Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Safiyah Bharwani April 10, 2018

Deriving a Metric to Compare Solutions of Malarial Strain Identification Problems and Performing Network Analysis of Disease Outbreaks Across Time

by

Safiyah Bharwani

Ymir Vigfusson, Ph.D.
Adviser

Department of Mathematics/Computer Science

Ymir Vigfusson

Adviser

Bree Ettinger, Ph.D.

Committee Member

Jeremy Jacobson, Ph.D.

Committee Member

2018

Deriving a Metric to Compare Solutions of Malarial Strain Identification Problems and Performing Network Analysis of Disease Outbreaks Across Time

By

Safiyah Bharwani

Ymir Vigfusson, Ph.D.

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Department of Mathematics/Computer Science

2018

Abstract

Deriving a Metric to Compare Solutions of Malarial Strain Identification Problems and Performing Network Analysis of Disease Outbreaks Across Time
By Safiyah Bharwani

This text will build upon the research conducted by Mustonen et al. to use a Bayesian method to identify strains of the *P. falciparum* species of malaria from mixed diagnostic samples. In their *StrainRecon* algorithm, a single weight vector used to measure the presence of malaria in an infected individual is utilized in order to infer the quantity of strains of malaria, the identity of each strain, and the proportion in which each strain is present. This information is grouped into matrix-vector combinations, with matrices containing information on the identity of each strain and the corresponding vector containing information on the proportion in which each strain is represented. Due to the fact that this inference problem is under-determined, there are multiple matrix-vector pairs presented as possible solutions. This work will build upon this prior research by deriving a novel method to compare the solutions produced by the *StrainRecon* algorithm. We will rigorously justify this metric and find an efficient implementation before performing hierarchical clustering over real-world data from the Centers for Disease Control and Prevention (CDC). In particular, we will focus our analysis on understanding how disease outbreaks of malaria have changed over time and attempt to track how the number of strains of malaria has changed in the field. This analysis is of key importance to researchers at the CDC, since there is a sparsity of information on how the number of strains of malaria has changed over time. Throughout this work, an emphasis will be placed on making mathematical results consumable to practitioners at the CDC.

Deriving a Metric to Compare Solutions of Malarial Strain Identification Problems and Performing Network Analysis of Disease Outbreaks Across Time

By

Safiyah Bharwani

Ymir Vigfusson, Ph.D.

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Department of Mathematics/Computer Science

2018

Acknowledgements

To my parents, who have always sought to make the world a better place

To my sister, for her unwavering support

To Grant, for all of the late-night walks

To Naman, for all of the inspirational conversations

To Rifat, for all of the adventures so far and all of the adventures to come

To Ymir, for giving me the tools and the mentorship to find new analytical tools to save lives

To James and Lauri, for all of the help and support they have offered

To Dr. Mitchell and Dr. Ruthotto, for always providing me with a bigger picture

Table of Contents

Deriving a Metric to Compare Solutions of Malarial Strain Identification Problems and Performing Network Analysis of Disease Outbreaks Across Time

Safiyah Bharwani[1], Ymir Vigfusson[2], Rebecca Mitchell[3], and Lars Ruthotto[3]

[1] Dep. of Math/Computer Science, Emory University, Author

[2]Dep. of Math/Computer Science, Emory University, Advisor

[3]Dep. of Math/Computer Science, Emory University,

Co-Researcher

April 10, 2018

1 Abstract

This text will build upon the research conducted by Mustonen et al. to use a Bayesian method to identify strains of the *P. falciparum* species of malaria from mixed diagnostic samples [19]. In their *StrainRecon* algorithm, a single weight vector used to measure the presence of malaria in an infected individual is utilized in order to infer the quantity of strains of malaria, the identity of each strain, and the proportion in which each strain is present. This information is grouped into matrix-vector combinations, with matrices containing information on the identity of each strain and the corresponding vector containing information on the proportion in which each strain is represented. Due to the fact that this inference problem is under-determined, there are multiple matrix-vector pairs presented as possible solutions. This work will build upon this prior research by deriving a novel method to compare the solutions produced by the *StrainRecon* algorithm. We will rigorously justify this metric and find an efficient implementation before performing hierarchical clustering over real-world data from the Centers for Disease Control and Prevention (CDC). In particular, we will focus our analysis on understanding how disease outbreaks of malaria have changed over time and attempt to track how the number of strains of malaria has changed in the field. This analysis is of key importance to researchers at the CDC, since there is a sparsity of information on how the number of strains of malaria has changed over time. Throughout this work, an emphasis will be placed on making mathematical results consumable to practitioners at the CDC.

Contents

2 Introduction

2.1 Background: Malaria

Over the course of human history, the infectious disease that has killed the most humans is malaria [21]. In 2016, there were an estimated 216 million cases of malaria in 91 countries, indicating an increase of 5 million cases in one year. Even more, the malaria parasite caused 445,000 deaths in 2016, with a disproportionately large percentage of deaths among children under 5 years of age [20].

Although malaria is curable, the risk posed by the parasite is compounded since it can hide undetected in the body for days, weeks, months, and in some cases, even decades [12]. As such, in order to treat the disease effectively upon diagnosis, the implementation of a targeted treatment as early as possible is critical to save lives.

However, finding the most effective treatment is made more difficult by the fact that there are five different parasite species of *Plasmodium*, the parasite which causes malaria, that are deadly to humans. Further, each of these species have multiple strains which can infect humans. In particular, strains of the species *P. falciparum* are the deadliest, and it has been shown that most human deaths due to malaria are caused by infection from *P. falciparum* [20]. Treatment of malaria is made more difficult due to the fact that *P. falciparum* quickly mutates and that infected persons often have more than one strain of the parasite [3, 13].

During a 2016 study conducted by Emory University, researchers drew blood from 1,300 untreated children with malaria from Angola, Ghana, and Tanzania. Then, after extracting the DNA of malaria parasites from the blood and using polymerase chain reaction (PCR), they found that approximately 15 percent of the blood samples contained mixtures of both drug-sensitive and drug-resistant strains of malaria [5]. Even more, this study also showed that instead of a general treatment for malaria, more tailored approaches, such as those based upon the specific strains present in the infected individual, might be more effective. This approach could even help prevent parasites from acquiring resistance to malarial drugs while increasing the benefits from treatment.

However, the most common method to identify the identity of each strain of malaria and the proportion in which each strain is present—namely, DNA extraction and amplification using PCR—is expensive both in terms of time and cost. Even more, this method has even been shown to be susceptible to "persistent problems" such as "inadvertent contamination of one strain with another" and "confusion about the identity" of common strains [22]. Further, another disadvantage with this approach is that PCR-based diagnoses provide information focused on the species level as opposed to specific genetic patterns such as strain.

As such, we identify a clear need for a cheaper and more accurate method to find the number, identity, and proportion of strains of malaria an individual infected with *P. falciparum* contains. Such a method would allow for tailored approaches to malaria treatment in such a way that both increases the upside

of treatment and limits the downside of parasites gaining resistance.

2.2 Formalizing the Problem

In the research conducted by Mustonen et al., the *StrainRecon* algorithm was developed in response to the problem above [19]. Let us begin by reviewing the methods of this algorithm.

Suppose that $\mathbf{d} \in \mathbb{R}^m$ represents the given mixed diagnostic sample measurement, which contains information on the proportion of mutations in the sample at each of the m Single Nucleotide Polymorphism (SNP) sites. We would then formalize $\mathbf{d}$ as:

$$\mathbf{d} = \mathbf{M}\mathbf{w} + \mathbf{e}, \tag{1}$$

where $\mathbf{M} \in \{0,1\}^{mxn}$ is a binary matrix that encodes the presence/absence of a mutation at each SNP site for each strain (as 1 or 0, respectively). Further, $\mathbf{w} \in \mathbb{R}^n$ represents the relative frequency of each of the n strains (such that $\sum_{i=1}^{n} \mathbf{w}_i = 1$). Lastly, $\mathbf{e} \in \mathbb{R}^m$ represents noise, which we assume is drawn from a multivariate Gaussian distribution.

The task now is to find $\mathbf{M}$ and $\mathbf{w}$, given $\mathbf{d}$. Due to the presence of noise, the algorithm searches for the solutions to $||\mathbf{M}\mathbf{w}–\mathbf{d}||_2$. We can formalize this as the following Bayesian inverse problem:

$$\pi[\mathbf{M}, \mathbf{w}|\mathbf{d}] \approx \exp(-\frac{1}{2}||\mathbf{M}\mathbf{w}–\mathbf{d}||_{\Gamma}^2 - \lambda||\mathbf{M}||_0) \cdot \pi\mathbf{w}. \tag{2}$$

Then, using Maximum-A-Posteriori (MAP) estimation based block coordinate descent, the algorithm estimates the largest mode of the posterior. Finally,

using Gibbs sampling, which is a kind of Markov-Chain Monte Carlo (MCMC), we quantify the uncertainty of the solutions.

Note that the aforementioned *StrainRecon* algorithm can be generalized to applications of other diseases, such as E. coli and salmonella.

3 Motivation

According to the CDC, it is of vital importance that all treatments of malaria must be informed by a laboratory diagnosis of the strains of malaria that are present in the infected individual [8]. From the *StrainRecon* algorithm, we see that there exists a potential alternative method to laboratory diagnosis that would allow for a more resource-efficient method to detect the identity and proportions of the strains of malaria. However, note that for any given measurement vector $\mathbf{d}$, the *StrainRecon* algorithm finds many possible solutions, since the problem is not injective.

On the other hand, while the *StrainRecon* algorithm does not produce a unique solution, we do in fact find that it is effective at detecting the strains with the highest associated weights. In other words, the algorithm infers the identity of the dominant strains with the greatest degree of confidence and the identity of the least dominant strains with the lowest degree of confidence. In fact, when the *StrainRecon* algorithm was tested on CDC pilot data, it was able to find the most likely number of strains n 80% of the time. Further, of these, the algorithm was able to recover the identities of nearly 100% of the primary strain (which was present in greater than 80% concentration), 92% of the secondary strain (which was present in approximately 2-10% concentration), almost 80% of the tertiary strain (which was present in approximately 0.5-2% concentration), and approximately 70% of the quaternary strain [19]. Putting all of this together, we can see that in order to provide the maximum value for biostatisticians at the CDC, we find that there is a need to derive a method

to compare the various solutions produced by the *StrainRecon* algorithm with the purpose of narrowing the solution set and, if possible, isolating the "most likely" solution. We expect that this narrowed set of solutions will capture the information of the dominant strains.

Furthermore, beyond treatment of individuals infected with malaria, the CDC is also focused on understanding disease outbreaks on a broader population level in order to control and eventually eradicate the disease [9]. As such, it would be of interest to conduct a population level analysis of malaria strains across time in the hope that such an analysis would provide feedback on the success of past interventions and guide future responses to outbreaks of malaria.

Currently, the CDC responds to outbreaks of malaria by increasing surveillance, conducting case investigations, implementing vector control measures, and prescribing antimalarial drugs when an outbreak occurs [1]. Researchers have been able to measure the impact of these interventions through analyzing patient outcomes. Even more, another important metric for success is analyzing how the number of strains of malaria has changed across time. In order to control and eventually eradicate the disease, it is of vital importance to successively reduce the number of strains of *P. falciparum* that infects individuals. However, since the CDC uses a binary test for malaria in the field that simply finds the presence or absence of the parasite, there is a sparsity of information on the number of strains of malaria that are prevalent [4]. This piece of information is particularly important, since the knowledge of strain-level data could help inform targeted preventative measures such as a vaccine for *P. falciparum*. In

fact, the CDC has noted that "the search for a vaccine [for malaria] is considered to be one of the most important research projects in public health," since other methods of treating the disease are not sufficient for elimination [11]. However, the development of such a vaccine is complicated by the fact that the vaccine must factor in the genetic diversity of both the parasite and the infected individual [10]. As such, it would be of vital interest for researchers to be aware of the levels of genetic diversity in *P. falciparum* over time in order to develop an effective vaccine.

Therefore, we again find the need to develop a method to compare various $\mathbf{M}$, $\mathbf{w}$ solutions, since a population-level analysis of *P. falciparum* would necessarily require a comparison of each infected individual's most likely $\mathbf{M}$, $\mathbf{w}$ solution.

