Correlation between Staphylococcal protein A types and geographic location

Nina Romanow

A thesis presented for the degree of
Bachelor of Science

Algorithmic Bioinformatics
Heinrich Heine University Düsseldorf
Germany
3rd October, 2022

Acknowledgments

I want to express my gratitude to Prof. Dr. Gunnar Klau for creating the opportunity to write this thesis, suggesting this fascinating topic, and nudging me towards implementing python scripts within my thesis - which made me even more enthusiastic about Python and coding in general. I am also very grateful to Philipp Spohr for assisting and guiding me during the implementation of the methods and for providing lots of helpful feedback and impulses. In addition, I want to offer my special thanks to Simon Z. for answering all my coding-related questions and being of great help during Python emergencies. For valuable emotional support and patience throughout the process of implementing and writing, I want to express heartfelt gratitude to Maxim O. Finally, I want to thank Rafael C. and Maxim O. for taking the time to proofread this thesis.

Abstract

Staphylococcus aureus is a methicillin-resistant human pathogen (MRSA) causing severe infections with high mortality globally. Treatment options are limited, and the high adaptability of the pathogen leads to a fast-growing variety of MRSA strains. Molecular typing of a polymorphic X-region of the protein A gene (spa) has helped to control and document the variety of resulting spa-types. Spa-types consist of a repeat succession, in which each repeat represents a DNA sequence. The Ridom SpaServer provides strain records for about 20.000 different spa-types, naming the spa-type, their respective repeat succession, and geographic origin. Newly found spa-types are added to the database frequently, and some research has already been done by evaluating the data. However, no study explored the possible link between spa-types' genetic similarity and their origin location. We outlined groups of genetically related spa-types by aligning and clustering the repeat successions and used those groups to investigate their locations further. Fundamentally this thesis presents methods and their application to study the correlation between spa-types and geographic location.

Contents

1 Introduction

Staphylococcus aureus is a methicillin-resistant human pathogen (MRSA) causing severe infections with high mortality, such as pneumonia, septicemia, and other invasive diseases. Spreading globally, it has become one of the leading causes of bacterial infections in hospitals and other healthcare settings. Treatment options for MRSA are currently limited, and the pathogen itself can adapt to a changing environment, which leads to a fast-growing variety of MRSA strains [1]. Molecular typing of MRSA has helped control and document the spreading of various strains. A polymorphic X-region of the protein A gene (spa) was used for the typing; it is made up of a variable number of small repeats [2]. Repeats represent a specific DNA sequence and have a unique ID. In July 2022, over 800 different repeat sequences were recorded in the Ridom SpaServer Database [1]. For every spa-type entry in the database, a strain record is given, which consists of additional data such as the isolation/submission year and the original location of every type accession.

The collected data might be used as a tool to comprehend the spreading and evolution of MRSA strains. Evaluating the spreading of Staphylococcus aureus in different locations with regard to the genetic similarity of the strains may give us some insight into new strains emerging. This information can contribute additional input for epidemiologists. Understanding how genetically similar types of MRSA are spreading could influence the development of treatment options by finding solutions that work for a group of strains.

In this thesis, we proposed and applied a method to measure the genetic distance between different strains (types) of MRSA, using repeat sequences, and analyzed the potential correlation between their genetic and geographic distance. Classifying similar spa-types in complexes has improved the research on MRSA in the past [3]. We have built on this knowledge and examined MRSA locations of origin within the formed complexes.

Essentially we want to examine whether genetically similar MRSA types emerge in geographically close locations and if a correlation between the two variables is measurable.

To approach this question, we use the existing alignment algorithm calculating the homo-edit distance to measure the genetic similarity of spa-types. We align a selection of n spa-types in a $n \times n$ matrix, to further group spa-types with a low score - indicating a close genetic similarity - into clusters, using varying parameters. The origin locations of the spa-types inside the created clusters are evaluated and presented in an interactive map, showing the spa-types found within a country and the cluster they belong to.

[1] *https://spa.ridom.de*

Finally, we will discuss the results our research has produced and the limitations of our method. There will also be an outlook on possible further implementation strategies, to refine and evolve our work.

2 Background Information

This chapter will provide essential background information on this thesis. There will be an introduction to Staphylococcus Aureus and a description of spa-typing with a few examples. Further, we will take a short excursion to the Ridom SpaServer Database to understand how the data given for the upcoming methods is presented.

2.1 Staphylococcus Aureus

Methicillin-resistant Staphylococcus aureus is one of the leading causes of clinical infections. Around $20-30\%$ of the population carry the MRSA in their body asymptomatically, and $20-60\%$ can be intermittent carriers [4]. The pathogen is mostly nasally colonized and spread by person-to-person transmission or contact with contaminated items [2]. The number of infections is constantly growing, increasing the burden on health care resources. Staphylococcus aureus infection mortality was high during past influenza pandemics and can also occur as an additional bacterial infection in patients infected with COVID-19. Especially Staphylococcal pneumonia has been complicating an existing COVID-19 infection in patients [5]. Since MRSA is difficult to treat, further research on its spreading and evolution is important.

2.2 Spa Typing

Spa typing supports infection control measures and provides us with more information about MRSA strains. The polymorphic X region of the protein A gene (spa) appears in every strain of Staphylococcus aureus[6]. It has been shown that DNA sequence analysis of this region delivers an accurate and rapid method to discriminate between different strains of MRSA [7].

In this region, individual repeats have an average length of 24 base pairs. Every repeat is assigned a numeric repeat ID. In September 2022, 838 different repeats, and 20683 spa-types were sequenced and stored in the Ridom Spa Server. A spa-type consists of multiple repeats, the amount varies between $1-20$. A numeric ID is assigned to every type, in form of tx, with x being a number between $001-20683$. Typing and concatenating repeats allows us to perform further analysis on relatively short numeric sequences of max. 25 repeats instead of working with long DNA sequences, which could add up to ≈ 600 base pairs. An example of a spa-type is given in Figure 1.

2.3 Ridom SpaServer

The company Ridom, located in Münster, Germany, provides and maintains a public database in which the spa-type records are stored. Researchers can submit additional sequence records for existing spa-types or add an entirely new spa-type. The SpaServer contains the repeat

[2] *https://www.cdc.gov*

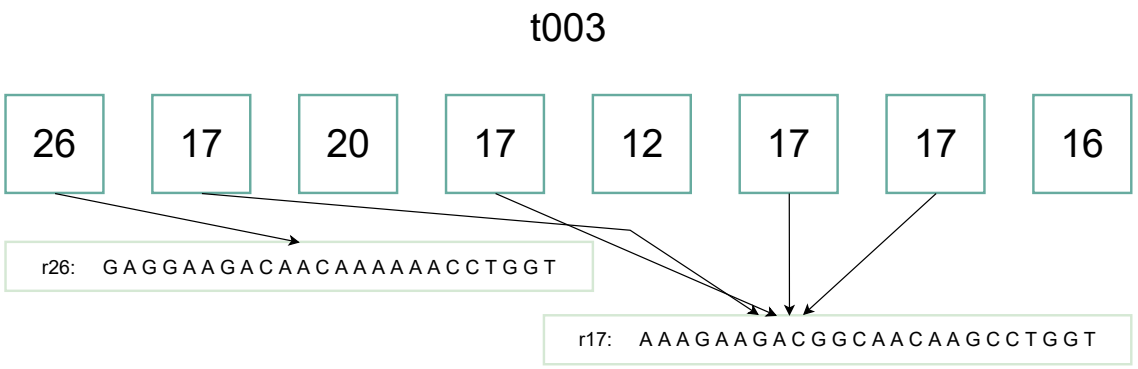

Figure 1: Spa-type $t003$ and its repeat succession. Each repeat number represents a DNA sequence, the same repeat can appear multiple times.

