Claremont Colleges
Scholarship @ Claremont

HMC Senior Theses HMC Student Scholarship

2016

Building a History of Horizontal Gene Transfer in E. Coli

Matthew Wilber

Recommended Citation

Wilber, Matthew, "Building a History of Horizontal Gene Transfer in E. Coli" (2016). *HMC Senior Theses*. 75.
https://scholarship.claremont.edu/hmc_theses/75

This Open Access Senior Thesis is brought to you for free and open access by the HMC Student Scholarship at Scholarship @ Claremont. It has been accepted for inclusion in HMC Senior Theses by an authorized administrator of Scholarship @ Claremont. For more information, please contact scholarship@cuc.claremont.edu.

Building a History of Horizontal Gene Transfer in E. Coli

Matt Wilber

Eliot Bush, Advisor

Darryl Yong, Reader

HARVEY
MUDD
COLLEGE

Department of Mathematics

April, 2016

Copyright © 2016 Matt Wilber.

The author grants Harvey Mudd College and the Claremont Colleges Library the nonexclusive right to make this work available for noncommercial, educational purposes, provided that this copyright statement appears on the reproduced materials and notice is given that the copying is by permission of the author. To disseminate otherwise or to republish requires written permission from the author.

Abstract

Bacteria's ability to pass entire genes between one another, a process called Horizontal Gene Transfer (HGT), has a major impact on bacterial evolution. In an ongoing project at Harvey Mudd, computational methods have been used to catalogue the HGT events that have impacted a group of closely related bacteria.

This thesis builds on that project, by improving our ability to identify gene families — groups of genes in different strains that are related. Previously, similarity was measured only by comparing two genes' DNA sequences, ignoring their positions on the organism's DNA. Here, we leverage genes' relative position to make a better measurement of gene similarity. These improved similarity measurements will improve the existing pipeline's ability to identify HGT events.

Contents

Abstract **iii**

Acknowledgments **ix**

1 Introduction **1**

2 Background **3**
- 2.1 Mechanisms of HGT 3
- 2.2 Detecting HGT Events 3
- 2.3 Harvey Mudd College Pipeline 4
- 2.4 Variation of Information 8

3 Methods **9**
- 3.1 Evaluation of the Harvey Mudd College Pipeline 9
- 3.2 Improvements to the Harvey Mudd College Pipeline 11

4 Results **15**

5 Conclusions and Future Work **19**

Bibliography **21**

List of Figures

2.1 Visualization of the variance of information metric. Adapted from Figure 2 in (Meilă, 2007). 7

3.1 The species tree used as input for the ALF evolutionary simulation. The tree has two main subtrees, with a large amount of evolutionary time in separating the two. The resulting genomes analyzed are those in the more recent subtree: SE001, SE010, SE011, SE012, SE013, and SE014. 10

3.2 Two sequences of genes, with gene similarity represented by similar colors. Genes *B* and *G* appear similar, but by color alone it is unclear whether they are homologous. However, the evidence that their neighbors *C* and *F*, as well as *A* and *H*, are highly similar, is indicative of *B* and *G* being homologous as well. 11

4.1 A plot of distances from the true clustering of gene families to the estimated clustering of gene families, based on unmodified gene similarity measures (x–axis) and modified gene similarity measures (y–axis). For all samples, $\sigma = .8$ was used as the threshold parameter. The dotted red line is a best–fit line with intercept fixed at 0 and slope equal to 0.64. 16

4.2 The distance from the estimated gene families to the actual gene families, using the same evolutionary simulation, for various values of the threshold parameter σ. When gene similarity scores were not modified, the distance was 0.92. . 17

Acknowledgments

I'd like to thank my advisor Eliot Bush for his insight and support throughout my research. I'd also like to thank my labmate Kevin Heath for his help acquainting me with the project's code base.

Chapter 1

Introduction

Today, there is increasing pressure on scientists to understand the means by which bacteria evolve. Antibiotic treaments for bacterial illnesses are becoming increasingly ineffective, as populations of bacteria develop resistance traits. Many strains are becoming resistant to multiple types of antibiotics, causing difficulties or sometimes even an inability to treat bacterial disease (Levy and Marshall, 2004).

When antibacterials are used to treat illnesses, there is strong selective pressure for bacterial populations to develop antibacterial resistance. Bacteria have shown the ability to quickly adapt to this pressure, with affected populations quickly developing resistance to one or even multiple drugs. High percentages of bacterial strains are gaining resistance to commonly used antibiotics (Wenzel and Edmond, 2000). Understanding exactly how this resistance develops has become a necessity for the medical field.

