

The Ribosomal Database Project II

Expanding the Database and Enhancing User Services

T.G. Lilburn, J. R. Cole, C. T. Parker, P. R. Saxman, R. P. Farris, B. L. Maidak, T. M. Schmidt, G. M. Garrity, S. Pramanik and J. M. Tiedje

Michigan State University, East Lansing, Michigan

Introduction

The Ribosomal Database Project (RDP-II) (1) provides data, programs and services related to ribosomal rRNA. The RDP-II is a value-added database available to the research community through the RDP-II web site (<http://rdp.cme.msu.edu>). The RDP-II obtains the bulk of its rRNA sequences from the public nucleotide databases, but is more than a collection of sequences. The RDP-II organizes sequence data into alignments, annotates the sequence data, provides a phylogenetic context for the data and offers a suite of services and tools to assist in the handling and analysis of the data. Annotation includes up to date names, strain and culture deposit information, sequence length and quality information and references. An ongoing effort at the RDP-II is the improvement of alignments in view of recent research on the ribosome. RDP-II is also working on methods to incorporate higher taxonomic information. In order to provide a phylogenetic context for the data, RDP-II now has over 100 trees

that span the phylogenetic breadth of the data. With release 8.1 two major improvements to the tools provided to users will appear. The first is a revamped hierarchy browser, which will allow searching as well as visualization of sequence length and quality (the latter is calculated as the number of ambiguous bases in a sequence divided by the total number of bases). The second new tool will be an interface to PHYLIP that will allow users to calculate a distance matrix using DNADIST and to infer a tree from this matrix using NEIGHBOR, a neighbor-joining algorithm. In addition, release 8.1 will more than double the number of aligned SSU rRNA from Eukaryotes to over 5000 sequences, the bulk of which are microbial in origin. Finally, an on-line tutorial has been developed to introduce new users to the RDP-II and phylogenetic inference.

The Hierarchy Browser

The Function Buttons
Options - users can turn off the sequence snapshot and sequence count displays.
Browse - users select sequences from the hierarchical display using the mouse.

Search - Users find sequences by searching on the short identifier and the full sequence description. Found sequences are automatically added to the selected sequences set.

Download - Users can download the selected set of sequences in aligned or unaligned format and in one of four formats (GenBank, Fasta, Phylip interleaved, Phylip sequential).

The Hierarchy Number
 Each sequence is assigned to a position in the RDP hierarchy. Each level of the hierarchy is numbered. The hierarchy is not strictly taxonomically or phylogenetically organized, but we strive for consistency.

The Sequence Snapshot
 The icon to the right of the select button represents the length and quality of the sequence. White regions contain no sequence data and a red icon color indicates that the sequence has more than 5% ambiguities in its sequence.

Sequence View and Add
 The two buttons to the right of the short identifier allow one to add that sequence to the user's selection file (green) and to view the sequence file (blue). Once a sequence is selected the button reads Del and is red.

User Enhancements

The Hierarchy Browser

Shown above is a screen shot from the new Hierarchy Browser. This tool allows users to select a set of sequences from the RDP-II database for downloading or for use with some of the on-line tools provided by the RDP. Users can select sequences either by browsing the hierarchy or by doing a key word search on the RDP-II short identifier and full sequence description. The sequence snapshot icon gives the user an idea of the length and quality of the sequence, which can be important if the sequences are to be used in phylogenetic analyses. Users can also keep track of how many sequences they have selected. The information in the Hierarchy Browser also helps users avoid selecting more than one sequence from the same organism, select sequences from type material only and so on. It is worth noting that users can add their own sequences to the selected set of sequences, too. It is possible to download the sequences as is, that is, the RDP alignment can be preserved (this will allow the alignment to be maintained between sets of sequences downloaded at different times), with gaps common to all of the sequences in the selected set removed, or with all gaps removed. Finally, one of four data formats can be chosen for the downloaded sequences: GenBank, Fasta, Phylip interleaved or Phylip sequential.

