Meta-Storms Users' Manual

Version 1.2

Sept 12, 2013

Meta-Storms is a metagenomic database system toolkit to systematically and efficiently organize and search metagenomic data. It includes the following components: (i) creating a database of metagenomic samples based on their taxonomical annotations, (ii) efficiently indexing all samples in the database based on taxonomical structures, and (iii) searching for a metagenomic sample against the database by a fast scoring function based on quantitative phylogeny. Meta-Storms also has modules to efficiently manage the database, which is especially useful when the database is huge.

1 PACKAGE DEPENDENCY

GCC 4.2 or higher

2 ENVIRONMENT VARIABLE

This step is **very important.**

```
$export MetaStorms=Directory of Meta-Storms
```

3 INSTALL

```
$tar –xzvf Meta-Storms-1.2.tar.gz
$cd Meta-Storms
$make
$make comp_sam
```

All binary file mentioned in the following sections can be find at "bin" of Meta-Storms directory

4 METAGENOMIC DATABASE

Meta-Storms can build database(s) with metagenomic samples and make index for fast query. Samples in the database must be pre-computed Parallel-META (Su, et al., BMC Systems

Biology, 2012. http://www.computationalbioenergy.org/parallel-meta.html) with its default reference database. Samples in the database can also be divided into different groups, and query in these groups.

5 MAKE DATABASE

5.1 Pre-computing

To build a database, all samples must be pre-computed by Parallel-META, and insert the analysis results as database entries. Sample names are indicated by the directory names of their analysis results, in which “classification.txt” must be included. In the following content, “sample” represents the “analysis result of sample”.

There are two method to build a metagenomic database: making all samples in one directory as a new database (make_index), or adding all samples in one directory to an exist database (add_index). Once the database is built, samples’ location could NOT be modified.

5.2 make_index

The “make_index” can build a new database with all samples of one single input directory. The input directory path appointed by parameter –i must be the full absolute path. All samples of a new database will be in group 0.

Usage :

```
make_index [-option] value
        option :
        -i samples input path, must be the full absolute path
        -o index output path
        -n database name
        -h help
```

E.g. to make a new database named “DB1” using all samples located at /opt/data/samples1, and save the database at ./DB1.

```
$make_index –n DB1 –i /opt/data/samples1 -./DB1
```

5.3 add_index

The “add_index” can add all samples of one single directory to an exist database. The input directory path appointed by parameter –i must be the full absolute path.

Usage:

```
add_index [-option] value
        option :
        -i samples input path, must be the full absolute path
        -d database index name
        -o index output path
        -g to assign group, defaults is 0
        -h help
```

E.g. to add all samples located at /opt/data/samples2 to DB1, and make a new group of Group 1.

```
$add_index –d ./DB1 –i /opt/data/samples –o ./DB1 –g 1
```

6 QUERY

The query samples should also be pre-computed by Parallel-META as samples in the database. Query path appointed by parameter –i is the path of “classification.txt” in analysis result path of each query sample. Meta-Storms supports both indexed query and exhaustive search in the database. For fast fetch to the results we suggest the indexed query. To only search in specific group please appoint the group number using parameter –g. Meta-Storms will return both the match samples and their similarity value to the query. Based on our experiments, two samples could be considered as significant similar if their similarity value is equal to or above **85%**.

```
query_index [-option] value
        option :
        -d database index name
        -i query sample path
        -o result output file, default is to output on screen
        -n hit number, default is 5
```

-e exhaustive search

-g to assign group number, default is off

-t cpu core number, default is 1

-h help

E.g. to find the top 10 hits of a sample located at /opt/data/query_samples/sample_1 in database DB1 using 8 CPU cores.

$query_index –d ./DB1 –i /opt/data/query_samples/sample_1/classification.txt –n 10 –t 8

7 SIMILARITY VALUE AND SIMILARITY MATRIX

Meta-Storms also supports to directly calculate the similarity value of two samples or the similarity matrix of several samples. Samples to calculate similarity values should also be pre-computed by Parallel-META. Based on our experiments, two samples could be considered as significant similar if their similarity value is equal to or above **85%**.

comp_sam [-option] value

option :

-i two samples path for single sample pair comparison, conflict with -l

-l sample name list table for multi-sample comparison, confilct with – i

-o result output file, default is to output on screen

-t cpu core number, default is 1

-h help

E.g. to calculate two samples' similarity value. These two samples are located at /opt/data/query/sample_1 and /opt/data/query/sample_2

$comp_sam –i /opt/data/query/sample_1/classification.txt /opt/data/query/sample_2/classification.txt

E.g. to calculate 10 samples' similarity matrix using 8 CPU cores, and save the results in ./results.txt These 10 samples' address are in a plain-text file "list.txt" with each file in a row like：

/opt/data/query/sample_0/classification.txt

/opt/data/query/sample_1/classification.txt

/opt/data/query/sample_2/classification.txt

```
……
/opt/data/query/sample_9/classification.txt
```

The command line is

```
$comp_sam –l ./list.txt –o ./results.txt –t 8
```

8 NOTICE

1. Please set the environment variable of section 2 before the installation.

2. All metagenomic samples for Meta-Storms must be pre-computed by Parallel-META using default database.

3. For make_index and query_index, input path appointed by parameter –i must be a directory of results analyzed by Parallel-META. For each sample, its result at least contains the “classification.txt” file. After building a database, all samples’ location could **NOT** be modified.

4. For query in small database (eg. sample number <= 200) or database with very similar samples (eg, samples from the same source) we recommend the exhaustive.

5. To calculate the similarity value matrix of several samples, the list file which contains the sample number and sample path should in the correct format mentioned in section 7.

9 CONTACT

For software usage, reporting bugs, further development or any other problems please feel free to contact:

Dr. Kang Ning (ningkang@qibebt.ac.cn) or

Xiaoquan Su (suxq@qibebt.ac.cn)