Getting aligned protein tree for multiple genomes

Yu-Wei Wu yuwei.wu@tmu.edu.tw

Graduate Institute of Biomedical Informatics

Taipei Medical University, Taiwan

This tutorial is for the script "get_protein_align.pl" that I wrote for getting the alignment of multiple marker genes from a number of genomes. You need perl to execute this script. My perl version is 5.18.2, but any version should work (unless it is too old).

Download script

The script can be accessed at https://github.com/yuwwu/get-marker-align.

Prerequisite

Several software packages are needed to run this software. Please download the software and enter the full path of the executable in the perl script.

1. HMMER 3.0 (http://hmmer.org/download.html)
2. MUSCLE (http://www.drive5.com/muscle/downloads.htm)
3. Gblocks (http://molevol.cmima.csic.es/castresana/Gblocks.html)
4. Prodigal (http://prodigal.ornl.gov/ or https://github.com/hyattpd/prodigal/releases/)

Follow the installation instruction to install the above software. If you do not know how to add them into system paths, simply modify line 18-21 of the script to indicate the actual paths. The script will also check the existence of the programs before it starts running.

Download PFAM Hidden Markov Model file

Please go to ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam30.0/ and download the "Pfam-A.hmm.gz" file. Place the file under "data" folder and unzip it using "gzip -d Pfam-A.hmm.gz" command. You can also modify line 5 to specify the actual location of the hmmer file.

The next step is to compress the Pfam hmm file. The command is "`hmmpress Pfam-A.hmm`" as the following screenshot:

```
ywwei@ywwei-u:~/get_align/data$ hmmpress Pfam-A.hmm
Working...    done.
Pressed and indexed 16306 HMMs (16306 names and 16306 accessions).
Models pressed into binary file:   Pfam-A.hmm.h3m
SSI index for binary model file:   Pfam-A.hmm.h3i
Profiles (MSV part) pressed into:  Pfam-A.hmm.h3f
Profiles (remainder) pressed into: Pfam-A.hmm.h3p
ywwei@ywwei-u:~/get_align/data$
```

How to run the script

Usage: perl get_protein_alignment.pl (list file) (output alignment file)

The script takes in two parameters, the first is a list file consisting of all genomes you want to analyze, and the second is the output alignment file.

Specifying the number of thread

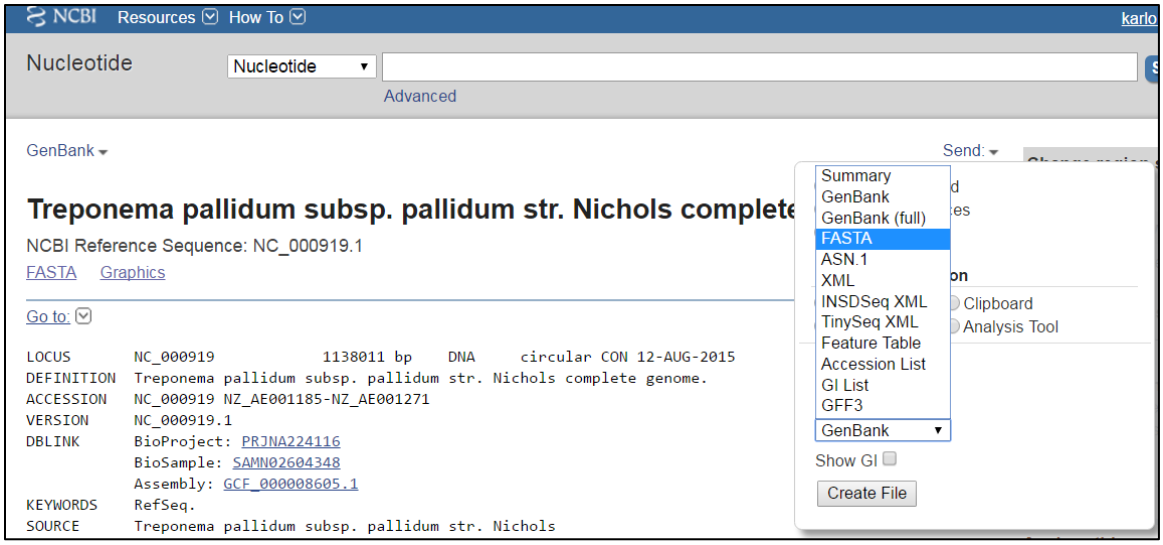
The number of threads can be specified by changing line 6 “my $CPU = 4;” of the script. Default CPU number is 4. Change it to whatever number you like.

An example run

Below I will demonstrate how to build a protein tree from a set of four genomes. Here I use complete genomes for demonstration, but draft genomes or genomes just recovered from metagenomes can be used as well.

Step 1: Download Genomes

Please go to NCBI website to download the following genomes for test purpose. Warning: please download the genome sequences (i.e. complete record) in fasta format. Do not download the coding sequences—in a lot of cases we need to process the gene annotation by ourselves. The script will not work if you provided only coding sequences.



I will use the following genomes for demonstration purpose. Please go to the NCBI website and download the genomes.

