The MMD Method for Efficient Source Attribution

Anurag Thilakchandra Kojapady
Gongbo Zhang

April 19, 2020

Abstract

This case study project is inspired by the paper ***"Mining Whole Genome Sequence data to efficiently attribute individuals to source populations"***[1] by ***F.J. Perez-Reche et al***. The paper presents a minimal multilocus distance (MMD) method which rapidly deals with these large datasets as well as methods for optimally selecting loci. In this project we will mainly focus on the MMD method for mining Whole Genome Sequence (WGS) data. We will first explain the MMD method and demonstrate in determining the source of human campylobacteriosis, explore the results corresponding to the different number of loci in genotypes. Also, we will compare the method's precision with Mathematica's **Classify** function, discuss the difference and possible problems.

Contents

1 Introduction

Attributing or assigning individuals to source populations is very important in many disciplines, for example, we could determine the source of infection, animal trading, and geographical human or plant origins. Consider Figure 1.1, we have three sources including C/S Asia, Middle East and Europe. We are trying to assign the unknown individual $\boldsymbol{u}$ to one of the sources, where $\boldsymbol{u}$ comprises unique genetic markers, and each number in $\boldsymbol{u}$ represents one locus. We need to compare $\boldsymbol{u}$ with the genetic profile of various known source populations which consists more than one individual, hence assign $\boldsymbol{u}$ to the most likely source population. To achieve this, one can use supervised machine learning to train a classifier with predefined sources, and the loci in each individual are features. We will use the MMD method proposed by ***F.J. Perez-Reche et al***[1], which is a very fast and efficient method dealing with large datasets.

