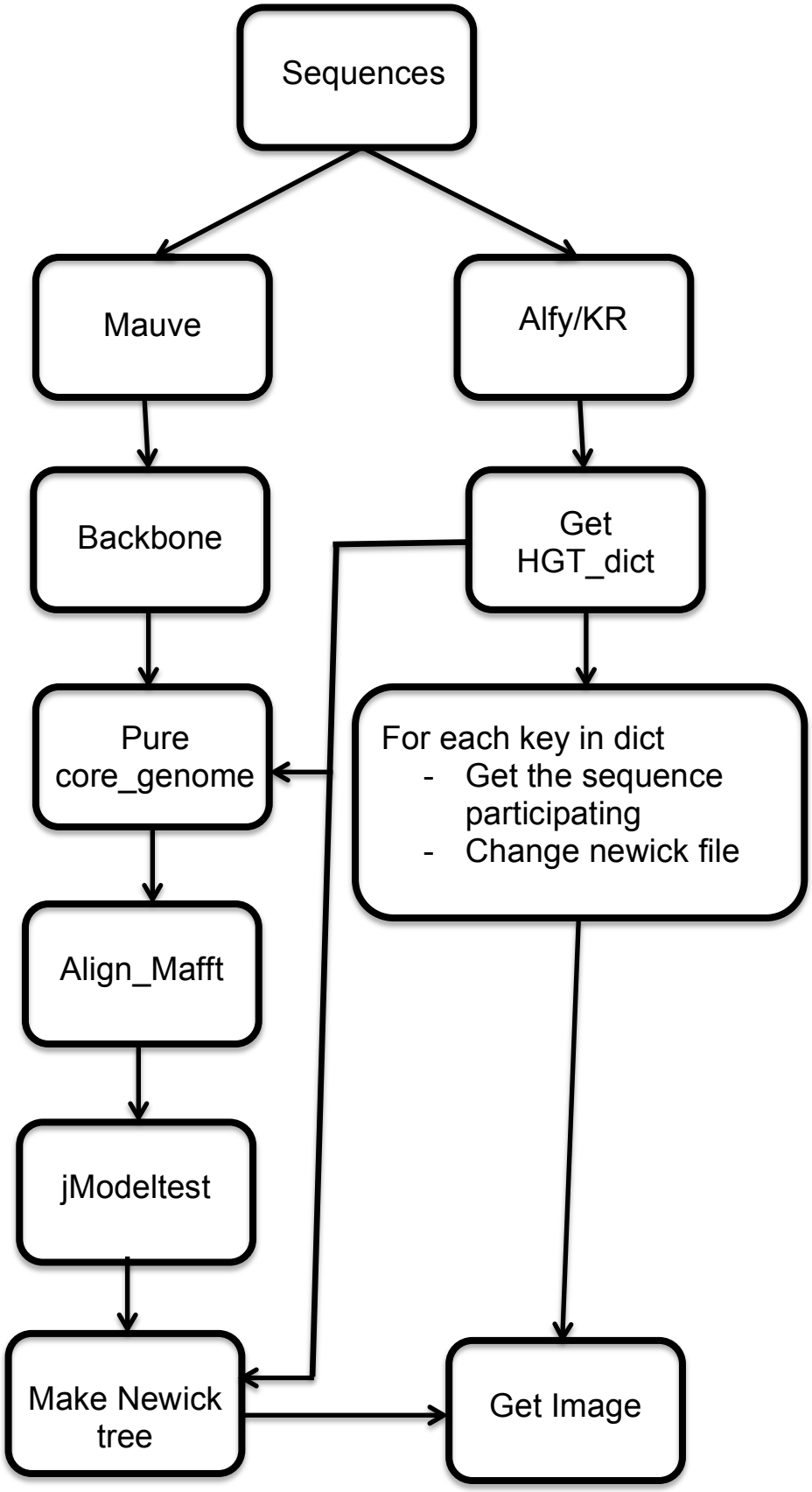
Pipeline for Analysis of Horizontal Gene Transfer in Bacterial Genomes

Stuti Agrawal
Rebecca Elyanow
Luigi Leung
Prateek Tandon
Yiming Xin

Contents


Sequences
Mauve
Alfy/KR
Backbone
Get
HGT_dict
Pure
core_genome
For each key in dict
- Get the sequence participating
- Change newick file
Align_Mafft
jModeltest
Make Newick tree
Get Image

1 Required software

mauve Aligner - http://gel.ahabs.wisc.edu/mauve/
jModelTest - https://code.google.com/p/jmodeltest2/downloads/list
Biopython - http://biopython.org/wiki/Main_Page
IPython(Recommended) - http://ipython.org/
Alfy - http://guanine.evolbio.mpg.de/alfy/alfy_1.5.tgz

2 Running the Pipeline

The pipeline can be run by calling:

aligned_sequence(data_dirpath, data_output, data_backbone, location_mauve, location_jModelTest, output_model)

data_dirpath = path to directory that holds to sequences you wish to analyze
data_output = the name of the output alignment file from mauve
data_backbone = the name of the output backbone file from mauve
location_mauve = path to mauve program
location_jModelTest = path to jModelTest program
output_model = name of model chosen by jModelTest

3 Alignment

Use progressive MAUVE [1] to align all sequences of the same species in the given directory. The standard defaults of progressive MAUVE are used.

4 Creation of the core genome

MAUVE's .backbone file is used to identify regions that are conserved among all genomes and the regions that are not. The regions conserved among all genomes are extracted using this file and concatentated together to create the core genome (including HGTs).