Synthesizing all of the above, we summarize the questions of interest as follows:

- **Question 1:** How can we facilitate a comparison of various $\mathbf{M}$, $\mathbf{w}$ solutions?
- **Question 2:** Further, given a blood sample from an individual infected with malaria, what is the most likely $\mathbf{M}$, $\mathbf{w}$ diagnosis?
- **Question 3:** Finally, given population-level data of infected persons across time, how have the disease outbreaks of malaria changed across time?

3.1 Outline

To address the above questions, we will begin by describing the various types of data that we will use in our analysis. Next, we will derive the *SR* metric, which will allow us to compare multiple **M**, **w** solutions. This will be supplemented by a discussion of how to efficiently implement the *SR* metric. Subsequently, we will turn to applying our metric to pilot data in order to attempt to isolate the most likely solutions provided by the *StrainRecon* algorithm. Finally, we will use CDC field data from Asembo, Kenya to perform network analyses that will characterize how disease outbreaks of *P. falciparum* have changed since 1996.

4 Dataset Description

For this research, we work with both real world data from the CDC as well as synthetic data generated in silico.

4.1 Real World Data

We will work with two sets of data from the CDC. Specifically, we will work with measurements from 24 reference SNP sites where there tend to be many mutations specific to *P. falciparum*. The pilot data contains measurements on individual blood samples, while the field data contains measurements from a sample of infected individuals in Kenya in the years 1996, 2001, 2007, and 2012.

4.1.1 Pilot Data

From blood samples which had multiple strains of malaria mixed together, traditional methods were used to first isolate the malarial strains and then amplify them using PCR. From this, we acquired our measurement vector $\mathbf{d}$, which was almost always of $\mathbb{R}^{24}$ dimension (though there were a small minority of samples which had less than 24 SNP sites).

In addition, we were supplied with an approximation for the true binary matrix $\mathbf{M}$ and weight vector $\mathbf{w}$ for each sample. However, it is important to note that $\mathbf{M}$ is only an approximation and in some cases is not fully accurate due to mutation and/or contamination. Even so, it is useful to work with this data as it resembles data from real world conditions in the field.

Nine general types of experiments were performed over three categories of

samples, as seen in the following table:

Experiment	**True Proportions**	**Number of Strains**
A-1, B-1, C-1	0.975, 0.02, 0.005	3
A-2, B-2, C-2	0.95, 0.04, 0.01	3
A-3, B-3, C-3	0.88, 0.1, 0.02	3

Note that each experiment is also tagged with one of the strings BLO (Blood), DBS (Dried Blood Samples), or DNA. Any of the A-1, A-2, and A-3 experiments that are tagged with BLO or DBS are anticipated to have mutations. As such, there may be the same or more strains as expected, since a subset of a single strain may have mutated to form a new strain. However, these mutated strains are expected to be very similar to at least one of the original strains. We can see that this situation would benefit from the use of a metric which would cluster together solutions with small differences, such as those caused by mutations. The development of such a metric will be discussed later in the paper.

Furthermore, observe that each of the proportions listed above include one strain that is present in a much smaller amount relative to the others. As such, as we noted earlier, we will be much less confident about finding the identity of these less dominant strains since noise during the measurement process may affect the algorithm's ability to detect them. Conversely, however, we will have a much greater confidence in the algorithm's ability to identify the more dominant strains, and we will use this key fact during our analysis later in the paper.

4.1.2 Field Data

To analyze how disease outbreaks of malaria have changed over time, we will turn to measurements collected by the CDC in Kenya in 1996, 2001, 2007, and 2012 [13]. This data is valuable because it allows us to draw insights on the overall changes in *P. falciparum* over time by studying measurements from a sample of infected persons. Similarly to the pilot data, we work with measurements from 24 SNP sites specific to *P. falciparum* [6]. In each of the aforementioned four years, we are provided with measurements from 455, 445, 279, and 374 infected individuals, respectively.

However, in contrast to the pilot data, we are not aware of the true number of strains, the true proportions, or the true identities of the strains in this dataset. Instead, in order to analyze the quantity and characteristics of the strains that have infected the population, we must use the *StrainRecon* algorithm to infer the number, relative weights, and identities of the strains. Using this output, we will be able to conduct network analyses on the disease outbreaks over time.

4.2 Synthetic Data

While the pilot data is used to test the *StrainRecon* algorithm under 'field' conditions, the synthetic data is used to fine-tune the algorithm under a variety of different conditions. In this data generated in silico, we first look to create our $\mathbf{M}$ matrix and $\mathbf{w}$ vector by specifying:

- m, the number of SNP sites (typically 24)

- k, the true number of strains

- e, the standard deviation of the multivariate Gaussian distribution from which we randomly sample for noise.

In this way, we are able to create the binary matrix $\mathbf{M} \in \{0,1\}^{mxk}$, the weight vector $\mathbf{w} \in \mathbb{R}^k$ where each of the k elements are randomly found under the constraint that $\sum \mathbf{w}_i = 1$, and the noise vector $\mathbf{e} \in \mathbb{R}^k$ where each of the k elements are found by sampling from N(0, $\mathbf{e}$). Now, using the same formula for our measurement vector, we find our synthetic measurement $\mathbf{y} \in \mathbb{R}^m$:

$$\mathbf{y} = \mathbf{M}\mathbf{w} + \mathbf{e}. \tag{3}$$

Note that each of the elements in $\mathbf{y}$ satisfies $0 \leq \mathbf{y}_i \leq 1$.

5 Developing the *SR* Metric

5.1 Motivation

In order to answer our motivating questions, it is clear that a method to compare multiple **M**, **w** solutions would be useful. In particular, we are interested in grouping together solutions that are similar, since this may allow us to both narrow the set of possible **M**, **w** solutions produced by the *StrainRecon* algorithm as well as identify important changes in malarial strains over time. Thus, we find the need to define a metric space in order to perform clustering on the **M**, **w** solutions.

Defining Differences in Sets of M, w Solutions

In order to understand how the *SR* metric should behave, let us analyze the following sets of **M**, **w** solutions.

Example 1 Each of the weight vectors and strains in the **M** matrices are sorted (from highest weight to lowest) when they are produced by the *StrainRecon* algorithm. However, it may be the case that the same strain is represented in slightly different proportions in the **M**, **w** pairs we are comparing, as in Figure 1.

In such a case, we would expect the *SR* metric to reflect only a small level of difference (i.e., a small distance), since the same strains are present in slightly different proportions.

Example 2 In the above example, we compared two matrices with the same strain identities. However, what if we encounter a case such as Figure 2 in which

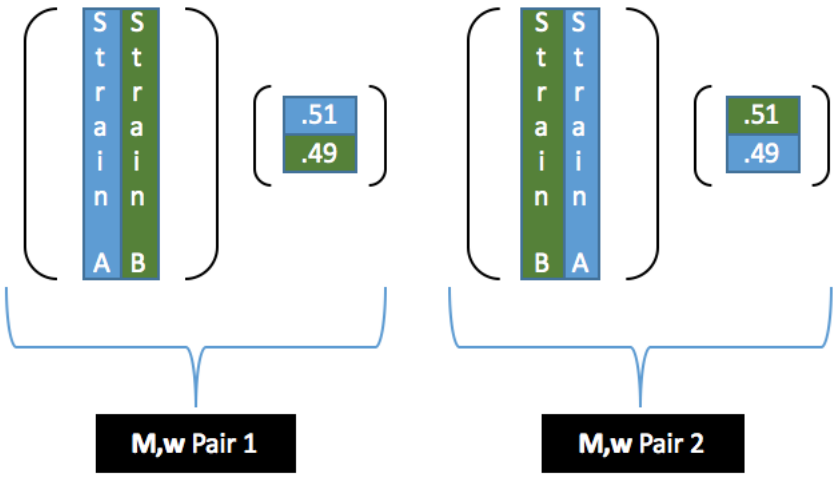

Figure 1

the dominant strains are identical but the least dominant strains are different?

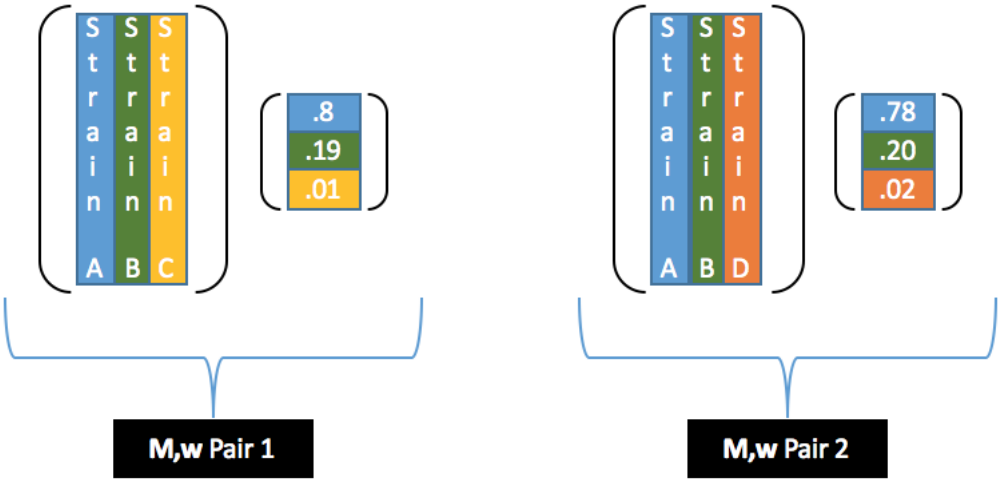

Figure 2

Observe that Strains A and B, the dominant strains, are both present in each **M**, **w** pair, but with slightly different weights. In addition, the least dominant strain in each **M**, **w** pair is different.

From Example 1, we know that we would still expect a high level of similarity between Strains A and B despite small differences in weights. However, the question remains, how do we treat the fact that the identities of Strain C and Strain D are different? Since Strains C and D have the lowest weights in solution 1 and solution 2 respectively, we can intuitively see that the two **M**, **w** pairs are still fairly similar overall. Thus, differences in the least dominant strain are not significant in determining the distance between the a set of **M**, **w** pairs.

Conversely, this also implies that differences in the dominant strains will lead to a large distance between a set of **M**, **w** pairs. We will explore this further in the next example.

Example 3 Finally, what would we expect to see if there existed a small number of mutations in a strain that was present in both solutions?

In Figure 3, we see that the dominant strain in the first solution has a small number of SNP site differences (represented as grey boxes) compared to the dominant strain of the same identity in the second solution. From Example 2, we know that since these differences are present in a dominant strain, we would expect a higher level of distance compared to if they were present in less dominant strains. However, since there exist only a relatively small number of SNP site differences between the dominant strains in these two otherwise identical solutions, we would expect that the distance would still be relatively

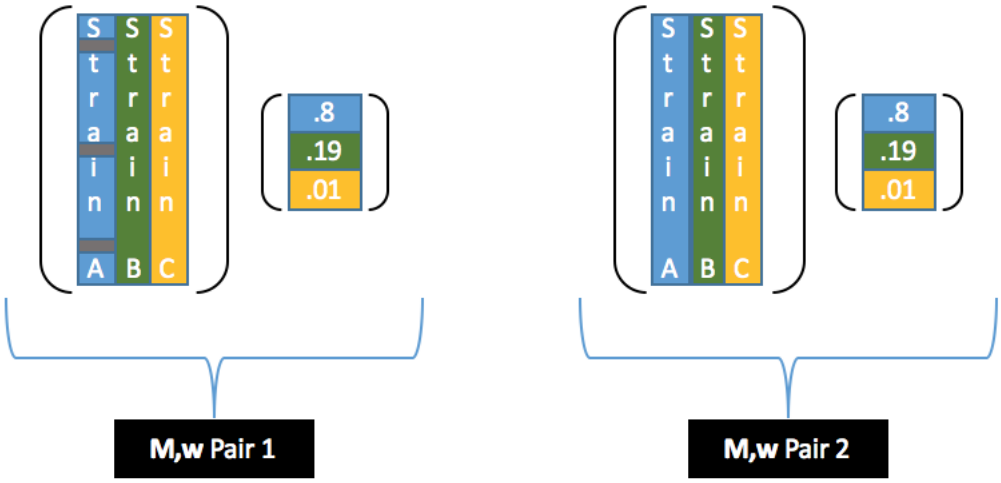

Figure 3: Example 3

small. In other words, mutations that are present in the dominant strain will lead to a higher level of distance compared to mutations present in strains that are lower weighted. Further, a small number of mutations will lead to a smaller distance compared to a large number of mutations.

