Human ethnicities classification with kernel functions on single nucleotide polymorphisms

Course: Machine Learning 2

Víctor Adell
Pau Autrand
Miquel Escobar

November 29, 2019

Contents

1 Introduction 3

2 A brief on single nucleotide polymorphisms 3

3 Description of the dataset 5

3.1 Data preprocessing 5

3.2 Some insights on the dataset 7

4 Classification by means of k-SVM 8

4.1 Dirac kernel 8

4.2 Spectrum (n-gram) kernel 10

4.3 Allele match kernel 13

5 Visualization by means of k-PCA 14

6 Conclusions 14

References 15

1 Introduction

The purpose of this project is to classify the single nucleotide polymorphisms (also known as SNPs) of our dataset into 4 classes, each corresponding to a different ethnicity. The dataset consists of humans' genome sequences, which is useful to observe variation and mutations among different ethnicities. What we are looking for is, precisely, to detect these variations and mutations of each of the ethnicities in order to classify the observations.

2 A brief on single nucleotide polymorphisms

This section intends to provide a succinct outline of the foundations in genetics necessary for this project. According to *Genetic Mutation* in *Nature Science* [1], the haploid human genome consists of 3 billion nucleotides. While a mutation is defined as any alteration in the DNA sequence, biologists use the term *single nucleotide polymorphism* (SNP) to designate **a single base pair alteration that is common in the population**.

A base pair refers to two bases which form a ”rung of the DNA ladder”. A DNA nucleotide is made of a molecule of sugar, a molecule of phosphoric acid, and a molecule called a base. The bases are the *letters* that spell out the genetic code. In DNA, the code letters are `A`, `T`, `G`, and `C`, which stand for the chemicals adenine, thymine, guanine, and cytosine, respectively. In base pairing, adenine always pairs with thymine, and guanine always pairs with cytosine [3].

Specifically, a polymorphism is any genetic location (or locus) at which at least two different sequences are found, with each sequence present in at least 1% of the population. Note that the term *polymorphism* is generally used to refer to a normal variation, or one that does not directly cause disease. Moreover, the cutoff of at least 1% prevalence for a variation to be classified as a polymorphism is somewhat arbitrary; if the frequency is lower than this, the *allele* (i.e. the variant form of a gene) is typically regarded as a mutation [9]. Moreover, all common SNPs have only two alleles [4].

For example, two sequenced DNA fragments from different individuals, `AAGCCTA` to `AAGCTTA`, contain a difference in a single nucleotide. In this case we say that there are two alleles: C and T. This particular case extracted from [4] and is portrayed in the Figure 1 below.

In a narrow sense, the term *genotype* can be used to refer to the alleles, or variant forms of

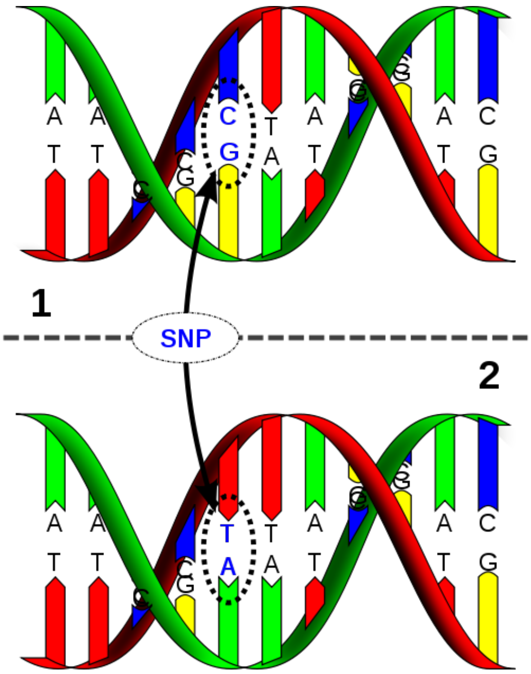

Figure 1: Diagram of a SNP [4]

a gene, that are carried by an organism [2]. Humans are diploid organisms, which means that they can have two possible alleles at each genetic position, or locus, with one allele inherited from each parent. Each pair of alleles represents the genotype of a specific gene.