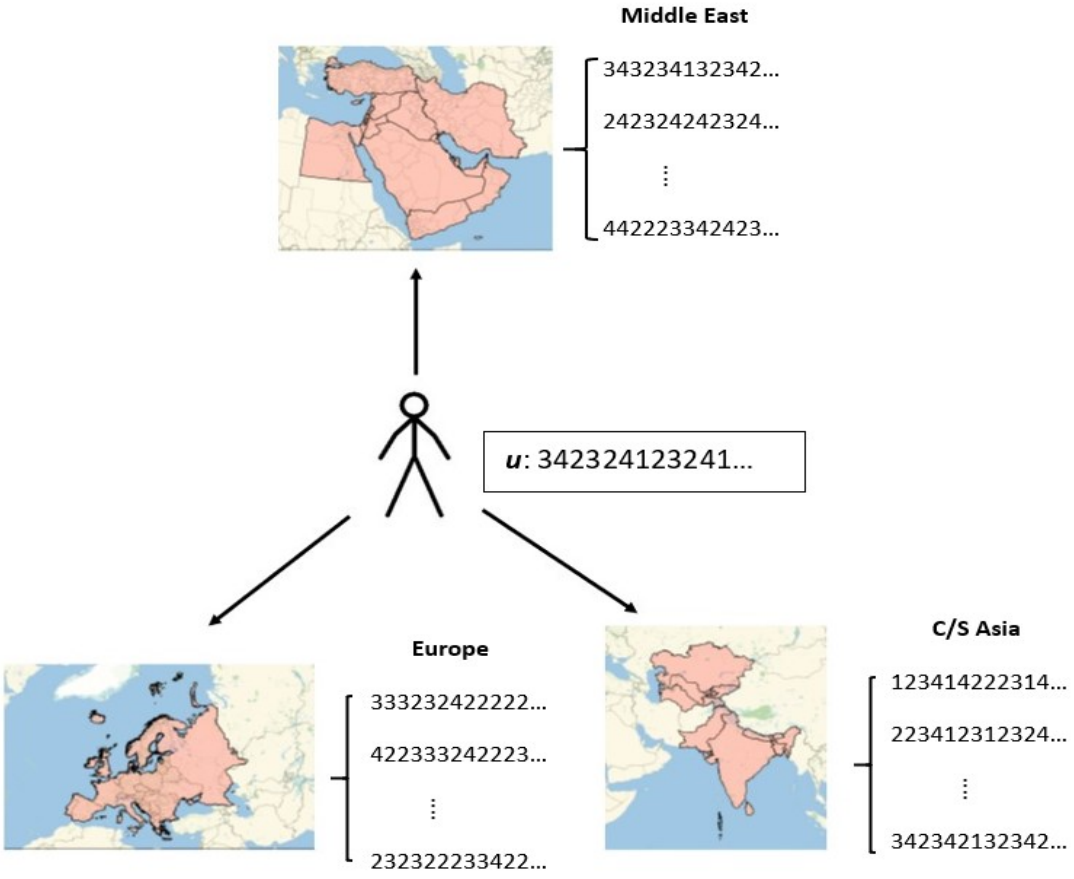

Figure 1.1: Human geographical origin attribution

Whole Genome Sequencing[2] is an integrated method for analysing the entire genome at a single time, which is becoming available to discover large number of markers, and the next-generation sequencing technology makes it equally useful for sequencing any species ranging from viruses to humans. Existing methods for source attribution such as STRUCTURE[3] deals with relatively short genotypes consisting a few to tens or hundreds of loci. However, to achieve

more accurate results, one would like to use extended genotypes with more than 1000 loci. The computation time for STRUCTURE method increases at least linearly with the number of loci contained in genotype, which is impractical in this case. In contrast, the MMD method will be a much more efficient way for source attribution dealing with extended genotypes.

The main aim of our work is to explain how the MMD method constructs and proceeds, and apply the method to real world applications to draw conclusions. We will also discuss results such as accuracy and computation time for analysing different size of datasets, as well as proceeding the same datasets with Mathematica for comparison.

2 Method and Datasets Used for Analysis

2.1 The MMD Method

In order to clearly explain the MMD method, we start by giving some key definitions.

Definition 1 (Hamming Distance)**.** *Given two vectors* $\boldsymbol{u}$, $\boldsymbol{v} \in F^n$ *where* F^n *is the field up to* n *dimensions, and the Hamming distance between* $\boldsymbol{u}$ *and* $\boldsymbol{v}$*, or* $\boldsymbol{d(u,\ v)}$ *is defined as the number of places where* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *is differed.*

For example, consider vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, and the length for both vectors is 10

$$\boldsymbol{u} = 1122333312$$
$$\boldsymbol{v} = 3322322311$$

the places where the numbers in $\boldsymbol{u}$ and $\boldsymbol{v}$ differed are labelled red, and there are 5 places differed in total, hence $\boldsymbol{d(u,\ v)} = 5$ in this case.

We have introduced $\boldsymbol{u}$ before, which is a vector containing L loci charactering the unknown individual, where $\boldsymbol{u} = \{u_i\}_1^L$. We will need to calculate the Hamming distance between $\boldsymbol{u}$ and all individuals in the defined source $\boldsymbol{s}$, the source $\boldsymbol{s}$ consists more than one individual, hence a probability distribution for the Hamming distance between the individual and the source could be defined. Moreover, one could deduce the cumulative probability distribution (c.d.f) $F_{u,s}(\lambda)$ according to probability distribution for the Hamming distances.

Definition 2 ($F_{u,s}(\lambda)$)**.** *$F_{u,s}(\lambda)$ is the cumulative probability distribution which gives the probability that the Hamming distance between* $\boldsymbol{u}$ *and any genotype of the source* $\boldsymbol{s}$ *is smaller than* λ*, i.e.*

$$F_{u,s}(\lambda) = P(\boldsymbol{d(u,\ s)} < \lambda) \tag{1}$$

Definition 3 (q-quantile)**.** *The q-quantile of a probability distribution is defined as the value* $\lambda = \lambda_q$*, such that* $F_{u,s}(\lambda_q) = q$*.*

For instance, 0.05-quantile will give the value λ such that $F_{u,s}(\lambda) = 0.05$, and we call this specific $\lambda = \lambda_{0.05}$. If q is very small, the distance λ_q gives an estimate for the minimal distance between $\boldsymbol{u}$ and the source $\boldsymbol{s}$.

Assume we are trying to attribute unknown $\boldsymbol{u}$ to its source $\boldsymbol{s}$, and there are n sources altogether. The MMD method proceeds as follows:

(1). Calculate the Hamming distance $\boldsymbol{d(u,\ s)}$ between the unknown individual $\boldsymbol{u}$ and every genotype in $\boldsymbol{s}$;

(2). Construct a probability distribution for the Hamming distances obtained in (1), and thus the cumulative probability distribution $F_{u,s}(\lambda_q)$;

(3). Repeat procedure (1) and (2) for every source, and we will obtain n corresponding cumulative distribution functions;

(4). For a given probability q, we will obtain a set of n λ_q values, say $\{\lambda_q\}$ for each c.d.f. calculated in (3) as there are n sources, and we want to find $\lambda_{min} = \min\{\lambda_q\}$. Once λ_{min} is calculated, $\omega_{u,s} = F_{u,s}(\lambda_{min})$ will also be calculated. $\omega_{u,s}$ gives the probability that Hamming distances between $\boldsymbol{u}$ and every genotype in source $\boldsymbol{s}$ is less equal than λ_{min}, which represents proximity between $\boldsymbol{u}$ and the source $\boldsymbol{s}$. In other words, $\omega_{u,s}$ will give a higher value if $\boldsymbol{u}$ is closer to a particular source $\boldsymbol{s}$.

