Pipeline for Analysis of Horizontal Gene Transfer in Bacterial Genomes

Stuti Agrawal
Rebecca Elyanow
Luigi Leung
Prateek Tandon
Yiming Xin

Contents

1 Pipeline Overview

The purpose of the pipeline is to take in as input a set of whole genome sequence data, identify the core and distributed genomes as well as regions of horizontal gene transfer, create phylogenetic trees for the regions of horizontal gene transfer, and annotate the genomes. This is accomplished using four publicly available software packages: MAUVE, Alfy, jModelTest, and Phylip. MAUVE is used to align the sequences as well as to identify the conserved regions among all sequences (after removing HGTs, this makes up the core genome). Alfy is used to identify regions of horizontal gene transfer and input them into a dictionary. For each entry in the dictionary, a phylogenetic tree is constructed. The most appropriate model is chosen using jModelTest and the tree is created using Phylip. If the region overlaps with a region in the core genome, that region will be removed. Once all the HGTs are removed from the core, a tree will be generated in the same manner.

1.1 Graphical User Interface

The purpose of the graphical user interface (GUI) is to provide users an intuitive and user-friendly experience. Due to the diverse number of operating systems and devices on the market, it is best to minimize cross-platform issues by offering a web-based GUI that is accessible from any environment using the tools (like an internet browser) that are already installed on the user's system. For a user-friendly interface, the Twitter Bootstrap library is used to style the web app with large buttons and legible fonts. For communications between the pipeline and the GUI, JavaScript, PHP and CGI scripts are used to minimize server setup and increase hosting compatibility if users decide to host the program themselves.

2 Testing MAUVE

3 Testing Alfy

4 Testing Phylip

5 Comparison to other software

Through our research, we have not found another piece of software that takes in raw sequence data and performs whole genome alignment, HGT identification, tree building, and annotation. There are many programs that solve pieces of the problem (some of which we have implemented in our solution), but none that solve every part of our problem.

For our comparison, we will explore available software packages that offer solutions to parts of our pipeline: whole genome alignment, HGT identification, tree building, and annotation.

5.1 Whole Genome Alignment

There are many software packages that deal with whole genome alignment including: BLAST [8], MUMMER [6], LAST [7], and Mugsy [5].

We choose to use MAUVE [1] over these other software for a few reasons:

1. It runs from the command line on Mac, Linux, and Windows
2. It does not require a reference sequence
3. It can take both genbank and fasta files as input
4. It outputs a .backbone file which can be utilized for identification of the core genome

5.2 HGT identification

We looked into two HGT identification packages, RDP and Clonal Frame, before settling on the package we are currently using, Alfy. We decided not to use RDP [9] because 1. it only runs on Windows machines and 2. The current version is still unstable.We decided not to use Clonal Frame [10] because 1. it only runs on Mac/Linux machines and 2. it has a history of being very slow and difficult to use (according to discussion with Linda Hiller). We choose Alfy [4] because it can run on all platforms and, testing with 15 genomes, it runs in under 5 minutes.

5.3 Tree building

5.4 Sequence annotation

6 GUI

Initially, we decided to use Django to serve our web app due to its language similarity with our pipeline scripts. However, though testing and learning more about the framework's offerings, we decided to abandon it in preference to a more server-compatible traditional framework using JavaScript, PHP and CGI.

We choose to abandon Django in preference to the alternative for a few reasons:

1. We had trouble hosting Django on the school's AFS server due to install permissions
2. JavaScript, PHP and CGI had been around for over 18 years and is compatible on almost all commercial servers. This will help lower users' costs if users decide to rent server space.
3. Django is picky about code correctness/formatting and therefore getting a draft preview of how the web app will look can be time consuming.
4. We had no experience with Django and going through the tutorials made us realize developing on it will be completely different from what we had learned in the past.

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.

[5] Angiuoli SV and Salzberg SL. Mugsy: Fast multiple alignment of closely related whole genomes. Bioinformatics 2011 27(3):334-4

[6] "Versatile and open software for comparing large genomes." S. Kurtz, A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S.L. Salzberg, Genome Biology (2004), 5:R12.

[7] Adaptive seeds tame genomic sequence comparison. SM Kielbasa, R Wan, K Sato, P Horton, MC Frith, Genome Research 2011.

[8] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

[9] Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26, 2462-2463.

[10] Inference of Bacterial Mi-croevolution Using Multilocus Sequence Data" by X. Didelot and D. Falush, Genetics, Vol. 175, 1251-1266, March 2007.