sequences available as a FASTA download and extensive records for differing types. A downloadable .txt file with every spa-type sequenced, including its type ID and repeat succession, is also provided. Spa-types are sorted by name or relative global frequency; each spa-type has its table of strain records. For every SpaServer accession, a record with the isolation year, submission year, origin country, and other information is created. Frequent types like $t003$ have up to ≈ 20.000 records containing different origin locations; less frequent types mainly consist of just one entry, sometimes without an origin location. In this thesis, we will work with the provided file containing all spa-types and use the individual strain records to acquire the repeat sequence and the geographic location of spa-types.

3 Methods

Multiple steps will be taken to examine the correlation between spa-types and geographic location. The methods and algorithms used in this thesis will be described, as well as a short explanation of how the provided data had to be modified to be evaluable. Furthermore, the decision-making process during the implementation of the methods will be pointed out and their possible impact on the result will be discussed further in section 5.

3.1 Data preparation

Since we want to examine the geographic location of all spa-types given, we wrote a Python script to clean up the provided .txt file containing all spa-types and their repeat succession. We've checked all strain record entries for every spa-type and searched for existing country entries with the help of a Web Scraper. Most of the entries contained a valid country name, but some had to be eliminated due to empty entries or abbreviations which could not be assigned to a country. A small number of entries contained city names that were successfully mapped to their respective country name. After cleaning up the data, 17871 different spa-types were left, which had to be sorted numerically in a new .txt file.

Frequent types such as $t003$ or $t032$ have many different country entries, partially up to 20.000, while other types only have one record saved. In this thesis, we have chosen to work with one location per spa-type only, so we have created a second local list, selecting the most occurring country for a spa-type and assigning it to its ID.

3.2 Homo Edit Distance

Alignment algorithms describe a process of comparing and detecting similarities or differences between genetic sequences. Simple sequence analysis can consist of calculating the number of matching symbols in two different sequences of the same length. This value measures the degree of similarity and is commonly known as the alignment score of a pairwise alignment. The number of dissimilarities between the sequences is called the Hamming distance.

The Hamming distance does not contain the possibility to emulate biological events such as deletions and insertions within a string, so we want to consider a different approach. The homo-edit distance (HED) problem calculates the minimum number of homo-deletions or homo-insertions, converting one string into another. A homo-insertion inserts a string of equal characters, called a block, into another string; the inverse operation is called homo-deletion. A homo-deletion can be used to merge identical characters into blocks. The overall alignment score is reduced when using blocks consisting of multiple characters instead of focusing on single characters. The Hamming distance of the sequences "CTC" and an empty string is 3, considering every single character. With the HED, the score is 2, after deleting "T" and then calculating the distance between the empty string and the block "CC" [8].

So far, there have not been many applications of the HED in bioinformatic problems, but the sequence analysis of problems including tandem repeats has been proposed, mentioning the Staphylococcus aureus protein A gene. Many spa-types contain tandem repeats - sequences that are repeated numerous times, as shown in spa-type $t1260$: $14-44-12-17-17-17-17-23-18$. During the alignment process, every number in a repeat succession is treated as a single character, but using HED, the tandem repeat is treated as a continuous block. To measure the genetic distance between the different spa-types, we have applied the HED algorithm to a selection of n repeat successions using a pairwise alignment strategy. The output is a $n \times n$ distance matrix which can be used as an input for a clustering algorithm to create clusters based on the calculated HED.

The length of the spa-type repeat successions varies between 1 and 20 repeats, which has to be considered when calculating a score between two sequences of different lengths. To approach a more realistic score, we divided the computed HED score of two sequences by the sum of their length.

Figure 2 shows a selection of six spa-types and three pairwise calculated and modified HED scores. Similar repeat successions have a low score of around 0.25; the score increases when the sequences are less alike.

The runtime of the HED of two strings $s = s_1, ..., s_n$ and $t = t_1, ..., t_m$ is $\mathcal{O}(\max(n, m)^3)$.

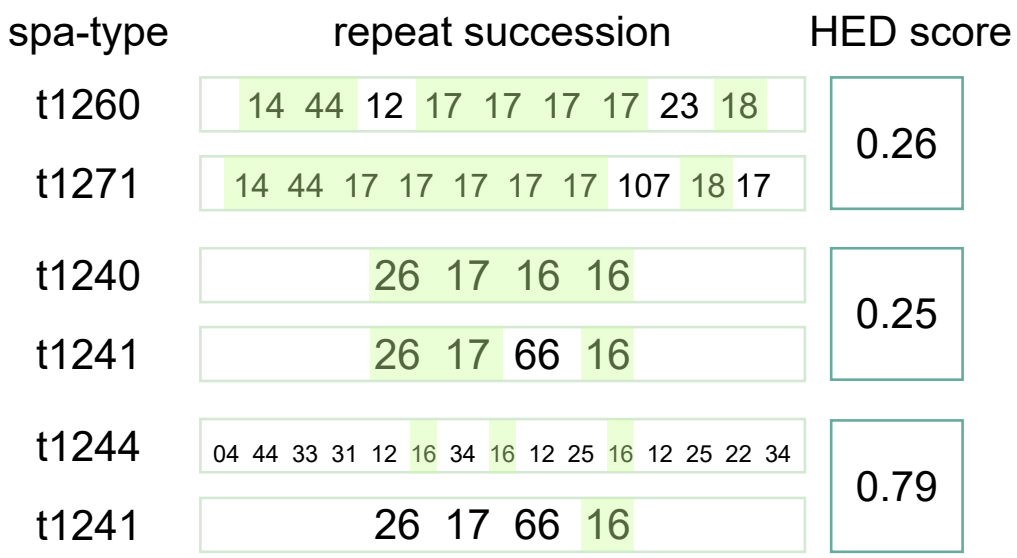

Figure 2: Six spa-types, their respective repeat successions, and three HED scores calculated. The score inside the right box is calculated by aligning the two repeat sequences shown to its left.

Since the number of calculations is increasing exponentially for n spa-types, the runtime increases accordingly. Hence we calculated the HED matrix for a selection of up to $n = 500$ spa-types. More sequences can be used, but the runtime of calculating a 1000×1000 matrix takes about 4 hours on a regular computer while growing exponentially. The results of selecting up to $n = 500$ spa-types are transferable on a higher n. Thus we can predict accurate results with an n lower than 1000.

3.3 GeoPandas

GeoPandas is a data science library that adds support for geospatial data. It can be installed with common package managers like Conda and used in Python scripts. We have used GeoPandas in a script to retrieve the geographical geometry of the previous selection of n spa-types.

