Pipeline for example data

To provide a clear demonstration for operating the tool of dbHT-Trans by general biologists, one example data is also attached, which contains 10, 000 assembled transcripts of RNA-Seq. Here, you can filter these sequences according to the pipeline as below step by step.

Installation

- **MySQL**

  In cases of Red Hat, Fedora and CentOS system, you can install MySQL with Yum:

  ```
  $ sudo yum install mysql-community-server
  $ sudo service mysqld start
  ```

  In cases of Debian and Ubuntu system, you can install MySQL with APT:

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

  For others, please refer to the official manual of mysql. (http://dev.mysql.com/doc/refman/5.6/en/linux-installation.html).

- **Python**

  If Python has been installed, you can check the version information:

  ```
  $ python --version
  ```

  Otherwise, you can install Python:

  ```
  $ sudo apt-get install python2.7
  ```

  or

  ```
  $ sudo yum install python
  ```

- **Python package for MySQLdb**

  One third-party library of MySQLdb is required for operating dbHT-Trans (https://pypi.python.org/pypi/MySQL-python/). Please follow its official instruction to install.

Download and use

- **Download**

  You can download dbHT-Trans from https://github.com/chengroup/dbHT-Trans and decompress the file of '**dbHT-Trans-master.zip**'. Please then change directory to 'dbHT-Trans-master'.

  ```
  $ tar xfz dbHT-Trans-master.zip
  $ cd ./dbHT-Trans-master
  ```

- **Configuration**

You must first edit the file of '**config.txt**' (./dbHT-Trans-master/config.txt) by updating the required MySQL information, which includes the following fields:

```
host=localhost
username=root
password=123456
database=dbht_trans
port=3306
```

After updating these fields, please enable this configuration by executing the following command:

```
$ chmod a+x dbHT-Trans-Extractor dbHT-Trans-Operator
$ chmod a+x ./model/usearch7.0.1090_i86linux32
$ python config.py
```

- ## Operating the module of 'dbHT-Trans-Operator'

Example data is provided in the example directory, which include a fasta file (example_data.fa) and a gene list file (example_data.list). A reference protein database file should be additional needed for homologous sequence search. You could download the mouse protein database from Uniprot.

```
$ gunzip MOUSE.fasta.gz
$ ./dbHT-Trans-Operator -T ./example/example_data.fa -g ./example/example_data.list -d MOUSE.fasta –L 200
```

- ## Operating the module of 'dbHT-Trans-Extractor'

After this processing step by 'dbHT-Trans-Operator', you can output results based on the threshold values of both sequence identity (-i) and query coverage (-c).

```
$ ./dbHT-Trans-Extractor -i 0.6 -c 0.5 -T transcript -S stat.txt -M meta_table.txt -P protein.fasta -C orf.fasta - F filtered_gene.list
```

Here, you will get six result files (*stat.txt*, *meta_table.txt*, *protein.fasta*, *orf.fasta*, *transcript.retained.fasta*, and *transcript.discarded.fasta*), which include the fasta files of both retained and discarded transcript sequences, the file of statistical results and others. Please refer to the detailed description for these files in dbHT-Trans manual.