The Tree Builder

To the left is a series of screen shots of the new tree building tool. The tool is a web interface for the programs DNADIST and NEIGHBOR, which were written by Joe Felsenstein as parts of his PHYLIP package (2). Users can infer trees using their own sequences, sequences from the RDP-II or a combination of both. After selection of a sequence data set, an evolutionary distance matrix is calculated, using a correction based on one of several evolutionary models. The tree is then built using UPGMA or neighbor joining. The completed tree is displayed on the web page and may be edited and then downloaded to the user's local machine in PDF, PS, EPS, or Newick file format.

The Start Page

Function Buttons

Users go to different steps in the tree building process using these buttons.

Data Set

Select the starting data set.

Tree File Number

Trees built by users are saved under a unique identifier for about one week after the last use. This allows users to easily return to the same tree session.

The Distance Matrix Page

Alignment Mask

The positions included in the calculation are graphically represented as blue bars and the number of positions is given.

Format the File for Download

Format the Matrix
 The matrix can be square or upper or lower triangular. Evolutionary distances or similarities can be displayed.

Evolutionary Models

The evolutionary model can be chosen and two of the parameters can be set.

The Tree Builder Page

Edit the Tree

It is possible to rotate nodes on the tree and re-root it as well. Distances can be displayed on the branches and the size of the tree in the viewer can be changed.

