

TreeRefiner User Manual

Download and Installation:

- 1) Download the TreeRefiner source code from
<http://treerefiner.stanford.edu/download.html>
- 2) Decompress the files. Use the following command:
tar xvzf treerefiner_v1.tgz
- 3) This will create a sub-directory named treerefiner_v1/ inside the current directory
- 4) Change to the treerefiner_v1/ sub-directory and make the Treerefiner executable.
cd treerefiner_v1
make
- 5) This will create the executable file called treerefiner.

Using TreeRefiner:

To run TreeRefiner please use the following command.

```
./treerefiner <Input Alignment Filename> <Phylogenetic Tree Filename> <Substitution  
Score Filename> <radius>
```

Each input parameter is explained in detail below:

Input Alignment File:

Treerefiner requires files to be text files. Hence it will not work with .doc files from Word or other word processors. Most processors allow the user to save as a text file by selecting "Save As" from the file menu.

The Alignment file must be in the MFA format. The MFA format is described below:

- The MFA format consists of multiple sequences.
- Each sequence begins with a single line descriptor followed by lines of sequence data
- The sequence descriptor line begins with the '>' character

An example of such a file would be:

```
>human_aligned:  
AAA---GGGGTTCGCGCGC-----GTCTCT-GT
```

```

>baboon_aligned:
AAAA---GGGTTC--CGCGGGG---TCTCTGG
>mouse_aligned
TTCTAA---GGTTCCTCTC---AAATTTCTG
>rat_aligned
TTCTAAAGGG-----CGCGCGAAATT---CTG

```

Phylogenetic Tree File:

Treerefiner requires that the tree be specified in Newick format. For example the tree for the above alignment file would look like:

```
((human, baboon)(mouse, rat));
```

Also, please note that the species names specified in the tree must be substrings of the sequence descriptor names in the alignment file. Thus you cannot have something like '>human' in the alignment file and 'human-being' in the tree file.

Substitution Score File:

Treerefiner requires that the substitution scores be read from a file. For your convenience, a default substitution score file called 'nucmatrix.txt' is already provided. The file looks as follows:

	A	C	G	T	N
A	91	-114	-31	-123	-43
C	-114	100	-125	-31	-43
G	-31	-125	100	-114	-43
T	-123	-31	-114	91	-43
N	-43	-43	-43	-43	-43
	-500	-25			

The numbers above represent the substitution scores between every pair of nucleotides. The matrix is required to be in the format shown above. Specifically the order A,C,G,T,N should not be changed. The two numbers below represent the gap open/close penalty and the gap extension penalty respectively.

Radius:

This represents the radius around the input alignment in which you want Treerefiner to perform its optimization. Radius must be a positive value (> 0). Higher radii will result in slower running times.

Output:

The output file will be named InputAlignmentFile.rfn. The extension '.rfn' would be appended to the input filename. Treerefiner generates its output in MFA format. The order of sequences in the output file would be the same as the input file. The sequence descriptors also would be the same.

For example, the command

```
./treerefiner input.mfa input.tree nucmatrix.txt 2
```

will produce the output file input.mfa.rfn