(5). As there are n sources altogether, we will obtain n corresponding $\omega_{u,s}$ values, say $\{\omega_{u,s}\}$, and let $S = \sum\{\omega_{u,s}\}$. Hence, we will be able to estimate the probability that $\boldsymbol{u}$ is attributed to a specific source $\boldsymbol{s}$ as

$$p_{u,s} = \omega_{u,s}/S \tag{2}$$

Note that an individual $\boldsymbol{u}$ is necessarily attributed to at least one source by the methodology.

Consider the diagram 2.2 referenced *Fig S8.* from [1], this is an example for attributing individual unknown human origin to different sources, including C/S Asia, Europe and Middle East, same as in Figure 1.1. According to the MMD method, for an unknown individual $\boldsymbol{u}$, first we will need to compare $\boldsymbol{u}$ with all individuals in the defined sources. In this case, we have three sources, after algorithm steps (1), (2), and (3) we will obtain three c.d.f. corresponding to C/S Asia, Europe and Middle East as shown below.

Fig S8. from ***Mining Whole Genome Sequence data to efficiently attribute individuals to source populations***

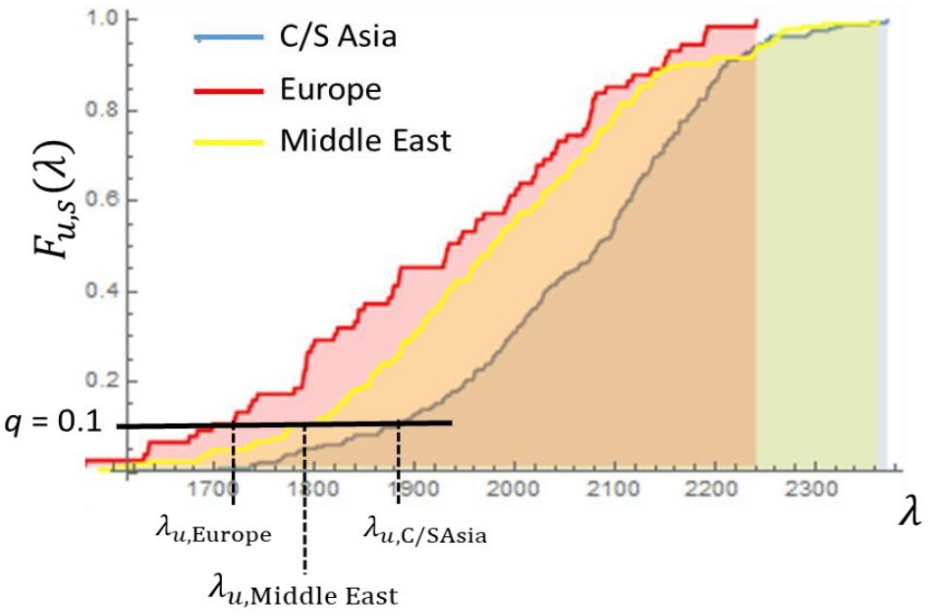

Figure 2.2: MMD example for human geographical origin attribution

Next, we will need to find the q-quantile in procedure (4), here $q = 0.1$ and we obtain three λ_q values for each c.d.f.. Since $\lambda_{0.1}$ for Europe is the smallest, this will become our λ_{min} value, and $\omega_{u,s} = F_{u,s}(\lambda_{min})$ will also be calculated, which represents proximity between $\boldsymbol{u}$ and a particular source $\boldsymbol{s}$.

Finally, for step (5), we will estimate the probability that $\boldsymbol{u}$ is attributed to each source, and in conclusion, the genotype of $\boldsymbol{u}$ is closest to Europe, followed by Middle East and C/S Asia.