Using GeoPandas and the list with the most occurring country entries for every type, we have fetched the coordinates of a point located in the country named. The information provided by the SpaServer was limited to the country name only, so we have decided to use the center point coordinates for each given country. A dataframe containing the spa-type ID, the repeat succession, the country name, and the center point coordinates was created.

The coordinates were used to calculate the geographical distance between the different spa-types. For this, we have used the haversine formula, which determines the great-circle distance between two points on a sphere [9]. We calculated the pairwise distance for our selection of n spa-types and got another $n \times n$ distance matrix, with the geographical distance in kilometers. Figure 3 shows the geographic distance for spa-types named in Figure 2.

Both distance matrices, geographical and genetic, can be used to cluster the spa-types into groups according to their distances to examine whether there are patterns in the genetic

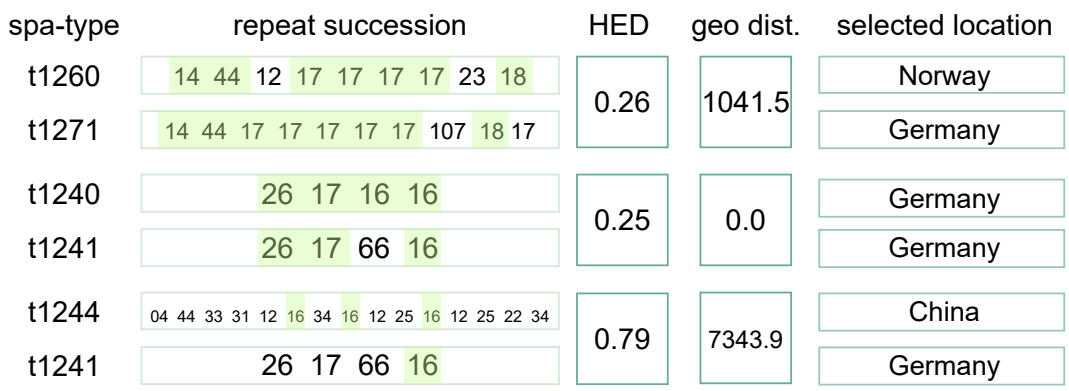

Figure 3: Additional information about the selected locations of the spa-types in Figure 2 and their geographic distance calculated.

sequence or the locations of grouped types.

3.4 Hierarchical clustering

Clustering algorithms work well for large amounts of information by organizing them into smaller clusters which then can be examined subsequently [10]. In this thesis we want to group our n selection of spa-types into clusters based on their genetic distance (HED), to determine a possible connection between their geographic occurrences. The spa-types can also be visualized based on the groups they have been clustered into.

Hierarchical cluster analysis (HCA) is a method of clustering that builds a hierarchy of clusters without specifying a fixed number of clusters before. The objects inside one cluster will be broadly similar, based on the selected data feature, which allows us to examine other features for possible correlations [11]. HCA is performed with a distance matrix or raw data. Since we have already calculated the HED matrix, it will be the input for HCA. At first, different measures of distance, like 'euclidian', 'hamming' or 'cosine' can be used to compute the distance matrix made up of observations. Each observation is treated as a separate cluster. We have $n \times n$ different HED observations given. Next, a linkage matrix is created: two observations with a certain distance are merged into a new cluster, this process is repeated iteratively until all clusters are merged. Multiple linkage options are available, for our data we have chosen the single linkage method, where the merging of two clusters is based on their minimum distance. Single linkage methods control nearest neighbor similarity [12], which will be especially useful for genetic clustering, to group genetically similar spa-types. The output of HCA is a dendrogram, which shows the relationship between the different clusters. In the script clustering.py we use the HCA functions given by SciPy, an open-source Python library for scientific and technical computing [3].

[3] *https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html*

3.4.1 Genetic distance clustering

After computing the genetic distance matrix for n spa-types, we condensed the square-form distance matrix into a vector-form distance vector and performed the single linkage function. Given a $n \times n$ distance matrix, a $n*(n-1)/2$ sized vector is returned, where:

$$v[\binom{n}{2} - \binom{n-i}{2} + (j-i-1)]$$

is the distance between points (spa-types) i and j. Through hierarchical clustering, this input is transformed into a $(n-1)$ by 4 matrix Z. The linkage method computes the distance $d(s,t)$ between two clusters s and t. The method 'single' uses the Nearest Point Algorithm and assigns:

$$d(u,v) = min(dist(u[i], v[j]))$$

for all points i in cluster u and j in cluster v.

For the dendrogram cluster coloring and further grouping, a cluster cutoff value was set to 0.25. Spa-types in a cluster with a maximum HED of 0.25 are color-coded accordingly in the resulting dendrogram. We experimented with different values here and decided to choose 0.25, the spa-types grouped with a bigger cut-off (0.5) were too dissimilar, and few clusters were created. The dendrograms created will be shown and discussed in section 4.

3.5 Correlation of genetic and geographic distance

Clustering the distances gave us an insight into where genetically related spa-types are located and how geographically close spa-types differ in their repeat successions. To further examine the possible correlation between the two distances, we have used the Pearson correlation coefficient to measure the strength of the relationship between the two data sets. Pearson's R measures the linear correlation between the genetic and geographic distance. The correlation coefficient can be applied to our sample, represented by R_{xy}, where $x \in X =$ genetic distance and $y \in Y =$ geographic distance. Pearson's R is calculated by:

$$R_{ij} = \frac{C_{ii}}{\sqrt{C_{ii}C_{jj}}}$$

C_{ij} is the covariance of x_i and x_j, C_{ii} is the variance of x_i.

X and Y are the $1-D$ arrays containing our observations, having the same shape and indexing. The result can take on a value in the $[-1,1]$ range. The maximum value $R = 1$ confirms a perfect linear relationship between x and y, indicating a strong correlation between the two datasets. Any value greater than 0 indicates a positive correlation between x and y,

and values below 0 indicate a negative correlation. Table 1 interprets the meaning of the R value[13]:

Interval correlation	Level of correlation
0.9 to 1.0 (-0.9 to -1.0)	Very high positive (negative) correlation
0.7 to 0.9 (-0.7 to -0.9)	High positive (negative) correlation
0.5 to 0.7 (-0.5 to -0.7)	Moderate positive (negative) correlation
0.3 to 0.5 (-0.3 to -0.5)	Low positive (negative) correlation
0.0 to 0.3 (.0 to -0.3)	negligible correlation

Table 1: Interval of the R-value and the corresponding level of correlation.

The correlation indicator is calculated for our n selection of spa-types and their observations.

3.6 Implementation

The implementation of the methods and supplementary resources can be found at:
https://gitlab.cs.uni-duesseldorf.de/albi/albi-students/ba-nina-romanow/

4 Results

The results were computed for a max amount of $n = 500$ spa-types, to keep the figures overseeable. In this section we will present the results for two different selections of $n = 200$ spa-types, results for $n = 500$ spa-types will be shown in the appendix.

4.1 Database

In our thesis, we used the spa-type and repeat records from September 2022, containing 832 repeats and 20686 different spa-types. Some spa-type entries had an empty location record and were not considered in our evaluation. As explained in section 3.1, we were left with 17871 different spa-types to evaluate.

