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 GUI

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 [9], MUMMER [7], LAST [8], and Mugsy [6].

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 [10] because 1. it only runs on Windows machines and 2. The current version is still unstable.We decided not to use Clonal Frame [11] 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.

We use the methods mentioned in the papers [4],[5] to detect HGT events. If we want to identify the HGT regions of a genome(say seq1), we run the Alfy with seq1 as query and all the other strains genome as subjects. Alfy will output all the potential HGT position intervals with the strains where seq1 might gets HGT from. We run Alfy for all sequences in this way. Then we use kr [5], another software working as an estimator of the pairwise number of substitutions between long DNA sequences, to generate a correlation matrix of all sequences. After that, we form a closest mates group for all the sequences we test by choosing the corresponding sequences in the matrix with a value below the threshold(we set it to 0.0045 temporally). Then we just filter the output of Alfy by removing the regions whose subject sequences are in the closest group of query sequence.

kr estimates global distances between genomes while Alfy estimates the local homology to detect horizontal gene transfer. So if there are two strains found to have local homology but no global homology, then it is reasonable to infer that it is a HGT event.

Also, the accuracy of this method seems acceptable given the result reported in the paper[4],[5]. However, in the process we test our code, we found that some sequences have no closest-mates group output, which means the sequence is not close to any of other sequences according to the matrix we generate. In this situation, the whole sequence of such strain would be identified as HGT regions which makes no sense. So we should choose the sequences which we can assume have relatively high homology as inputs in order to avoid such situation.

5.3 Tree building

5.4 Sequence annotation

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] Domazet-Lošo M, Haubold B. Alignment-free detection of horizontal gene transfer between closely related bacterial genomes. Mobile Genetic Elements 1:3, 230–235; September/October 2011;G2011 Landes Bioscience.

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

[7] "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.

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

[9] 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.

[10] 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.

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