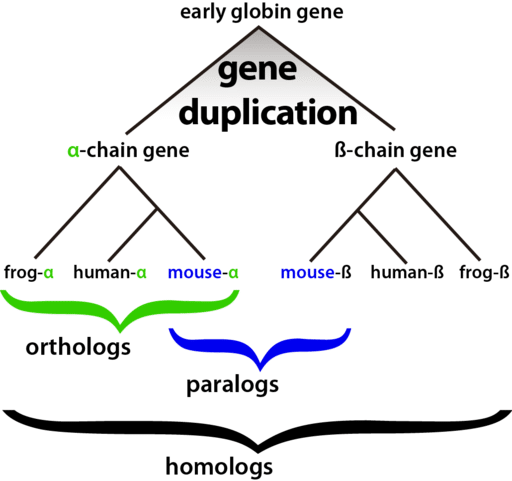
owPReMark User's Guide



version 2.0

by Sunshin Kim



1. **Installation**

We developed and designed this program in very portability, Efficient, reliable and flexible way. The program run in all environment (Linux, Win, Mac OS) in same way. This program can download from [www.github.com/abc](http://www.github.com/abc) at free of cost for research and non profit purpose. The developer is not liable for any kinds of lose during this program running. We defined this Program as owPReMark (One Way Parallel Markov Clustering Method).

To run the owPReMark(one-way best hit Parallel ReMark), you need to make sure that following programs is on your computer.

① Python 3

② Python modules: NumPy, Bio, alive\_bar

③ blast 2.2.26 ∼

**1) Python 3**

Python is an interpreted, high-level, general-purpose programming language. Most versions of Linux comes with Python 2 preinstalled. We developed and run our Code in Python 3.7 so we recommend to use python version greater than 3.5. Python can be download from the following sites.

<http://www.python.org/download/>

You can also down load Python from Anaconda sites which is very popular among Data Scientists.

**2) Python modules**

Python module can be easily installed with pip installer which in inbuilt in Python3.Any Python Module can be install by typing

pip install <module\_name> in your terminal

* + pip install numpy
  + pip install Bio
  + pip install alive\_bar

**3) blast**

In the following URL, download a blast file of OS version related on your system.

<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>

\* The blast version 2.2.26 is faster than version 2.2.27.

Following example is how to download and install the blast in the Ubuntu 64bit.

(Ubuntu 32bit: ncbi-blast-2.2.27+-ia32-linux.tar.gz)

**Sudo apt-get install blastp** in linux environment

4) setting an environment variable for blastp

① system-wide environment variable

Open the environment file located in the /etc/ by administrator's authority and input the file path in the end of line.

$sudo gedit /etc/environment



To apply to your system, reboot system or type following command.

$source /etc./environment

**To check installation of blastp type blastp in terminal or shell.**

② the edition of program code

Opening the program file using editor program, there is a part of program code to edit the values in the top. In the following part, input the path of blast and program name(blastp) to the default variable.



③ The use of command mode

Set the path of blast and program name(blastp) using -b or --blastp in the command mode.

$python owPReMark.py -M 1 3 -g AAE CAC -b home/jungwon/Downloads/ncbi\_blast-2.2.27+/bin/blastp

After Complete Setup Clone or Download our owReMark tools from github and unzip the folder.

**This program can be run in two way.**

Like python this program is has two basic modes: script and interactive mode. The normal mode is the mode where the scripted and finished.

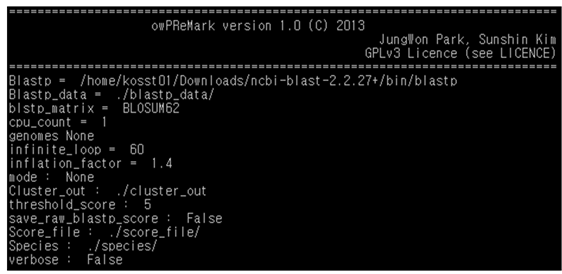
1. **Interactive mode**

Interactive mode is a command line shell which give immediate feedback for each statement. The interactive mode is to easily and quickly run the owPReMark using default value. It is available to set the detailed environment and parameter by changing the default value of program code. You cannot change the environment values and parameters when program is running. Unchanged values are used to default value. You can check for the parameters and environment values. To run the interactive mode, use the following command in the directory existed for program file.

