CMSC829A Final Report

Angela Jiang

December 6, 2023

1 Introduction and Background

In horizontal gene transfer (HGT), genes are not inherited from parent to offspring but through other mechanisms. Horizontal gene transfer plays a critical role in both prokaryotic and eukaryotic evolution. Hence, it is important to identify horizontally transferred genes. However, there are few methods available.

One recent method relies on synteny index. Synteny is the idea that gene positions are mostly conserved.

2 Methods

2.1 Implementation

Sevillya et al. (2020) did not make their code publicly available. Therefore, in order to conduct a simulation study, their method was replicated to the best of the author's ability.

Given two genomes, $G1$ and $G2$, we wish to obtain a list of HGT-suspected genes. That is, genes with an exceptionally low synteny index (SI). SI is defined as the number of shared genes in a neighborhood of k genes. Genes don't have identical nucleotide sequences across different genomes, but they can evolve from a common ancestor. Orthology detection methods, such as OrthoFinder, can find gene families through clustering.

We cannot rule out that these HGT-suspected genes evolved through gene loss or neighborhood rearrangement. Neighborhood rearrangement is defined as neighborhoods swapping positions. Hence, Sevillya et al. adopted a probabilistic approach to determine whether an HGT-suspected gene represents a true horizontal gene transfer event, or whether it evolved by chance.

2.2 Simulations

Simulations were run with the following procedure, similar to Sevillya et al. (2020): a genome $G1$ was created with $|G1| = 1000$, and each gene was 1000 base pairs in length. At each position in each gene, a nucleotide was chosen at random, with each base pair A, T, C, and G having equal probability. Each gene had an associated mutation rate, which was chosen randomly via a normal distribution with a given mean and standard deviation. Then, genome 2 $G2$ was created as an identical copy to $G1$, along with genomes 3 and 4, $G3$ and $G4$, which served as witness genomes.

3 Results

Note that method failed when genes have a hamming distance of 0 due to DivisionByZero error, which is unlikely, but possible in a biological context of a very conserved or very closely related species.

4 Discussion

ayayayay

5 Conclusion

We blabalba