When bacterial populations develop antibiotic resistance, it is frequently gained through gene transfer (Gyles and Boerlin, 2013). More specifically, the resistance is likely to originally be transferred from organisms of a different strain or species, or even a bacteriophage. This type of transfer is known as horizontal gene transfer, or HGT.

Horizontal gene transfer turns out to be a significant cause of evolution in bacteria. It has been frequently observed to introduce genes that increase the fitness of a population, or change the way the population interacts with its environment. While biologists are identifying more and more examples of HGT, and developing a better understanding of how HGT occurs, little work has been done to develop an exhaustive list of HGT events that have introduced novel genes to a set of closely related strains of bacteria.

In this paper, we seek to build upon previous work to write a framework that allows us to construct such a list, given current knowledge of bacterial genomes. In particular, by using computational methods in genetics to identify groups of similar genes in related strains of bacteria, we hope to determine which genes have been horizontally transferred from organisms external to our set of strains. The ability to determine exactly where horizontal gene transfer has occurred may give us a better understanding of the situations in which it occurs, so that we may tackle problems like the growing number of occurrences of antiobitic resistance with greater information.

Chapter 2

Background

2.1 Mechanisms of HGT

There are various mechanisms through which HGT can occur. When a cell takes up genetic information from its surroundings and incorporates the information into its own genome, it has undergone *transformation*. A cell can also obtain genetic information directly from other organisms. When two bacteria have contact between their cell membranes, it is possible for genetic information (such as a plasmid) to be transferred across the boundary. This process is known as *conjugation*. Bacteriophages can also transfer genes to bacteria after attaching to the cell membrane through a process known as *transduction*.

Understanding these mechanisms can be important for solving the problem of detecting past HGT events, given a genome. For example, horizontally transferred genes may be flanked on both sides in the genome by repeating sequences (Ochman et al., 2000). Identifying these flanking sequences is one signal that may be characteristic of genes that originate from phages or plasmids.

2.2 Detecting HGT Events

Past researchers have attempted to identify HGT events in various ways. One simple way is to search for regions of a genome where the nucleotide or amino acid composition (e.g. the GC content) is significantly different from the rest of the genome. These regions are more likely to have originated from another species. However, it is still possible for regions to end up with an unusual composition due to other factors, such as selective processes. This method can also fail to identify HGT events when the transfer is between closely related species, or if the event happened far in the past and the gene's composition has changed to reflect that of the recipient's genome (Eisen, 2000). Therefore, this method, when used alone, has the potential for many false positives and false negatives in its identification of horizontal transfer events.

Other studies have aimed to identify HGT events by identifying the closest match for each gene in a set of genomes. If the closest match is with a gene from another species, we can infer that horizontal transfer has occurred. These algorithms can work relatively quickly, but again there is the risk of false positives. For example, if the gene in question is actually a vertically inherited gene that both species received from a common ancestor, but two strains are evolving particularly slowly, they might be identified as having undergone a horizontal gene transfer event. Therefore, it is difficult to use this strategy alone to accurately identify HGT events.

Another way biologists have attempted to detect HGT events is by comparing the phylogenetic trees of various genes. When two genes that share a common ancestor have different trees, it's possible that the difference resulted from horizontal gene transfer. However, this relies on the accuracy of phylogenetic reconstruction methods. Furthermore, for closely related strains, computed gene trees may not be well resolved. Phylogenetic reconstruction may also fail to identify the novel genes that have been horizontally transferred from outside the clade. Therefore, we must also be careful when using gene trees alone to identify HGT events.

2.3 Harvey Mudd College Pipeline

The work discussed in this report is part of an ongoing project at Harvey Mudd College under the supervision of Professor Eliot Bush, aiming to use computational techniques to identify all HGT events that have occurred within the genomes of a set of closely related strains. Specifically, the lab has focused on working with strains of *E. Coli*. Previous work in the lab has focused on using unusual genome content patterns to identify HGT events, but in 2014–2015, Madison Hansen began laying the groundwork for an algorithm that takes advantage of both sequence alignments and gene tree comparisons (Hansen, 2015).

2.3.1 Identifying Gene Homology

The pipeline set up by Hansen takes the genomes of a clade of closely related strains, and tries to identify all HGT events that affected the clade. The pipeline first identifies homologous genes as genes with high similarity. It computes pairwise similarity scores between all genes in these genomes using the BLAST local alignment software (Altschul et al., 1990), and uses them to determine homologous pairs of genes.

2.3.2 Reconstructing a Phylogeny

Once we have determined pairs of homologous genes, we can determine which of these pairs are orthologous genes (genes that have been vertically inherited from a common ancestor) that are common to all strains in the clade. We refer to such genes as *core genes*. These genes represent a set of genes that we can use to directly compare the evolutionary progress of each strain in the clade.