In the case of a *single nucleotide polymorphisms*, there are two possible alleles which can be represented by the uppercase letters A and B for the sake of generality. A varied population could therefore feature three possible genotypes at this locus: AA, AB, or BB. A particular genotype is described as homozygous if it features two identical alleles and as heterozygous if the two alleles differ.

On average, SNPs are found every 1,000–2,000 nucleotides in the human genome. They are important as markers, or signposts, for scientists to use when they look at populations of organisms in an attempt to find genetic changes that predispose individuals to certain traits, including disease [1].

Within a population, SNPs can be assigned a minor allele frequency — the lowest allele frequency at a locus that is observed in a particular population. This is simply the lesser of the two allele frequencies for single-nucleotide polymorphisms. There are variations between human populations, so a SNP allele that is common in one geographical or ethnic group may be much rarer in another [4].

3 Description of the dataset

The dataset for this project has been extracted from the *11 Million SNP Profiles Datasets* hosted in *Harvard Dataverse* [7]. The observations selected are those corresponding to the founding generation of each of the ethnicities, which are *African American*, *Estonian*, *Korean*, *Palestinian.* Each data record is labeled as follows:

- **Id:** Unique identifier of the individual.
- **Gender:** Male (M) or female(F).
- **Father:** Unique identifier of the father. NA for 1st generation.
- **Mother:** Unique identifier of the mother. NA for 1st generation.
- **African American ethnicity:** Individual's fraction of American African ethnicity, from 0 to 1.
- **Estonian ethnicity:** Individual's fraction of Estonian ethnicity, from 0 to 1.
- **Korean ethnicity:** Individual's fraction of Korean ethnicity, from 0 to 1.
- **Palestinian ethnicity:** Individual's fraction of Palestinian ethnicity, from 0 to 1.
- **Father's ordered SNP alleles:** the alleles for each of the locus corresponding to the individual's father. Each allele is represented with just one nucleotide, as its pair is implicit (the options are CG and AT).
- **Mother's ordered SNP alleles:** the alleles for each of the locus corresponding to the individual's mother. Each allele is represented with just one nucleotide, as its pair is implicit (the options are CG and AT).

3.1 Data preprocessing

Data editing

The downloaded files have a total of 45000 rows combined. For computation purposes, we sampled a total of $N = 500$ observations, $\frac{N}{4} = 125$ for each ethnicity, that is, individuals with a fraction of 1 of the corresponding ethnicity. Each of the observations contains a total of $P = 39108 \times 2 = 78216$ characters, one for each locus and allel. The reason for which the data has been sampled is the execution cost. The computation of a kernel matrix implies the application of the kernel function on $N \times N$ pairs of values, for each of the P locus, which leaves us with a total cost of $O(N^2 \times P)$. In this case, $P > N$, which means that the cost is actually $O(N^3)$. Wrapping it up, this means that a greater N would translate into a much slower execution due this *cubic cost*, which would not be useful in the scope of this project, as it does not intend to analyze large volumes of data but instead seeks to analyze the performance of the selected kernels.

Population sample	**Population size**	**Female/Male**	**Origin**
EthA	125	70 / 55	African American
EthE	125	67 / 58	Estonian
EthK	125	63 / 62	Korean
EthP	125	68 / 57	Palestinian

Table 1: Description of size and origin of each population sample for the randomly sampled data with $N = 500$

Data wrangling

The aim of this process of data wrangling is to transform both *Allele* columns into several columns, each one corresponding to the combination of the locus of the parents at a given position. As explained before, this combination is expressed as either AA, AB or BB. For memory purposes, simplicity and because it would have to be computed later anyways, this string expression has been mapped into an integer. We used the following encoding f, where we defined:

- a_1 as the father's allele
- a_2 as the mother's allele
- a_n as the normal allele
- a_a as the altered allele

Then,

$$f(a_1, a_2) = \left\{ \begin{array}{ll} AA, & \text{if } a_1 = a_2 = a_n \\ AB, & \text{if } (a_1 = a_n) \oplus (a_2 = a_n) \\ BB, & \text{if } a_1 = a_2 = a_a \end{array} \right\}$$