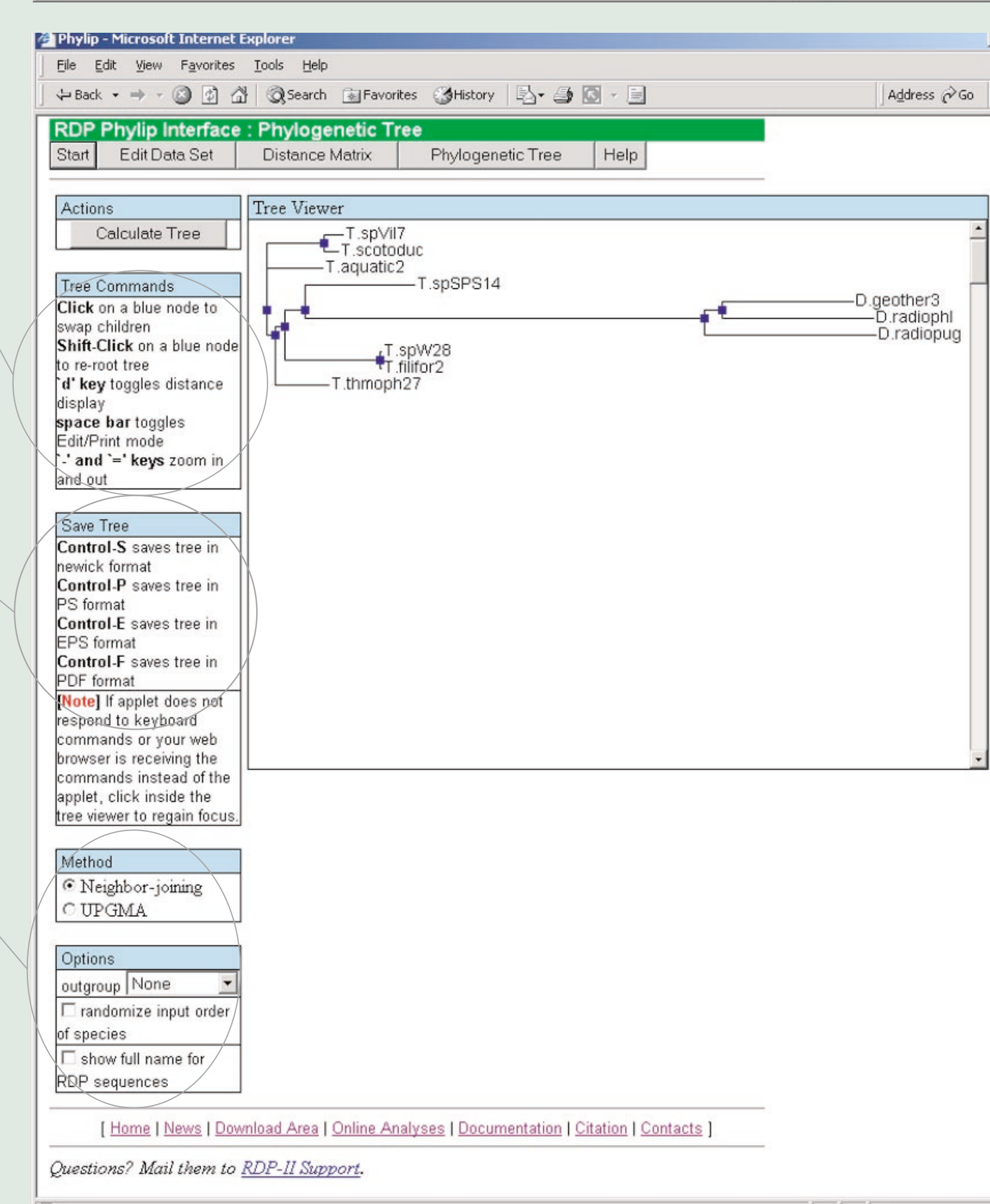
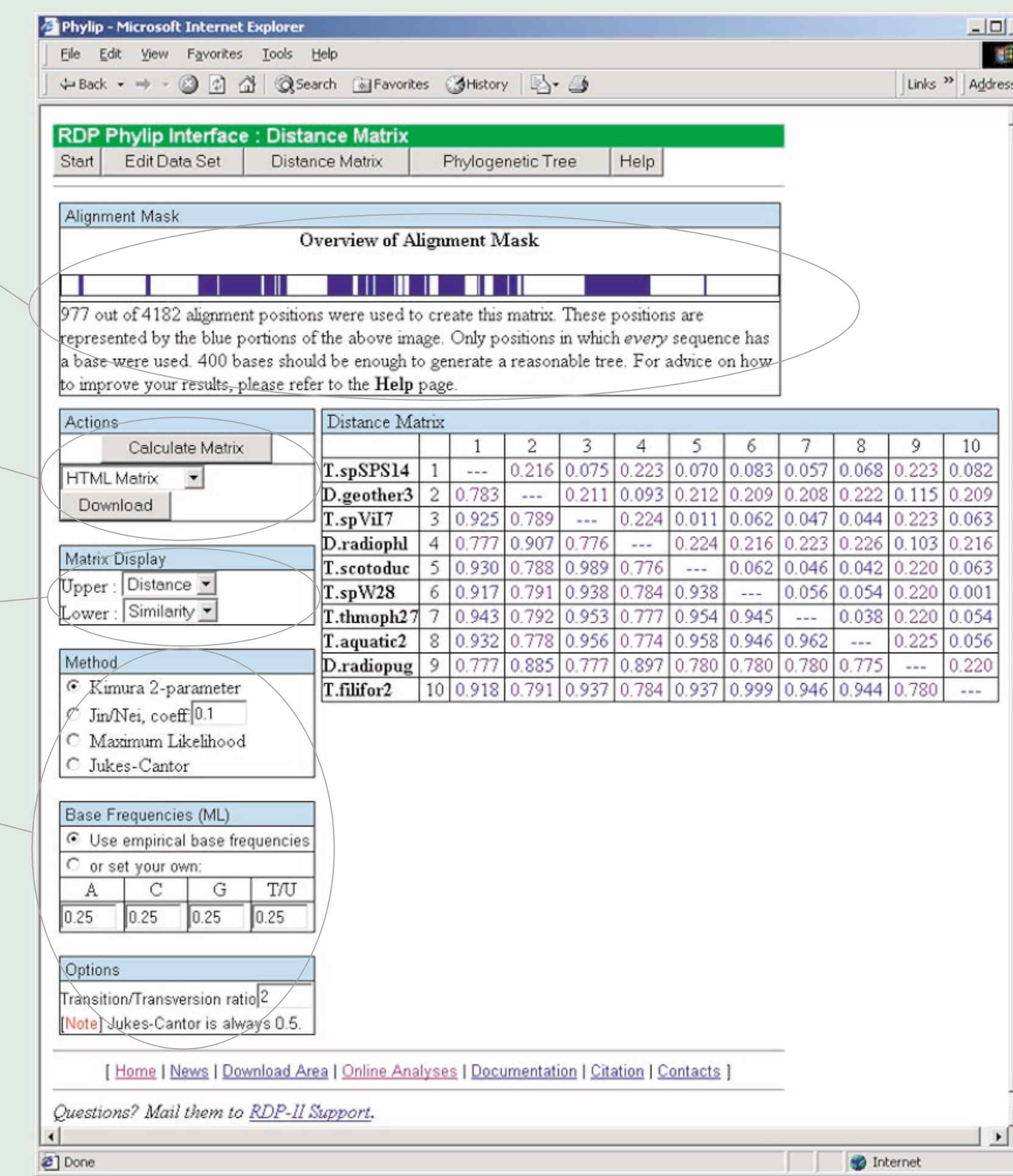