2.2 Description of datasets used in analysis

As previously mentioned, the computation time of the method depends on types of input, hence it is necessary to use varying degrees of information. If exceeding information is used to investigate something simple, then most of it becomes redundant and consequently more time is required to perform the computations. Similarly, if less information is used to investigate something more complicated, it might lead to a loss of performance or accuracy. To find the balance between performance and computation time, it is imperative to distinguish the right amount of information required to draw satisfactory conclusions. In this section, we use three different datasets for human campylobacteriosis attribution in order to investigate the threshold of information required to obtain conclusive results with reasonable accuracy.

Since it is necessary to compare performances, all three datasets describe the same genomes but with varying resolutions. The genome isolates are of the *Campylobacter* bacterium which can infect animals and humans to cause disease, and isolates in the datasets have been collected from pigs, wild birds, chicken, sheep and cattle. Along with these sources, *campylobacter* isolates are also collected from humans. Using the MMD method we try to attribute the human *campylobacter* isolates to one of the five possible sources, and the computation time for each dataset is recorded and compared against the accuracy of attribution. It is assumed that when more information is used, the attribution accuracy is higher.

The datasets used in analysis are described as below, and all datasets contain isolates including five mentioned sources and humans:

(1). A MLST (Multi-Locus Sequence Typing)[4] dataset with 1173 isolates and 7 genetic markers for each isolate.

(2). A rMLST (Ribosomal Multi-Locus Sequence Typing)[5] dataset with 1089 isolates and 53 genetic markers for each isolate. Some rMLST genotypes were not available for the isolates considered in the other two datasets, which is the reason there are 1089 isolates for rMLST and 1173 for the other two datasets.

(3). An extended genome with 1173 isolates and 25937 SNP markers in each isolate, SNPs (Single Nucleotide Polymorphisms)[6] occur at high frequencies in both animal and plant genomes.

3 Results

We have tested all datasets mentioned previously for human *campylobacter* attribution with the MMD software proposed in [1], and Figures 3.3, 3.4 and 3.5 below are also generated from the software.

Figure 3.3 shows the probability of attributing human isolates to any one of the five sources using the first dataset with 25937 SNPs. Since this dataset contains the highest resolution of information and the conclusions agree with several other studies conducted previously, these results will be considered as a performance baseline to compare the remaining datasets against. The results show that the probability of attribution of human isolates to the source of chicken is the highest, being around 60% which is significantly higher than the other sources. After chicken,

sheep has 20% probability and cattle has around 15% probability for attribution. The time taken to perform source attribution using this dataset and the MMD method is 3.64 minutes.

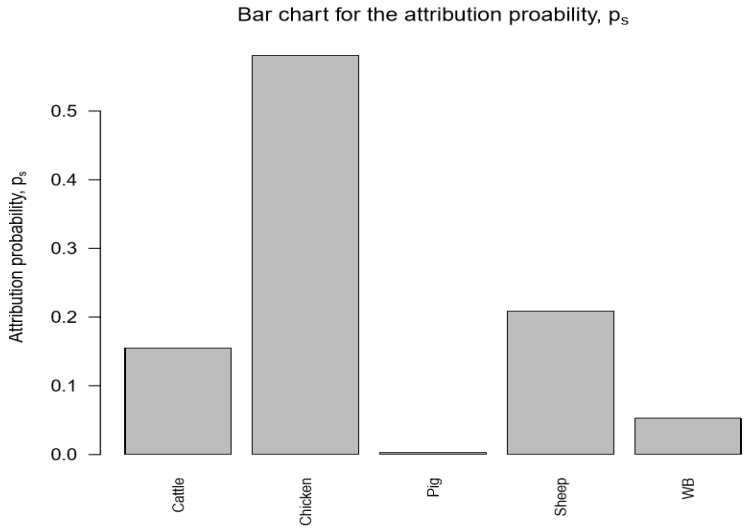

Figure 3.3: SNP data with 25937 data points obtained from MMD software proposed by [1]

In Figure 3.4 for rMLST data, shows that there is the highest probability of attribution to chicken as the source again. Also, there is more confusion between sheep and cattle as no significant difference of probability shown between the two sources. The probability of attribution to wild birds and pigs agree with the results of the SNP dataset although there are slight variations. The time taken to perform source attribution using this dataset and the MMD method is 1.2 minutes .

