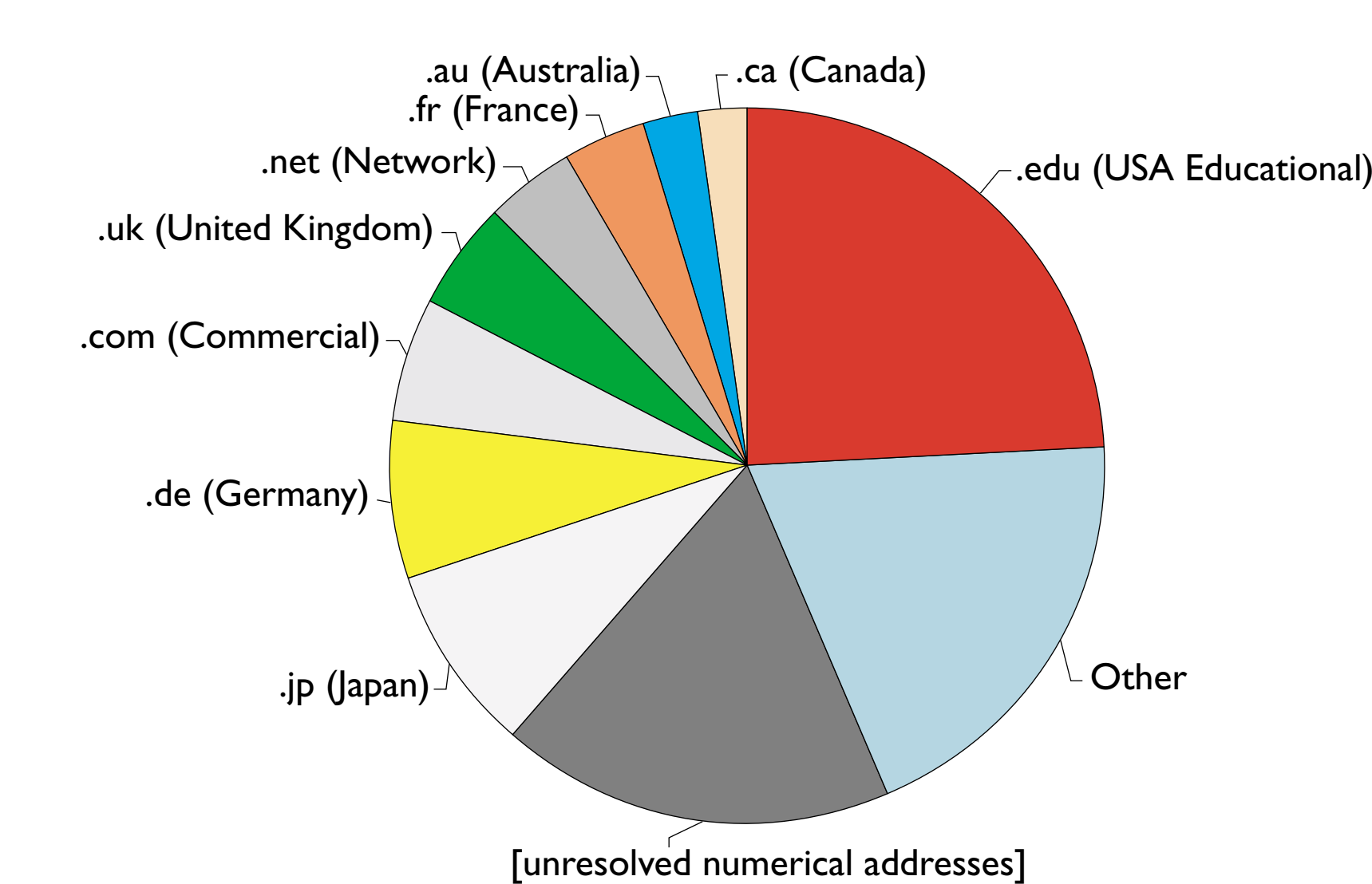


Figure 7. Breakdown of user requests to the RDP-II web site by domain name.

From the RDP home page (shown below), users can access online analysis tools, documentation pages that explain how to use the analysis tools and provide links to other useful WWW sites or go to the download area to get trees, alignments (including the prokaryotic alignment in an ARB-friendly format) or other data. The online analysis tools include similarity searches on the database, a chimera detection program, a probe match program and a Java-based facility for designing experiments using the terminally labeled restriction fragment length polymorphism (TRFLP) technique.