For a gene to a be considered a core gene, it must satisfy the following criterion. For each pair of strains in the clade, the gene must exist in both strains, and be a *best reciprocal hit* (BRH) according to the BLAST analysis. A best reciprocal hit occurs when the gene in one strain matches the gene in the other strain better than any other gene in the other strain's genome, and vice versa.

This gives a strict requirement on genes that we consider core genes. After identifying core genes, we can reconstruct a phylogenetic tree for the clade, using the core genes as a standard to measure a distance between different strains in the clade. To measure these distances, we use the software MUSCLE to generate a multiple sequence alignment of all the core genes of every pair of strains in the clade (Edgar, 2004).

There exist several different algorithms that allow us to create a phylogenetic tree once we have an alignment of core genes between pairs of strains or species. For large datasets where speed is a requirement, neighbor-joining is a greedy algorithm that runs quickly but can give sub-optimal solutions, if we use the alignments to find a distance metric between species. In our research, we use maximum likelihood methods which are computationally intensive but guarantee an optimal solution. We use the RAxML implementation of these methods to determine a phylogenetic tree (Stamatakis, 2014). Once we have created a phylogenetic tree for our clade, we have a framework within which we can identify horizontal transfer.

2.3.3 Finding Families of Homologous Genes

In addition to their use in constructing a phylogenetic tree as above, we can also use the homologous genes identified by the BLAST analysis to identify families of homologous genes, with less strict requirements than those we impose on core genes. Once we find these families of homologous genes, we can investigate their distribution within the phylogenetic tree for our clade, to determine whether a HGT event introduced this gene family to our clade. For example, a tree that contained the gene tree for this gene family as a strict subtree of the full tree could be indicative of HGT.

To identify gene families, we simply cluster all the genes in our genome based on the similarities given by BLAST. These similarities are scores between 0 and 1 that measure the percentage of identity between two genes. This percentage identity is calculated using gene alignment methods. Genes with high percentage identity, above an input threshold, are considered homologous genes and are placed in the same gene family cluster.

To achieve these clusters, we use the SiLiX program, which, as its name suggests, implements *single linkage clustering* on genes specified by BLAST output (Miele et al., 2011). The ability of SiLiX to directly take BLAST analysis as input makes it a natural choice for our pipeline.

While other clustering methods may use some kind of centroid or representative of a family to determine whether different genes belong in a family, single linkage clustering is much less strict in its requirements. If a gene is relatively close to *any* gene already in a family, it will be added to the family. This makes sense in the context of homologous genes, since evolutionary processes may drive two genes to evolve in different manners that make them relatively far apart when directly compared, but still similar to a version of the gene that has not undergone evolutionary pressures, and we would still want to consider all three to be in the same gene family.

As an input, SiLiX takes a similarity threshold, based on the similarity scores output by BLAST. Above this threshold, two genes are “linked” and are placed in the same cluster. In our work, we use an 80% similarity threshold of amino acid sequences, the same as the threshold used by Hansen.

The output of SiLiX then gives us gene families, which we can use to identify where in the clade’s phylogenetic tree a gene family can be found. This should give us a subtree, or a union of subtrees, which we can compare to the overall tree structure to determine whether HGT has occurred.