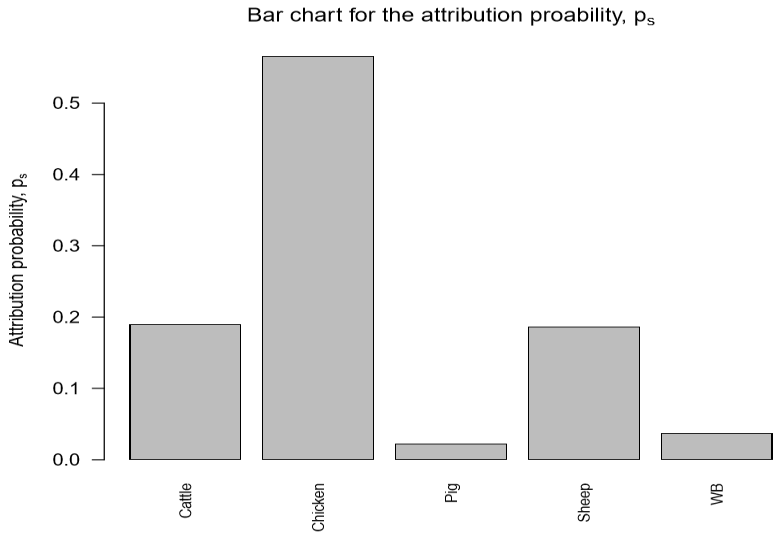

Figure 3.4: rMLST data with 53 data points obtained from MMD software proposed by [1]

Finally, for the third MLST dataset, Figure 3.5 shows an increase of confusion significantly. The probability of attribution to chicken as a source has reduced compares to Figure 3.3 and Figure 3.4. There is a higher probability to be attributed to cattle than there was in either of the other two results. The time taken to perform source attribution using this dataset and the MMD method is 1.3 minutes.

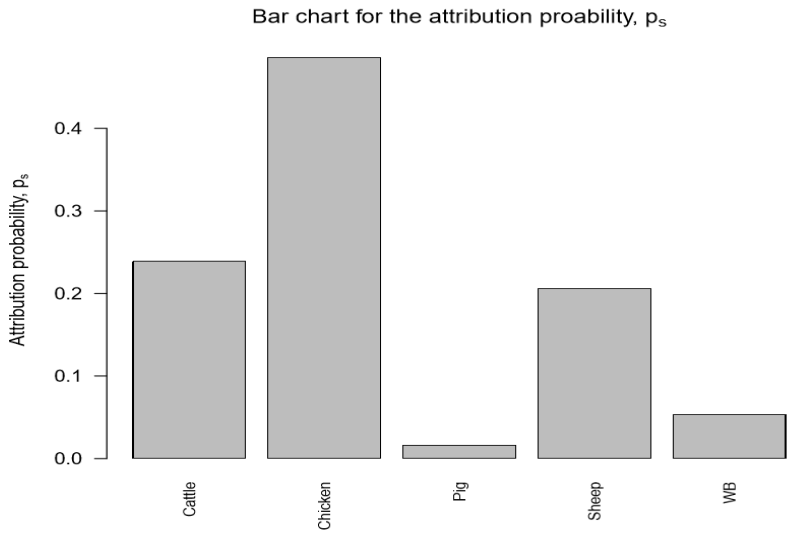

Figure 3.5: MLST data with 7 data points obtained from MMD software proposed by [1]

We note that, all the time measurement for the above analysis includes the time taken to load the data and perform miscellaneous pre-processing, hence it might not be appropriate to use it for comparison when the size of dataset is small, as the time for proceeding with relatively short genotypes is mostly dominated by the time for software takes to pre-process and read the data instead of the calculation itself. (MMD implementation in **R**[7], times based on an Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz processor.)

3.1 Effect of genotype size on attribution accuracy and time

In Figure 3.6 below the rMLST dataset is 0.4% the size of the SNP dataset and it reduces the time taken by the SNP dataset for computation by 67%. The MLST dataset is 0.035% the size of the SNP dataset and it reduces the time taken by the SNP dataset for computation by 64%. The significant reduction in dataset size and computation time for the rMLST dataset is apparent.

However, it must be pointed out that the code takes almost as long for the 7 MLST dataset as for the rMLST dataset. This illustrates the fact that for small datasets as the ones used here, most of the computation time is dominated by the start-up time (reading data and preparation for the actual MMD calculation). In addition, from [1], we also know *"The running time remains essentially constant for runs with less than* 10^5 *loci since the start-up time dominates over the MMD algorithm computation time."*

Moreover, although the size of the MLST dataset and computation time are low, it does not seem to maintain the accuracy of classification when compared to the rMLST dataset. This hints to a possibility that the rMLST dataset achieves a satisfactory trade-off between accuracy of classification, computation time and size.