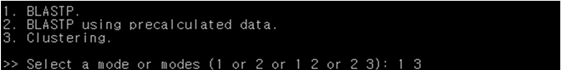
$python3 owPReMark.\_V1.0.py or

./owPReMark\_V1.0.py

# !! Most of Linux distribution come with python 2 preinstalled so check weather python will run python 3 or python2



Running the program, the environment values and parameters are printed.



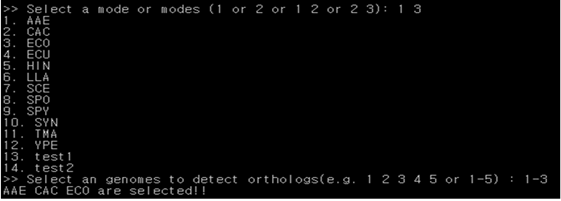
After that, input the mode which you want. The following mode is available to make a choice.

**1 or 2 or 1 3 or 2 3**

When you use the mode 2, if all of or a part of precalculated data in regard of selected genomes are not existed, the additional calculation will be conducted to make the blastp score data. Inputting the mode 3 alone, it will be not worked properly. If you want to use the blastp score data which are made newly, move the file of new blastp score data, is \*\_\*\_one\_way\_best\_hit\_\*, to the directory of blastp\_data from the directory of score\_file. And then, use the mode 2 3. (See 4)

In the mode 1 or 2, Log file will be named to "Log". But log file will named to "name of file\_S5\_1.4.log (reverse score : 5, inflation factor : 1.4)" in the mode 1 3 or 2 3.

**Tip** If you want to analyze by adding a new genome to the genome of the already analyzed data, you can **save time** by using mode 2.



The upper picture is screen to select the genomes. If you want to include the new genomes, move the files of protein sequences you chosen to directory of species. The file list which is included new genomes will print on screen. (See 5)



Input the matrix number to use in the blastp.

**About Matrix**

In bioinformatics, the BLOSUM (**BLO**cks **Su**bstitution **M**atrix) matrix is a substitution matrix used for sequence alignment of proteins. BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences. They are based on local alignments.BLOSUM matrices are also used as a scoring matrix when comparing DNA sequences or protein sequences to judge the quality the quality of the alignment.

**BLOSUM80 🡪** more related proteins (the blocks of amino acid sequence are 80 % or more identical: the substitution frequencies are based on the remaining 20% of the amino acids that have changed.)

**BLOSUM62 🡪** midrange (is the most frequently used matrix for detecting the majority of weak protein similarities)

**BLOSUM45🡪** distantly related proteins (very Suitable for detecting long and weak alignments)

The BLOSUM62 used for so many years as a standard is not exactly accurate according to the algorithm described by Henikoff and Henikoff.



Input the number of CPU to use in the blastp. If you want to use parallel computation, input the number of CPU more than 2. The number of CPU on your system will be printed on the screen.

**Inflation factor** 🡪 In Statistics, the variance inflation factor (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone.



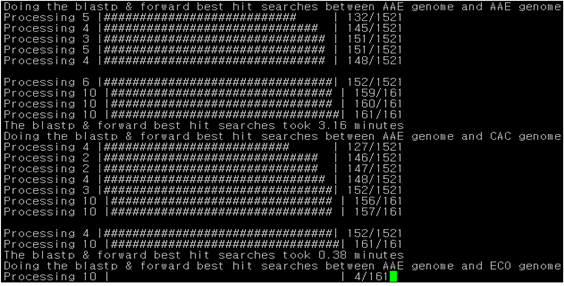
If you chose the mode 3, this screen will be printed. But if you select the mode 1 or 2 will not print the this screen.