Formulating Distance in Mathematical Terms

Given the above explanation of how we would expect the SR metric to work, let us now turn to trying to express the SR metric mathematically.

A large challenge encountered with clustering in high dimensional spaces is that even within clusters, there are often differences on a few dimensions. An effective way to deal with this problem is to weight the inputs differently, which as motivated previously, also seems to make intuitive sense for our analysis [7]. In addition, we must find some way to compare each strain of the first $\mathbf{M}$, $\mathbf{w}$

pair with each strain of the second $\mathbf{M}$, $\mathbf{w}$ pair. Finally, the *SR* metric must include some method to quantify the number of mutations.

Given these properties, we propose the following formula for the *SR* metric:

$$SR = \frac{1}{(n_1 \cdot n_2)} \cdot \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} [d(s_i, t_j) \cdot f(s_i, t_j)], \tag{4}$$

where $s_i, 1 \leq i \leq n_1$, represents each strain (i.e. each column) in $\mathbf{M}_1$ and where $t_i, 1 \leq i \leq n_2$, represents each strain (i.e. each column) in $\mathbf{M}_2$. Further, let us suppose that $d(s_i, t_j) = \|(s_i - t_j)\|_1$. This would simply count the number of differences between s_i and t_j. In addition, we define $f(s_i, t_j) = \sqrt{s_i \cdot t_j}$, which would give us a geometric average of the weights.

Let us check if this form would be valid by verifying that the distance between $\mathbf{M}$, $\mathbf{w}$ and a permutation of $\mathbf{M}$, $\mathbf{w}$ would equal to 0. A visualization of this example is seen in Figure 4.

We can see that this formation of the *SR* metric would in fact violate this check, since all of the terms in the summation will be positive, leading to a positive overall distance. Thus, since the *SR* metric will not equal 0, this formulation of the metric is invalid.

However, from this initial formulation, we can draw the following insights:

- Firstly, we can see that any metric that relies upon summing some function of distance multiplied by some function of weight will give us a nonzero overall distance in a case such as Figure 4. As such, we instead consider comparing $\mathbf{M}_1, \mathbf{w}_1$ with the permutations of $\mathbf{M}_2, \mathbf{w}_2$.

- Secondly, although there is value in including the weights to determine

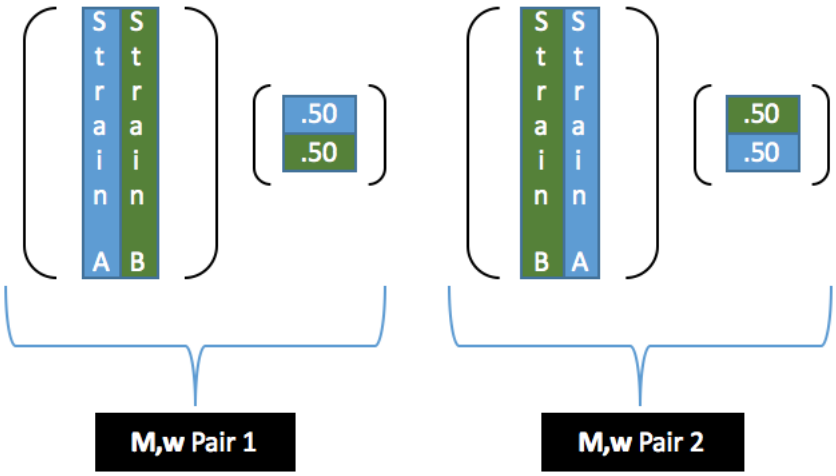

Figure 4

the distance, including the function of weights as a geometric average does not have a clear interpretation. Instead, we suggest comparing $\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1)$ and $\mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2)$. In this way, each strain has been multiplied by its appropriate weight, and therefore both the weights and the SNP site differences can be compared by $\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) - \mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2)$. Additionally, in order to ensure that the distance between $\mathbf{M}$, $\mathbf{w}$ pairs is always positive, we suggest encompassing the *SR* metric in the Frobenius norm (also known as the Euclidean norm), since this norm is commonly used in clustering [14]. Note that the Frobenius norm is defined as $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$.

Therefore, we reformulate the *SR* metric as the following:

$$SR = \min_{P \in \mathbb{P}} \|\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) - \mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) \cdot \mathbf{P}\|_F \,, \tag{5}$$

where $\mathbf{P}$ represents the permuation matrix that minimizes the distance between each pair of $\mathbf{M}$, $\mathbf{w}$ solutions.

We can see that the *SR* metric reflects the properties we listed above, since:

1. The metric is invariant to the reordering of columns.

2. The metric is most influenced by strains with a high weight and least influenced by strains with a low weight.

3. The metric is robust to small changes in the rows. In other words, if there exists a small number of SNP site differences between the pair of $\mathbf{M}$ matrices, the distance is also relatively small.

One potential drawback of the *SR* metric is that it is computationally expensive since we must take the permutation of $\mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) \cdot \mathbf{P}$. We will address these efficiency concerns later in the Implementing the Metric section.

5.2 Deriving the *SR* Metric

$$SR = \min_{P \in \mathbb{P}} \|\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) - \mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) \cdot P\|_F \tag{6}$$

Theorem 5.1 *SR is a metric if the following conditions are satisfied:*

1. $d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{M}_2, \mathbf{w}_2)) \geq 0$

2. $d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{M}_2, \mathbf{w}_2)) = 0$ *iff* $\mathbf{M}_1 = \mathbf{M}_2, \mathbf{w}_1 = \mathbf{w}_2$

3. $d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{M}_2, \mathbf{w}_2)) = d((\mathbf{M}_2, \mathbf{w}_2), (\mathbf{M}_1, \mathbf{w}_1))$

4. $d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{M}_2, \mathbf{w}_2)) \leq d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{N}, \mathbf{v})) + d((\mathbf{N}, \mathbf{v}), (\mathbf{M}_2, \mathbf{w}_2))$

To complete these proofs, let us first prove the following lemmas.

Lemma 5.2 *Permutation matrices are always invertible.*

Proof: Permutation matrices are by definition square matrices obtained from permutations of the identity matrix. By the Invertible Matrix Theorem, since permutation matrices are row-equivalent to the identity matrix, they are always invertible.

Lemma 5.3 $||A||_F = ||AP||_F$ *for permutation matrix P.*

Proof:

We perform the proof for a permutation matrix P that permutes over columns. We will denote $\pi(n)$ as some permutation of the columns, as specified by the P matrix that minimizes the metric distance.

$$||A||_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}|a_{ij}^2| = \sum_{i=1}^{\pi(n)}\sum_{j=1}^{m}|a_{ij}^2| = ||AP||_F^2$$

Thus, the Frobenius norm of a matrix is equivalent to the Frobenius norm of that matrix multiplied by any permutation matrix. While we performed this proof for a permutation matrix P that permutes over columns, a nearly equivalent proof can be performed to show that the result is valid for permutation matrices P that permute over rows.

5.2.1 Proofs for Metric

1. $d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{M}_2, \mathbf{w}_2)) \geq 0$

   This condition is easily satisfied, since the entire *SR* metric is encompassed within the Frobenius norm, which is by definition greater than or equal

to 0.

2. $d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{M}_2, \mathbf{w}_2)) = 0$ iff $\mathbf{M}_1 = \mathbf{M}_2, \mathbf{w}_1 = \mathbf{w}_2$

   - Forward direction:

     If $d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{M}_2, \mathbf{w}_2)) = 0 \rightarrow \mathbf{M}_1 = \mathbf{M}_2, \mathbf{w}_1 = \mathbf{w}_2$

     – Recall that the *StrainRecon* algorithm outputs solutions such that the strains in each $\mathbf{M}, \mathbf{w}$ pair are listed in order of non-increasing weights. As such, we will consider a $\mathbf{M}$ matrix to be valid only if its strains are arranged in terms of non-increasing weights. With this assumption, there exists a unique condition that $\mathbf{M}$ matrices that follow this format are equivalent if $\mathbf{M}_1 = \mathbf{M}_2$.

     – In addition, we assume that $\mathbf{w}_{1_i}$, $\mathbf{w}_{2_i} \neq 0, 1$. This assumption holds true in the *StrainRecon* algorithm.

     Let us continue with our proof of the forward direction.

     Recall that $0 < \mathbf{w}_{1_i}, \mathbf{w}_{2_i} < 1$ and $\mathbf{M}_{1_{i,j}}, \mathbf{M}_{2_{i,j}} \in 0, 1$.

     We need to prove that: $m_{1_{i,j}} w_{1_i} = m_{2_{i,j}} w_{2_i}$ iff $m_{1_{i,j}} = m_{2_{i,j}}, w_{1_i} = w_{2_i}$.

     Let us examine each of the possible values:

     – $m_{1_{i,j}} w_{1_i} = m_{2_{i,j}} w_{2_i} = 0$ iff $m_{1_{i,j}} = m_{2_{i,j}} = 0$

     – $m_{1_{i,j}} w_{1_i} = m_{2_{i,j}} w_{2_i} = 1$ iff $m_{1_{i,j}} = m_{2_{i,j}} = 1$

     – $m_{1_{i,j}} w_{1_i} = m_{2_{i,j}} w_{2_i}$, where $0 < c < 1$ iff $w_{1_i} = w_{2_i}$

Each of these must be true given the assumptions we have made, so the forward direction of this proof is valid.

- Backward direction:

  If $\mathbf{M}_1 = \mathbf{M}_2, \mathbf{w}_1 = \mathbf{w}_2 \rightarrow d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{M}_2, \mathbf{w}_2)) = 0$

  Recall from above that if $\mathbf{M}_1$ and $\mathbf{M}_2$ are listed in order of non-increasing weights, $\mathbf{M}_1$ and $\mathbf{M}_2$ are equivalent if $\mathbf{M}_1 = \mathbf{M}_2$. If $\mathbf{M}_1 = \mathbf{M}_2$, then the minimizing permutation matrix is the identity matrix.

$$|| \begin{bmatrix} m_{1_{1,1}} & m_{1_{1,2}} \cdots & m_{1_{1,n}} \\ m_{1_{2,1}} & m_{1_{2,2}} \cdots & m_{1_{2,n}} \\ \vdots & \vdots \quad \ddots & \vdots \\ m_{1_{m,1}} & m_{1_{m,2}} \cdots & m_{1_{m,n}} \end{bmatrix} \begin{bmatrix} w_{1_1} & 0 \cdots & 0 \\ 0 & w_{1_2} \cdots & 0 \\ \vdots & \vdots \quad \ddots & \vdots \\ 0 & 0 \cdots & w_{1_n} \end{bmatrix} -$$
$$\begin{bmatrix} m_{2_{1,1}} & m_{2_{1,2}} \cdots & m_{2_{1,n}} \\ m_{2_{2,1}} & m_{2_{2,2}} \cdots & m_{2_{2,n}} \\ \vdots & \vdots \quad \ddots & \vdots \\ m_{2_{m,1}} & m_{2_{m,2}} \cdots & m_{2_{m,n}} \end{bmatrix} \begin{bmatrix} w_{2_1} & 0 \cdots & 0 \\ 0 & w_{2_2} \cdots & 0 \\ \vdots & \vdots \quad \ddots & \vdots \\ 0 & 0 \cdots & w_{2_n} \end{bmatrix} \begin{bmatrix} 1 & 0 \cdots & 0 \\ 0 & 1 \cdots & 0 \\ \vdots & \vdots \quad \ddots & \vdots \\ 0 & 0 \cdots & 1 \end{bmatrix} ||_F$$

$$= || \begin{bmatrix} m_{1_{1,1}} w_{1_1} & m_{1_{1,2}} w_{1_2} \cdots & m_{1_{1,n}} w_{1_n} \\ m_{1_{2,1} w_{1_1}} & m_{1_{2,2}} w_{1_2} \cdots & m_{1_{2,n} w_{1_n}} \\ \vdots & \vdots \quad \ddots & \vdots \\ m_{1_{m,1}} w_{1_1} & m_{1_{m,2}} w_{1_2} \cdots & m_{1_{m,n}} w_{1_n} \end{bmatrix} - \begin{bmatrix} m_{2_{1,1}} w_{2_1} & m_{2_{1,2}} w_{2_2} \cdots & m_{2_{1,n}} w_{2_n} \\ m_{2_{2,1} w_{2_1}} & m_{2_{2,2}} w_{2_2} \cdots & m_{2_{2,n} w_{2_n}} \\ \vdots & \vdots \quad \ddots & \vdots \\ m_{2_{m,1}} w_{2_1} & m_{2_{m,2}} w_{2_2} \cdots & m_{2_{m,n}} w_{2_n} \end{bmatrix} ||_F$$

$$= || \begin{bmatrix} 0 & 0 \cdots & 0 \\ 0 & 0 \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 \cdots & 0 \end{bmatrix} ||_F = 0$$

Thus, we see that the backward direction of the proof is also valid. As such, the second condition is also satisfied since both the forward and backward directions of the proof are true.