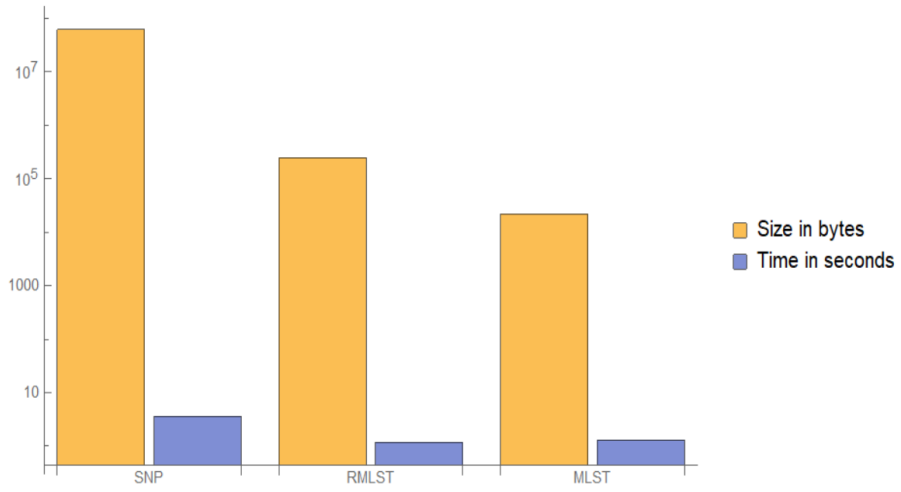

Figure 3.6: Data size and computation time comparison for SNP, rMLST and MLST datasets

3.2 Use of Mathematica for attribution

Mathematica is used to attribute the human *campylobacter* isolates and the probability of attribution is shown in Figure 3.7 with the rMLST dataset compares to the MMD method. The results are very similar between the two algorithms, and it takes 1.625 seconds in Mathematica to proceed the data including the time taken to load the data, transform it and train the classifier. This is a significant reduction in time taken when compared to the MMD method implemented in R. The super-function **Classify** in Mathematica is used for classification, which automatically chooses the appropriate machine learning model in order to get the best results. In this case, it uses the Nearest Neighbours algorithm for attribution[8]. We have also tried to train a classifier with Mathematica using the SNP dataset, but it takes too long to train and kernel quits automatically.

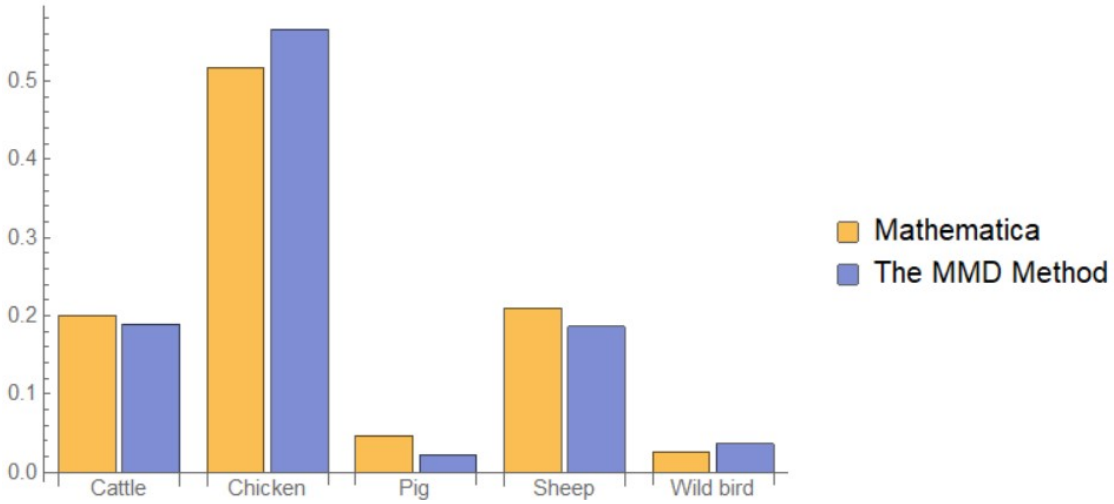

Figure 3.7: rMLST data attributed using Mathematica

3.3 Self-attribution performance between MMD and Mathematica

We try to attribute all the isolates from all five sources to themselves (self-attribution) using the rMLST dataset. The data is separated into training and test data, where 70% is taken as training data and the remaining 30% as test data, in both the MMD method and Mathematica.