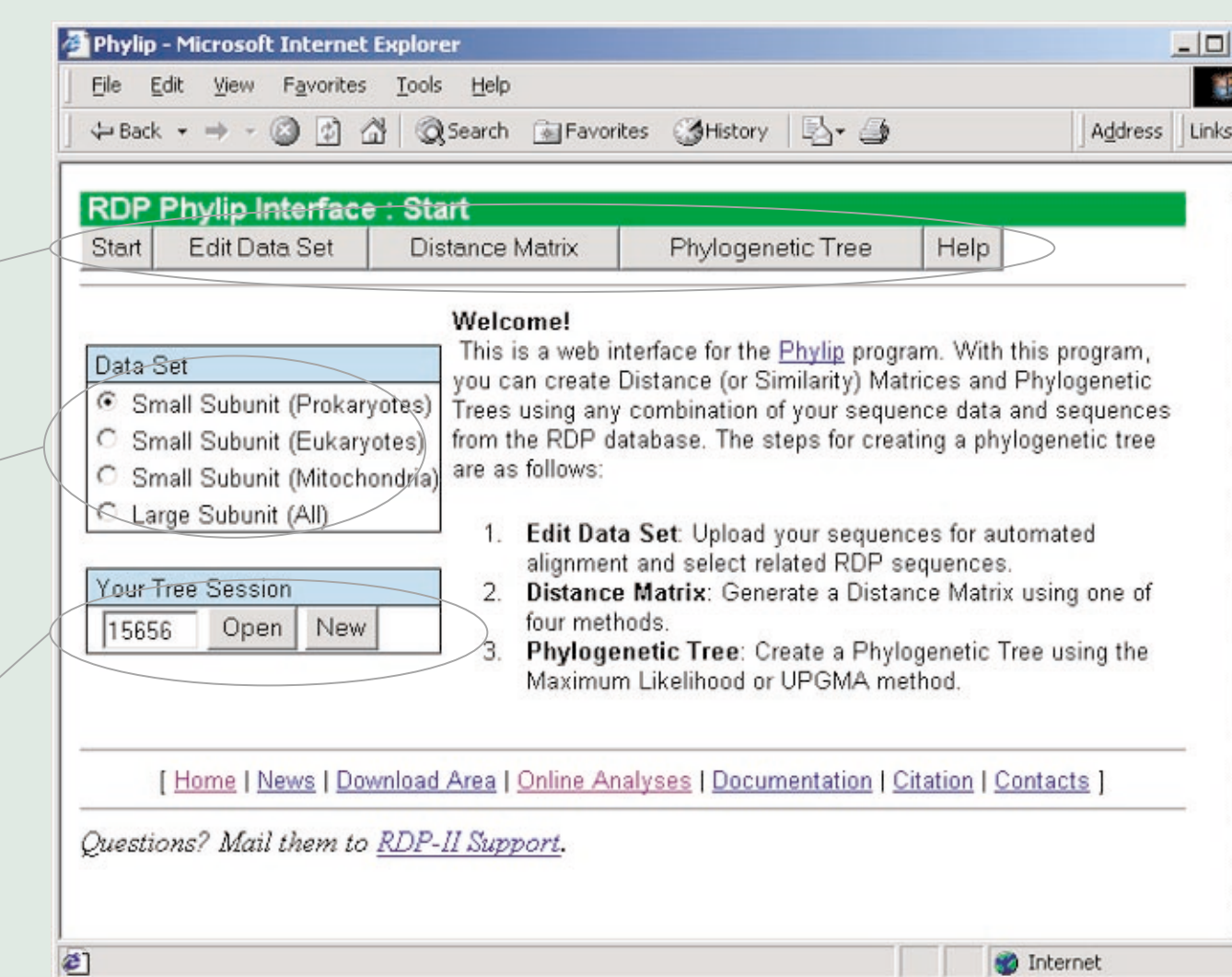
Save the Tree