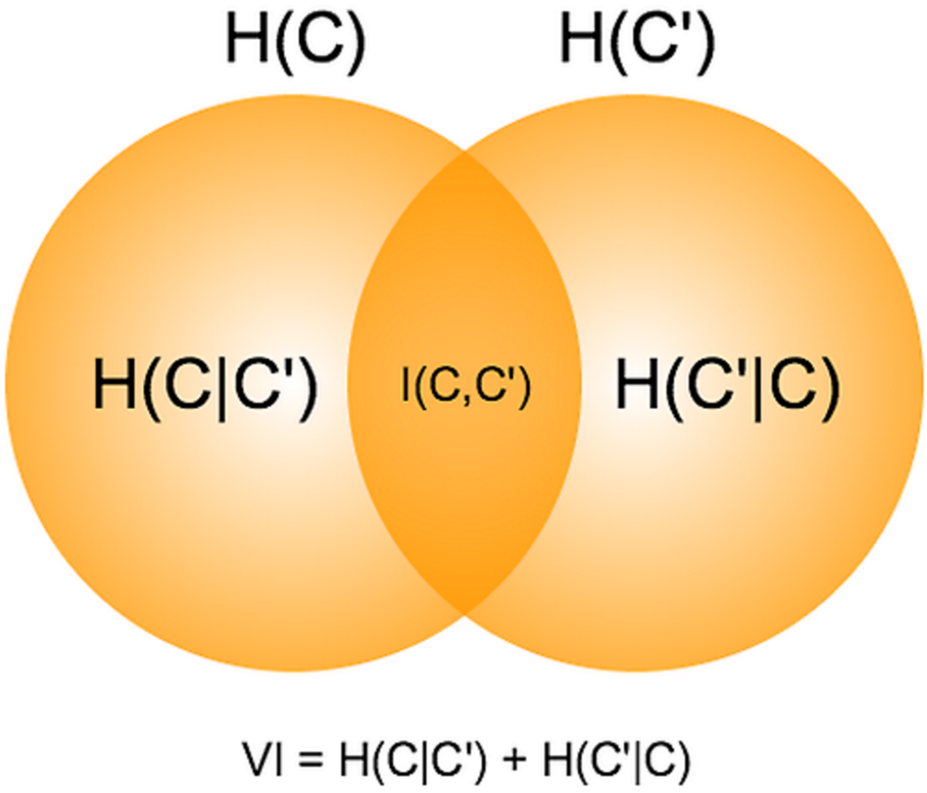

Figure 2.1 Visualization of the variance of information metric. Adapted from Figure 2 in (Meilă, 2007).

2.3.4 Identifying HGT Events

The work by Kevin Heath and Zunyan Wang in the summer of 2015 picked up from the groundwork of Hansen in the spring of 2015 and developed a method to identify HGT events. The method they developed, on a high level, computes a score that measures the extent to which duplication and deletion events had to occur for a gene tree to exist (assuming no horizontal transfer). The greater the extent of these events, the less likely it is that the gene tree actually results from vertical descent. For more information, read the thesis of Kevin Heath, Harvey Mudd College class of 2016.

2.4 Variation of Information

To compare gene families that result from clusterings by SiLiX to actual gene families, we require a metric that can measure distance between clusterings. A clustering is a partition C of a set D into clusters $c_1, c_2, \ldots, c_k$, so that

$$c_i \cap c_j = \emptyset \qquad i \neq j \tag{2.1}$$

and

$$c_1 \cup c_2 \cup \cdots \cup c_k = D. \tag{2.2}$$

The Variation of Information metric is a distance metric between different clusterings of the same objects (Meilă, 2007). It is based on the idea of entropy, the amount of information that is in an object, or how many bits it takes to store that object in computer memory. The entropy of a clustering C is denoted $H(C)$, and is defined

$$H(C) = -\sum_{c \in C} P(c) \log P(c), \tag{2.3}$$

where the probability $P(c)$ of a cluster $c \in C$ is the proportion of elements in D that are in the cluster c. We can also define the *mutual information* of C and C' as

$$I(C, C') = \sum_{c \in C} \sum_{c' \in C'} P(c, c') \log \frac{P(c, c')}{P(c)P(c')}. \tag{2.4}$$

Here, $P(c, c')$ is the probability an element in D is in both the cluster c (in the clustering C) and the cluster c' (in the clustering C'). We can think of $I(C, C')$ as how much information is shared by the clusterings C and C'. The variation of information metric, on the other hand, can be thought of as the amount of information that is *not* shared by C and C',

$$\begin{aligned} VI(C, C') &= [H(C) - I(C, C')] + [H(C') - I(C, C')] && (2.5) \\ &= H(C) + H(C') - 2I(C, C'). && (2.6) \end{aligned}$$

The variation of information metric can be visualized as in Figure 2.1. The conditional entropy $H(C|C') = H(C) - I(C, C')$ can be used to write the variation of information metric as $VI(C, C') = H(C|C') + H(C'|C)$, but the interpretation is still the same. $VI(C, C')$ is the amount of information in the clusterings C and C' that is not shared by the two.

Variation of information is provably a metric and has several other properties that are valuable in the context of clustering distances (Meilă, 2007). As a result, it is a good choice as a way to evaluate the performance of gene family clustering.

Chapter 3

Methods

Following the work of Heath and Wang, the work done in the fall of 2015 and the spring of 2016 has aimed to serve the following goals. Firstly, we developed a framework for the objective evaluation of gene family clustering in the Harvey Mudd College pipeline. We then developed modifications to gene similarity scores based on gene order data to improve inferences of gene homology. Finally, the evaluation framework was used to measure the improvement of the gene family clustering algorithm.