5 Building the phylogenetic tree

1. Select the best model to fit the phylogenetic tree using JModelTest [2]
2. Generate hundred bootstraps (seqboot) with Phylip [3]
3. Create a phylogenetic tree using Phylip (default = Maximum Likelihood Tree, DNAML)

This is done for the core genome (without any HGT regions that are a part of the core) to identify the evolutionary relationship between the species and for the pool of the HGT regions that are a part of the core as well as the distributed genomes so as to identify the relationships between the organisms with reference to the HGTs (This enables the identification of the current relationship between the species).
The generated phylogenetic trees for the core as well as the HGT regions will be in the phyloxml format.

6 Identifying HGT regions

Regions of Horizontal Gene Transfer will be identified using the distributed genome of each strain and the program Alfy [4]

7 Annotate the genes within the horizontally transferred regions

The sequence regions where HGT has occurred will be queried against the Antibiotic Resistance Gene Database (ARDB), and for regions that are not in ARDB, they will be queried against the NCBI database.

8 Graphical User Interface

The GUI is web based. The following are instructions for setting up to locally host the GUI via Apache web server, as well as, instructions for hosting on CMU's AFS server with limited permissions and access.

8.1 Hosting Locally

Setup
The following instructions apply to OSX and unix-based operating systems. Windows instructions will be in [brackets]. And instead of `nano` [or notepad], feel free to use any other text editor.
On OSX, Apache is already installed. On other unix-based OS, if it is not already installed, install it using the OS's packages utility or via terminal command `$ sudo apt-get install apache2`.
[For Windows, install with the downloaded `httpd-versionNumber-win32-src.zip`
from `http://httpd.apache.org/docs/2.2/platform/windows.html`]

1. To get the GUI running on the local computer, please enable php in Apache by going into its httpd.conf by typing into the terminal:

   ```
   $ sudo nano /etc/apache2/httpd.conf

   [ Open C:\\Program Files\Apache Software Foundataion\Apache2.2\
     and open the httpd.conf in notepad. ]
   ```

2. Uncomment the line (delete the # character) , save and exit:

   ```
   LoadModule php5_module libexec/apache2/libphp5.so
   ```

3. Enable Apache web server by typing:

   ```
   $ sudo apachectl start
   (To stop the web server, $ sudo apachectl stop )

   [ Click on the httpd.exe in \Apache 2.2\bin\ folder to start the service. ]
   ```

8.2 Hosting on AFS

Setup
The following instructions apply to hosting on Carnegie Mellon University's server as a student.

To enable CGI and PHP scripts, open in the internet browser:
`https://my.contrib.andrew.cmu.edu/index.cgi`

Enter your AndrewID and password, then under "`CGI services`" click on:
`(Re)enable authenticated AFS (CGI AFS-write) support`

Upload the .zip contents that you downloaded previously to the `www` folder in your afs space, or, for your convenience, we have a Github repository for easier uploading and for syncing any possible future updates. To do this, type the following:

Connect to school's clusters (afs):

```
$ ssh unix.andrew.cmu.edu
```

Clone the Github repository to your `www` directory:

```
$ git clone git://github.com/713/project.git ~/www/teamB
```

Give permission to contrib web server to run CGI scripts:

```
$ cd ~/www/teamB/
$ fs sa . contrib.[Your AndrewID]@club.cc.cmu.edu rlidwk
```

Give write permissions to the CGI, PHP scripts, and output folder:

```
$ chmod +rw user_email.txt
$ chmod 755 user_results
$ chmod 755 *.php
$ chmod 755 *.cgi
```

Edit the location of your server URL in `config.py`:

```
$ nano ~/www/teamB/config.py
```

Bypass other limitations:
Because AFS's Python version 2.6.6 breaks the smtplib for sending emails, we can use Enthought Python Distribution (EPD)'s Python to bypass this problem. For your convenience in not having to install EPD add an alias to your shell. Assuming you are using bash, the most common shell,

```
$ echo 'alias
    python="./afs/andrew.cmu.edu/usr23/lleung/epd_free-7.3-2-rh5-x86_64/bin/python"'
    >> ~/.bashrc
$ source ~/.bashrc
```

Change the "`HOST_713PROJECT=`" variable to:

```
HOST_713PROJECT="http://www.contrib.cmu.edu/~YourAndrewID/teamB/"
```

8.3 Accessing Web App

1. If hosting locally, open the `index.html` in browser by dragging it out of its folder and into a browser
   or by typing its location in the browser's URL box starting with "`file://`"
   If hosting on AFS, type into the URL box:

   `http://www.contrib.andrew.cmu.edu/~YourAndrewID/teamB/`

2. Select sequences to process.
3. Provide an email address for receiving an email to view the results when job is finished.
   The email is from `03713.project@gmail.com`
   and with the title "`Job Completed: A message from 03-713 Team B's web app`"
4. Click on the blue "Process" button.

Web App Results
After the pipeline is done processing, an email is sent. Within the email, there is a link that redirects the user to the results webpage.

References

[1] Aaron E. Darling, Bob Mau, and Nicole T. Perna. 2010. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss, and Rearrangement. PLoS One. 5(6):e11147.

[2] Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9(8), 772.

Guindon S and Gascuel O (2003). A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood". Systematic Biology 52: 696-704.

[3] Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.

[4] Domazet-Lošo M, Haubold B. Alignment-free detection of local similarity among viral and bacterial genomes. Bioinformatics. 2011;27:1466–1472. doi: 10.1093/bioinformatics/btr176.