The tree can be saved in four file formats, including the Newick format that allows the tree to be imported into other tree viewers or into other phylogenetic inference programs such as PAUP®.

Tree Building Options

The method used to build the tree is selected here. Outgroup selection and randomization of sequence addition is also selected here. Users may also choose the style of leaf label here: full name (when available) or RDP-II short identifier.

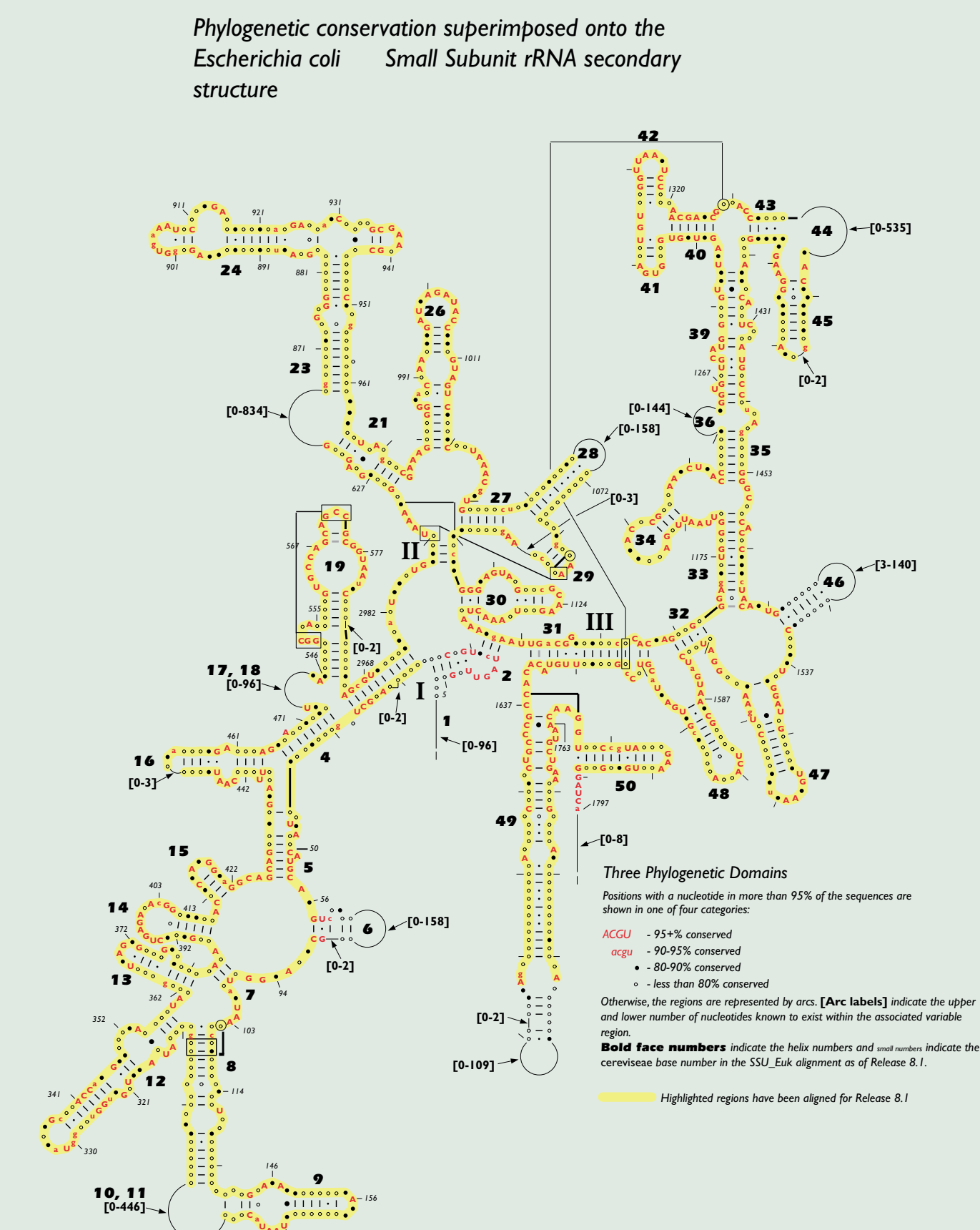
The Tree Builder



New Data for Release 8.1

Through the efforts of Dr. Scott Dawson at the University of California at Berkeley, the number of aligned Eukaryotic small subunit rRNA sequences in the RDP-II database has increased from 2,055 to 5,201. Due to uncertainties in the secondary structure of Eukaryotic SSU rRNA, the sequences have been aligned only in the most conserved regions. The aligned regions are highlighted in the secondary structure diagram below. In