To apply this encoding to our data, we had to define the normal and the altered allele for each locus. In fact, it could be chosen randomly and it would not have any connotations for the kernels we use, but for formality purposes we will select the most frequent allele as the *normal*, and the most *infrequent* as the *altered*. We do this because the normal allele would be the most frequent over all humans: thus it is trivial to see that the most frequent in the dataset has the higher chance of actually being the most frequent over all humans. [1]

3.2 Some insights on the dataset

As an example, we show the SNP's in the 10 last positions of the first 10 individuals in the database, which are all part of the ethnic group EthA:

	r39098	**r39099**	**r39100**	**r39101**	**r39102**	**r39103**	**r39104**	**r39105**	**r39106**	**r39107**	**Population**
Ind. 0	2	0	1	2	0	2	0	0	1	0	EthA
Ind. 1	0	1	1	1	1	2	1	0	1	0	EthA
Ind. 2	1	1	1	2	0	2	0	1	1	0	EthA
Ind. 3	2	0	2	2	1	2	1	2	1	0	EthA
Ind. 4	0	0	0	2	1	2	1	0	1	1	EthA
Ind. 5	1	1	1	2	1	1	0	0	0	0	EthA
Ind. 6	1	1	2	2	1	2	1	0	2	0	EthA
Ind. 7	1	1	1	2	0	2	1	0	0	0	EthA
Ind. 8	2	1	2	1	2	1	0	0	0	0	EthA
Ind. 9	2	0	2	1	1	2	1	1	1	0	EthA

Table 2: First rows of the last 10 single-nucleotide polymorphishms in the dataset

To get familiarized with the data distribution, we show each the population's alleles fraction for each of the last 10 locus in the dataset:

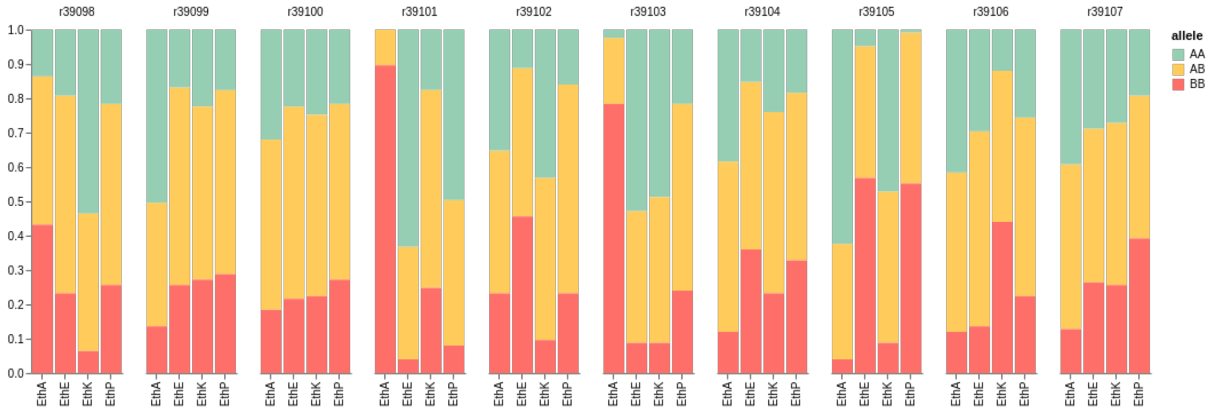

Figure 2: Proportion of each allele in different SNP's by population

We observe that some SNPs have clearly distribution differences between populations, which allows the distinction between them, while the locus with similar allele distributions do not

[1] It is worth noting that the encoding process is very slow, as it has to access $N \times P \times 2$ values in order to encode each of the cells, which executed in python (refer to `pre_processing/` directory) takes some minutes. We could have used a compiled language, but considering this step has to be executed just once, it is not a critical issue.

contribute for the classification process. To wrap up the data preprocessing and exploration, and with all the knowledge acquired up to this point, we want to remark that the use of a kernel method is an adequate approach to generate a classifier. Kernel methods are exceptionally well suited for situations in which the number of dimensions (referred as P in this document) is very large, specifically when the number of dimensions P is much greater than the number of observations N ($P >> N$). The reason for this is that the size of the kernel matrix will always be $N \times N$, regardless of the value of P. Consequently, when training the data, the computation cost is incredibly lower than in the case, for instance, that we trained a neural network using all P dimensions.