3.1 Evaluation of the Harvey Mudd College Pipeline

In order to evaluate a system that identifies gene families, we require some source of ground truth, a dataset of a genome for a clade that also has all gene families documented. We have opted to use evolution simulation software that can generate both the genome and the gene families we seek to use for such an evaluation. This allows for fast, repeated testing on varied sample data.

For this purpose, we chose the ALF simulation engine, since it appears to be the most thorough evolutionary simulation (in terms of the various evolutionary processes taken into account) that provided the output desired (Dalquen et al., 2012). ALF allows us to control the genome of the initial population, as well as the number of strains desired, and various other evolution parameters. This allows us to evaluate our pipeline for various types of populations that can be modeled on different species or clades of interest.

Since we are mainly interested in identifying Horizontal Gene Transfer events that introduce genetic material from species outside out our clade,

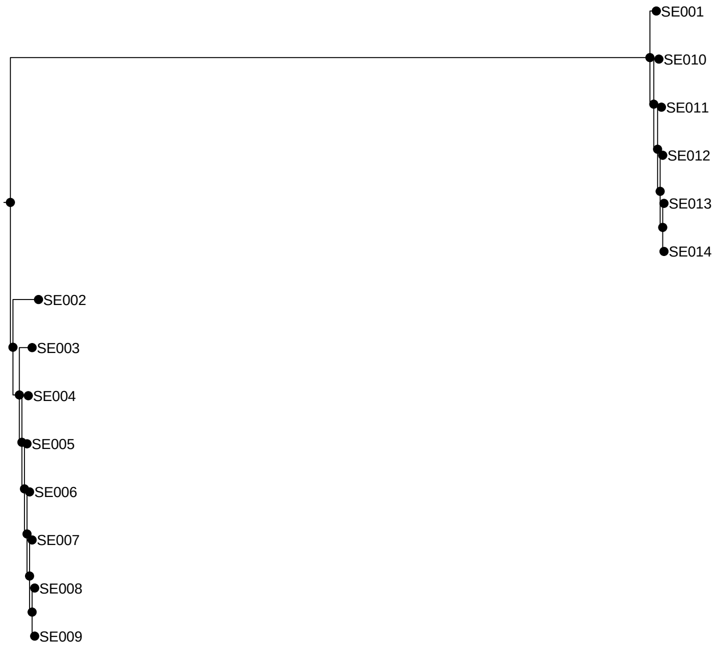

Figure 3.1 The species tree used as input for the ALF evolutionary simulation. The tree has two main subtrees, with a large amount of evolutionary time in separating the two. The resulting genomes analyzed are those in the more recent subtree: SE001, SE010, SE011, SE012, SE013, and SE014.

we need ALF to simulate both populations in our clade, and outside our clade. To do so, we used the tree in Figure 3.1, which has two subtrees that are very distantly related. This allows the older subtree to effectively introduce novel genes to the newer subtree. Therefore, when we analyze the newer subtree's genomes, we emulate the situation we have with the E. Coli clade, where we aim to identify novel horizontally transferred genes.

One of the outputs of the ALF engine is a database of the genome for all the strains evolved in the simulation. This database can be used as the input to the Harvey Mudd College pipeline. We use the pipeline to produce a set of gene family estimates which can be compared to the actual gene families generated by the ALF simulation, which can be extracted from the simulation log.

Mathematically, we can consider the estimated and actual gene families to be two *clusterings* of the genes in the clade into gene family clusters. We can measure the distance between these clusterings using the variation of information metric described above. This distance measured be-

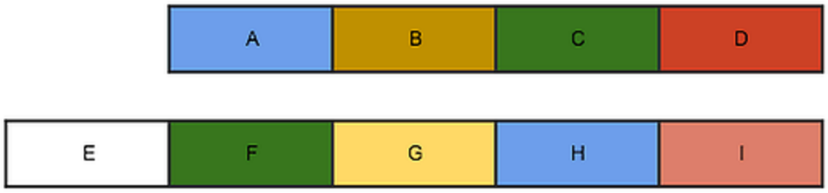

Figure 3.2 Two sequences of genes, with gene similarity represented by similar colors. Genes B and G appear similar, but by color alone it is unclear whether they are homologous. However, the evidence that their neighbors C and F, as well as A and H, are highly similar, is indicative of B and G being homologous as well.

tween our estimated families and the actual families objectively evaluates the pipeline's ability to identify gene families. As changes to the pipeline are made, we can recalculate this distance and see if it decreases to a statistically significant degree.

3.2 Improvements to the Harvey Mudd College Pipeline

With an objective metric to tell us how well our gene family clustering is performing, we gain the ability to make incremental changes to the gene family clustering algorithm, and objectively measure the impact on our results.