4.2 Alignment

Table 2 and Table 3 show the HED matrix of five different spa-types and their repeat successions, giving a basic overview of how the HED differs between similar and entirely dissimilar sequences. Strongly similar types have a lower HED score, $t3749$ and $t6076$ have the common sub-sequence $17-25-17-25-16-28$ and also share the repeats 23 and 05. The HED score here is 0.3 with the sequences being similar, $t12161$ and $t6076$ only share the repeat 17 and have a higher score of 0.71.

spa-type	repeat succession
t5463	08-25-24-25
t9584	07-23-12-12-12-12
t3749	07-23-13-23-31-05-05-17-25-17-25-16-28
t12161	07-17-34-34
t6076	26-23-20-05-17-25-17-25-16-28

Table 2: Spa-types and their respective repeat successions.

	t5463	t9584	3749	t12161	t6076
t5463	0.0	0.6.	0.65	0.75	0.71
t9584	0.6	0.0	0.53	0.4	0.63
t3749	0.65	0.53	0.0	0.53	0.30
t12161	0.75	0.4	0.53	0.0	0.71
t6076	0.71	0.63	0.30	0.71	0.0

Table 3: HED calculated between the 5 spa-types in table 2.

For the following clustering, we have chosen a cut-off score of 0.25, so the sequences clustered would imply a strong genetic alignment. HED served as a solid base for clustering, as the scores appeared to be accurate when inspecting different spa-types, their similarity and their resulting HED score.

4.3 Clustering

Hierarchical clustering analysis was performed on different selections of $n = 200$ spa-types. The results can be replicated with the implementation provided by setting the random.seed value to 20 in the script selection.py. The dendrogram (1) in Figure 4 based on the HED matrix, shows 200 types clustered into different groups. 17 different clusters were formed, spa-types that did not fit into one of the clusters were marked as additional individual clusters. Several spa-types were clustered, meaning we are able to form groups with genetic similarity. The spa-types within a single cluster have similar repeat successions since the HED clustering cut-off was set relatively low to 0.25. For our research question, we want to compare the genetic similarity to the location of spa-types, so we created two pie charts for the two biggest formed clusters to inspect the distribution of locations. Cluster 27 and 17 were evaluated and are shown in Figure 5 and Figure 6.

Looking at the pie charts, we can see that there seems to be no dominating location. The spa-type locations are spread across the globe evenly. The results lead us to the hypothesis that there may not be a strong correlation between the genetic and geographic distance, meaning that genetically similar spa-types do not imply being geographically close. To inspect the lo-

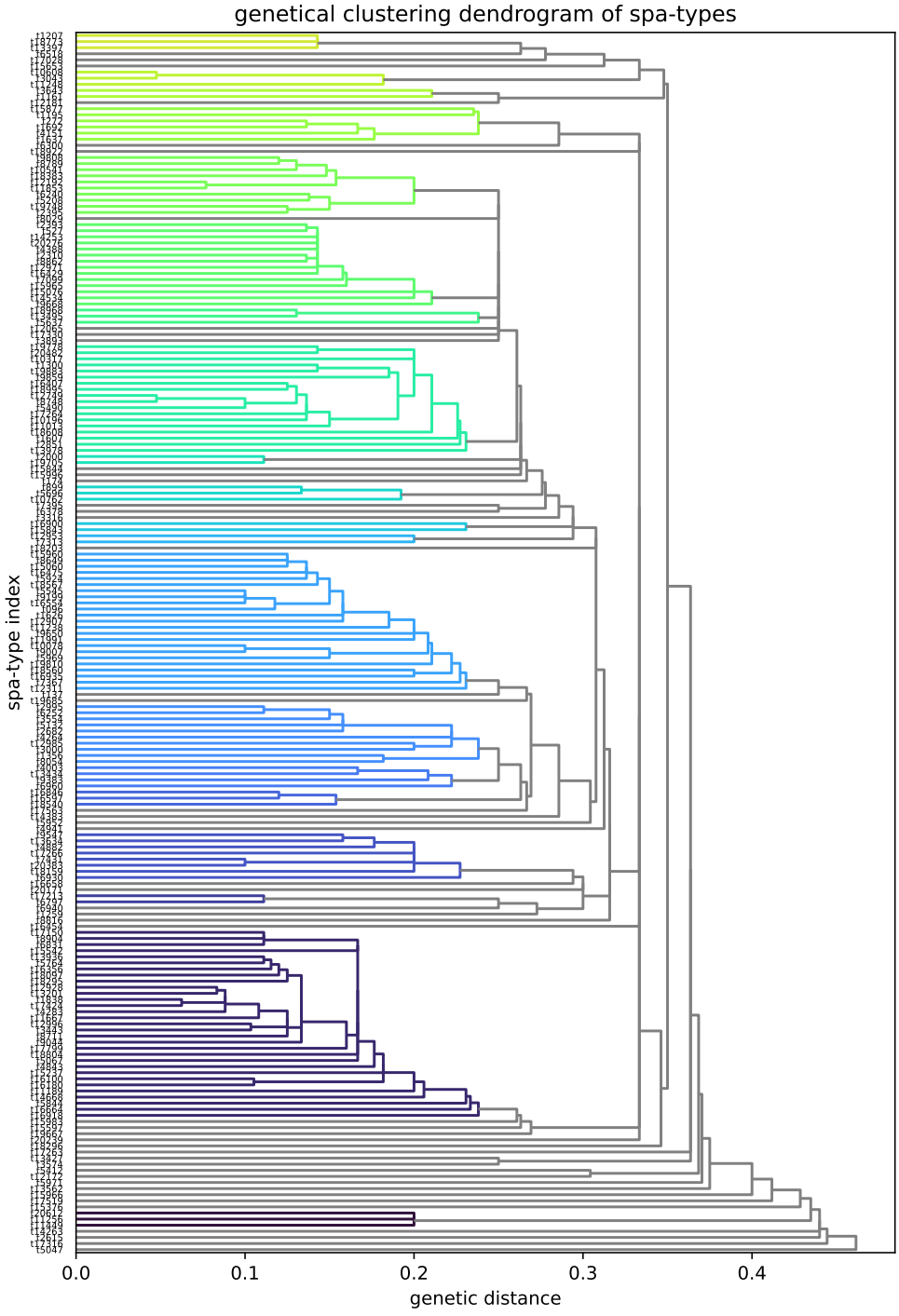

Figure 4: Dendrogram (1) based on the selection of $n = 200$ spa-types, with HED cut-off = 0.25

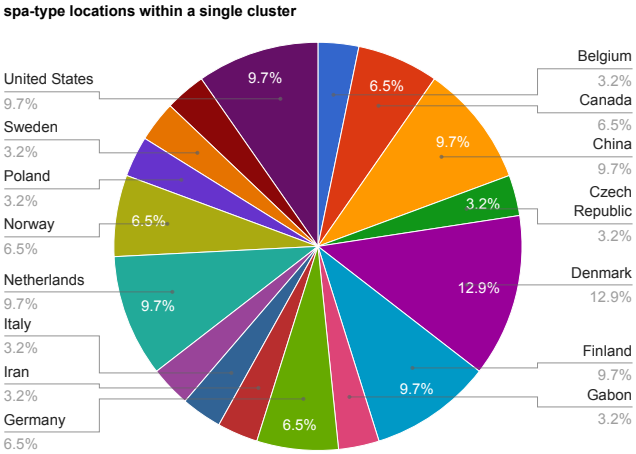

Figure 5: Distribution of countries in cluster number 27 in dendrogram (1)