4 Classification by means of k-SVM

For scoring genetic similarity between two individuals, we have considered several kernels. We have started with general kernels for multivariate categorical vectors such as *Dirac* and *n-gram* and we have finally tried *ad-hoc* ones such as the *allele match kernel.*

The kernel matrices have been programmed in C++ for efficiency reasons and the code is available in the `build_kernels/` folder of the repository. As for the kernel methods, we have relied on the `kernlab` package in R [5].

4.1 Dirac kernel

Let $\mathbf{x}, \mathbf{x}'$ be the genotypes of two individuals at P different locus. The Dirac or overlap kernel is defined as

$$k(\mathbf{x}', \mathbf{x}') = \frac{1}{P}\sum_{k=1}^{P} \mathbb{I}_{\{x_k = x'_k\}},$$

where

$$\mathbb{I}_{\{z\}} = \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{if } z \text{ is false} \end{cases}$$

Therefore, the similarity between two genotypes is then equivalent to the number of locus in which they match. To show that it is indeed a positive semi-definite kernel, we can focus on the comparison at a given locus between x_l and x'_k. We now let $\phi_{\text{locus}}(x_k)$ be a vector of length

3 where each position is an indicator of whether x_k falls into each of the alleles AA, AB, BB [8],

$$\phi_{\text{locus}}(x_k) = \begin{cases} (1,0,0) & \text{if } x_k = AA \\ (0,1,0) & \text{if } x_k = AB \\ (0,0,1) & \text{if } x_k = BB \end{cases}$$

These two locus have a match score of 1 if they have the same allele and zero otherwise, which is captured by the dot product $< \phi_{\text{locus}}(x_k), \phi_{\text{locus}}(x'_k) >$ so that for a single locus, the dot product is a valid kernel. Summing these kernels across all P locus and then multiplying by the positive scalar $\frac{1}{P}$ will also result in a valid kernel. Hence, the overall kernel function is a mapping between these two spaces: $\phi : \{AA, AB, BB\}^P \rightarrow \{0,1\}^{3 \cdot P}$

The Figure 3 below displays the computed Dirac kernel matrix for our dataset composed of 500 individuals.

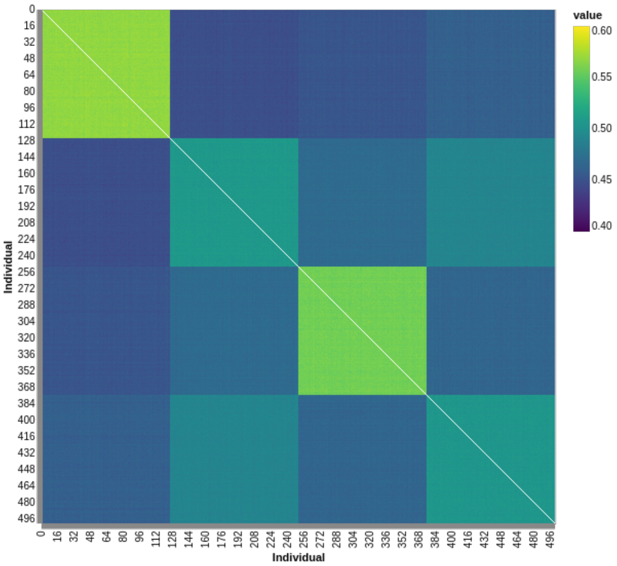

Figure 3: Kernel matrix generated by Dirac kernel

We first note that the matrix is essentially composed of square sections which are approximately homogeneous in their values. These sections perfectly match up with each population of 125 individuals, which are ordered alphabetically in the dataset: *EthA, EthE, EthK, EthP*. Moreover, the brightest colors in the kernel matrix, which correspond to the highest values of the kernel function are in the diagonal which means that members of the same population are more similar among them. It is also worth mentioning that the *Estonian* and *Palestinian* ethnicities entail the most notorious similarity with respect any other pair.

