Pipeline for Analysis of Horizontal Gene Transfer in Bacterial Genomes

Stuti Agrawal
Rebecca Elyanow
Luigi Leung
Prateek Tandon
Yiming Xin

Contents

1 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

2 Alignment

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

3 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).

4 Building the phylogenetic tree

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

5 Identifying HGT regions

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

6 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.

7 Graphical User Interface

The GUI is web based. During development, we hosted it via Apache web server locally and CMU's AFS server remotely.

Setup
The following instructions apply to OSX and unix-based operating systems. And instead of `nano`, 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`

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
   ```

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`)

4. 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://`"

Web App

Now that `index.html` is open in the browser,

1. Select sequences to process.
2. 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`"
3. 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.