

1. Extracting species

`./extract_relevant_species.pl`

Extract relevant species from a set of FASTA files. For this study, we obtained sequences from the following data sources:

http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/current_GREENGENES_gg16S_unaligned.fasta.gz

http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/current_RDP_gg16S_unaligned.fasta.gz

http://www.arb-silva.de/typo3conf/ext/myth_repository/secure.php?u=0&file=fileadmin/silva_databases/release_104/Exports/SSURef_104_tax_silva.fasta.tgz&t=1314717517&hash=8d9aa2264895a810350c9a84772cf1bc

The list of species can be specified by modifying the perl script `extract_relevant_species.pl`.

2. Filtering regions

`./isolate_multiregions.pl`

Isolates (multiple) regions where endpoints are best matches of user-specified k-mers. In particular, the user can isolate regions using the parameters: `--leftstr`, `--midstr`, and `--rightstr`. Their corresponding coordinates can be easily specified through:

<code>leftcoord</code>	literature left coordinate
<code>midcoord</code>	literature mid coordinate
<code>rightcoord</code>	literature right coordinate

The lengths of the regions are specified by the following:

<code>mid-trimneg=<INT></code>	count len from left of mid and then truncate
<code>mid-trimpos=<INT></code>	count len from right of mid and then truncate
<code>right-trimneg=<INT></code>	count len from left of 'right marker' and then truncate
<code>right-trimpos=<INT></code>	count len from right of 'right marker' and then truncate
<code>left-trimneg=<INT></code>	count len from left of 'left marker' and then truncate
<code>left-trimpos=<INT></code>	count len from right of 'left marker' and then truncate

The resulting regions extracted are then written to files specified using:

<code>mid-fsa=<FILE></code>	writes to file the filtered seqs of mid-trim
<code>left-fsa=<FILE></code>	writes to file the filtered seqs of left-trim
<code>right-fsa=<FILE></code>	writes to file the filtered seqs of right-trim

There are two additional parameters that can be used to enable flexibility/fuzzy matching of the left and right k-mers:

offsetleft=<INT> number of offset positions allowed for leftstr
offsetright=<INT> number of offset positions allowed for rightstr

In this study, we used the following parameters:

```
isolate_multiregions.pl --midstr='ACTCCTACGGGAGGCAGCA' --  
rightstr='GTCGTCAGCTCGTGYYG' --rightcoord=1061 --midcoord=338 --right-trimneg=258 --  
right-trimpos=0 --mid-trimneg=270 --mid-trimpos=0 --offsetleft=100 --offsetright=100
```

3. Removing exact duplicates within intra-species

`./filter_duplicates_within_intra_species.pl`

This script removes exact duplicates within a particular intra-species set. The *species-of-interest* is given through `--species` parameter. Use `--fsa` to specify the FASTA input from the previous step. In our study, this filtering is performed for both V2 and V6 regions across each of the 15 intra-species set.

4. Computing intra-species PID

`./compute_intra_species_pid_distrib.pl`

Compute the intra-species percentage-identity distribution using Needleman-Wunsch. This script parametrized the alignment algorithm with the default `blastn` parameters: match/mismatch score of 1/-2 and affine gap penalty open/extension of -5/-2.