We have divided our 500 × 500 kernel matrix into a training set ($\frac{2}{3}$), in which we have modelled a k-SVM, and a test set ($\frac{1}{3}$). The accuracy on the test set has been of 100% as the confusion matrix below depicts:

```
          truth
pred    EthA EthE EthK EthP
  EthA    41    0    0    0
  EthE     0   41    0    0
  EthK     0    0   41    0
  EthP     0    0    0   44
```

Table 3: Confusion matrix on test dataset with Dirac kernel

4.2 Spectrum (n-gram) kernel

In the context of the alleles in single nucleotide polymorphisms, we have a finite alphabet $\mathcal{A} = \{AA, AB, BB\}$. We have that an n-gram is a block of n adjacent characters from the alphabet $\mathcal{A}$ and we define the kernel

$$k(\mathbf{x}', \mathbf{x}') = \sum_{s \in \mathcal{A}^n} |s \in \mathbf{x}| \cdot |s \in \mathbf{x}'|$$

The implicit feature map represents the genotype $\mathbf{x}$ as a multiset of its subparts of length n. Therefore, the dimensionality of this space is $|\mathcal{A}^n| = 3^n$. The applications of the n-gram kernel on genomic similarity measures have been studied by [8] and we will first analyze the 1-gram (or unigram) kernel in which we are solely comparing the distribution of the alleles AA, AB and BB for all possible SNPs. Its corresponding kernel matrix is displayed in the Figure 4 below.

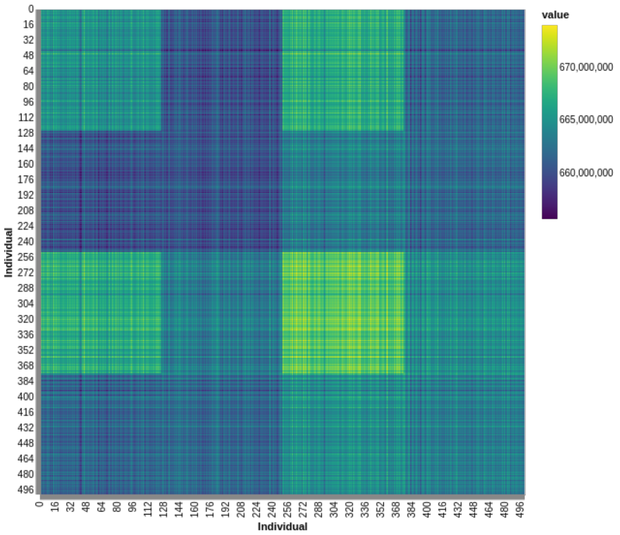

Figure 4: 1-gram kernel matrix

The results on the test set have an error rate of 26.95% and its reason can be unveiled from the confusion matrix below:

```
          truth
pred    EthA EthE EthK EthP
   EthA   40    0    0    0
   EthE    0    0    0    0
   EthK    0    0   43    0
   EthP    0   45    0   39
```

Table 4: Confusion matrix on test dataset with 1-gram

Essentially, all Estonian individuals are being classified as though they were *Palestinians*. In order to check if this tendency is stable for different values of n, we have computed the kernel matrices for the $\{3, 5, 7, 11\}$-gram:

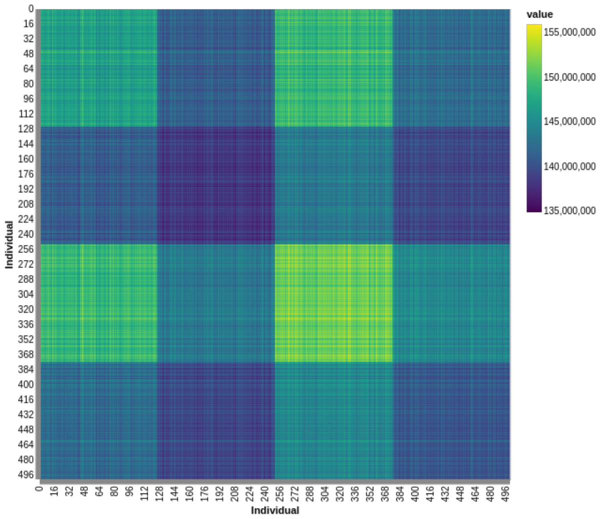

Figure 5: 3-gram kernel matrix

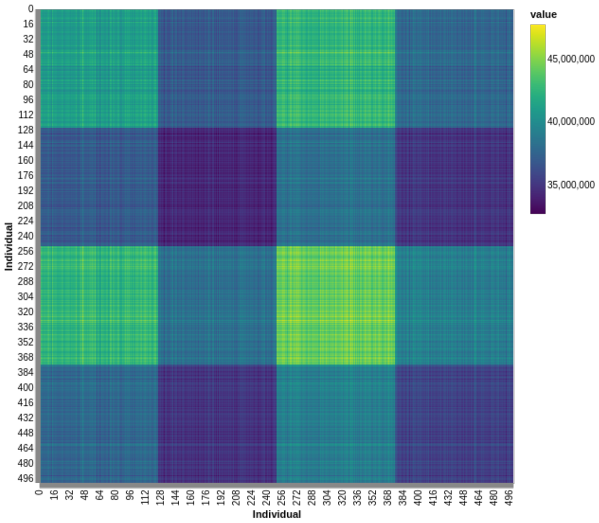

Figure 6: 5-gram kernel matrix

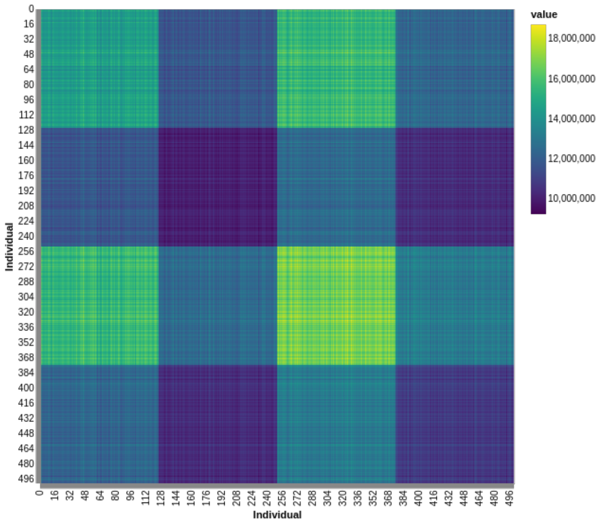

Figure 7: 7-gram kernel matrix

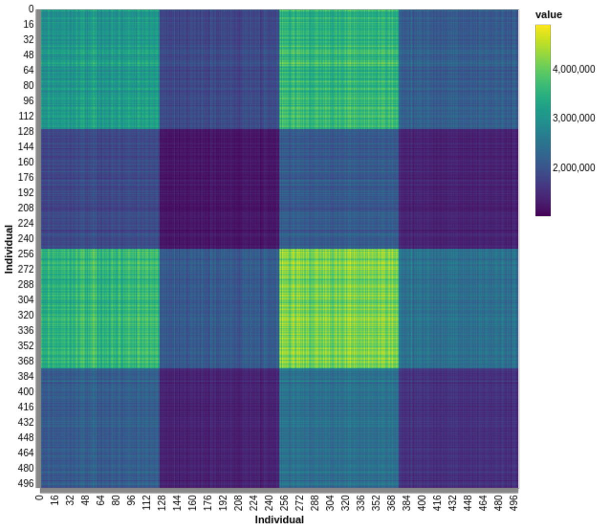

Figure 8: 11-gram kernel matrix

As we increase n, a certainly surprising pattern seems to gain relevance. The Korean-Korean square has the highest values of the kernel function and two groups of rows seem to emerge, one composed by the first and third row (*African American* and *Korean*) and the other composed by the second and fourth row (*Estonian* and *Palestinian*). Moreover, this second group has the particularity that the values of the kernel function are higher when they are compared with the members of the first group with respect to themselves. This fact leads us to hypothesize that the *African American* and *Korean* ethnicities have a more stable distribution of alleles in consecutive SNPs.

