**Adding new reference datasets to MLTreeMap:**

MLTreeMap is a software framework, designed for phylogenetic and functional analysis of metagenomic data. It searches for instances of marker genes on nucleotide sequences and deduces their most likely origin in a set of reference phylogenies. Part of this set are tree-of-life phylogenies and several functionally important gene families. This guide explains how to add additional reference phylogenies to MLTreeMap.

To produce sequence placements in a reference phylogeny, MLTreeMap needs 4 types of reference files:

- **alignment files**: these files contain the aligned reference sequences (proteins, not DNA).

- **hmm files**: a corresponding hmm file belongs to each alignment file.

- **tree files**: the files with the reference phylogeny in Newick-tree format.

- **translation file**: sequence names within the alignment files and the tree files have to be represented by numbers. The translation file allows MLTreeMap to translate those numbers back to names.

To create these files and integrate them into MLTreeMap follow the instructions below:

**Step 1:**

Choose a name for your new alignment. It has to be 7 characters long.

In the following examples I will use “markers” as name. You can exchange it with any wording (as long as it is 7 characters long), but if filenames are concerned, the rest should be kept as it is given in the guide.

Choose your marker genes. Create a file called tax\_ids\_markers.txt. In this file you assign a number to each marker gene name (separate them with a tab).

e.g.

1 marker1

2 marker2

etc.

Now rename your marker genes as follows:

marker1 becomes 1\_markers (i.e. its number, followed by an underline and the name of the entire alignment. MLTreeMap will need this information for parsing later on).

**Step 2:**

Align the marker genes in FASTA format (we use MUSCLE to do so).

As a result you should get a file somewhat like this:

>1\_markers

----MATNNVV---SELYQLA

>2\_markers

---MMATTNVV----ELYQLA

etc.

Name the file according to the alignment name, which you have chosen in step 1 and ad “.fa”. In our example the name would be “markers.fa”.

Format this file, using formatdb (available from NCBI). This is necessary for the BLAST step of MLTreeMap.

./formatdb –i markers.fa

**Step 3:**

Use hmmbuild (available at <http://hmmer.org/>) to create the hmmfile:

./hmmbuild –s markers.hmm markers.fa

**Step 4:**

Use a software (for consistency with the MLTreeMap pipeline preferably RAxML) to create a phylogenetic tree based on your alignment. Save this tree in Newick format.

Let us call it “markers\_tree.txt”. It should look somewhat as follows:

((1:0.333, 2:0.323),…);

**NOTE:** The tree has to be rooted. If you have an unrooted tree, you can easily root it manually (e.g. (A,B,C); becomes ((A,B),C);) or by using iTOL (http://itol.embl.de/upload.cgi).

**NOTE 2**: The sequence names within the tree should be represented by their number, which you defined in step 1 (i.e. marker1 is 1, marker2 is 2 etc…).

**Step 5:**

Copy the files to the following places in the MLTreeMap directory:

cp markers.fa MLTreeMap\_home/data/alignment\_data/

cp markers.fa MLTreeMap\_home/data/geba\_alignment\_data/

cp markers.hmm MLTreeMap\_home/data/hmm\_data/

cp markers.hmm MLTreeMap\_home/data/geba\_hmm\_data/

cp markers\_tree.txt MLTreeMap\_home/data/tree\_data/

cp tax\_ids\_markers.txt MLTreeMap\_home/data/tree\_data/

Further, an entry is needed in the file MLTreeMap\_home/data/tree\_data/cog\_list.txt:

To the last section (#functional\_cogs) you add a line somewhat like this:

markers y

This information tells MLTreeMap, that there is an additional analysis to be done, based on the “markers” alignment. Any results for this analysis will begin with a “y\_” (e.g. y\_concatenated\_RAxML\_outputs.txt).

**NOTE**: This guide describes how to add a reference phylogeny, which is based on only one alignment file.

If you want to add a reference phylogeny based on several alignment files (as is the case for our “tree-of-life” phylogenies), it will be easiest to replace the phylogenetic cogs in “cog\_list.txt” with your new cogs and abolish the traditional “tree-of-life” analysis of MLTreeMap.