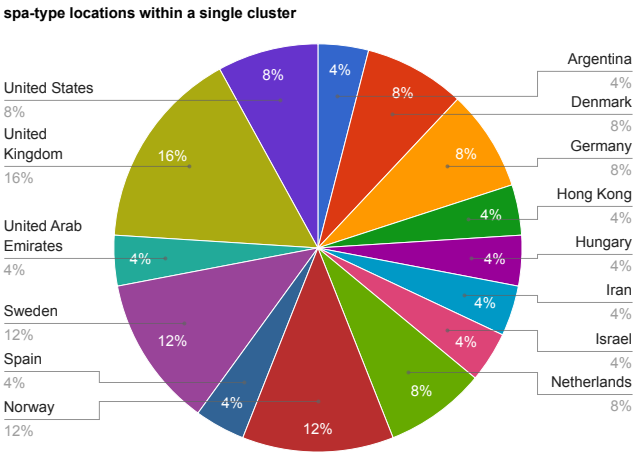

Figure 6: Distribution of countries in cluster number 17 in dendrogram (1)

cations of the spa-types even further, we created an interactive HTML map based on the HED and HAC results. The different maps can be accessed in our repository's map folder, and we invite you to explore the results[4]. The map extracts in Figure 7, corresponding to dendrogram (1), show the spa-types located inside a country and their respective cluster. Germany appears to be the location of many different spa-types from cluster 5.

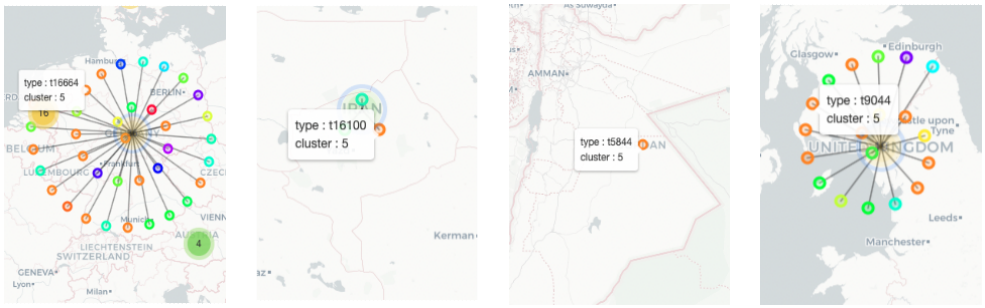

Figure 7: spa-types from dendrogram (1) located in Germany, Iran, Jordan and the UK; each color represents a cluster

Still, spa-types belonging to this cluster were also found in the United Kingdom, the United States, Iran, Jordan and Sweden. Some countries, like Iceland, only have spa-types from one cluster located, which could imply a strong correlation, but it should be noted that there are only 3 out of 200 spa-types sequenced in Iceland. Countries with more spa-types sequenced represent multiple cluster groups.

Since the clustering with a maximum HED score of 0.25 did not show a correlation between genetic and geographical distance, we tried to compute a dendrogram with a higher and lower HED cut-off. The following figures show dendrogram results by clustering $n = 200$ types with the cut-off set to 0.15 and 0.3.

In dendrogram (2), shown in Figure 8, the clusters are smaller and the locations inside the clusters are primarily in Europe.

Cluster 43 of the dendrogram (2) consists of spa-types found in Spain, Sweden, Norway, the UK, Denmark, Germany, and the Netherlands. This could imply a correlation, but there are also clusters given with spa-types located in Ireland, Denmark, and New Zealand. The results for the computed dendrograms are also saved in "genclusterresult.txt" files in our repository. Setting the HED cut-off to 0.3 resulted in bigger cluster sizes and more geographical diversity inside of them, which is shown in Figure 10, representing cluster 19 and Figure 9 showing the dendrogram (3).

The HED cutoff is essential in calculating the clusters and which spa-type locations are represented inside. The evaluation of the different clustering results shows no clear correlation. But as they are strongly dependent on the size of the generated clusters, we will apply another

[4] *https://gitlab.cs.uni-duesseldorf.de/albi/albi-students/ba-nina-romanow/*

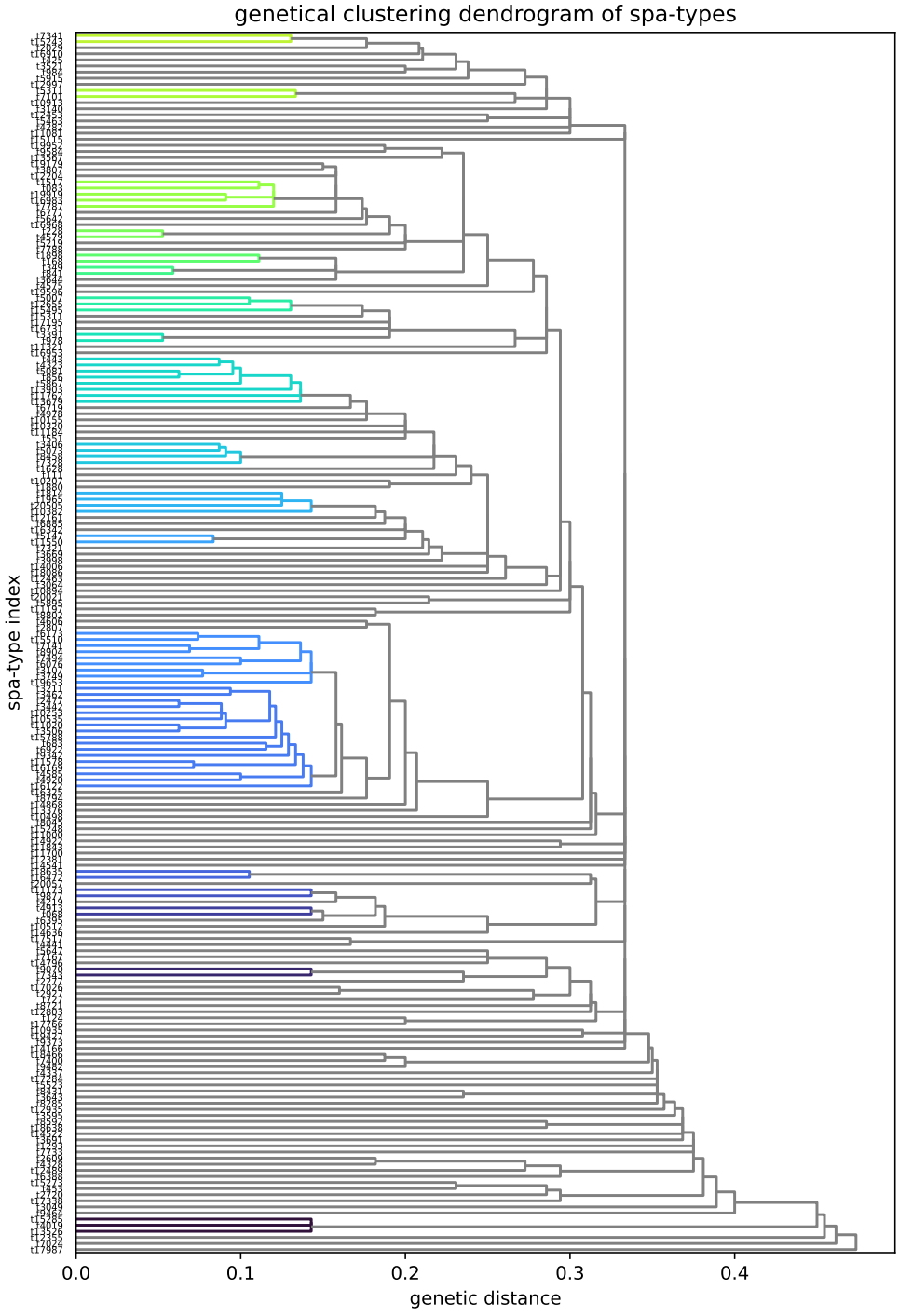

Figure 8: Dendrogram (2) based on the selection of $n = 200$ spa-types, with HED cut-off = 0.15