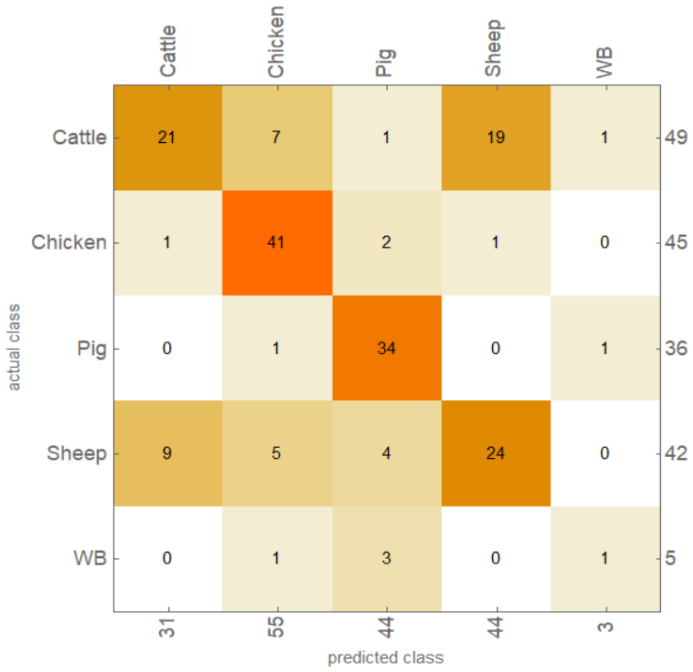

Figure 3.8: Confusion matrix for all sources

We start with Mathematica, and obtain the confusion matrix plot in Figure 3.8 which could thus deduce the true positive rate illustrated in Fig 3.9, representing the probability of correctly attributing itself for each source, and we could see the best classified class is pig and the worst classified class is wild bird.

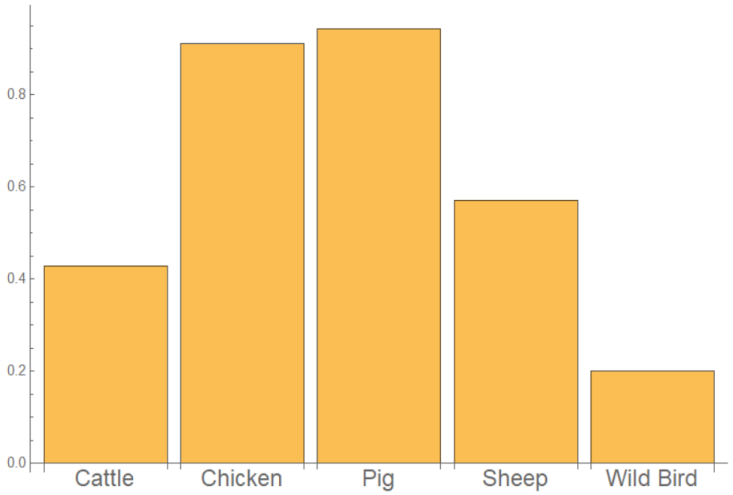

Figure 3.9: TPR of self-attribution for all sources

Also, from the confusion matrix plot 3.8, we could also deduce the false positive rate which represents the probability of incorrectly attributing to each source as illustrated below in Figure 3.10. The class that gets mistaken for other classes the most is the sheep class, and the wild bird class has the lowest false positive rate, this is likely due to a small number of examples for the wild bird class present in the dataset.