3.2.1 Leveraging Gene Order Data

In the current pipeline, the methods used to cluster gene families compare pairs of genes the same way, regardless off their position on a chromosome or plasmid. One improvement to the clustering we aim to implement going forward is to leverage unused gene order data.

To leverage gene order data, we can define a metric to measure the amount of *synteny* between two genes, or how similar the genes' neighbors are to one another. A high degree of similarity between neighboring genes is indicative that the original genes are more likely to be homologous. Figure 3.2 shows an example of how syntenic information may clarify cases where it is unclear, simply based on BLAST similarity, whether a pair of genes is homologous.

Other research has investigated the use of syntenic correlation scores to improve the performance of a general pipeline for gene family identification (Ali et al., 2013). The correlation score is relatively complex, however, and more difficult to interpret and implement than a simple measure of synteny. In this initial exploration of how synteny impacts our results, we instead develop a simple synteny measurement which we can use to modify similarity scores between genes.

We can modify the similarity scores resulting from the BLAST comparison of pairs of genes to reflect the similarity of their adjacent neighbors. While it would be possible to reimplement SiLiX's clustering methods to take this information into account, it may be easier for investigation purposes to simply modify the similarity scores output by BLAST to be scaled towards 1, with the modified score increasing as the similarity of neighbors increases.

Suppose we have a gene A with neighbors B and C, and we aim to determine a modified similarity score $s^*(A, X)$ for some gene X with neighbors Y and Z. If the original similarities $s(B, Y)$, $s(B, Z)$, $s(C, Y)$, or $s(C, Z)$ are greater than some threshold σ, then we want to have $s^*(A, X)$ increase in a manner that scales with the neighbors' similarity. A simple way to do this is through linear interpolation. Let G and G' be two genes. Then define

$$f(G, G') := \begin{cases} \frac{s(G,G')-\sigma}{1-\sigma} & \text{if } s(G, G') > \sigma \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

This maps a similarity score to a value between 0 and 1. We can then define an interpolation parameter,

$$\alpha_{A,X} = \max\left(\frac{f(B, Y) + f(C, Z)}{2}, \frac{f(B, Z) + f(C, Y)}{2}\right) \tag{3.2}$$

which is also stricly between zero and one, and takes into account possible reversals of gene order. We can then interpolate to define our modified similarity score,

$$s^*(A, X) := s(A, X)(1 - \alpha_{A,X}) + 1 \cdot \alpha_{A,X}. \tag{3.3}$$

This gives us a modified similarity score equal to $s(A, X)$ when $\alpha_{A,X} = 0$, and equal to 1 when $\alpha_{A,X} = 1$. Note that we can also write this as

$$s^*(A, X) := s(A, X) + \alpha_{A,X}(1 - s(A, X)), \tag{3.4}$$

a form that makes it clear that the modified score is always at least as large as the old similarity score. This modified score allows us to mathematically encode our intuition that two genes are more likely to be homologous if their neighbors are homologs.

Note that one possible drawback of this modified score is that gene pairs with a BLAST similarity score near 0 may have their modified similarity near 1 if their neighbors are sufficiently similar. This situation seems unlikely if the gene pair is not homologous, but it is possible.

Chapter 4

Results

Since the ALF evolution engine includes random elements, we tested the impact of modified similarity scores by running multiple evolutionary simulations, and using the resulting genomes as input for both the original pipeline and the modified pipeline. A plot of the new and old distances from our gene families to the actual gene families can be found in Figure 4.1.

We can see that the clusterings resulting from modified gene similarity scores in every case are closer than the corresponding clusters using unmodified gene similarity scores. In addition, the decrease was by a relatively consistent percentage of the score. Fitting a best fit line with intercept 0 (since when the unmodified data results in perfect gene family identification, we expect the same of the modified data) results in a line with slope 0.64 (95% confidence interval $[0.62, 0.66]$). This indicates that on average, according to this metric, the modified method improved our gene family identification by 36%.

It is important for us to verify that these improvements are statistically significant, given the randomness inherent in the evolutionary models used. To check this, we ran a Student T–test to check whether the old distances from the correct gene families were significantly different from the new distances, with a significance level of 95%. The T–test resulted in a p–value of $1.35 \cdot 10^{-9}$, indicating that the new and old data were significantly different. Therefore, our improvement is statistically significant.