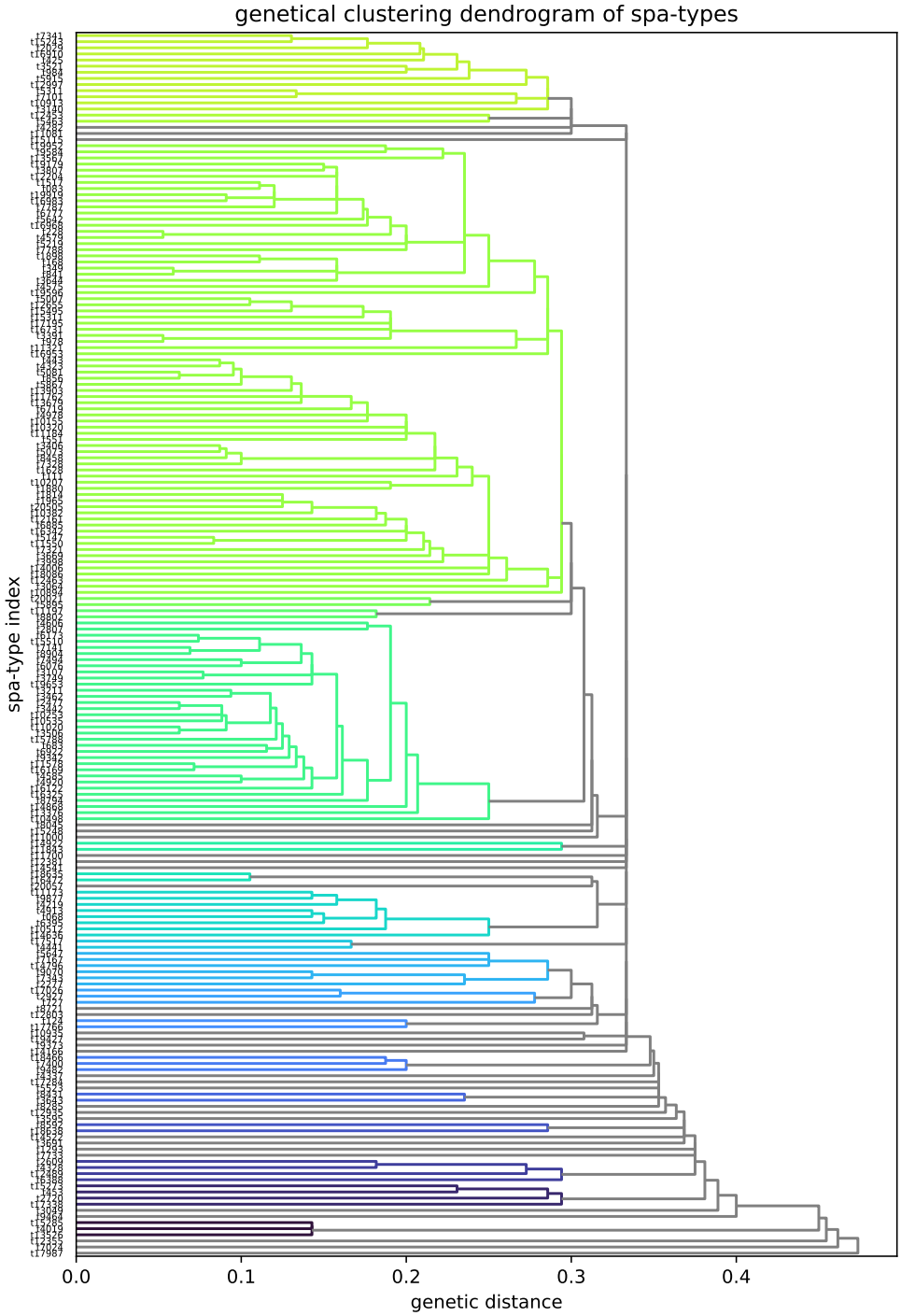

Figure 9: Dendrogram (3) based on the selection of $n = 200$ spa-types, with HED cut-off = 0.3

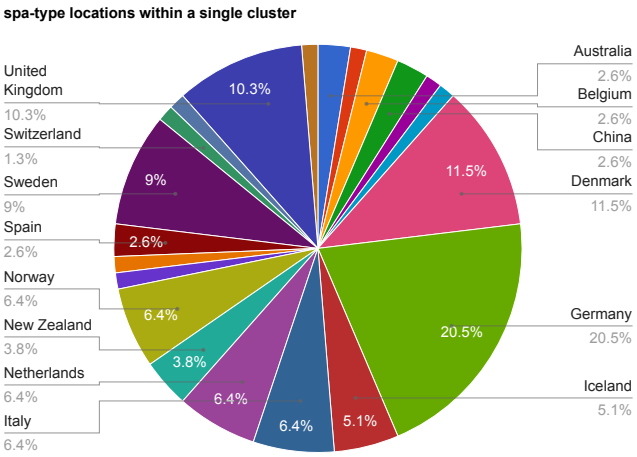

Figure 10: Distribution of countries in cluster number 19 in dendrogram (3)

method in the next section to examine the correlation without an additional factor such as the cluster cut-off.

4.4 Correlation

For the selections of $n = 200$ and $n = 500$, we have created scatterplots that display the relationship between the variables x and y, x being the genetic distance = HED and y the geographic distance between the spa-types, which we calculated before. Each point in the scatterplot represents an observation. Since we have n spa-types, the distance matrices are $n \times n$, and the condensed matrices have the form $n \cdot (n-1)/2$, we will have $n \cdot (n-1)/2$ pairs of observations available. Each genetic distance between two spa-types is paired with the geographic distance of their location.

A linear relationship between two variables can be identified by looking at the pattern in the scatterplot. Both scatterplots computed do not show a specific direction nor shape. Many points are displayed with a geographic distance of 0, since 70.46 of all spa-types analyzed were sequenced in Germany. Thus the probability of having many spa-types originating in Germany in any n selection is high. When calculating a distance matrix, those spa-types would have different genetic distances, but the geographic distance would be 0. The results of the scatterplots suggest that there is no strong relationship between the distances. To support this suggestion, we computed the correlation coefficient described in section 3.5 for both selections. The R-value for 200 spa-types (Figure 11) is -0.003 and 0.018 for 500 spa-types (Figure 12). Both values can be classified as "negligible correlation" (see Table 1 in section 3.5).