3. $d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{M}_2, \mathbf{w}_2)) = d((\mathbf{M}_2, \mathbf{w}_2), (\mathbf{M}_1, \mathbf{w}_1))$

   Let P_1 be the minimizer of a such that

$$a = \|\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) - \mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) \cdot \mathbf{P_1}\|_F \,, \tag{7}$$

   and let P_2 be the minimizer of b such that

$$b = \|\mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) - \mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) \cdot \mathbf{P_2}\|_F \,. \tag{8}$$

   We will try to show that a = b.

   - <u>Forward Direction</u>

     a $= \|\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) - \mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) \cdot \mathbf{P}_1\|_F$

     $= \left\|-\mathbf{P}_1(-\mathbf{P}_1^{-1}\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) + \mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2))\right\|_F$

     (By Lemma 5.2, we can factor $-\mathbf{P}_1$ as shown above.)

     $= \left\|-\mathbf{P}_1(\mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) - \mathbf{P}_1^{-1}\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1))\right\|_F$

     (Rearrange terms algebraically as shown above.)

     $= |-1| \left\|\mathbf{P}_1(\mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) - \mathbf{P}_1^{-1}\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1))\right\|_F$

(By the property of the Frobenius norm, we factor out -1 as shown above.)

$= ||\mathbf{P}_1||_F||\mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) - \mathbf{P}_1^{-1}\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1)||_F$

(By Lemma 5.3, we can remove $\mathbf{P}_1$ from the norm as shown above.)

$= ||\mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) - \mathbf{P}_1^{-1}\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1)||_F$

So, we are left with $||\mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) - \mathbf{P}_1^{-1}\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1)||_F$, where $\mathbf{P}_1^{-1}$ is some matrix that permutes the columns of $\mathbf{M}_1$ to find the distance between $(\mathbf{M}_1, \mathbf{w}_1)$ and $(\mathbf{M}_2, \mathbf{w}_2)$.

By the definition of *SR* metric, the expression $||\mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) - \mathbf{P}_1^{-1}\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1)||_F$ is greater than or equal to $||\mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) - \mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1)\mathbf{P}_2||_F$, where $\mathbf{P}_2$ is the minimizing permutation matrix that also permutes the columns of $\mathbf{M}_1$.

Therefore, a $\geq ||\mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) - \mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1)\mathbf{P}_2||_F$ = b.

- Backward Direction We can show an equivalent proof for the backward direction, which will demonstrate that b $\geq$ a.

Therefore, we have satisfied the third criterion, since a = b

4. $d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{M}_2, \mathbf{w}_2)) \leq d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{N}, \mathbf{v})) + d((\mathbf{N}, \mathbf{v}), (\mathbf{M}_2, \mathbf{w}_2))$

   We need to prove:

   $\min_{P \in \mathbb{P}} \|\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) - \mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) \cdot \mathbf{P}\|_F \leq \min_{P_1 \in \mathbb{P}} \|\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) - \mathbf{N} \cdot \text{diag}(\mathbf{v}) \cdot \mathbf{P}_1\|_F +$

   $\min_{P_2 \in \mathbb{P}} \|\mathbf{N} \cdot \text{diag}(\mathbf{v}) - \mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) \cdot P_2\|_F$.(9)

Let us begin by examining the left hand side.

$d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{M}_2, \mathbf{w}_2)) = \min_{P \in \mathbb{P}} \|\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) - \mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) \cdot \mathbf{P}\|_F$

$\leq \|\mathbf{M}_1 \cdot diag(\mathbf{w}_1) - \mathbf{M}_2 \cdot diag(\mathbf{w}_2)\mathbf{P}_2\|_F$

(Since $\mathbf{P}$ is the minimizer, distance calculated with any other permutation matrix will necessarily be greater than or equal to the distance calculated with $\mathbf{P}$.)

$= \|\mathbf{M}_1 \cdot diag(\mathbf{w}_1) - \mathbf{N} \cdot diag(\mathbf{v}) + \mathbf{N} \cdot diag(\mathbf{v}) - \mathbf{M}_2 \cdot diag(\mathbf{w}_2)\mathbf{P}_2\|_F$

(Add and subtract equivalent terms.)

$\leq \|\mathbf{M}_1 \cdot diag(\mathbf{w}_1) - \mathbf{N} \cdot diag(\mathbf{v})\|_F + \|\mathbf{N} \cdot diag(\mathbf{v}) - \mathbf{M}_2 \cdot diag(\mathbf{w}_2)\mathbf{P}_2\|_F$

(We complete this step using the properties of the Frobenius norm.)

$= \left\|\mathbf{P}_1^{-1}(\mathbf{M}_1 \cdot diag(\mathbf{w}_1) - \mathbf{N} \cdot diag(\mathbf{v}))\right\|_F + \|\mathbf{N} \cdot diag(\mathbf{v}) - \mathbf{M}_2 \cdot diag(\mathbf{w}_2)\mathbf{P}_2\|_F$

(We complete this step using Lemma 5.3.)

$= d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{N}, \mathbf{v})) + d((\mathbf{N}, \mathbf{v}), (\mathbf{M}_2, \mathbf{w}_2))$

So, we can see that $d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{M}_2, \mathbf{w}_2)) \leq d((\mathbf{M}_1, \mathbf{w}_1), (\mathbf{N}, \mathbf{v})) + d((\mathbf{N}, \mathbf{v}), (\mathbf{M}_2, \mathbf{w}_2))$.

$\therefore$ Thus, we have now proved that SR is a metric.

6 Implementing the Metric

Now that we have defined the *SR* metric, we move on to the next stage of finding the pairwise distances between each set of **M**,**w** solutions. However, given that we must compare a matrix with each permutation of another matrix, we find that coding the *SR* metric as is would cost O(n!), where n represents the number of strains. By Stirling's Approximation, we know that O(n!) is bounded both above and below by O(n^n) [18]. Thus, coding the *SR* metric as stated above would have a cost of approximately O(n^n). While this is not extremely prohibitive, since we expect the number of strains to be less than 10, we still find that this is an inefficient implementation.

As such, we begin by trying to find any redundancies in the computations that can be eliminated. In the *SR* metric, we compare the weighted strains of $\mathbf{M}_1$ with all of the permutations of the weighted strains of $\mathbf{M}_2$, as shown in Figure 5.

So, we can see that for each strain in $\mathbf{M}_1$ we are simply trying to find the minimum of the normed differences between itself and each of the strains in $\mathbf{M}_2$. For instance, consider the following example:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, a = \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} \text{ and } B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, b = \begin{bmatrix} 0.3 \\ 0.5 \\ 0.2 \end{bmatrix}$$

.

Let us compute the normed difference between each weighted strain in **A**

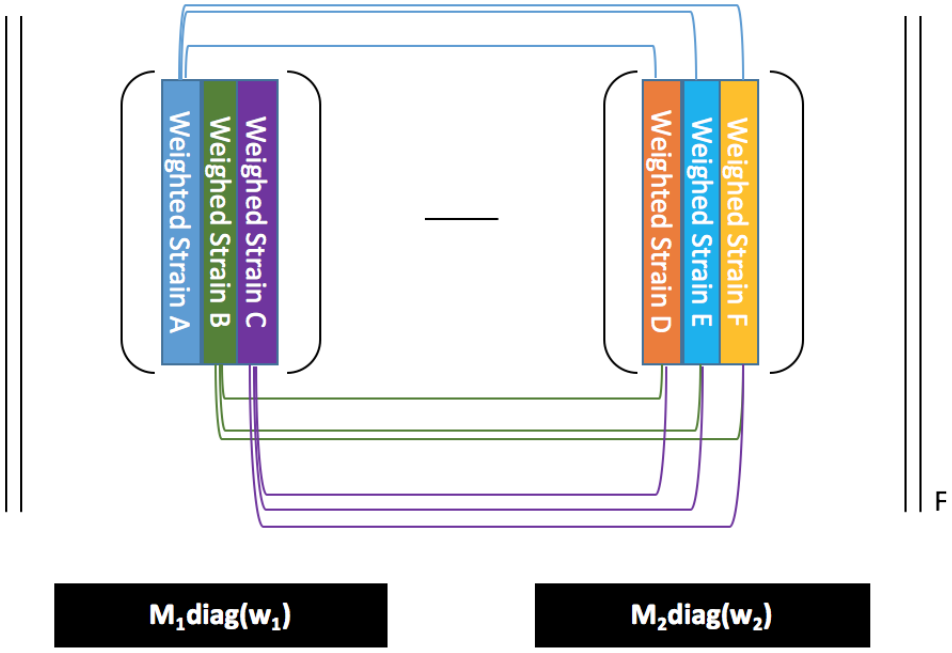

Figure 5

and each weighted strain in $\mathbf{B}$. We can visualize this as a graph, as seen in Figure 6.

In the *SR* metric, we include a permutation matrix that will re-arrange the columns of $\mathbf{M}_2$ in order to find the minimum distance between ($\mathbf{M}_1$, $\mathbf{w}_1$) and ($\mathbf{M}_2$, $\mathbf{w}_2$). We can see that this would be equivalent to finding sum of the minimum normed distances between each weighted strain of $\mathbf{M}_1$ and $\mathbf{M}_2$. Therefore, we can now reframe our problem as a minimum cost maximum matching problem.

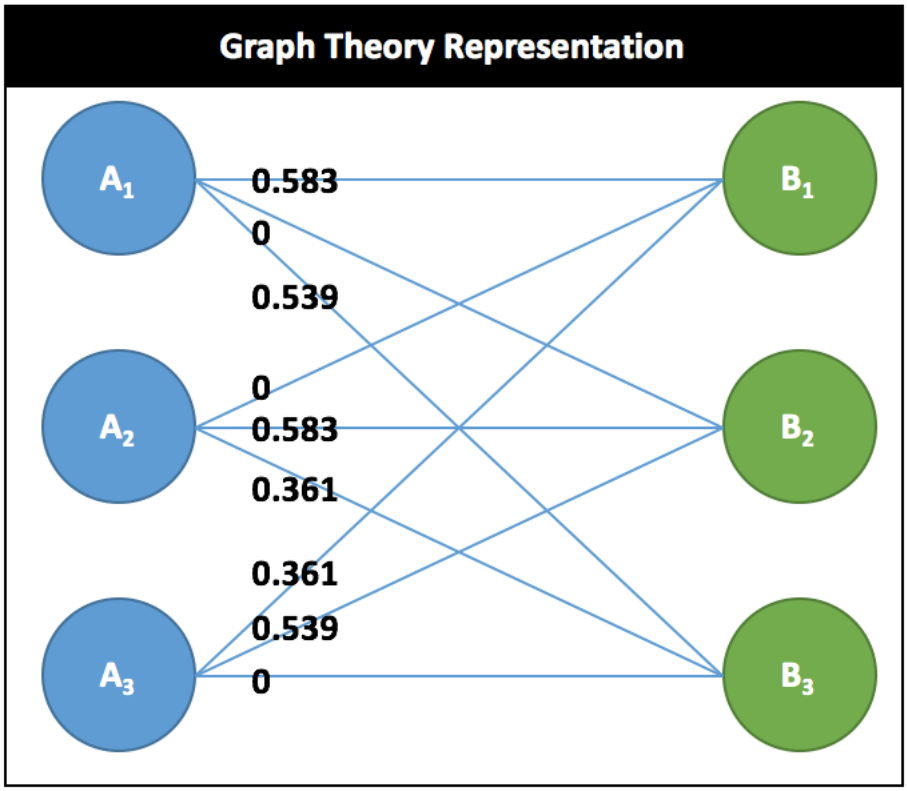

Figure 6

Let us justify this assertion. To begin, define a full bipartite nxn graph G where each edge (i,j) has weight $\sqrt{\sum_{i=1}^{n_{M_1}} \sum_{j=1}^{n_{M_2}} (\mathbf{M}_1[:,i]\mathbf{w}_1[i] - \mathbf{M}_2[:,j]\mathbf{w}_2[j])^2}$.