The results obtained on the test set of the 3-gram kernel are identical with respect to the unigram and the rest of cases are outlined in the tables below:

```
      truth
pred   EthA EthE EthK EthP
  EthA   40    0    0    0
  EthE    0   39    0    3
  EthK    0    0   43    0
  EthP    0    6    0   36
```

Table 5: Confusion matrix on test dataset with 5-gram kernel (Error rate of 5.39%)

```
      truth
pred   EthA EthE EthK EthP
  EthA   40    0    0    0
  EthE    0   37    0    1
  EthK    0    0   43    0
  EthP    0    8    0   38
```

Table 6: Confusion matrix on test dataset with 7-gram kernel (Error rate of 5.39%)

```
      truth
pred   EthA EthE EthK EthP
  EthA   37    0    2    0
  EthE    0   38    0    3
  EthK    3    0   41    0
  EthP    0    7    0   36
```

Table 7: Confusion matrix on test dataset with 11-gram kernel (Error rate of 8.98%)

When analyzing the confusion matrices of the k-SVM, we can corroborate what we could visually see on the kernel matrices. Essentially, the only misclassifications occur between the pairs *EthA-EthK* and *EthE-EthP*.

4.3 Allele match kernel

The allele match kernel is suggested by [6] and can be regarded as a modification of the *Dirac* kernel. Essentially, for each pair of locus entails a weight of 4 if they are the same, 2 if one is a heterozygote and the other is a homozygote and 0 if they don't share any common alleles. In this case we would have the following kernel function on a locus-per-locus basis,

$$\phi_{\text{locus}}(x_k) = \begin{cases} (2, \sqrt{2}, 0) & \text{if } x_k = AA \\ (\sqrt{2}, \sqrt{2}, \sqrt{2}) & \text{if } x_k = AB \\ (0, \sqrt{2}, 2) & \text{if } x_k = BB \end{cases}$$

As in the *Dirac* kernel, the overall kernel function results in adding up all of the locus-per-locus comparisons and then dividing by P. The resulting kernel matrix is displayed in the Figure 9 below.

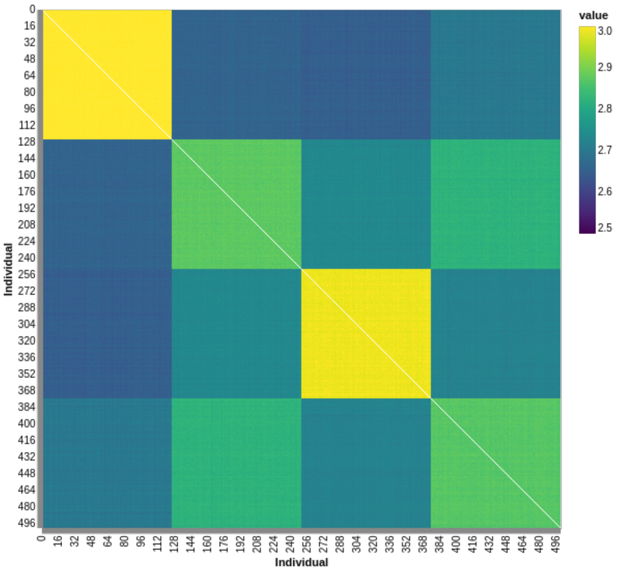

Figure 9: Allele match kernel matrix

Its appearance is certainly similar to the *Dirac* kernel and we have also obtained a 100% accuracy on the test set.

5 Visualization by means of k-PCA

In order to have some insight, we can compare a kernel with perfect accuracy (*Dirac*) with respect to *11-gram* kernel by visualizing the 2-dimensional kPCA of their kernel matrices.

