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 further reference phylogenies to MLTreeMap.
This guide can be downloaded from http://mltreemap.org/treemap_cgi/show_download_page.pl.

To produce sequence placements in reference phylogenies, 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**: these files contain the reference phylogenies in Newick-tree format.
- **Translation files**: sequence names within the alignment files and the tree files have to be represented by numbers. The translation files allow MLTreeMap to change 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 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 numbers, 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) add a line similar to this (tab delimited):

markers z

This information tells MLTreeMap that there is an additional analysis to be done, based on the 'markers' alignment. The names of the files containing results for this analysis will begin with 'z_' (e.g. z_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 (similar to our 'tree-of-life' phylogenies), it will be easiest to replace the phylogenetic cogs in 'cog_list.txt' with your new cogs and thus to abolish the traditional 'tree-of-life' analysis of MLTreeMap.