Every permutation matrix $\mathbf{P}$ *has a one-to-one correspondence with a maximum matching B of G. Moreover, the value of*

$\mathrm{m} = \min_{P \in \mathbb{P}} \|\mathbf{M}_1 \cdot \mathrm{diag}(\mathbf{w}_1) - \mathbf{M}_2 \cdot \mathrm{diag}(\mathbf{w}_2) \cdot \mathbf{P}\|_F$ *is equal to the cost of the corresponding maximum matching B of G.*

Proof: Every permutation matrix $\mathbf{P}$ can be characterized exactly by one

permutation group $(p_1, p_2, .., p_n)$ out of n! possibilities. Let us call this group sigma($\mathbf{P}$).

Forward Direction

For each $\mathbf{P}$, look at sigma($\mathbf{P}$). The value of $\min_{P \in \mathbb{P}} \|\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) - \mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) \cdot \mathbf{P}\|_F$ corresponds to the sum of the values $\sum_{i=1}^{n!} [\mathbf{M}_{1i} \cdot \text{diag}(\mathbf{w}_{1i}) - \mathbf{M}_{2i} \cdot \text{diag}(\mathbf{w}_{2i}) \cdot \text{sigma}(\mathbf{P})_i]$. This is equal to the value of the edges in the matching $((1, \text{sigma}(\mathbf{P})_1), (2, \text{sigma}(\mathbf{P})_2), \dots)$. This is a maximum matching B since it contains n edges, and each side of G has n vertices.

Backward Direction

Each maximum matching B = $(1, a_1), (2, a_2), ..., (n, a_n)$ has the property that the a_i's are all distinct values in 1, 2, ..., n. They can thought of as a permutation group corresponding to permutation matrix $\mathbf{P}$. The values of these edges are $\sum_{i=1}^{n!} [\mathbf{M}_{1i} \cdot \text{diag}(\mathbf{w}_{1i}) - \mathbf{M}_{2i} \cdot \text{diag}(\mathbf{w}_{2i}) \cdot \text{sigma}(\mathbf{P})_i]$, which equals $\min_{P \in \mathbb{P}} \|\mathbf{M}_1 \cdot \text{diag}(\mathbf{w}_1) - \mathbf{M}_2 \cdot \text{diag}(\mathbf{w}_2) \cdot \mathbf{P}\|_F$, thus showing that B has a corresponding permutation matrix $\mathbf{P}$.

$\therefore$ Consequently, the minimizer $\mathbf{P}*$ of the norm corresponds exactly to the minimum cost maximum matching B* of G.

Thus, we have shown that we can reformulate our original problem into a min-cost max matching problem. As such, we can now solve our problem using the Hungarian algorithm, which will find the min-cost max matching of G in O(n^3) time [15, 17]. According to the argument above, that exactly equals the value of the *SR* metric.

Description of the Hungarian Algorithm We have rigorously shown that

we can reformulate our original problem as a min cost max matching assignment problem, which is a type of network optimization problem. The assignment problem is typically used to match n persons with n tasks on a one-to-one basis such that the person-task pairs are distinct and the total cost is minimized [15]. In this context, our task is to find the distance between a pair of matrices, such that the distance between two matrices is equal to the minimum distance between the first matrix and all of the permutations of the second matrix. As such, we are trying to match the n strains of the first matrix solution with the n strains of the second matrix solution such that the each of the strains from the first matrix are paired uniquely and distinctly with each of the strains from the second matrix.

In order to find an optimal assignment for a nxn matrix, we complete the following steps [15]:

1. Subtract the smallest entry in each row from all the entries of its row.
2. Subtract the smallest entry in each column from all the entries
3. Place a line through each of the rows and columns that contain zero entries such that the minimum number of lines is used.
4. Check if the Optimality Condition is met:
    - If the minimum number of lines used in the previous step is n, we have found an optimal assignment and the algorithm can be terminated.
    - If the minimum number of lines used is less than n, find the smallest entry not covered by a line. Subtract this entry from each uncovered

row, and then add it to each covered column. Begin again at Step 3, and continue iterating until the optimality condition has been met.

Example

As an example of how the algorithm works, consider **A, a** and **B, b** (described at the start of the section), which are identical except for the ordering of the strains.

First, we preprocess our data as shown in Algorithm 1 to form the initial Hungarian matrix.

Algorithm 1 Transforming Inputs into Hungarian Matrix

```
for i in 0:n_A
    for j in 0:n_B
        Aa = A[:,i] a[i];
        Bb = B[:,j] b[j];
        dif = Aa-Bb;
        norm = norm(dif, 2);
        Hungarianmatrix[i][j] = norm;
```

Thus, we form our initial Hungarian matrix **H**:

Initial Hungarian Matrix			
	B_1	B_2	B_3
A_1	0.583	0	0.539
A_2	0	0.583	0.361
A_3	0.361	0.539	0

Now, we begin the process of finding the minimum permutation using the Hungarian algorithm:

As per Step 1, we first subtract row minima. However, since each row contains a zero, subtracting row minima has no effect. Similarly for Step 2, subtracting column minima has no effect since each column also contains a zero.

Now, we attempt Step 3 and find that we would require 3 lines to cover all zeros, indicating that we have satisfied the Optimality Condition.

Hungarian Matrix After Step 3

	B_1	B_2	B_3
A_1	~~0.583~~	~~0~~	~~0.539~~
A_2	~~0~~	~~0.583~~	~~0.361~~
A_3	~~0.361~~	~~0.539~~	~~0~~

Thus, we have found an optimal assignment. We find the optimal assignment as the sum of the bolded elements of **H** as shown below:

Hungarian Matrix Optimal Assignment

	B_1	B_2	B_3
A_1	0.583	**0**	0.539
A_2	**0**	0.583	0.361
A_3	0.361	0.539	**0**

Note that the bolded positions correspond to the optimal assignment in the original cost matrix. Since in this case the Hungarian matrix has not changed,

the optimal assignment shown is the same as the optimal assignment in the original cost matrix. To find the optimal value, we simply sum the each of the bolded positions and find a minimum cost of 0.

Recall that the **M**, **w** pair being compared contain the identical strains and weights, but in a different order. Thus, before we computed the metric distance, we would expect to see a distance of 0. We can see that both our formation of the *SR* metric and our implementation of it are as expected, since we do in fact find a minimum distance of 0 between the two matrix-vector solutions being compared.

6.1 Proof of Big-O Complexity of the *SR* Metric

We now have described two main steps to implement the *SR* metric. First, we use the preprocessing algorithm to transform the inputs into the Hungarian matrix. Secondly, we implement the Hungarian algorithm, which will find an optimal assignment of the strains and thereby find the distance between two **M**, **w** solutions.

In the preprocessing algorithm, note that our cost is $O(n^2)$, since we must perform $n_A \cdot n_B$ computations. Further, the Hungarian algorithm is well known to have a cost of $O(n^3)$ [17].

Thus, we can see that implementing the metric using the steps described above will now have an overall complexity of $O(n^3)$, which is significantly more efficient than our initial cost of $O(n!)$.

7 Clustering Methodology

We are now equipped with a metric that will allow us to perform clustering. In order to best answer our questions of interest, however, we must first try to understand which method of clustering to use.

Let us begin by describing the properties that our ideal clustering method will reflect.

1. *The clustering method will allow for an arbitrary number of clusters.*

   The motivating question behind the work conducted by Mustonen et al. is the following: given a blood sample from an infected person that contains a measurement of the proportion of mutations from a reference strain of malarial DNA, can we predict the number of strains of malaria with which the individual is infected, the identity of those strains, and the proportion in which those strains are present? Inherent in this question is a level of uncertainty about the number of different strains of malaria that are present in an individual's blood sample. Thus, to answer Question 2, we find that we cannot use a clustering algorithm that requires the number of clusters to be specified beforehand.

   In addition, we are also unaware of how many strains we expect to see in the field in each year. In fact, this is one of the questions we hope to address through our analysis of the field data. Therefore, to answer Question 3, we similarly find that we must use a clustering algorithm that accepts an arbitrary number of clusters.

As such, since we do not have enough knowledge of our data a priori to state with certainty the correct number of clusters (i.e., the correct number of strains), we choose not to use centroid clustering algorithms (such as K-means).

2. *The clustering method will be able to describe key differences between multiple solutions.*

   In addition, we hope that our ideal clustering algorithm will be able to effectively describe the key differences between solutions on multiple levels. As previously mentioned, the *StrainRecon* algorithm is significantly better at predicting the identity of the strains with the highest weight compared to the strains with the lowest weight. Therefore, our clustering method must be able to describe differences in strains on multiple levels of granularity corresponding to the weight of each strain.

3. *The clustering method will be easily interpretable.*

   Finally, let us recall that the ultimate objective of this research is to inform biostatisticians at the CDC as they analyze the blood samples of individuals infected with malaria. As such, we wish to produce a form of output from our clustering that is both easy to interpret and informative.

Thus, based upon the above criterion, we turn to hierarchical clustering as our ideal method, since it will allow us to clearly view how different solutions diverge from one another in a structured and clear manner.

However, let us also note the drawbacks of this approach. Firstly, hierarchi-

cal clustering has a time complexity of $O(n^3)$ and requires $O(n^2)$ memory. As a result, this method can be inefficient when dealing with large datasets [2]. In addition, the merges in hierarchical clustering are computed in a greedy manner, which implies that errors that occur when merging clusters are immutable will affect the final output.

While the aforementioned drawbacks are important to note, we find that the advantages conferred to us by the use of hierarchical clustering are unique.

In addition, we now address the question of which linkage method to use in order to conduct the hierarchical clustering. Based on comparisons of the dendrogram outputs from the various linkage methods, we choose to select the WPGMA (Weighted Pair Group Method with Arithmetic Mean) method, since it consistently produces the most easily interpretable and balanced trees. The WPGMA algorithm merges the nearest two clusters, for instance a and b, into a higher-level cluster such that $a \cup b$ has a distance from another cluster c according to the following formula: $d((a \cup b), c) = \frac{d(a,c)+d(b,c)}{2}$.

Thus, we have now found an ideal clustering method to use with our metric to understand the data.

8 Exploring CDC Pilot Data

8.1 Calibrating the *SR* Metric

Let us begin our analysis by confirming that the *SR* metric performs as expected. In order to do this, we must deal with data for which we already know the true solution. Further, we want to ensure that the *SR* metric performs well in real world conditions to ensure the validity of our analysis of the field data later in this paper. As such, we choose to perform hierarchical clustering upon the pilot data.

Recall that the pilot data was categorized into nine general types of experiments. There were three different mixes of strains used (i.e., three distinct **M** matrices with different strains in each matrix), which were labeled A, B, or C. In addition, there were three different ratios used to mix the strains (i.e., three distinct **w** vectors with different proportions), which were labeled by 1, 2, or 3. Further, a level of background noise was introduced by including blood cells, dried blood cells, or human DNA into the mixtures.

In an analysis that clusters the pilot data, we would therefore expect to see three large clusters corresponding to the distinct **M** matrices. Further, within these clusters, we would expect to see smaller clusters for each different **w** vector. In other words, if the *SR* metric performs the way we expect, we would expect to see three large clusters for A, B, and C and three sub-clusters within each of these for the different proportions.

To perform this analysis, let us begin with the input **d** vector. For this

calibration to be most useful given real world conditions, we must only use the **d** vector and the *StrainRecon* algorithm, since in true field conditions we would not be provided with the true **M**, **w** solutions. However, recall that the *StrainRecon* algorithm solves an undetermined inverse problem and therefore outputs multiple possible **M**, **w** solutions given a single **d** vector. As such, we must develop some method to isolate a representative **M**, **w** pair from all of the possible solutions.

Developing a Heuristic for the "Most Likely" Solution

It is not immediately clear how to select a representative solution from all of the outputs of the *StrainRecon* algorithm. We begin approaching this problem by performing hierarchical clustering on the outputs from the *StrainRecon* algorithm's attempt to solve C-1 (DBS) for 3 strains. (Note that one of the unknowns is the true number of strains, and therefore the *StrainRecon* algorithm must attempt to find the true solution for multiple different numbers of strains. As such, an elbow chart is used to select the correct number of strains by measuring the marginal returns of the percentage of variance explained.)

