MEGAPTERA internal description

Christoph Heibl
Harztalstr. 29,
83714 Miesbach, Germany

August 5, 2016

Contents

1 Description of `step*` functions

1.1 stepA

1.

1.2 stepB

1. If `update.seqs = "all"` and if table `acc_<locus>` exists, i.e. if `stepB` has been run before, delete table `acc_<locus>` thereby triggering a thoroughly new search.

2. (Re-)create table `acc_<locus>` .

3. Create a list of taxon names to be passed to `downloadSequences` (either serial or parallel). All downloaded sequences will be written to table `acc_<locus>` with the attribute `status` set to `'raw'`.

4. Search the attribute `spec_ncbi` for a set of regular expressions indicating sequences that stem from samples that are undetermined at the level of species and set their attribute `status` to `'excluded (indet)'`.

5. Crop subspecies names in attribute `taxon` with `strip.infraspec()`; the full names are still available in attribute `spec_ncbi`.

6. Exclude sequences that are too long to align, i.e., sequences exceeding the the number of `max.bp` base pairs (default is 5000 bp), by tagging their attribute `status` as `'excluded (too long)'`.

7. Run `dbMaxGIPerSpec` to chose the `max.gi.per.spec` longest sequences per species for alignment; the rest will be tagged as `'excluded (max.gi)'`.

8. Run `dbUpdateTaxonomy` to detect species with sequences that have no entry in the table `taxonomy`. Sequences of species that cannot be classified are tagged as `'excluded (unclassified)'` in the attribute `status`.

9. Issue summary on screen and exit.

1.3 stepC

1. Check if table `acc_*` exists, i.e. if `stepB` has been run. If not, exit with error.

2. Clear results from previous runs of `stepD`.

3. Clear results from previous runs of `stepE`.

4. Produce table of species counting numbers of sequences and assessing if species are aligned with the `char_length() function`. If the table is empty, exit without error.

5. Mark single-sequence species in the `status` column with `'single'`.

1.4 stepD

1.

1.5 stepE

1.

1.6 stepF

1. Set threshold values for `min.identity` and `min.coverage`.

2. Open database connection.

3. Check if `stepE` has been run; if not stop.

4. Check if `stepF` has been run before, i.e. if MSA table exists. YES: go to 5. NO: go to xx.

5. Check if MSA table needs to be updated. This implies checking if the set of species names has changed, but also if the set of GIs has changed in any species. Maybe md5-checksums could be used to achieve this?

6. Erase downstream results before updating: `spec/gen_gene`, nexus and phylip files.

1.7 stepG

1. Check if MSA table exists. YES: go to next step. NO: break.

2. Check if any entry in the `status` column equals `'raw'`, which is the trigger for aligning the sequences in the MSA table.

2 Description of database

2.1 Table `taxonomy`

2.2 Table `reference`

2.3 Table `acc_*`

gi

taxon

spec_ncbi

status describes the status of the sequence along the pipeline; xx values are defined and are listed in the order of their appearence along the pipeline:

`'raw'` is the default status for every downloaded sequence.

`'excluded (indet)'` is set by `stepB`.

`'excluded (too long)'` is set by `stepB` or `stepF`.

`'excluded (max.gi)'` is set by `stepB`.

`'excluded (unclassified)'` is set by `dbUpdateTaxonomy`. For species names that are present in any one `acc_*` table but not in the `taxonomy` table, this function tries to derive their taxonomic classification using congenerics. If this is not possible, a species (and all its sequences) are tagged as unclassified and excluded from downstream steps of the pipeline.

`'single'` is set by `stepC` and marks all species that are represented with only one sequence, not requiring alignment.

`'aligned'` is set by `stepC` for all aligned species.

genom

npos

identity

coverage

dna

2.4 Table spec_*/gen_*

3 Checking pipeline status

4 Updating the database

The pipeline is designed to minimize computational costs when updating a MEGAPTERA project.

Step	Changes	Trigger
A		
B	new sequences	none; GenBank has to be searched
C	new sequences	conspecifics of unequal length
D		
E		
F		
G	new sequences	'raw' in status column
	new guide tree	*not implemented*
	species excluded	'raw' in status column
H	(re)alignment by G	'aligned' in status column
I		

Problems

- stepC deletes entries the attributes identity and coverage of the acc_* table. This triggers the rerunning of stepE.

5 Parallelization

These functions contain parallelized apply-like functions:

- ncbiLineage

6 Extended ingroup and surrogate species