It is also useful to see how the distance between the estimated gene clustering and the actual gene families is sensitive to the threshold parameter σ. Figure 4.2 shows how the distance between estimated gene families and actual gene families varies as σ changes from 0 to 0.8. It appears the

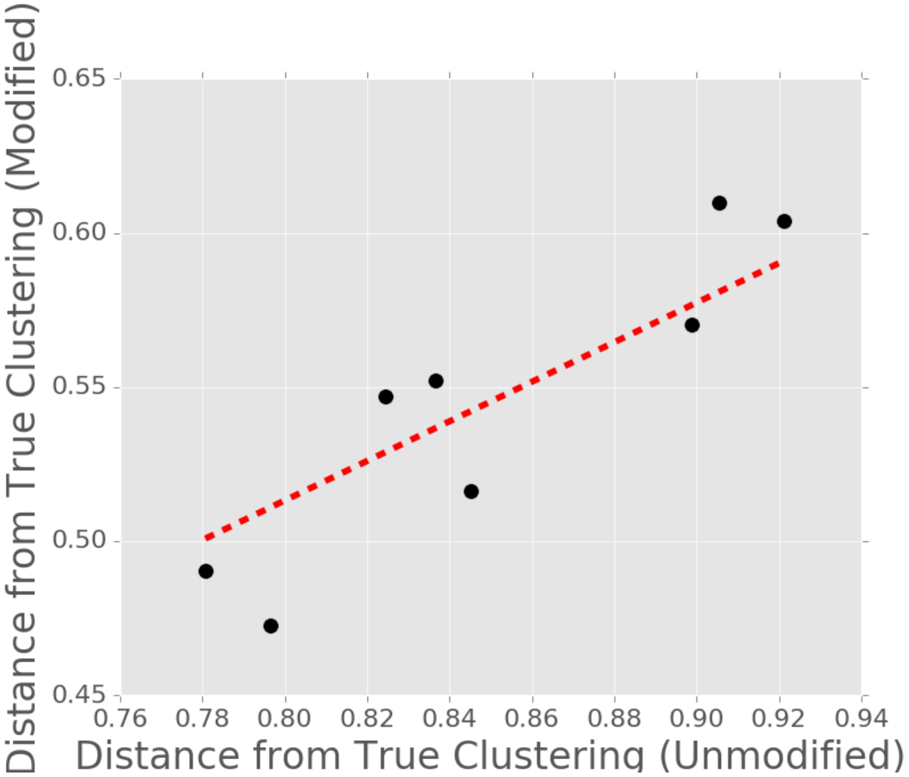

Figure 4.1 A plot of distances from the true clustering of gene families to the estimated clustering of gene families, based on unmodified gene similarity measures (x–axis) and modified gene similarity measures (y–axis). For all samples, $\sigma = .8$ was used as the threshold parameter. The dotted red line is a best–fit line with intercept fixed at 0 and slope equal to 0.64.

performance of the clustering is not very sensitive to σ, and for σ within the range of 0 to 0.8 the resulting distances only vary between about 0.50 and 0.61. Without modifying the gene similarity scores, the distance was 0.92.

The algorithm performed best when σ was near 0.4, indicating that there is a tradeoff between extremes for the σ parameter. If the threshold is too low, the algorithm may be using neighboring genes that are similar by chance as evidence that two genes are homologous. On the other hand, if the threshold is too high, we will be discarding relevant evidence that can be used to make better inferences about which genes are homologous. As a result, we may achieve better results in general by optimizing σ to be near 0.4.