Release 8.1, the Eukaryotic hierarchy is based on the NCBI taxonomic information found in the GenBank sequence record, rather than on the clustering seen in a phylogenetic tree. In the future, we hope to incorporate SSU rRNA-based phylogenetic information in the Eukaryotic hierarchy as well.

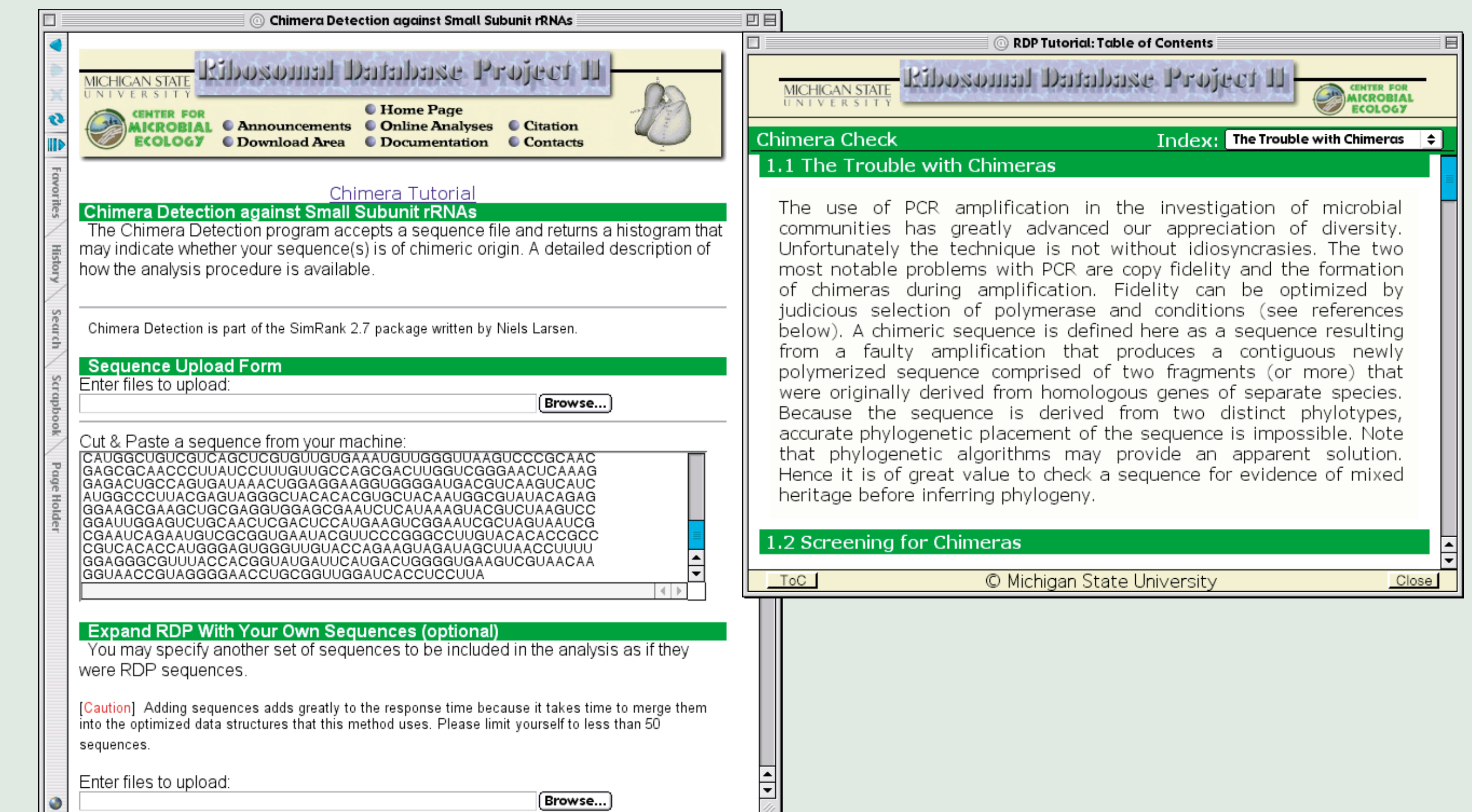
Release 8.1 also includes over 100 phylogenetic trees spanning the breadth of Prokaryotic diversity. The trees were inferred using Weighbor from distance matrices calculated using RDP-II alignments. The trees include only sequences from the database that are more than 1399 bases long and that contain less than 4% ambiguous nucleotides. The sequences used in each tree were selected according to their placement in the RDP-II hierarchy and it is apparent from some of the trees that a number of sequences are misplaced. As with the Eukaryotes, improving the consistency and reliability of the hierarchy is an ongoing concern.



