Stable operational taxonomic units for studying microbial ecology

He *et al.*

**Box 1: Definitions of OTU clustering algorithms and executing commands, as used in this paper.**

**complete-linkage (CL) clustering**

Sequences are clustered such that the distances between all pairs of sequences in a cluster is less than or equal to the similarity threshold.

cluster(column=unique.dist,name=names,cutoff=0.05,precision=1000,method=furthest)

**single-linkage (SL) clustering**

Sequences are clustered such that each sequence in a cluster has a distance less than or equal to the similarity threshold to at least one other sequence in the cluster.

cluster(column=unique.dist,name=names,cutoff=0.05,precision=1000,method=nearest)

**average-linkage (AL) clustering**

Sequences are clustered such each sequence in a cluster has an average distance less than or equal to the similarity threshold to all other sequences in the cluster.

cluster(column=unique.dist,name=names,cutoff=0.05,precision=1000,method=average)

**dereplication**

Sequences that are 100% identical in length and composition are clustered together.

***de novo* clustering**

Cluster centroids are chosen from the collection of query sequences. Sequences are clustered with a centroid if the distance between the sequence and the cluster centroid is less than or equal to a user-defined similarity threshold.

**distance-based greedy** ***de novo* clustering (DGC)**

An approach for *de novo* clustering where, when a sequence matches more than one cluster centroid at less than or equal to the similarity threshold, the sequence is clustered with the closest centroid.

pick\_de\_novo\_otus.py -i fasta –o output\_dir –p dgc.params.txt

content in dgc.params.txt:

pick\_otus:otu\_picking\_method usearch61

pick\_otus:max\_accepts 16

pick\_otus:max\_rejects 64

pick\_otus:enable\_rev\_strand\_match True

pick\_otus:minlen 30

**abundance-based greedy** ***de novo* clustering (AGC)**

An approach for *de novo* clustering where, when a sequence matches more than one cluster centroid at less than or equal to the similarity threshold, the sequence is clustered with the most abundant centroid (i.e., the centroid sequence which has been observed more times in the data set).

pick\_de\_novo\_otus.py -i fasta –o output\_dir –p agc.params.txt

content in agc.params.txt:

pick\_otus:otu\_picking\_method usearch61

pick\_otus:sizeorder True

pick\_otus:max\_accepts 16

pick\_otus:max\_rejects 64

pick\_otus:enable\_rev\_strand\_match True

pick\_otus:minlen 30

**closed-reference clustering**

Sequences are clustered against an external reference database of centroid sequences, and assigned to a cluster if the distance between a sequence and a reference centroid is less than or equal to a user-defined similarity threshold. Sequences that fail to hit a reference centroid are discarded. When a sequence matches more than one cluster centroid at less than or equal to the similarity threshold, the sequence is clustered with the centroid to which it has the smaller distance.

pick\_closed\_reference\_otus.py -i fasta -o output\_dir –r reference\_dataset -p closedref.params.txt

content in closedref.params.txt:

pick\_otus:otu\_picking\_method usearch61\_ref

pick\_otus:max\_accepts 16

pick\_otus:max\_rejects 64

pick\_otus:enable\_rev\_strand\_match True

pick\_otus:minlen 30

**open-reference clustering**

Sequences are clustered against an external reference database of centroid sequences, and assigned to a cluster if the distance between a sequence and a reference centroid is less than or equal to a user-defined similarity threshold. Sequences that fail to hit a reference centroid are subsequently clustered de novo. The present study uses AGC for de novo clustering, as it is the most stable methods compared to other *de novo* methods.

pick\_open\_reference\_otus.py -i fasta -o output\_dir -m usearch61 -r reference\_dataset -s 1 -p openref.params.txt --min\_otu\_size 1 --prefilter\_percent\_id 0.0

openref.param.txt:

pick\_otus:sizeorder True

pick\_otus:max\_accepts 16

pick\_otus:max\_rejects 64

pick\_otus:enable\_rev\_strand\_match True