We find the dendrogram in Figure 7 from this clustering. We begin with five possible solutions that are produced by the *StrainRecon* algorithm, which is a relatively small amount of possibilities since the algorithm is solving for n=3, which is the true number of strains. When the algorithm attempts to solve using a different number of strains, the number of solutions increases. However, we choose to look at the simplest case in order to extract the "best" possible solution.

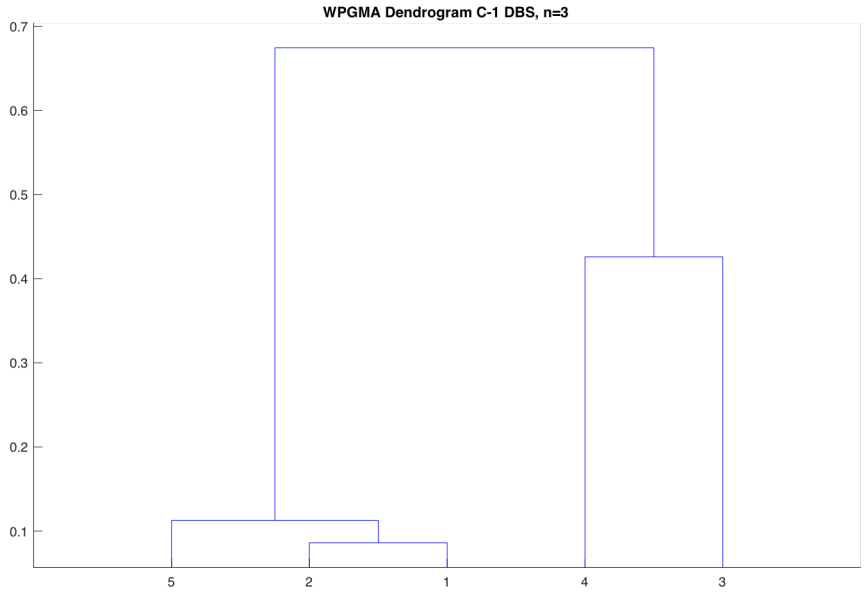

Figure 7

From Figure 7, we see that there are two main clusters of possible solutions. Even between these two main clusters, there exists only a small distance, which suggests that the dominant strain in each cluster is the same and the differences between the clusters are caused by other discrepancies in the less dominant strains. We validate this by examining the leaves in each cluster and comparing them with each other and with the true solution.

In addition, this dendrogram (as well as all the other dendrograms to be presented later in this paper) is plotted with an optimal leaf ordering, such that the sum of similarity of adjacent elements is maximized. So, we can see that solutions 5 and 3 are the most different solutions. Further, solutions 1 and 2

are the most similar since the height at which they are joined is the lowest.

While all of this information is interesting to note as an analysis of the accuracy of the *StrainRecon* algorithm, we do not find any clear method to translate this information into a single representative solution. We know that there are two clusters for the true solution and that these clusters have a low distance between them, but we do not know which cluster is more "correct" or which element within the cluster is the "most correct." From the above analysis, however, we have learned that we must find some way of comparing each $\mathbf{M}$, $\mathbf{w}$ solution to the true $\mathbf{d}$ vector, since this would help answer the question of which element is the "most correct" or most representative solution.

Thus, we turn instead to a different method for identifying the most representative solution. We propose using the misfit value suggested by Mustonen et al., which is calculated by the following formula [19]:

$$||\mathbf{Mw} - \mathbf{d}||_2. \tag{10}$$

Using this formula, we can calculate the misfit for each possible solution and then select the solution with the minimum misfit. We can see immediately that this method provides us with a clear solution that is representative of the true solution. (For Figure 7, the minimum misfit solution is represented by the leaf indexed at 1.) On the other hand, a potential drawback of this method is that the minimum misfit solution may not be the precise true solution. However, we find that the benefits of using the heuristic of the minimum misfit solution outweigh the drawbacks, since in general, the minimum misfit solution should correctly characterize the *StrainRecon*'s algorithm best guess for the true

solution.

Analysis of the Dendrograms for the Pilot Data

Using the minimum misfit solution for each observation, we now perform hierarchical clustering over the entire set of pilot data. Let us first examine the dendrogram for the minimum misfit solutions when the *StrainRecon* algorithm solves for n=3, which is the true number of strains.

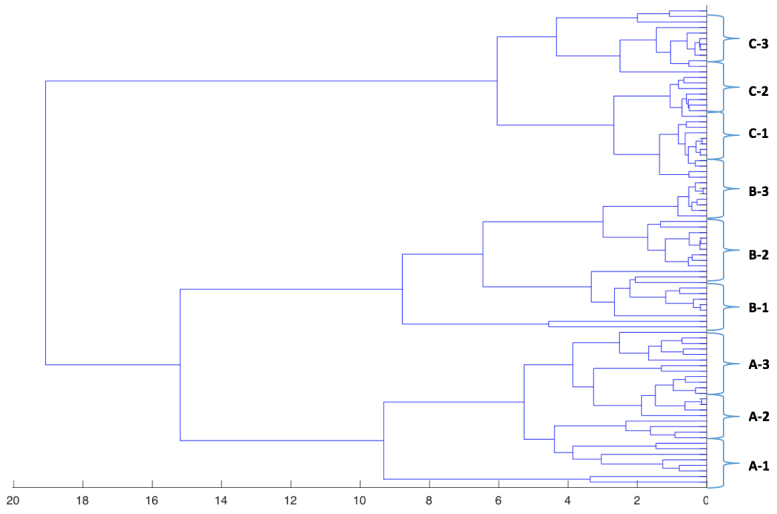

Figure 8

In Figure 8, which was also plotted with an optimal leaf ordering, we can see that the *SR* metric performs exactly as expected. We see that the A, B, and C groups are clearly separated. Further, within each group, we have three clusters for the weights, and the most similar weights are clustered together (for example, so that A-1 is closer to A-2 rather than to A-3).

Let us now make the analysis more complex by performing hierarchical clus-

tering over the minimum misfit solutions for the pilot data where the *StrainRecon* algorithm solves for n=4.

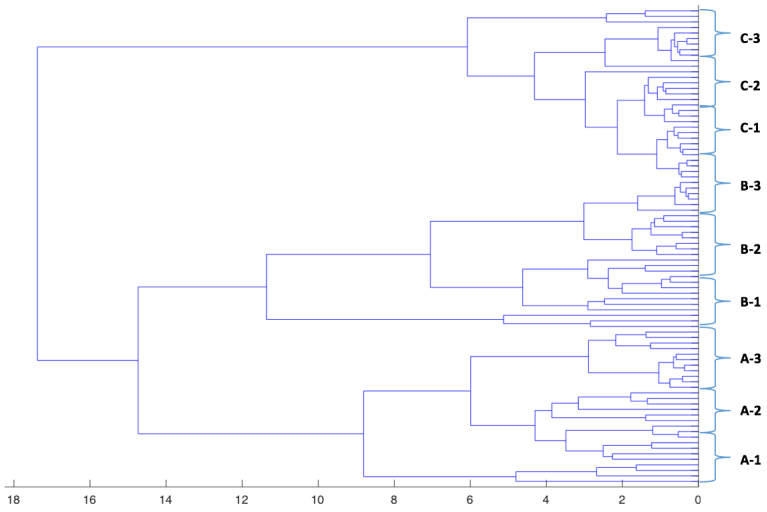

Figure 9

Again, we see that the *SR* metric performs as expected, which implies that the methods used to perform the hierarchical clustering are robust to some amounts of noise.

Even further, Figure 8 and Figure 9 motivate an interesting question. Since the ratios in the pilot data all contain one dominant strain, are the clusters from above determined mostly due to the emphasis that the *SR* metric places upon highly weighted strains? In order to answer this question, we will slightly modify our analysis such that our $\mathbf{M}$ matrix is only composed of the dominant strain and the $\mathbf{w}$ vector only contains a single weight of 1.

We can see from Figure 10 that the *SR* metric using just the dominant strain still performs well, since we again observe the expected grouping of A, B, and

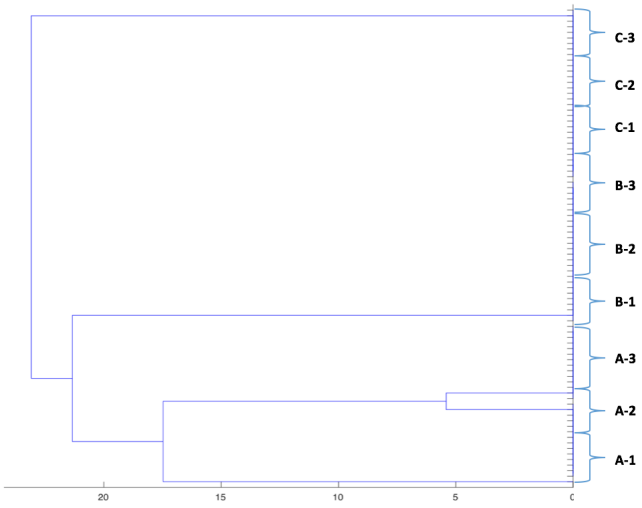

Figure 10

C matrices with similar weights placed more closely together. Furthermore, the dendrogram in Figure 10 keeps approximately the same shape as in Figure 8 at the higher levels. On the whole, we observe fewer clusters in this dendrogram, which is expected because several of the clusters on the lower levels were formed due to differences in the less dominant strains. Similarly to Figure 8, we do note that the observations for A-1 strains are clustered with the observations for the A-2 and A-3 strains at a relatively high height. We can now use our understanding from Figure 10 to conclude that this high height at which all the A observations are clustered is caused by differences in the dominant strain for A-1 compared to the dominant strains for A-2 and A-3. Digging deeper, however, why would we see such a large distance between the dominant strain for A-1 compared to the dominant strains for A-2 and A-3? Recall firstly that we expect the A strains to have more mutations compared to the B or C strains,

which explains why we see a less clear clustering for the A strains. Even further though, the least dominant weights in ratio 1 are extremely small, which means that the *StrainRecon* algorithm has more difficulty in finding accurate guesses for the **M**, **w** solution. Thus, due to the weights in A-1 and the fact that A strains are more likely to have mutations, we can understand the clustering pattern displayed in Figure 10.

Finally, let us use the *SR* metric on the dominant strain of the minimum misfit solutions when the *StrainRecon* algorithm solves for n=4.

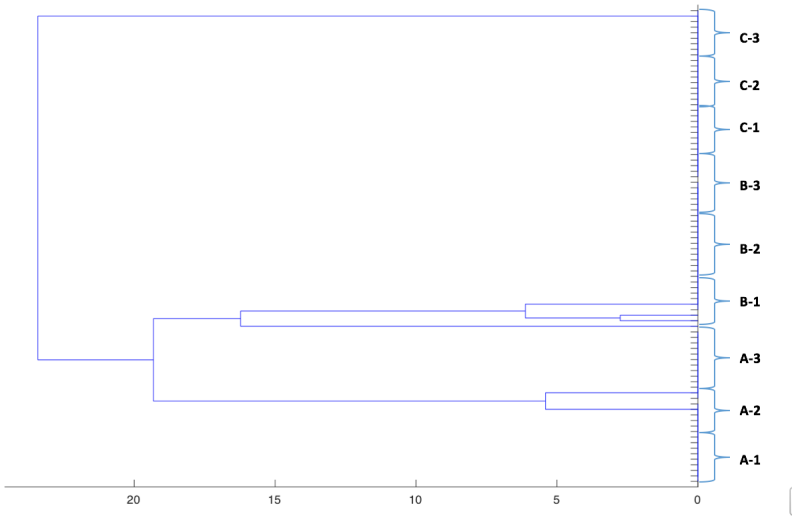

Figure 11

Again, we are able to confirm the effectiveness of the *SR* metric as well as the influence of the most dominant strain in determining clusters. In addition, as before, we also see fewer clusters on the lower levels. However, while we do note two lower level clusters in A, we now also find multiple lower level clusters in B. Why might this make sense? Going back to Figure 9, we see that just as

in this dendrogram, all of the B observations were joined at a higher distance compared to where all of the A observations or all of the C observations were joined. Therefore, we are again able to explain the comparisons of the full **M**, **w** solutions by analyzing the most dominant strain, since it seems as if discrepancies in the prediction of the dominant strain in B drove the clustering pattern in Figure 11.