Combining the clustering results and the correlation analysis, we can verify our assumption that there is no correlation between the genetic and geographic distance of spa-types.

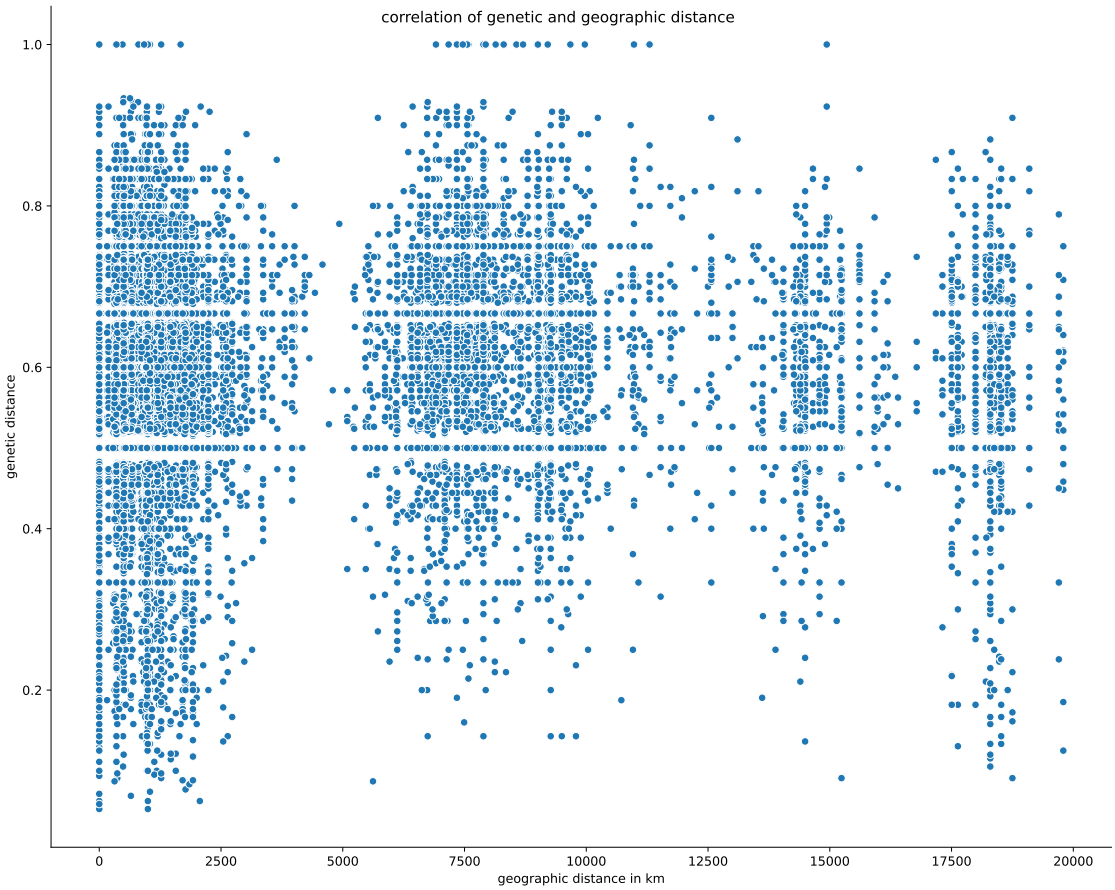

Figure 11: Correlation scatterplot (1) based on the selection of $n = 200$ spa-types

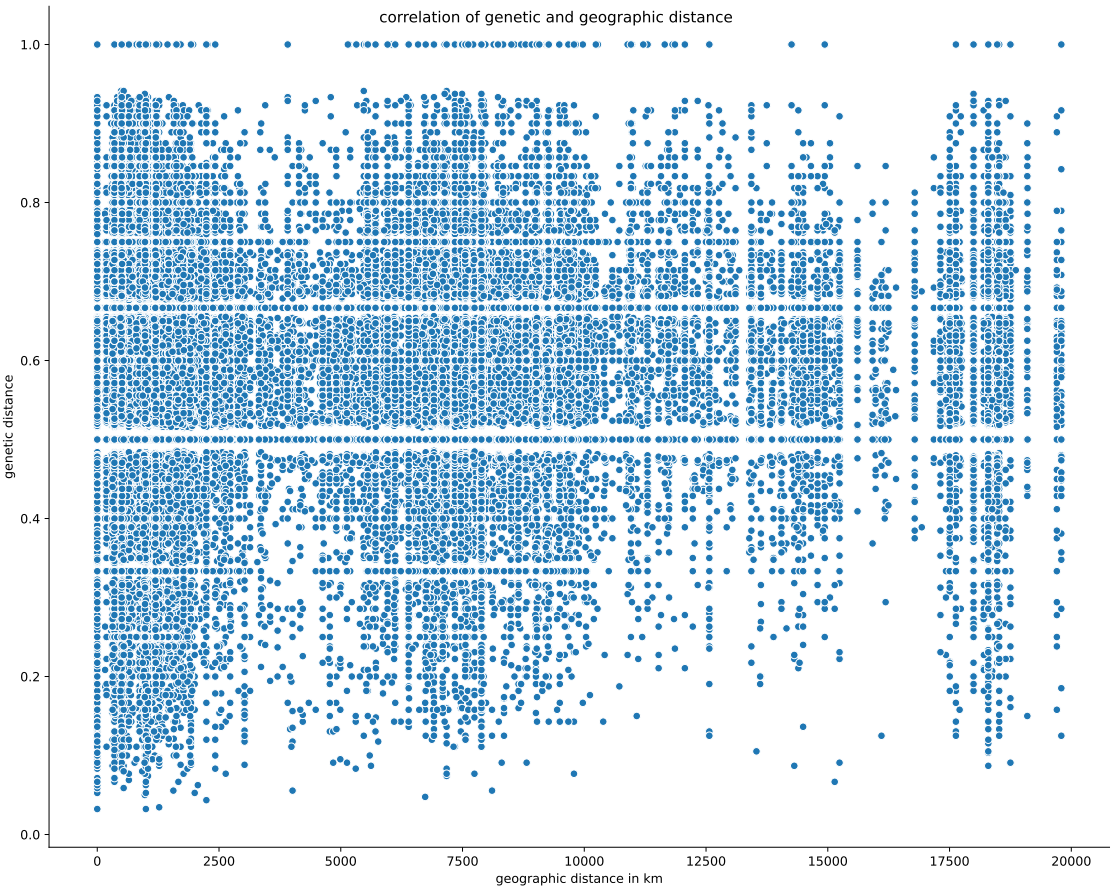

Figure 12: Correlation scatterplot (2) based on the selection of $n = 500$ spa-types

Patterns are visible inside the scatterplots, which can be explained by looking at the geographic distribution of all spa-types evaluated (Figure 13). There are 8 dominant countries, and other countries have a small number of spa-types sequenced. The geographic distances will be similar for any selection since they are mostly calculated between those 8 locations.

