Feb 04, 2020

Phylogenetic tree and ancestral sequence reconstruction

Matthew Kellom[1]

[1]University of California, Santa Barbara

In Development dx.doi.org/10.17504/protocols.io.6qrhdv6

Santoro Lab @ UCSB

Matthew Kellom
University of California, Santa Barbara

ABSTRACT

Protocol for the alignment, tree building, and ancestral sequence reconstruction of metagenomic protein coding sequences.

GUIDELINES

This protocol was used specifically for AMT ammonium transporter sequences from TARA Oceans and HOT/ALOHA datasets.

MATERIALS TEXT

Programs used:
DIAMOND
TranslatorX
MAFFT
RAxML
IQ-TREE
MrBayes
FastML (Web)
FigTree

BEFORE STARTING

Make sure all of metagenome sequences are in FASTA nucleotide format. Target amino acid sequences should also be in FASTA format.

1 Obtain protein-coding metagenome FASTA files.

2 Obtain a set of amino acid sequences that are known/trusted target homologs (Uniprot works well for this).

3 Use DIAMOND (blastx can also be used but will be slower for large datasets) to search for target hits from within the metagenome sequences. Recommend an e-value cutoff of at least 1e-20.

4 Translate protein-coding target hits to amino acid sequences with TranslatorX.

5 Use DIAMOND (or BLAST) to annotate hits against a local copy of the NCBI nr database.

6 Choose taxa of interest, and select sequences of appropriate length for the protein from the metagenome hits.

7 Align the collected sequences and closely related complete outgroup sequence(s) with MAFFT. Recommend settings of --maxiterate 1000 --localpair

8 Make tree with RAxML. Recommend settings -f a -k -m PROTGAMMAAUTO

Inspect the tree to see if it makes sense. If a node looks out of place, consult the alignment and figure out why and if you want to keep it in the dataset.

9 Check RAxML amino acid substition model and branch support with IQTree. Recommend settings -m TEST -nt AUTO -alrt 10000

10 Branch support with MR Bayes. Will need to create a nexus and mbatch file. Recommend mbatch settings:
lset nst=6 rates invgamma;
prset aamodelpr=fixed(lg); (The model choice here should match the model that RAxML and IQTree select.
mcmcp ngen=1000000 nruns=2 nchains=4 samplefreq=100 printfreq=10000 relburnin=yes burninfrac=0.25;
mcmc;
sumt;
sump;

11 Submit the MAFFT alignment and RAxML best_tree to the FastML web portal (http://fastml.tau.ac.il/). You can use the local version of this but I found it easier to just use the web portal. Be sure to select Amino Acids and the correct model from the tree building. Use Gamma Distribution and Maximum Liklihood. Deselect Optimize Branch Lengths and joint reconstruction.

12 Obtain the 25 most probable ancestral sequences at the nodes of interest from the FastML results page and visualize the tree results in FigTree.