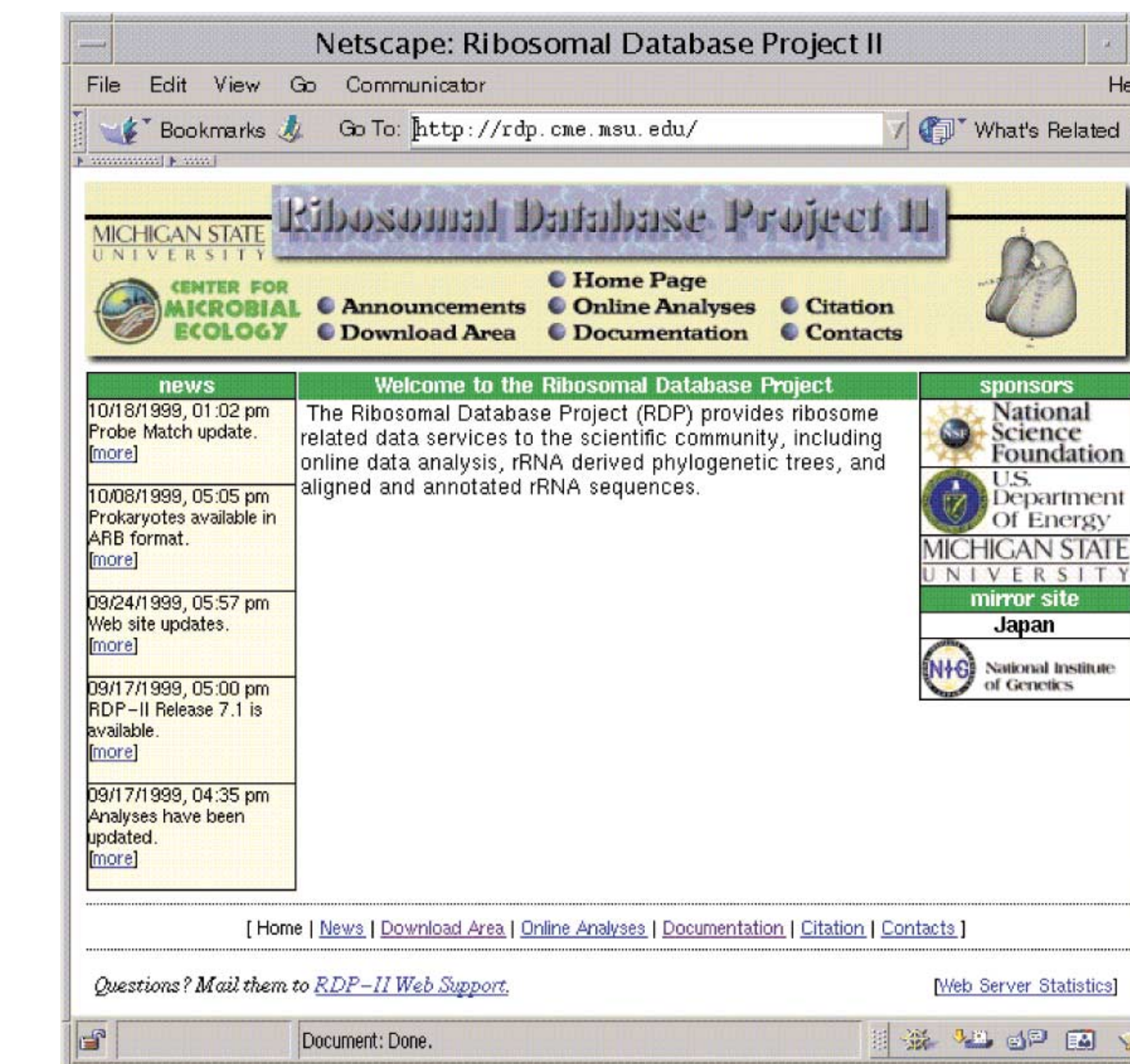


Figure 5. Exploring the aligned 16S sequence data. Evolutionary distance matrices were created for the roughly 4,500 sequences in Release 7.1 that were >1399 bases long and contained <4% ambiguities. The matrices were combined and subjected to a principal component analysis (PCA). The above figure shows the PCA plot of all the sequence data. Groups of prokaryotes can be located in the plot by overlaying the appropriate transparency.