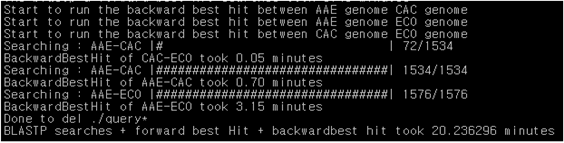


If you chose the mode 3, this screen will be printed. Input the output file name of clustering. Existing the same file name, program will be terminated. The output file name of clustering be comprised of inputted name, the values of reverse score and inflation factor. So, if the values of reverse score and inflation factor is different from existed file, the program will consider that file different.log file starting with this variable will start.



The program will always delete all of the file in the directory of score\_file every run.

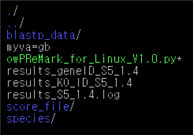




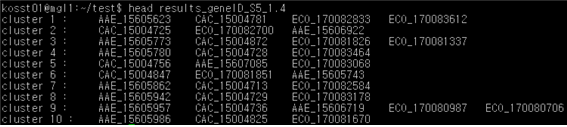
The program will compute the blastp score and the reverse best hit depended on the combination of genomes.



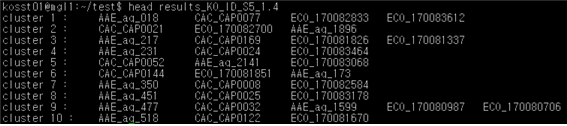
The start message of MCL and clustering will be printed. Finishing the calculation, the program will print the running time and terminate.



When the program is finished, two of result files and one of log file are created.



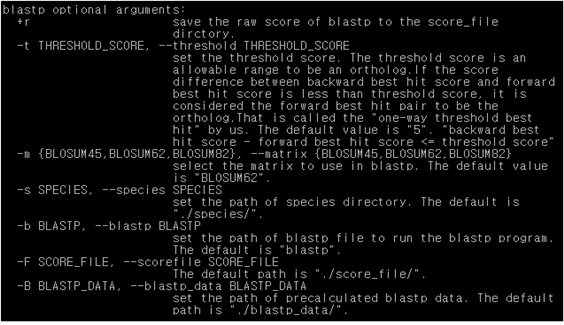
The one of results file, results\_geneID\_S5\_1.4, is printed the members of a cluster by gene ID. The S5 shows that the threshold score is 5. And the 1.4 is an inflation factor. The genes of cluster 1 are the analyzed genes which have a same function.

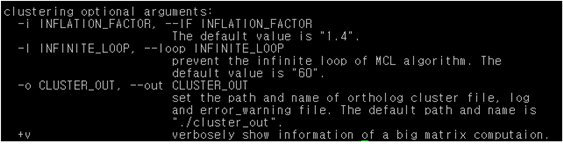


The KO of results\_KO\_ID\_S5\_1.4 is KEGG Orhtology (KO) database. The results of this file include KO Database ID instead of gene ID. If the ID corresponding to KO database does not exist, gene ID has not changed. The ID of the KO database is stored in a myva=gb file.

3. command mode

The commanmd mode can change the detailed environment values and parameters using command. You can see the arguments needed to run the command mode using -h or --help command.





There are 4 kinds of groups in command mode. They are optional argumens, essential arguments, blastp optional arguments and clustering optional arguments.

1) optional arguments

-c CPU\_COUNT, --CPU CPU\_COUNT

: The program will conduct a parallel computation using inputted number of CPU in the blastp. The number of CPU on your computer system will be printed on the help message. The parallel computation is applied to the only blastp.

2) essential arguments: If you don't input the arguments, the program will not run.

-M 1,2,3 [1,2,3 ...], --mode 1,2,3 [1,2,3 ...]

1: blastp

2: blastp using precalculated data

3: clustering

: Input the mode which you want. The following mode is available to make a choice.

1 or 2 or 1 3 or 2 3

When you use the mode 2, if all of or a part of precalculated data in regard of selected genomes are not existed, the additional calculation will be conducted to make the blastp score data. Inputting the mode 3 alone, it will be not worked properly. If you want to use the blastp score data which are made newly, move the file of new blastp score data, is \*\_\*\_one\_way\_best\_hit\_\*, to the directory of blastp\_data from the directory of score\_file. And then, use the mode 2 3. (See 4)

In the mode 1 or 2, Log file will be named to "Log". But log file will be named to "name of file\_S5\_1.4.log (reverse score: 5, inflation factor: 1.4)" in the mode 1 3 or 2 3.

