Spruce – Metadata Miner Walkthrough

1. Introduction

Spruce is a collection of tools to retrieve metadata and organize it according to a user-defined ontology of terms.

A primary application is retrieving metadata from NCBI GenBank format records.

For today's walkthrough, we will use a small set of GenBank records already retrieved from GenBank. If you would like to retrieve records for your own study, please see section titled "Advanced" below.

We use a set of terms to categorize the metadata. We will evaluate these terms in a comparative phylogenetic analysis by using Arbor

The sample files are from Mollicutes, a group of bacteria associated with plant and animal hosts. We will assign tags based on the host identity, as plant, vertebrate, or invertebrate. We can then use Arbor to test whether there is a phylogenetic signal in these host associations.

2. Download the required files

To complete this walkthrough, you will need to download the following files from https://thackerlab.weebly.com/phenomics.html

- mollicutes.example.metadata.txt
- sample.host.tags.txt
- tagit5.pl.txt
- mollicutes.matched.phy

Note: Usually, this last file would be "tagit5.pl" but weebly forced a ".txt" at the end of the filename.

Put all of these files together in a folder (or directory); for example, in a folder located on your desktop called "Phenomics".

3. Open a terminal window

To run the Perl scripts, first open a terminal window. On a mac, use the Finder > Go > Utilities > Terminal.app to reach the terminal window. On a pc, search for "cmd".

Next, navigate to the Phenomics folder, for example with "cd Desktop/Phenomics"

4. Add tags to the metadata.

Run the “tagit” script using Perl and the following format:

> “perl tagit5.pl.txt inputFileName tagListFileName outputFileName”

or

> perl tagit5.pl.txt mollicutes.example.metadata.txt sample.host.tags.txt example.output.txt

You might need to make the terminal window larger by dragging the corners.

The script will provide the number of each tag assigned to a record. In this case, 50 records were tagged as plant, 25 as vertebrate, and 17 as invertebrate. However, 68 records remain to be tagged.

The script provides the next 5 records that need to be tagged. What terms could you use to make sure these records are correctly tagged? The tagit script uses a tag list (“sample.hosts.tags.txt”) that you can edit to make sure every record is tagged.

To edit, simply open the tag list and add additional terms under the appropriate tags.

Be sure to save the tag list, then re-run the tagit script. The tags can also be inspected in the output file, “example.output.txt”.

Once you are satisfied with your tags, you can export a .csv file that can be read by Arbor.

5. Analyze the Tags in Arbor

To visualize the tags on a phylogeny of Mollicutes, we can use the Arbor app for ancestral state reconstruction: http://52.204.46.78/ancestral-state/

Once you have opened the Arbor app, drag the Mollicutes phylogeny (mollicutes.matched.phy) and your .csv file (created in step 4) onto the App. Drag the “tag” column to select it for analysis. Press go and you should find a visualization of your tags.

To test for phylogenetic signal, use this Arbor app:
http://52.204.46.78/phylogenetic-signal/

You will again need to drag your phylogeny and .csv file onto the app, then select the appropriate column for analysis. Press go to fit a maximum likelihood model to your data using the fitDiscrete algorithm.

For a full Arbor tutorial, look here: http://www.arborworkflows.com/tutorials/arborapps/

6. Make a New Tag File for Additional Analyses

You can create your own set of tags for other analyses, such as host tissues. Using the ***Advanced*** instructions below, you can first count all tissue types to determine a strategy for categorizing them.

7. Troubleshooting

The most common issues derive from the variability of line breaks across Unix/Linux, Mac, and PC platforms. These scripts work best with Unix line breaks.

Text editors that are useful for editing these files are TextWrangler for Mac and Notepad++ for PC. Both are free of charge.

8. Arbor Expert Mode

If you want to try Arbor's primary interface, look here: http://52.204.46.78/

Advanced

Preparing your computer

To run the Metadata Miner Perl scripts, mac users will be fine using the default Perl on their mac. Windows users will need to install a Perl compiler, such as Strawberry Perl, available for free here: http://strawberryperl.com/

The Perl script that extracts information from GenBank format records BioPerl. To run this script, users need to install BioPerl: http://bioperl.org/INSTALL.html

Download the following Perl scripts from https://thackerlab.weebly.com/phenomics.html

getmetadatagb.pl.txt
counttags_gb.pl.txt
counthostsandsources.pl.txt

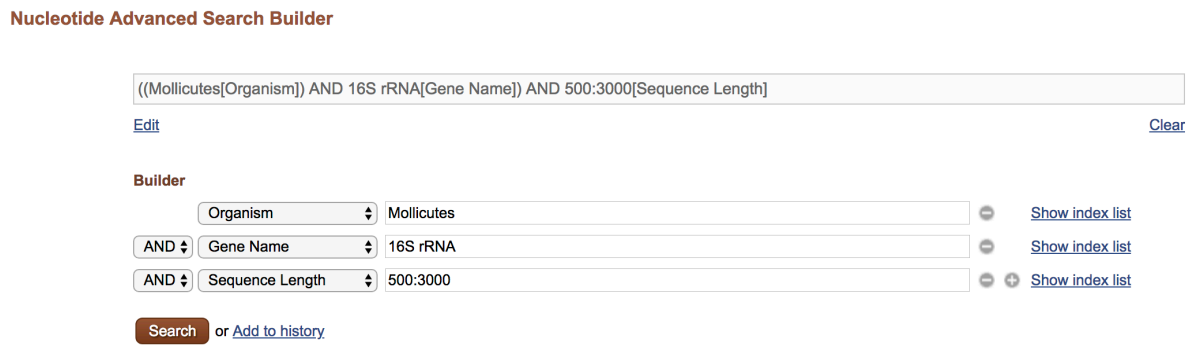
Retrieving your own records from GenBank