Thus, through this analysis of the pilot data, we are able to draw the following key insights. Firstly, our metric functions as predicted, even in real world conditions with additional noise added. Secondly, by using a reduced version of the solution and only comparing the dominant strains, we find that we can explain most of the significant differences in clusters in the full dendrogram. As such, even with noise and zero weights for the less dominant strains, we find that we are able to successfully capture the most significant aspects of the pilot data through clustering only on the dominant strains.

9 Analyzing Changes in Malaria Over Time

Recall that one of our primary motivations in this paper was to analyze how successful the CDC has been at reducing the prevalence of malaria. One method in which we can do so is by comparing the number of strains of malaria that are present in disease outbreaks over time. As such, we propose conducting hierarchical clustering using the minimum misfit solutions over the observations from the field data for each year in order to compare the number of clusters between years.

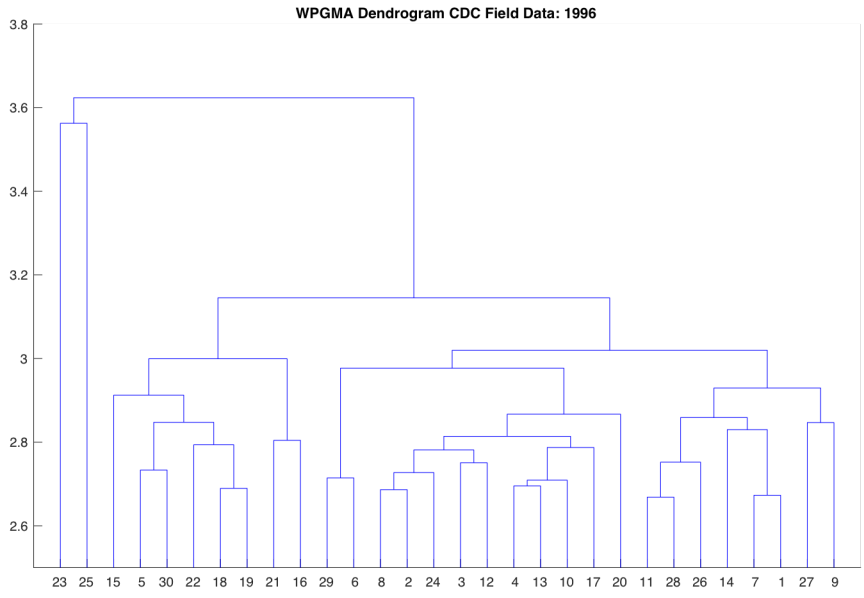

Figure 12

One method to answer the question of the genetic diversity of malaria across time would be to specify a distance value and find the number of clusters present

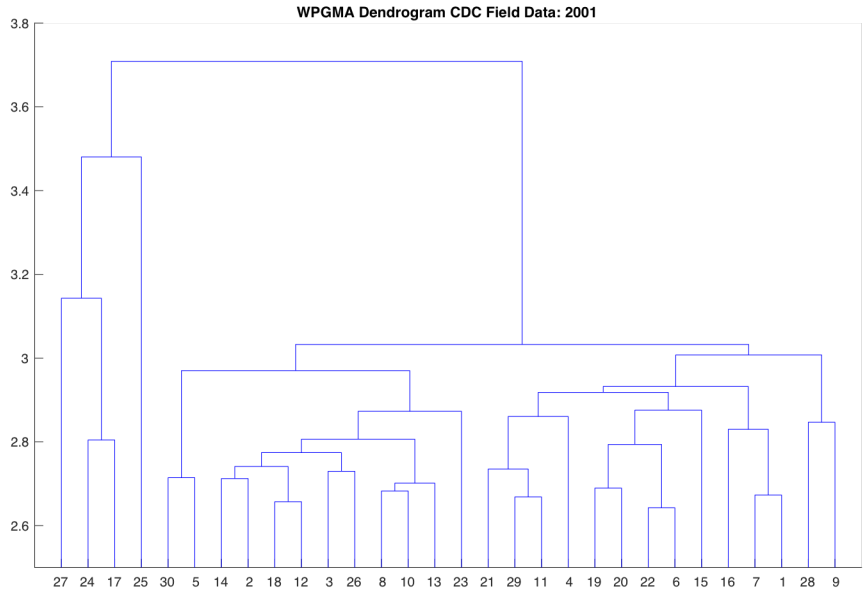

Figure 13

at this threshold in each year. However, as we can see by referencing the dendrograms for each year, the value we must specify is not clear beforehand. In fact, depending on the value of distance that is chosen, the true pattern of the change in number of clusters can be inaccurate. Thus, we turn instead to a package in R called Dynamic Tree Cut, which was developed for hierarchical clustering problems in bioinformatics. In this approach, the number of clusters in the dendrogram is determined by an "adaptive, iterative process of cluster decomposition and combination" and the algorithm terminates upon finding a stable number of clusters [16]. By this approach, we are able to more rigorously find the number of clusters than through a static threshold. Thus, we find that

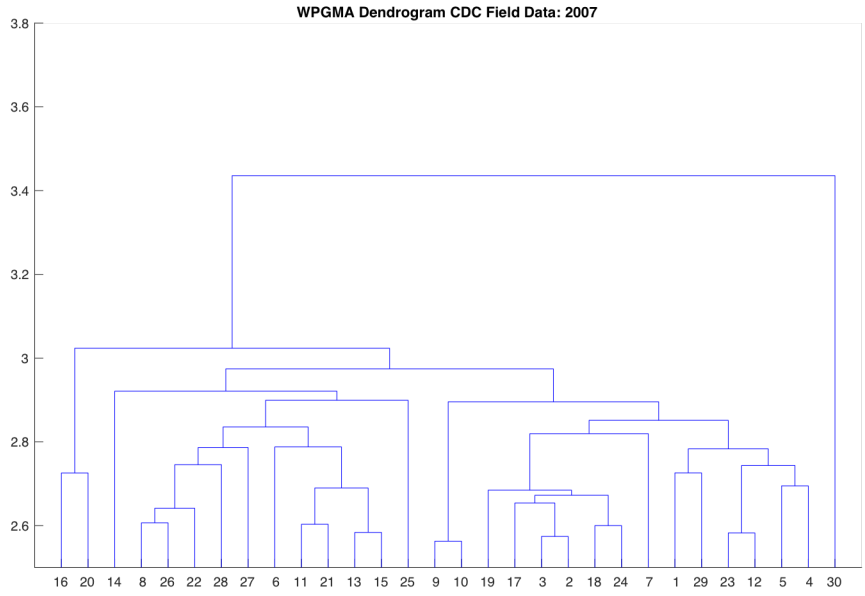

Figure 14

there are 8, 8, 4, and 6 clusters for the years 1996, 2001, 2007, and 2012 respectively. In turn, this indicates that the number of strains initially remained constant, then drastically decreased, and finally slightly increased.

However, it is not immediately clear as to what is driving the changes in the number of strains between years. One hypothesis is that the implementation of public health prevention and treatment methods has been inconsistent across time, owing to the poor infrastructure in Kenya. Alternatively, it could be the case that medicines to treat malaria led to a reduction in the number of strains by 2012. Then, due to the rapid evolution of malarial strains in response to these treatment measures, new strains emerged that were more resistant to

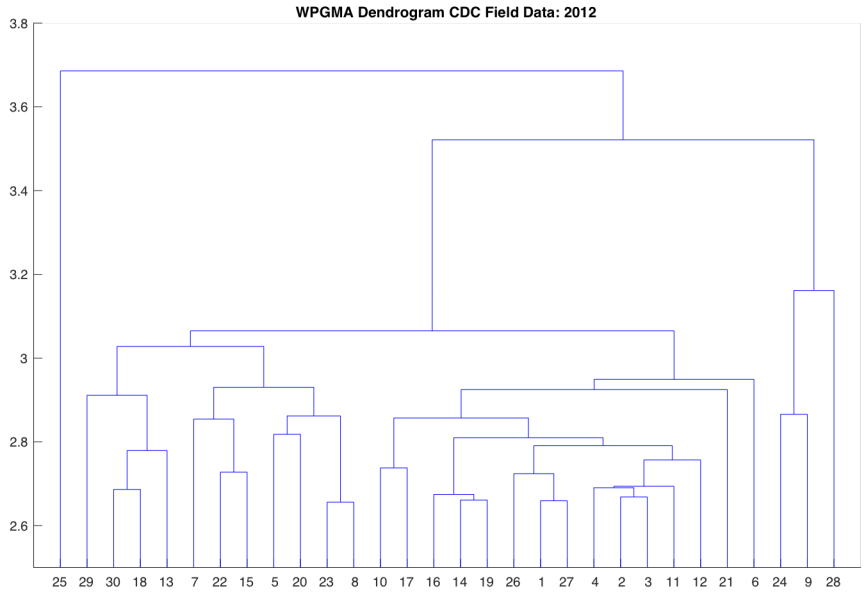

Figure 15

these drugs. While these theories are certainly plausible, we cannot conclusively state why this pattern is occurring as of yet.

An interesting question motivated by our analysis of the pilot data is: what would the clusters of the field data look like using just the dominant strains? In our previous analysis of the pilot data, we established that the dendrograms found by comparing the complete $\mathbf{M}$, $\mathbf{w}$ solutions and the dendrograms found by comparing the dominant strain with a weight of one looked extremely similar. This is because each of the mixing ratios included a highly dominant strain that drove the pattern for the clustering of the complete $\mathbf{M}$, $\mathbf{w}$ solutions. However, we do not know the true solutions for the field data. As such, we are not aware

beforehand whether there is some dominant strain that drives the clustering or if there is another dynamic that drives the process. However, we can set out to find an answer to this question by conducting a cluster analysis of the field data using just the dominant strains.

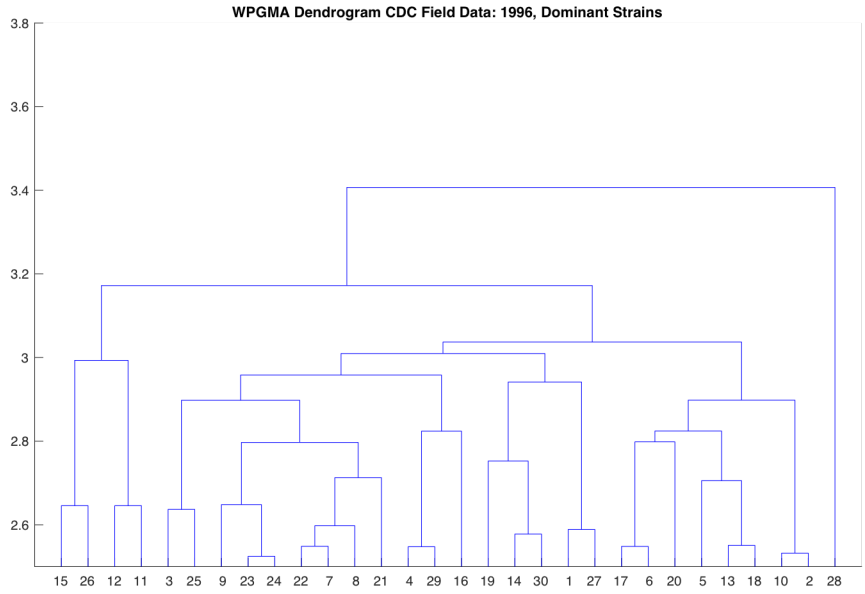

Figure 16

From comparing Figure 12 and Figure 16, we can see that the maximum distance between solutions is lower in the comparison of the dominant strains versus the complete solutions. This intuitively makes sense since the comparison of the complete solutions takes into account differences in the non-dominant strains as well as differences in the dominant strain. In addition, contrary to our expectations, we find that the number of clusters of the dominant strain in

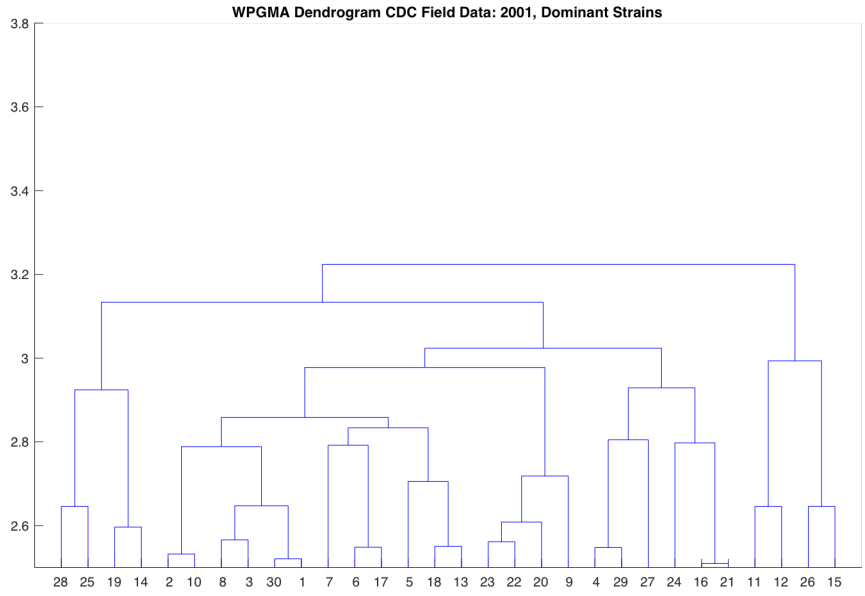

Figure 17

Figure 16 is comparatively large, implying that there is a great deal of variability between the dominant strains. In fact, when we conduct the dynamic tree clustering algorithm on Figure 16, we find that the number of clusters is 12. This suggests that the dominant strain reflects more variability than the overall **M**, **w** solutions.