**Tip** If you want to calculate addedly new genomes with already calculated the set of genomes, use the mode 2.

-g GENOMES [GENOMES ...], --genomes GENOMES [GENOMES ...]

: Input the names of genome. You can check the list of genomes in the directory of species.

3) blastp optional arguments

+r

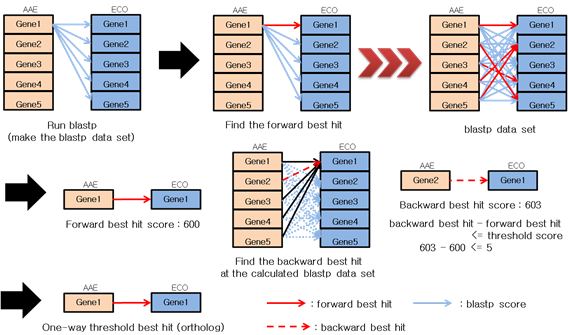
: If you want to save the score files made by blastp, use this command. The score files will save in the directory of score\_file.

ex) AAE\_AAE\_S5\_1, AAE\_CAC\_S5\_1 (The S5 is reverse score. The 1 is the number of CPU.)

-t THRESHOLD\_SCORE, --threshold THRESHOLD\_SCORE

: The threshold\_score is an allowable range to be an ortholog.

The orthologs are the genes related to the reciprocal best hit. The reciprocal best hit is two-way(bidirectional) best hit(A->B, A<-B). The best hits to determine the one-way best hit calculate the blastp scores to only one direction(A->B). The blastp scores of backward direction are calculated using the calculated blastp score of one direction. The one-way best hit should be called that the score difference between forward best hit and backward best hit, the minimum is 0 or greater than or equal to 0, is more less than or equal to the threshold score. If the threshold score is set to 0, concept of the one-way threshold best hit algorithm is the same as the algorithm using reciprocal best hit. However, performing the blastp once make it different from that.



-m BLOSUM45, BLOSUM62, BLOSUM82, --matrix BLOSUM45, BLOSUM62, BLOSUM82

: Input the matrix name as upper case to use in the blastp. The default value is "BLOSUM62".

-s SPECIES, --species SPECIES

: Specify the path where the amino acid sequence of the species you want to analyze. The default directory is ". / Species /".

-b BLASTP, --blastp BLASTP

: Set the path of blastp file to run the blastp program. The default is "blastp".

ex) /home/Downloads/ncbi-blast-2.2.27+/bin/blastp

-F SCORE\_FILE, --scorefile SCORE\_FILE

: It is the path to store the raw blastp score files (optional), best\_score files, oneway\_ threshold\_best\_hit files and blastp\_score\_split\_list file. The default path is “. /score\_file/".

-B BALSTP\_DATA, --blastp\_data BLASTP\_DATA

: Set the path of precalculated blastp data. The default path is "./blastp\_data/".

4) clustering optional arguments

-i INFLATION\_FACTOR, --IF INFLATION\_FACTOR

: Specify the value of the inflation factor used to inflation matrix during run of the MCL algorithm. The default value is "1.4".

-I INFINITE\_LOOP, --loop INFINITE\_LOOP

: Specify the maximum number of executions of the MCL algorithm to avoid an infinite loop during run of the MCL algorithm. The default value is "60". 60 is the value that is strictly enforced. We recommend that you do not exceed 100. There is a case of an infinite loop calculation during the execution of MCL algorithm. In this case, if the value set too high, the calculation time increases.

-o CLUSTER\_OUT, --out CLUSTER\_OUT

: Specify the path and name to save the name of the cluster output file. The default path and name are ". / Cluster\_out".

+v

: When calculating the matrix, information (time and the number of execution) computed the MCL algorithm will be printed.

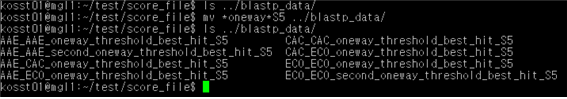
**4. Using and Addition of precalculated data**

If you want to use precalculated data, the one\_way\_threshold\_best\_hit file is required. The threshold value of one\_way\_threshold\_best\_hit files is all the same. When the file does not exist, that file will be generated at the runtime.

