Manual

Prerequisites

Perl 5.8 or later, with DBI module installed.
MySQL (version 5 or later).

For debian/ubuntu, Perl and DBI module should have been installed by default.

To check the installation status of those modules, run the following command in the terminal:

```
$ perldoc DBI
```

To install MySQL on debian/ubuntu:

```
$ apt-get install mysql-server
```

1. Setting up the MySQL database

Create the MySQL user and databases

a) Without root permission

In this section, MYSQL_DB_DIR is where you want to place the MySQL data.

```
mkdir -p MYSQL_DB_DIR/data MYSQL_DB_DIR/temp
```

Edit MYSQL_DB_DIR/data/my.cnf, make sure it has below contents:

```
[client]
user=root
socket=MYSQL_DB_DIR/mysql.sock

[mysqld]
datadir=MYSQL_DB_DIR/data
socket=MYSQL_DB_DIR/mysql.sock
tmpdir=MYSQL_DB_DIR/temp
log-error=MYSQL_DB_DIR/mysql.log
pid-file=MYSQL_DB_DIR/mysql.pid
#skip-networking
port=93306 #keep accordingly with get_species_taxids.pl and other .pl files
bind-address=192.168.4.98 #mgmt01
```

Start this MySQL daemon:

```
    mysqld_safe --defaults-file="MYSQL_DB_DIR/data/my.cnf" --no-auto-restart
```

Or shutdown it:

```
    mysqladmin --defaults-file="MYSQL_DB_DIR/data/my.cnf" shutdown
```

Log into MySQL daemon:

```
    mysql --defaults-file="MYSQL_DB_DIR/data/my.cnf"
```

b) With root permission

```
mysql -uroot -p
```

c) Once you have logged into MySQL, create a MySQL user and two databases for blast2hgt:

```
>insert into mysql.user(Host,User,Password)
  values("localhost","test",password("1234"));
>CREATE DATABASE taxondb;
>CREATE DATABASE taxnode;
>GRANT ALL ON taxondb.* TO 'test'@'localhost' IDENTIFIED BY '1234';
>GRANT ALL ON taxonnode.* TO 'test'@'localhost' IDENTIFIED BY
  '1234';
>flush privileges;
>QUIT;
```

Import required data to MySQL

a) Download `taxdump.tar.gz` (nodes.dmp and names.dmp are used):

```
$ cd $path_to_blast2hgt/install_database/
$ wget ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz
```

And decompress to $path_to_blast2hgt/install_database/;

b) Get all taxid from NCBI nr database:

```
$ install_database/extract_acc.taxid_from_nr.sh
```

c) Now import them into MySQL databases:

```
$ cd install_database && ./importdatabase.sh
```

This step may take hours to finish depending on the hardware performance; However, if it lasts for a week, please check.

2. Building (split) BLAST databases (nr or nt) for each taxonomy group

See `species_taxids/create_species_taxids.sh`.

Or use '`get_species_taxids.sh`' (officially released by NCBI), if you have a good internet connection to NIH.

3. BLAST against each taxonomy group

An example script is placed in `bin/blastp_each.taxids.final.sh`, please modify it according to your study design and HPC environment.

4. Running blast2hgt

See `0run.sh` for detail, and modify it when needed.

At this step, a table (*.rp.tsv) containing the genes (proteins), and their horizontal transfer probabilities (measured by alien index / bitscore difference / ratio of bitscore difference), as well as donor/recipient, will be produced. Users are required to manually screen HGT candidates from this table.

5. Reconstruction of phylogenetic trees to verify HGT

To confirm the HGT candidate, a phylogenetic tree is the golden standard.

Balanced sequence sampling from each taxonomy group

This part gets sequences from each taxonomy group, using a balanced sampling strategy. The top *n* best BLAST hits (based on bitscore) are selected.

lineage2idgroup3.batch.pl: get a maximum of 10 sequences per taxonomy group; and 1 sequence per species within self-group. Recommended to new users.

lineage2idgroup3.unlimited.pl: unlimited sequences per taxonomy group.

Command:

```
perl phy/lineage2idgroup3.batch.pl [representatives.txt]
  [*.lineage] [id_file]
```

`[representatives.txt]`: the defined taxonomy groups, could be slightly modified based on study design.

`[*.lineage]`: the merged `$query.nr.lin` and `$query.nr.taxid` produced by `0run.sh.` Here it could be referred as:

`` `cat $query.nr.lin $query.nr.taxid`. ``

`[id_file]`: A file containing HGT candidate IDs obtained from step 4. One ID per line.

At the end of this part, a *.id.org file will be generated for each HGT candidate listed in [id_file].

Sequence extraction and phylogenetic tree reconstruction

Examples can be found in below folders:

Sequence extraction: `phy/phy_verify.sh`.

Phylogenetic tree reconstruction: `phy/phy.sh`.

You may need to modify above two files according to your HPC environment.