Introduction

CAT is a pipeline for taxonomic classification of long sequences implemented in Python. It uses Prodigal software for gene prediction and Diamond alignment tool. CAT also requires database of reference sequences which have NCBI accession numbers in headers (contemporary NCBI databases) and NCBI taxonomy tree files. CAT allows to input files at two intermediate steps if files formatted accordingly (see examples of files).

Dependencies and where to get them

diamond	http://github.com/bbuchfink/diamond
prodigal	http://github.com/hyattpd/Prodigal

NCBI taxonomy tree files:
From ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/ :
from taxdump archive:

- names.dmp
- nodes.dmp

From ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/

- prot.accession2taxid

Getting started

Before start, please check that reference sequences have headers in the following format (accession number with version):

```
>WP_003131952.1 <and anything else>
```

As Diamond save into alignment file only name of reference sequence before the first space, to preserve information about functional annotation we can offer to change NCBI fasta headers with the following command:

```
$ perl -wple 's/^(>\S+\.\d+)\s(.+?) \[.*/$1:$2/g; s/ /_/g;' database > database_formatted
```

This will change headers from this variant:

```
>WP_003131952.1 30S ribosomal protein S18 [Lactococcus lactis]
```

to this one:

```
>WP_003131952.1:30S_ribosomal_protein_S18
```

When dependencies will be downloaded, you need to specify absolute paths to Prodigal and Diamond inside CAT:

```
diamond = '/absolute/path/to/executables/of/diamond'
prodigal = '/absolute/path/to/executables/of/prodigal'
```

CAT assumes that taxonomy tree files locate in the working directory, if this is not a case, please specify path to these files in CAT too:

```
path_to_taxonomy_files=’/path/to/files/folder/’
```

Generate Diamond database as described here http://ab.inf.uni-tuebingen.de/data/software/diamond/download/public/manual.pdf

If you added CAT into PATH environment variable, it could be run using the command like this:

```
$ CAT -f sequences.fna -db reference_database.dmnd -prefix library_one
```

To get help:

```
$ CAT -h
```

For more details about analysis algorithm, please see http://biorxiv.org/content/early/2016/09/01/072868