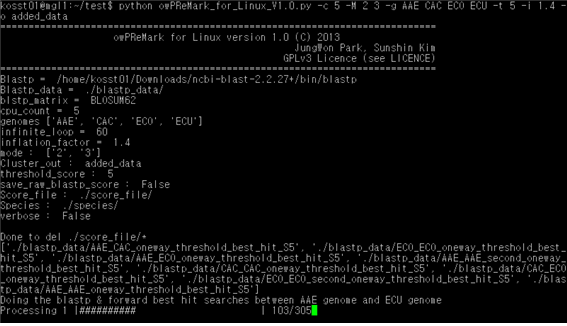
Here is an example to obtain the result of ortholog cluter which are added ECU using precalculated data of AAE, CAC and ECO. The precalculated data are processed using parallel computation with 5 of CPU.



First, move to the score\_file directory.

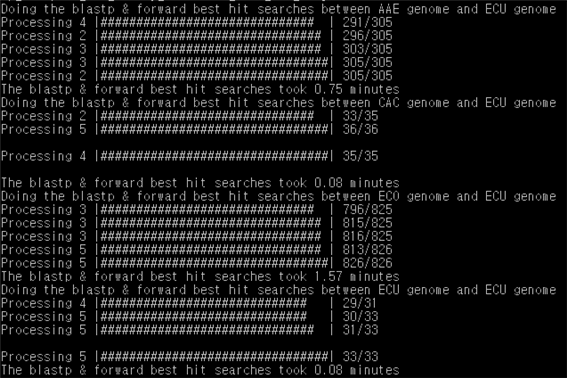


Move the one\_way\_threshold\_best\_hit\_S5 files of AAE, CAC and ECO to the blastp\_data directory.

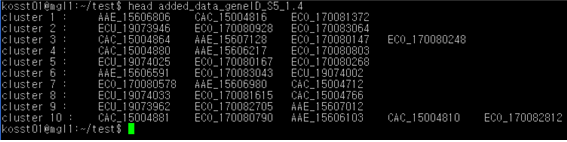


Use the mode 2 3. When you run the following command, further analysis is done.

python owPReMark\_for\_Linux\_V1.0.py -c 5 -M 2 3 -g AAE CAC ECO ECU -t 5 -i 1.4 -o added\_data



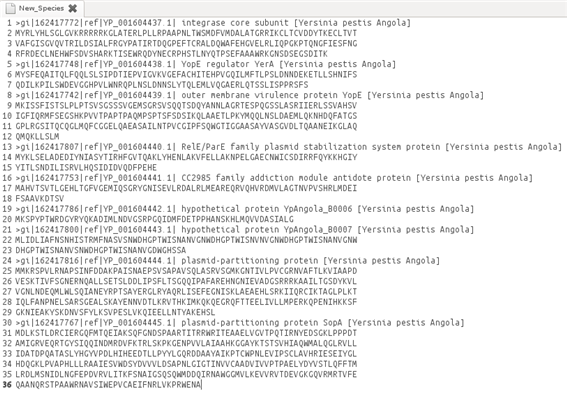
The data related ECU is not existing because of added species. Therefore, the data analysis about AAE-ECU, CAC-ECU, ECO-ECU and ECU-ECU combinations execute.

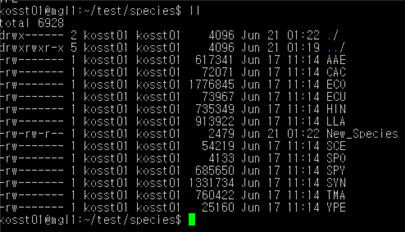


As shown in the figure above, the clusters of orthologs which include the data of ECU are generated among AAE, CAC, ECO and ECU.

5. Adding the genome (species)

Input the protein sequence into the file by fasta form. And put the protein sequence file created into the species directory. Name the file New\_Species





When you run the program, you can see the added species.