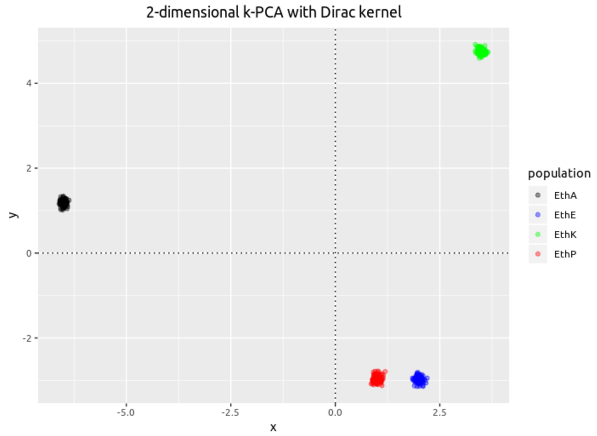

Figure 10: k-PCA with Dirac kernel

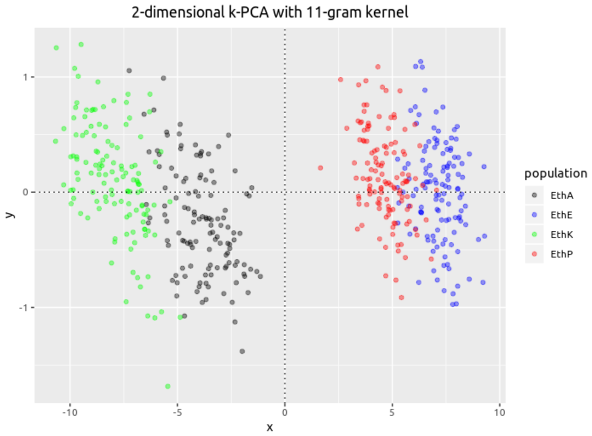

Figure 11: k-PCA with 11-gram kernel

We observe that the 2-dimensional kPCA representation classification of the *Dirac* kernel is actually linearly separable, while the 11-gram kernel clearly is not, which we could expect considering the corresponding accuracies.

6 Conclusions

The fact that we achieved a better performance with kernels on a *locus-per-locus* basis with respect to *n-gram* might be an indicator that the alleles in sequences of SNPs are not highly correlated.

Another observation is that the most similar ethnic groups out of the four, based on the results, are *Estonian* and *Palestinians*, while the most opposed are the *Korean* and the *African American*. This is what we could expect based on the geographic origins and on the physical appearance and attributes of each of the ethnicities, so the kernels' predicitions seem to align with our previous intuition.

As a general observation on kernels, we could say that a relevant property is that it is relatively easy to obtain general knowledge from a dataset with a generic kernel, but on the other hand it might difficult to find a good kernel for specific data. In this case have seen that a generic kernel such as *Dirac*'s was able to perform with perfect accuracy on the test set, even though the *Allele match* kernel obtained significant better results (it can be seen from the kernel matrices representation), as it is a kernel based on previous knowledge on this specific problem, which is not always present in prediction or classification problems.

References

[1] Suzanne Clancy. “Genetic Mutation”. In: *Nature Education* 1.1 (2008), p. 187. URL: https://www.nature.com/scitable/topicpage/genetic-mutation-441/.

[2] Nature Education. *Definition of genotype.* 2014. URL: https://www.nature.com/scitable/definition/genotype-234/.

[3] Renata Geer. *Genetics Review.* Ed. by National Center for Biotechnology Information. 1999. URL: https://www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Genetics/basepair.html.

[4] International Society of Genetic Genealogy Wiki. *Single-nucleotide polymorphism.* 2018. URL: https://isogg.org/wiki/Single-nucleotide_polymorphism.

[5] Alexandros Karatzoglou et al. “kernlab – An S4 Package for Kernel Methods in R”. In: *Journal of Statistical Software* 11.9 (2004), pp. 1–20. URL: http://www.jstatsoft.org/v11/i09/.

[6] Indranil Mukhopadhyay et al. “Association tests using kernel-based measures of multi-locus genotype similarity between individuals”. In: *Genet Epidemol* 34.3 (2010), pp. 213–221. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3272581/.

[7] Darrell Ricke. *11 Million SNP Profiles datasets.* Version V3. 2018. DOI: 10.7910/DVN/IDT8HZ. URL: https://doi.org/10.7910/DVN/IDT8HZ.

[8] Daniel Schaid. “Genomic Similarty and Kernel Methods”. In: *Human Heredity* 70 (2009), pp. 109–131.

[9] Richard Twynman. *Mutation or polymorphishm?* Ed. by Wellcome Trust. 2003. URL: http://fire.biol.wwu.edu/trent/trent/polymorphism.pdf.