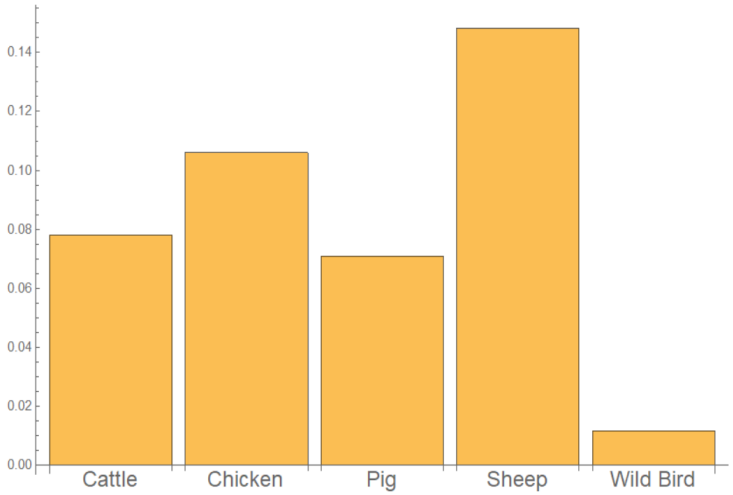

Figure 3.10: FPR of self-attribution for all sources

Figure 3.11 compares the probability of correct self-attribution between the MMD method obtained from [1] and Mathematica. Overall, their performances seem to be similar, with an exception of the wild bird class. The MMD method is more accurate in attributing wild birds when compared to Mathematica, and Mathematica has a slight edge over the MMD method for the chicken and sheep classes. The pig class seems to be the most distinguishable class out of all the sources for both MMD method and Mathematica.

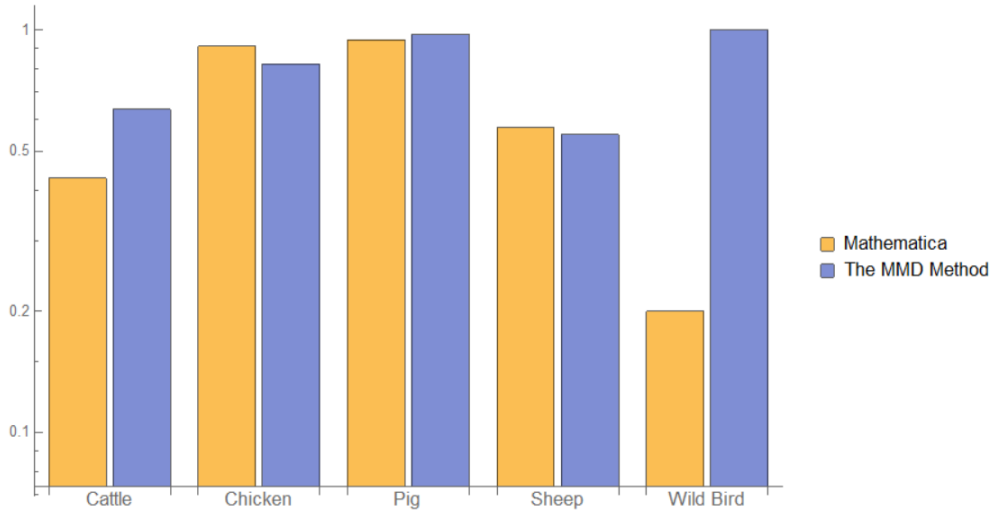

Figure 3.11: Self-attribution probability between MMD and Mathematica for all sources

4 Conclusion

In conclusion, the MMD method is very fast and accurate, and especially good at dealing with large datasets, and gives satisfactory results with high accuracy in a reasonable time. However, when dealing with medium and small size of datasets, the computation time of the MMD method has decreased but not significantly as most of the computation time is dominated by the start-up time, and the accuracy of the results also decreased. In contrast, Mathematica uses much less time for training a classifier with medium and small size of datasets, but Mathematica cannot analyse large datasets due to the oversize data and kernel quits automatically. Overall, the MMD method performs extremely well in analysing large scale datasets such as Whole Genome Sequence data, and it is more practical in the real world. In many scenarios such as medical diagnosis, it is crucial to include all the details which will give large data to proceed with, and certainly the MMD method will be a very efficient and accurate way to solve this kind of problems.

References

[1] Francisco Perez-Reche, Ovidiu Rotariu, Bruno Lopes, Norval J Strachan, and Ken J Forbes. Mining whole genome sequence data to efficiently attribute individuals to source populations. *BioRxiv*, 2020.

[2] Whole-genome sequencing. https://emea.illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing.html.

[3] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[4] Multilocus sequence typing (mlst). https://pubmlst.org/general.shtml.

[5] Ribosomal multilocus sequence typing (rmlst) - pubmlst.org. https://pubmlst.org/rmlst/.

[6] EK Khlestkina and EA Salina. Snp markers: methods of analysis, ways of development, and comparison on an example of common wheat. *Russian Journal of Genetics*, 42(6):585–594, 2006.

[7] The r project for statistical computing. https://www.r-project.org/.

[8] Dhilip Subramanian. A simple introduction to k-nearest neighbors algorithm, Jan 2020.