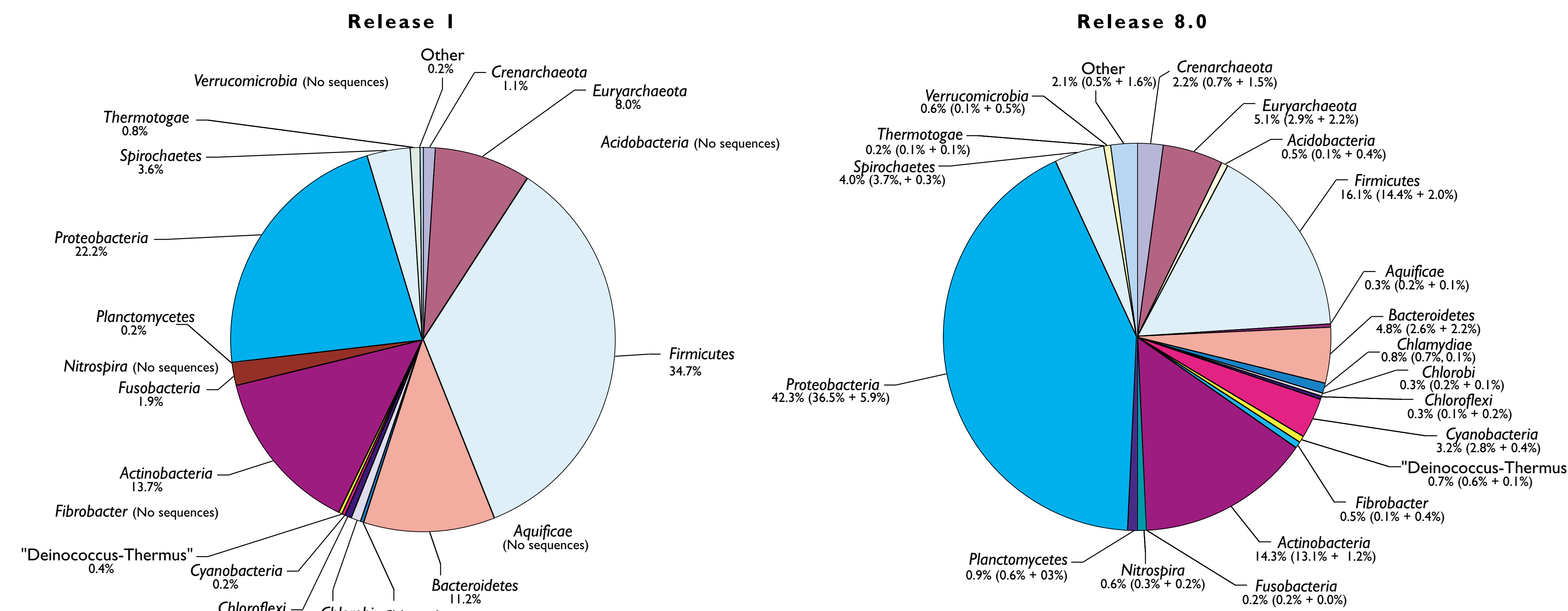


Figure 2. Charts showing the relative abundances (in per cent) of 20 phyla in RDP release 1.0 (left) and release 8.0 (right). In the brackets following the relative abundance, the relative abundance is broken down into the portion of sequences obtained from cultured organisms plus the portion of sequences amplified from the environment. Note that there were no environmental sequences in Release 1.0.

Release 8.0

RDP staff focused on aligning prokaryotic SSU rRNA sequences that were greater than 899 bp in length for Release 8.0. The number of new sequences (8955) more than doubles the number of sequences in the previous release and includes sequences for 130 new genera. To ensure that the greatest number of sequences would be aligned for Release 8.0, curation of the organism name, culture collection deposit information, type strain status, environmental clone characteristics, and reference was deferred until a later time. Until curation can occur, new sequences in Release 8.0 (for the most part) retain their GenBank accession number as the short identifier of the sequence.

The RDP serves the research community and thus relies on this community for its direction. We are interested in hearing from all of our users. If you have suggestions, comments or criticisms please contact us: curator@cme.msu.edu. Your ideas on everything from the look and feel of the web pages to the meaning of an alignment are welcome.

<http://rdp.cme.msu.edu/>

CONTACT INFORMATION