We observe similar results for each of the hierarchical clusterings of the dominant strains in 2001, 2007, and 2012. Again, we observe a large number of clusters for the dominant strains in each of the aforementioned years. Using the dynamic tree clustering algorithm, we find that there are 8, 8, and 10 clusters in each respective year.

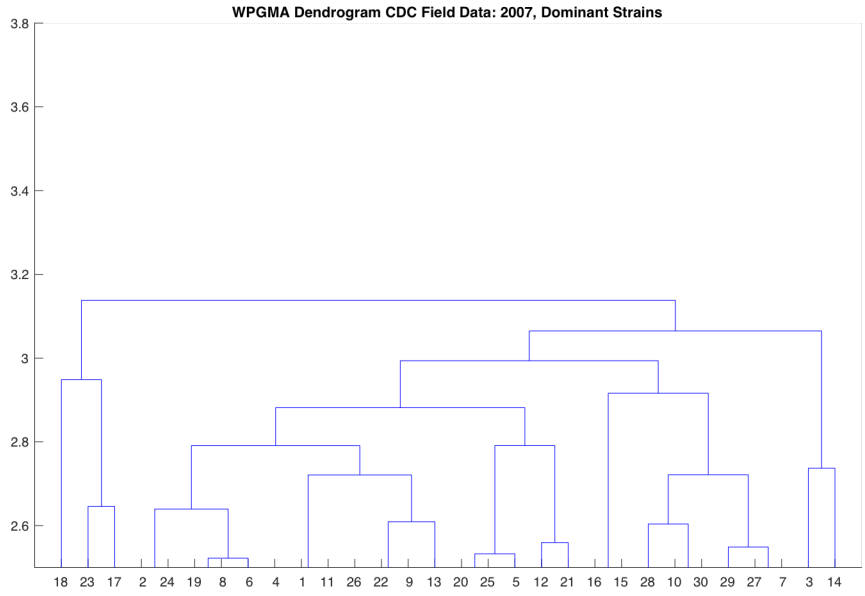

Figure 18

On the whole, we find that this analysis has revealed that the nature of the field data is fundamentally different than that of the pilot data. We find very weak evidence supporting the claim that the minimum misfit **M**, **w** solution for each field observation is characterized by a clear dominant strain. This has several implications. Firstly, we expect that the *StrainRecon* algorithm may have difficulty finding the true solution in field conditions. Secondly, although the main focus of our analysis of the field data is centered upon the strain identities as opposed to the weights, performing our clustering method with the full **M**, **w** solutions (i.e., including weights) seems to most faithfully characterize the data.

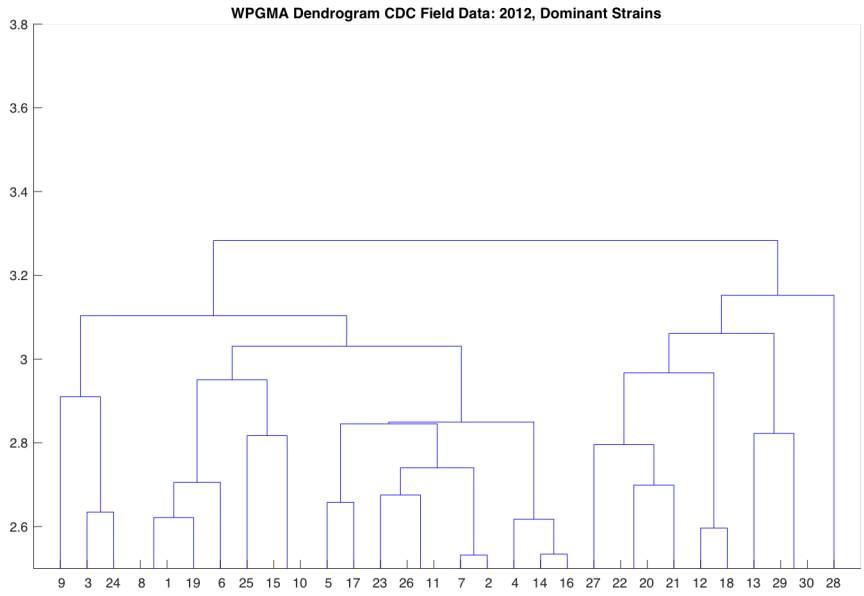
WPGMA Dendrogram CDC Field Data: 2012, Dominant Strains
3.8
3.6
3.4
3.2
3
2.8
2.6
9 3 24 8 1 19 6 25 15 10 5 17 23 26 11 7 2 4 14 16 27 22 20 21 12 18 13 29 30 28

Figure 19

10 Conclusion

10.1 Summary of Work

In this paper, we extended the analysis conducted by Mustonen et al. by deriving the *SR* metric to compare the various possible solutions produced by the *StrainRecon* algorithm. Next, we provided a rigorous mathematical justification proving that our measure of distance satisfied the definition of a metric, which allowed us to perform hierarchical clustering over multiple $\mathbf{M}$, $\mathbf{w}$ solutions. We first applied the *SR* metric to the pilot data from the CDC and displayed that it was robust under some amounts of noise. Further, we also provided evidence that showed that when a clearly dominant strain was present, a hierarchical clustering of the full $\mathbf{M}$, $\mathbf{w}$ solution was approximately equivalent to a hierarchical clustering of the dominant strain. Afterwards, we applied the *SR* metric to the field data and gathered the novel insight that the number of strains remained equivalent in 1996 and 2001 before dramatically decreasing in 2007 and subsequently increasing in 2012. Finally, we also developed an argument for the use of the full $\mathbf{M}$, $\mathbf{w}$ solution when clustering the field data, since we did not find strong evidence to support the assertion that there was a clearly dominant strain that was driving the dynamic behind the hierarchical clustering of the full $\mathbf{M}$, $\mathbf{w}$ solutions.

However, there are a few limitations in this work. Firstly, although using the heuristic of the minimum misfit solution provided us with a good approximation of the true solution, we find that the analysis could be improved upon by

finding a more sophisticated method to approximate the true solution. This is particularly important in field conditions, since there does not seem to be a clearly dominant strain, and the *StrainRecon* algorithm tends to produce a large number of possible solutions in this case. Further, while we are able to find the pattern of how the number of strains has changed across time, we find our analysis is limited by the fact that we are not able to explain the dynamic that is driving these changes.

10.2 Future Work

We can extend our analysis in three key ways. Firstly, we anticipate adding new data for 2017 to our analysis as soon as our collaborators are able to provide this information. (The receipt of new data has been delayed by volatile ground conditions in Kenya.) Adding in data for a new time point will allow us to obtain a more complete picture of how malaria is changing, particularly since we only currently have data from four time points. Secondly, it would be interesting to further our analysis of the changes in the number of strains of malaria over time by further investigating the dynamics that drive these changes. For instance, an analysis of which strains drop out over time and which strains re-emerge would be useful information to the CDC. Thirdly, we propose adding in other dimensions to this analysis, for instance, by factoring in spatial location through GPS data. This would help provide the CDC with a more nuanced understanding of how malaria is changing across each region, which may help inform more targeted treatments or preventative measures.

11 References

References

[1] Tarekegn A Abeku. Response to malaria epidemics in africa. *Emerging infectious diseases*, 13(5):681, 2007.

[2] Bill Andreopoulos, Aijun An, Xiaogang Wang, and Michael Schroeder. A roadmap of clustering algorithms: finding a match for a biomedical application. *Briefings in Bioinformatics*, 10(3):297–314, 2009.

[3] Frédéric Ariey, Benoit Witkowski, Chanaki Amaratunga, Johann Beghain, Anne-Claire Langlois, Nimol Khim, Saorin Kim, Valentine Duru, Christiane Bouchier, Laurence Ma, et al. A molecular marker of artemisinin-resistant plasmodium falciparum malaria. *Nature*, 505(7481):50, 2014.

[4] PB Bloland, DA Boriga, TK Ruebush, JB McCormick, JM Roberts, AJ Oloo, W Hawley, A Lal, B Nahlen, and CC Campbell. Longitudinal cohort study of the epidemiology of malaria infections in an area of intense malaria transmission ii. descriptive epidemiology of malaria infection and disease among children. *The American journal of tropical medicine and hygiene*, 60(4):641–648, 1999.

[5] Mary Bushman, Lindsay Morton, Nancy Duah, Neils Quashie, Benjamin Abuaku, Kwadwo A Koram, Pedro Rafael Dimbu, Mateusz Plucinski, Julie Gutman, Peter Lyaruu, et al. Within-host competition and drug resistance

in the human malaria parasite plasmodium falciparum. *Proc. R. Soc. B*, 283(1826):20153038, 2016.

[6] Rachel Daniels, Sarah K Volkman, Danny A Milner, Nira Mahesh, Daniel E Neafsey, Daniel J Park, David Rosen, Elaine Angelino, Pardis C Sabeti, Dyann F Wirth, et al. A general snp-based molecular barcode for plasmodium falciparum identification and tracking. *Malaria journal*, 7(1):223, 2008.

[7] Carlotta Domeniconi, Dimitrios Gunopulos, Sheng Ma, Bojun Yan, Muna Al-Razgan, and Dimitris Papadopoulos. Locally adaptive metrics for clustering high dimensional data. *Data Mining and Knowledge Discovery*, 14(1):63–97, 2007.

[8] Centers for Disease Control and Prevention. Treatment of malaria: Guidelines for clinicians, 2013.

[9] Centers for Disease Control and Prevention. How can malaria cases and deaths be reduced?, 2015.

[10] Centers for Disease Control and Prevention. Current and future research, 2017.

[11] Centers for Disease Control and Prevention. Frequently asked questions, 2017.

[12] Centers for Disease Control and Prevention. Malaria: Biology, 2017.

[13] Wangeci Gatei, John E Gimnig, William Hawley, Feiko Kuile, Christopher Odero, Nnaemeka C Iriemenam, Monica P Shah, Penelope Phillips Howard, Yusuf O Omosun, Dianne J Terlouw, et al. Genetic diversity of plasmodium falciparum parasite by microsatellite markers after scale-up of insecticide-treated bed nets in western kenya. *Malaria journal*, 14(1):495, 2015.

[14] Peter Grabusts et al. The choice of metrics for clustering algorithms. In *Proceedings of the 8th International Scientific and Practical Conference*, volume 2, pages 70–76, 2011.

[15] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2):83–97, 1955.

[16] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719–720, 2007.

[17] Eugene L Lawler. *Combinatorial optimization: networks and matroids*. Courier Corporation, 1976.

[18] Cristinel Mortici. On the stirling series of gamma function. *Thai Journal of Mathematics*, 9(1):153–159, 2012.

[19] Lauri Mustonen, Xiangxi Gao, Asteroide Santana, Rebecca Mitchell, Ymir Vigfusson, and Lars Ruthotto. A bayesian framework for molecular strain identification from mixed diagnostic samples. *arXiv preprint arXiv:1803.02916*, 2018.

[20] World Health Organization. Malaria, 2017.

[21] S. Shah. *The Fever: How Malaria Has Ruled Humankind for 500,000 Years*. Picador, 2011.

[22] J Wooden, S Kyes, and CH Sibley. Pcr and strain identification in plasmodium falciparum. *Parasitology today*, 9(8):303–305, 1993.