On-line Tutorial

An on-line tutorial on the use of the RDP-II web site has been developed. It is intended as an introduction to the RDP-II for new users and as an instructional module for teachers of senior and graduate level courses. The

tutorial appears as a pop-up window, shown below, and as users follow the tutorial they are able to use the analysis tools with supplied example data or with their own data.



Under Development

In order to keep pace with the increase in sequence data, RDP-II is developing the means for automating work flow. One of the most time-consuming tasks is sequence alignment. In collaboration with Michael Brown, we are adapting RNACAD (4), a stochastic context free grammar modeling system for SSU rRNA structure prediction, as an alignment builder. We have tested it on sequence data sets of up to 1,000 sequences and the initial results have been promising.

We are also investigating alternative methods of visualizing the relationships between the sequences in our database, both for our own and our

users purposes. An ordination method, principal components analysis, has been used to analyze distance matrices containing over 9,500 sequences. The clustering patterns are visualized in three dimensions and clustering in higher dimensions can also be determined. The plot at left is a top-down (center) and rotation about (petals) a PCA analysis result using the 7,000 plus sequences in the RDP-II that are more than 1399 bases long, have less than 4% ambiguous nucleotides and are associated with a named Prokaryote. This approach may also present an automated approach to the establishment of the RDP-II hierarchies.

References

1. B. L. Maidak, J. R. Cole, T. G. Lilburn, C. T. Parker Jr, P. R. Saxman, J. R. Farris, G. M. Garrity, G. J. Olsen, S. Pramanik, T. M. Schmidt and J. M. Tiedje. The RDP-II (Ribosomal Database Project). *Nucleic Acids Res.* 29:173-174 (2001).
2. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author: Department of Genetics, University of Washington, Seattle (1993).
3. Gutell, R. R., S. Subashchandran, M. Schnare, Y. Du, N. Lin, L. Madabusi, K. Muller, N. Pande, N. Yu, Z. Shang, S. Date, D. Konings, V. Schweiker, B. Weisner, and J. J. Cannone. Comparative Sequence Analysis and the Prediction of RNA Structure, and the Web. Manuscript in preparation.
4. Michael P. S. Brown. RNA Modeling Using Stochastic Context-Free Grammars. Ph.D. Thesis, University of California, Santa Cruz (1999).