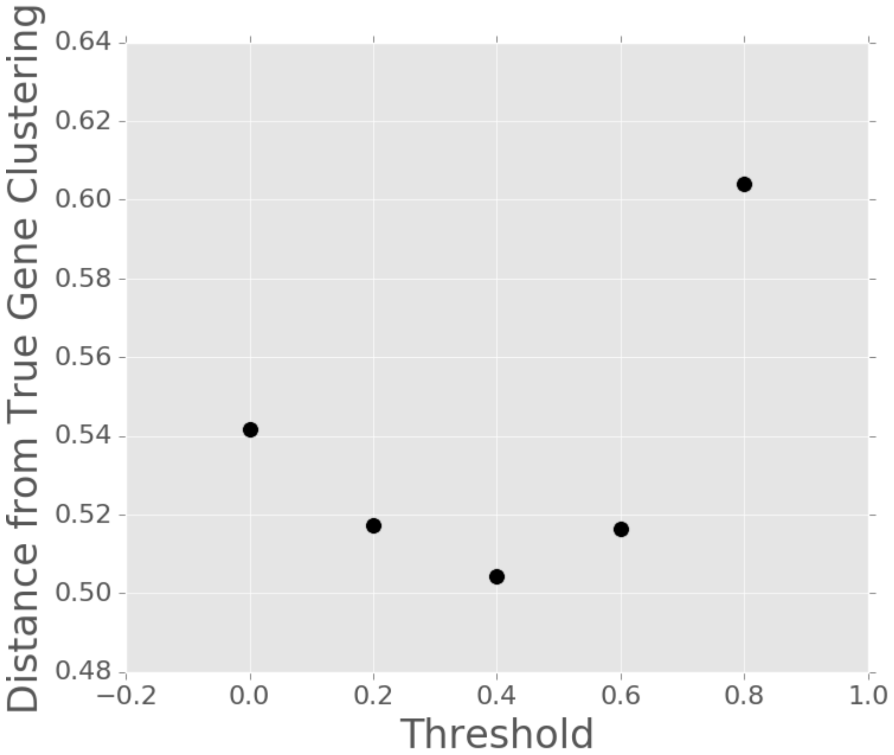

Figure 4.2 The distance from the estimated gene families to the actual gene families, using the same evolutionary simulation, for various values of the threshold parameter σ. When gene similarity scores were not modified, the distance was 0.92.

Chapter 5

Conclusions and Future Work

The project at Harvey Mudd College to exhaustively identify the set of HGT events that have impacted a set of closely related strains is one of the first of its kind, with much past work focusing on individual HGT events or bounds on the frequency of HGT events in given species. Thus far, the project has been able to successfully generate predictions of HGT events, but only in cases where the event is especially obvious or easy to detect. In this report, we have discussed much of what exists in the current pipeline, and our changes to improve gene family identification.

We investigated a simple model to use syntenic information to make better inferences about what pairs of genes are homologous. The method successfully produced gene families that were closer to the actual gene families produced by the simulation. These differences were also shown to be statistically significant.

However, this model is just a starting point. Now that we have shown syntenic information is valuable for making inferences about gene homology, we may be able to take advantage of ideas such as those in (Ali et al., 2013), where a syntenic correlation score that takes into account even more neighboring genes can be used to make better inferences about gene homology.

Furthermore, the idea of using a threshold to consider only gene similarity that is significant, rather than due to randomness, may be used to improve upon the syntenic correlation of Ali et al. In addition, the variation of information metric used here to measure gene family clustering quality can be used to re–evaluate the synteny measure in Ali et al.

As works continues, we hope this pipeline will become a powerful tool for identifying sets of HGT events in arbitrary clades. It is clear that HGT has a significant impact on the evolution of various species, as well as the exchange of key fitness traits. However, much is unclear, such as whether that impact varies between different bacterial groups. Learning more about the extent and impact of HGT is showing us that the evolutionary process of bacterial species is much more complex than was once believed. Hopefully, the Harvey Mudd College pipeline will eventually be able to open one window into viewing exactly how complex these processes are.

Bibliography

Ali, Raja H, Sayyed A Muhammad, Mehmood A Khan, and Lars Arvestad. 2013. Quantitative synteny scoring improves homology inference and partitioning of gene families. *BMC bioinformatics* 14(Suppl 15):S12.

Altschul, Stephen F, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. Basic local alignment search tool. *Journal of molecular biology* 215(3):403–410.

Dalquen, Daniel A, Maria Anisimova, Gaston H Gonnet, and Christophe Dessimoz. 2012. Alf-a simulation framework for genome evolution. *Molecular biology and evolution* 29(4):1115–1123.

Edgar, Robert C. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32(5):1792–1797.

Eisen, Jonathan A. 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Current opinion in genetics & development* 10(6):606–611.

Gyles, C, and P Boerlin. 2013. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Veterinary Pathology Online* 0300985813511131.

Hansen, Madison Hayley. 2015. Reconstructing historical horizontal transfer events in escherichia coli genome evolution. Harvey Mudd College Mathematics Senior Thesis.

Levy, Stuart B, and Bonnie Marshall. 2004. Antibacterial resistance worldwide: causes, challenges and responses. *Nature medicine* 10:S122–S129.

Meilă, Marina. 2007. Comparing clusteringsâĂŤan information based distance. *Journal of multivariate analysis* 98(5):873–895.

Miele, Vincent, Simon Penel, and Laurent Duret. 2011. Ultra-fast sequence clustering from similarity networks with silix. *BMC bioinformatics* 12(1):116.

Ochman, Howard, Jeffrey G Lawrence, and Eduardo A Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784):299–304.

Stamatakis, Alexandros. 2014. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Wenzel, Richard P., and Michael B. Edmond. 2000. Managing antibiotic resistance. *New England Journal of Medicine* 343(26):1961–1963. doi:10.1056/NEJM200012283432610. URL http://dx.doi.org/10.1056/NEJM200012283432610. PMID: 11136269, http://dx.doi.org/10.1056/NEJM200012283432610.