Species	Genome accession
E. coli	NC_002695
Salmonella	NC_003197
Treponema denticola	NC_002967
Treponema pallidum	NC_000919

Place the genomes at whatever place you like, for example at the same directory of the script.

Here is a screenshot about how I store and name the genomes.

```
ywwei@ywwei-u:~/get_align$ ls -l
total 14220
drwxrwxr-x 2 ywwei ywwei    4096 Nov 22 19:11 data
-rw-rw-r-- 1 ywwei ywwei 5577078 Nov 22 01:25 Ecoli.fasta
-rw-rw-r-- 1 ywwei ywwei    6434 Nov 22 01:33 get_protein_alignment.pl
-rw-rw-r-- 1 ywwei ywwei 4926931 Nov 22 01:26 Salmonella.fasta
-rw-rw-r-- 1 ywwei ywwei 2883892 Nov 22 01:26 Treponema_denticola.fasta
-rw-rw-r-- 1 ywwei ywwei 1154347 Nov 22 01:26 Treponema_pallidum.fasta
ywwei@ywwei-u:~/get_align$
```

Step 2: Build list file

It is very simple to build list file using "ls"—simply type "`ls *.fasta > mylist`" since I named all my genomes with ".fasta" extension. This list file "mylist" will be used for the next step. You can also manually type in all genome filenames using whatever text editor you like.

Step 3: Run the script

The command to run the script is

```
perl get_protein_alignment.pl mylist mylist.aln
```

in which the first parameter is the list filename, and the second parameter is the output. The gene prediction and the hmmer process will all take a while, so please wait patiently.

Here is the screenshot for running the script. Notice that the script dug up 178 single copy marker genes for building this protein tree.

```
ywwei@ywwei-u:~/get_align$ perl get_protein_alignment.pl mylist mylist.aln
Predicting genes for Ecoli.fasta
Predicting genes for Salmonella.fasta
Predicting genes for Treponema_denticola.fasta
Predicting genes for Treponema_pallidum.fasta
Running hmmscan on Treponema_denticola.fasta
Running hmmscan on Ecoli.fasta
Running hmmscan on Salmonella.fasta
Running hmmscan on Treponema_pallidum.fasta
Start Processing [Treponema_denticola.fasta]
Start Processing [Treponema_pallidum.fasta] remaining number of PFAMs: 337
Start Processing [Ecoli.fasta] remaining number of PFAMs: 186
Start Processing [Salmonella.fasta] remaining number of PFAMs: 178
Identified 178 marker genes for the genomes.
```

Step 4: Examine the outcome

Two files will be generated from the script is the run finishes successfully. The files are “mylist.aln” and “mylist.aln.pfam.” The former file consists of the alignment of the 178 proteins and the latter file is the complete list of all 178 proteins in terms of Pfams. Here is the screenshot of the top 10 lines of both files

```
ywwei@ywwei-u:~/get_align$ head mylist.aln
>Treponema_pallidum.fasta
FQTEVSQLLT LIIHSLYSHK EIFLRELISN ASDALDKLKY EALVEARIDI AFEEDAQRLV
VRDTGIGMNA EDLRANLGTI ARSGTKAFLS TLTRDQKQDS NLIGQFGVGF YSAFMVASKV
EVITKKAAWT SEGQNAYTGT CVVLHLSQEN SEFATRWRLE EVIKKYSDHI AFPTQKKVDQ
VNDAGALWKR PKSELKEEDY HRFYQTLTRD STPPLLYVHT KAEGTQEYVT LFYVPAKAPF
DLFHADYKPG VKLFVKRVFI TDDEKELLPV YLRFVRGVID SEDLPLNVSR EILQQNRVLA
AIKSASVKKL LGEFKRLAEC DGKKYDEFIT QYNRPLKEGL YSDYEHREQL LELVRFRTLS
FAEYVSRMKP DQKAIYYIAS PHAESYRLQG FEVLVMSDDI DGIVMPSVEL RPNEETDAAA
QREQGFKPLL ERLTHILSDS VKEVRLSKRL SDSVSCIVID ENDPTVQMER LMRATGQIKP
ILEINASHTL VQKLKESTDE AFVEDLAFVL LDQALLIEGM DVGSSVDFVK RVNRLLNTVF
ywwei@ywwei-u:~/get_align$ head mylist.aln.pfam
PF00183.16      HSP90, Hsp90
PF02390.15      Methyltransf_4, Putative
PF00889.17      EF_TS, Elongation
PF02639.12      DUF188, Uncharacterized
PF14622.4       Ribonucleas_3_3, Ribonuclease-III-like
PF02223.15      Thymidylate_kin, Thymidylate
PF01812.18      5-FTHF_cyc-lig, 5-formyltetrahydrofolate
PF02542.14      YgbB, YgbB
PF02410.13      RsfS, Ribosomal
PF00750.17      tRNA-synt_1d, tRNA
ywwei@ywwei-u:~/get_align$
```

Don’t worry about the spaces in the alignment file. These are created by Gblocks. Simply input the alignment into whatever tree-building software you like (I personally prefer MEGA5, 6, or 7) and there you go.