9. Search the GenBank Nucleotide database using the Advanced Search Builder

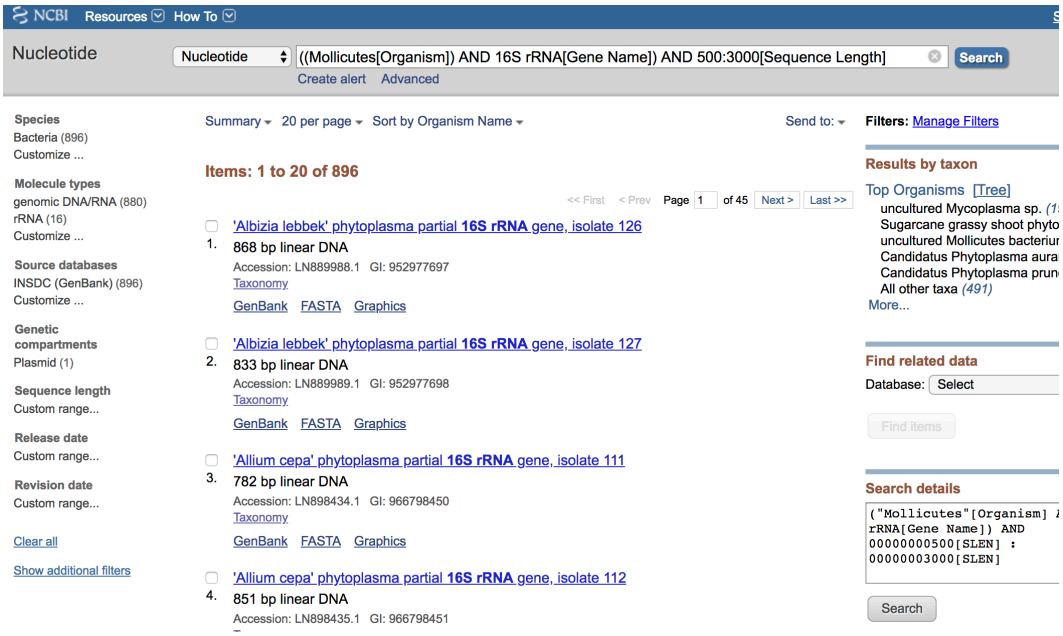
Use this link: https://www.ncbi.nlm.nih.gov/nuccore/advanced

Commonly used fields include Organism, Gene Name, and Sequence Length.

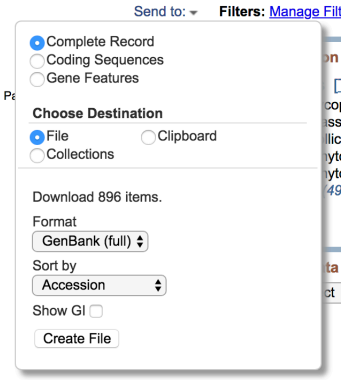
For example, you could search:
(Mollicutes[Organism] AND 16S rRNA[Gene Name]) AND 500:3000[Sequence Length]



This search will find 896 records (as of May 22, 2018).



If you click on “Send to:” in the upper right of your screen, you can see how to download the GenBank format records. After clicking “Send to:”, choose “File”, then choose Format “GenBank (full)” and choose your favorite “Sort by” (e.g. “Accession”). Then click “Create File”.



NCBI will create a file in your Downloads folder called “sequence.gb”. Feel free to rename the file something more informative, like “Mollicutes.May18.gb”

10. Examine the metadata provided in the GenBank records.

Each record contains metadata delimited by tags such as “/host” and “/lat_lon”, i.e. for the name of the host organism and the latitude/longitude of the collection location. If you are most interested in hosts and isolation sources (“/isolation_source”), you can obtain a count of these by using a Perl script.

The Perl script “countTags_GB.pl” will produce a list of all tags used in your GenBank format file. Run this script in your terminal window by entering:

```
perl counttags_gb.pl.txt Mollicutes.May18.gb
```

The Perl script “countHostsAndSources.pl” will produce a list of all hosts and isolation sources found in your GenBank format file.

```
perl counthostsandsources.pl.txt Mollicutes.May18.gb
```

Both scripts will give you the option of writing the output to a text file.

11. Extract the metadata from the GenBank records.

Note: this script requires BioPerl. See above under ***Preparing your computer*** for a link to install BioPerl.

To extract the metadata from GenBank format into tab-delimited text format, run the Perl script in your terminal window by entering:

```
perl getmetadatagb.pl.txt Mollicutes.May18.gb
```

This script will produce a file called “Mollicutes.May18.gb.metadata.txt”

12. Add user-specified tags to the extracted metadata.

By following steps 4 and 5 above, you can add your own customized tags to your metadata and analyze them in Arbor.