Elucidating The Effects Of CRISPR-Cas Systems On Lateral Gene Transfer Using Network Analysis

Siddharth Reed
400034828

October 26, 2018

Introduction

The focus of the project is to see whether strains with CRISPR are affecting LGT rates in congeneric species, using techniques from network theory. I found 48 genera, 39 with both CRISPR and Non-CRISPR strains, and 8 with only Non-CRISPR strains, all with > 15 genomes (with CRISPR presence defined by the same database Athena used `http://crispr.i2bc.paris-saclay.fr/`, all with GCF numbers).
For each set of genomes with both CRISPR,Non-CRISPR strains I will estimate insertion, deletion rates for the CRISPR and Non-CRISPR strains separately with Markophylo. Next I will construct an LGT network using the species tree and a set of gene trees (likely 50). Multiple replicate networks for a single set of genomes can be computed using different random subsets of gene trees. For each set of networks multiple statistics can be computed to compare the CRISPR node to the Non-CRISPR nodes, to see if the presence of CRISPR systems impact a node's effect on the network.
Networks statistics and Markophylo estimates will also be calculated for the genome sets with only Non-CRISPR strains. Different network statistics can be used to compare the mixed networks to the exclusively Non-CRISPR networks, as outlined below. Ultimately the questions are:

- Are LGT network dynamics different in networks with CRISPR-containing nodes than without?
- Are CRISPR-containing nodes different from Non-CRISPR nodes in the same network?
- Are maximum likelihood insertion,deletion rate estimates related to LGT network dynamics?

Steps

Outline of the project

- Pick sets of congeneric genomes
  - 40 Sets of congeneric genomes (each > 15 genomes) with CRISPR and Non-CRISPR strains
  - 8 Sets of congeneric genomes (each > 15 genomes) with only Non-CRISPR strains
  - All genomes have links to refseq
  - CRISPR presence as defined by `http://crispr.i2bc.paris-saclay.fr/crispr/BLAST/CRISPRsdatabase`, `http://crispr.i2bc.paris-saclay.fr/crispr/BLAST/noCRISPRsdatabase`, the same database that Athena used for her thesis
- For each set of congeneric genomes
  1. Filter mobile genetic elements
  2. Group genes into families by reciprocal blast hits

3. Build Presence/Absence matrix of gene families
4. Pick genes for species tree if they are present in all genomes (genefamily11.R)
5. Build species tree

- Use Markophylo to estimate gene insertion/deletion rates for CRISPR/Non-CRISPR taxa in genome sets with both, as well as for insertion,deletion rate for the whole tree
- Compare those whole tree estimates to whole tree estimated for those genome sets with only Non-CRISPR taxa
- Markophylo estimates also provide expectations for what the network statistics will say, lower insertion,deletion rates will presumably lead to sparser, lower total weight networks
- Network Analysis (for each set of congeneric genomes)
  - Building the network (using HiDe `http://acgt.cs.tau.ac.il/hide/`
    * Pick random subsets of genes to build gene trees
    * Build sets of gene trees, (50 genes per set) (bootstrap support for gene trees can also be incorporated into building the networks by specifying support a cutoff)
    * Each tree to build network must be rooted
    * HiDe authors also state that thier method is very robust to using gene trees lacking many taxa in the species tree as well as noise.
    * Using species tree from the Markphylo estimates and the sets of gene trees, use HiDe to produce LGT networks, (1 per set of gene trees). Each gene tree set acts as a replicate for building a network for a set of congeneric genomes. This will allow for statistical test when comparing network statistics between individual nodes or between entire congeneric genomes sets
  - Each HiDe network is an edge list, with a weight corresponding to an estimation of the proportion of genes that were transfered along that edge.
  - Each gene is scored between [0,1] of being transferred (1 is most likely a transfer) and each edge is the sum of all gene scores for that edge. The maximum possible edge score (i.e. more gene transfer) is the total number of input gene trees.
  - Compute network statistics (using all network replicates
    * For each node set (in networks with CRISPR and Non-CRISPR nodes)
      · Edge weight distribution (bias larger $\implies$ more transfer, more skewed $\implies$ certain nodes drive transfers)
      · Average total edge weght (larger $\implies$ more transfer)
      · Centrality (higher $\implies$ stronger driver of transfer)

        - Associativity (do CRISPR nodes transfer more with eachother or with Non-CRISPR nodes?)
    * For entire network (to compare Non-CRISPR only networks to mixed networks)
        - Total edge weight (Do CRISPR containing networks have higher edge weights?)
        - Average edge weight (Do CRISPR containing networks have higher edge weights?)
        - Density (do networks with more CRISPR nodes have transfer between more species?)
        - Clustering Coefficient (tight separate clusters vs sparser,network wide connections?)
        - Network Diameter (how many transfers on average are required for a gene to transfer between any two nodes)

- Do the expectations from Markophylo estimates meet the calculated network statistics (lower weight, sparser networks for genera with CRISPR nodes, if CRISPR is assumed to inhibit LGT)

- Are any of these network statistics correlated with insertion,deletion rates estimated by Markophylo?

Three sets of comparisons are going to be made:

- Difference between CRISPR and Non-CRISPR nodes in the same network

- Difference between networks with CRISPR nodes and ones with no CRISPR nodes

- Insertion,deletion rate estimates and whether they associate with trends in network properties

Time Considerations

From the steps outlined above, it seems that the most time consuming steps will be grouping the genes into gene families (all vs all BLAST) and constructing trees. Most of the scripts to build the Presence/Absence matrix were written already (either by you or me last summer). The authors of HiDe state that the running time for their network construction is < 6hours on a dataset of 22,430 genes tress, with a single core and 16GB RAM, which is much much larger than my networks. Further computing the networks statistics will also be fairly quick, as all the networks will be relatively small (> 100 nodes in most cases). igraph nad networkx are popular graph utility libraries that implement the calculations of the statistics I mentioned.

Further Considerations

Things I still need to consider:

- Double checking the CRISPR strain labels? (BLASTing against Cas1,Cas2 genes, genus specific?, CRISPRone Tool?, NCBI annotations?, MySacFinder HMMs?)
- Which statistical tests to compare network statistics?
- Bootstrap/weight cutoffs for edges in network?
- Does large edge variance $\implies$ variance in which genes transferred across that edge (function specific)?
- Create function specific gene tree sets for testing?
- What tree building program to use?
- How many replicates for networks, gene trees (bootstraps)?
- How to pick outgroups for trees?
- Which rooting program/method to use (for HiDe)?
- Annotate specific CRISPR system types (I,II,III)?
- build gene trees of Cas,Cpf genes, compare Cas trees network to transposases trees, ribosome trees (high and low expected transfer respectively)
- CRISPR array trees for networks?
- subsample PA matrix to get indel rates for specific gene tree networks, more data points for netstats against indel rate comparisons?