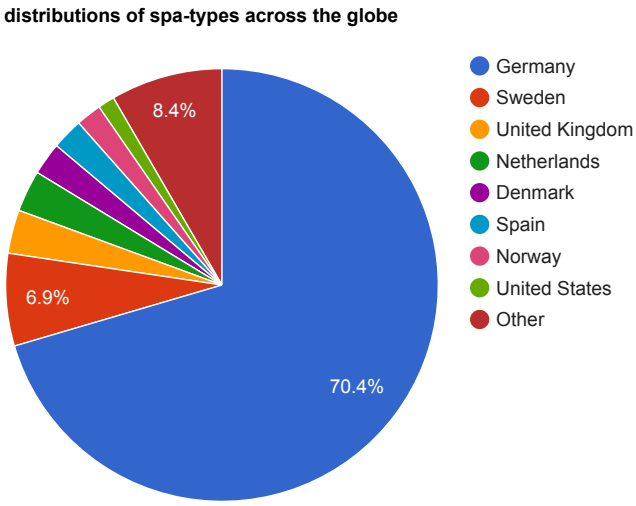

Figure 13: geographic distribution of all spa-types evaluated, countries having less than 1.25 percent of all spa-types located, were summarized into "Other"

5 Discussion

In this section, we will discuss the limitations of our methods and present further implementation ideas to examine the correlation more precisely for future work.

5.1 Further implementation

Homo Edit Distance The usage of the HED has proven to give an accurate score for genetic similarity. However, it is not entirely clear in which range the score indicates genetic similarities. Further research is also needed in comparing sequences of dissimilar length, for example, a spa-type with the repeat succession length of 2 and one with the length of 24. We approached this problem by dividing the scores by the sum of those lengths, but it is unclear whether this method accurately represents genetic relation. Comparing the results of HED with another alignment technique could also be considered.

Location data Some spa-types show thousands of sequencing records in different locations, others only had one sequencing location given. We chose only to consider the location recorded most for each type. This raises the question of whether the location of a spa-type like $t003$ that was recorded 20325 times can be weighed the same as $t400$ where the most prevalent location was given 7 times only. The locations' frequency should be considered for an accurate representation of spa-type spreading. We could evaluate every record given for a spa-type or choose the first record based on the Isolation year. Also, the locations provided are imprecise, consisting of country names only, but for larger countries like Russia, the distance between two cities can be significant and may influence the results. We obtained the coordinates for every country named, but since no cities were specified we decided to work with the geographic center of each country. All geographic distances calculated are based on those centers, more precise data could give us exact distances between the locations. Another idea would be to combine the locations given with flight data records to evaluate how much exchange is happening between them. Excluding records from Germany could be considered, they make up 70% of all location records, so the results are strongly biased.

We analyzed a relatively small amount of spa-types with our method, though the spa-types chosen have varied. For $n = 1000$ spa-types, the algorithms would take about 4 hours to calculate results, and the dendrogram would be challenging to read. Many clusters would have been created, making it more complicated to evaluate the different groups.

5.2 Conclusions

This thesis presented and implemented a method to analyze the correlation between the genetic and geographic distance of spa-types. The clustering of aligned spa-types leads us to conclude that there is no significant correlation between the two variables. To support this claim, the correlation coefficient was calculated. Still, some adjustments can be made for further examination of the initial question. Multiple factors can influence our results, and the possible improvements for future work were discussed in the previous section. Following factors outside of the method could influence the given outcome: Staphylococcus aureus has been around for a long time and 30% of the population carries the bacteria inside their bodies. The infection can be passed on asymptomatically. Traveling across countries and continents has never been easier, which might influence the fast spreading of different spa-types. The healthcare hygiene standards differ globally, and supply and shortages of medicinal products may amplify the spreading of spa-types. All countries considered during the evaluation are well visited by travelers and not wholly off-grid. The rapid spreading of spa-types makes it harder to comprehend the infection chains, and our research could not deliver significant insights. The implementation of the presented methods is freely available and briefly documented, the process can be retraced and modified. We look forward to seeing further work on the topic to get more insights into the spreading of spa-types globally.

References

[1] A. S. Lee, H. de Lencastre, and J. Garau et al. “Methicillin-resistant Staphylococcus aureus”. In: *Nature Reviews Disease Primers* (2018).

[2] H. M. E. Frénay, A. E. Bunschoten, and L. M. Schouls et al. “Molecular typing of methicillin-resistant Staphylococcus aureus on the basis of protein A gene polymorphism”. In: *European Journal of Clinical Microbiology and Infectious Diseases volume* (1996).

[3] W. Ruppitsch et al. “Classifying spa Types in Complexes Improves Interpretation of Typing Results for Methicillin-Resistant Staphylococcus aureus”. In: (2006). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1489472/.

[4] E. Danai, S. Corti, and C. Spirig et al. S. “Staphylococcus aureus Population Structure and Genomic Profiles in Asymptomatic Carriers in Switzerland”. In: *Frontiers in Microbiology* 11 (2020).

[5] S. A. Chandran et al. “COVID-19-associated Staphylococcus aureus cavitating pneumonia”. In: *BMJ Case Reports CP* 14.6 (2021).

[6] M. Hallin, A. W. Friedrich, and M. J Struelens. *spa Typing for Epidemiological Surveillance of Staphylococcus aureus*. Ed. by Dominique A. Caugant. Humana Press, 2009, pp. 189–202. ISBN: 978-1-60327-999-4. DOI: 10.1007/978-1-60327-999-4_15. URL: https://doi.org/10.1007/978-1-60327-999-4_15.

[7] L. Koreen, S. V. Ramaswamy, and E. A. Graviss et al. “spaTyping Method for Discriminating among Staphylococcus aureus Isolates: Implications for Use of a Single Marker To Detect Genetic Micro- and Macrovariation”. In: *Journal of Clinical Microbiology* 42.2 (2004), pp. 792–799. DOI: 10.1128/JCM.42.2.792-799.2004. eprint: https://journals.asm.org/doi/pdf/10.1128/JCM.42.2.792-799.2004. URL: https://journals.asm.org/doi/abs/10.1128/JCM.42.2.792-799.2004.

[8] M. Brand et al. “The Homo-Edit Distance Problem”. In: *bioRxiv* (2020).

[9] Rosetta Code. *Haversine formula — Rosetta Code*. Online; accessed 20-September-2022. 2022. URL: https://rosettacode.org/w/index.php?title=Haversine_formula&oldid=327674.

[10] Y. Zhao and G. Karypis. “Evaluation of Hierarchical Clustering Algorithms for Document Datasets”. In: *Proceedings of the Eleventh International Conference on Information and Knowledge Management*. CIKM ’02. McLean, Virginia, USA: Association for Computing Machinery, 2002, pp. 515–524. ISBN: 1581134924. DOI: 10.1145/584792.584877. URL: https://doi.org/10.1145/584792.584877.

[11] T. Bock. “What is Hierarchical Clustering?” In: *DisplayR* (2022). URL: https://www.displayr.com/what-is-hierarchical-clustering/.

[12] Vijaya, Shweta Sharma, and Neha Batra. “Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering”. In: (2019), pp. 568–573. DOI: 10.1109/COMITCon.2019.8862232.

[13] D.E. Hinkle, W. Wiersma, and S.G. Jurs. *Applied Statistics for the Behavioral Sciences*. Boston: Houghton Mifflin, 2003.

A Additional Figures and Tables

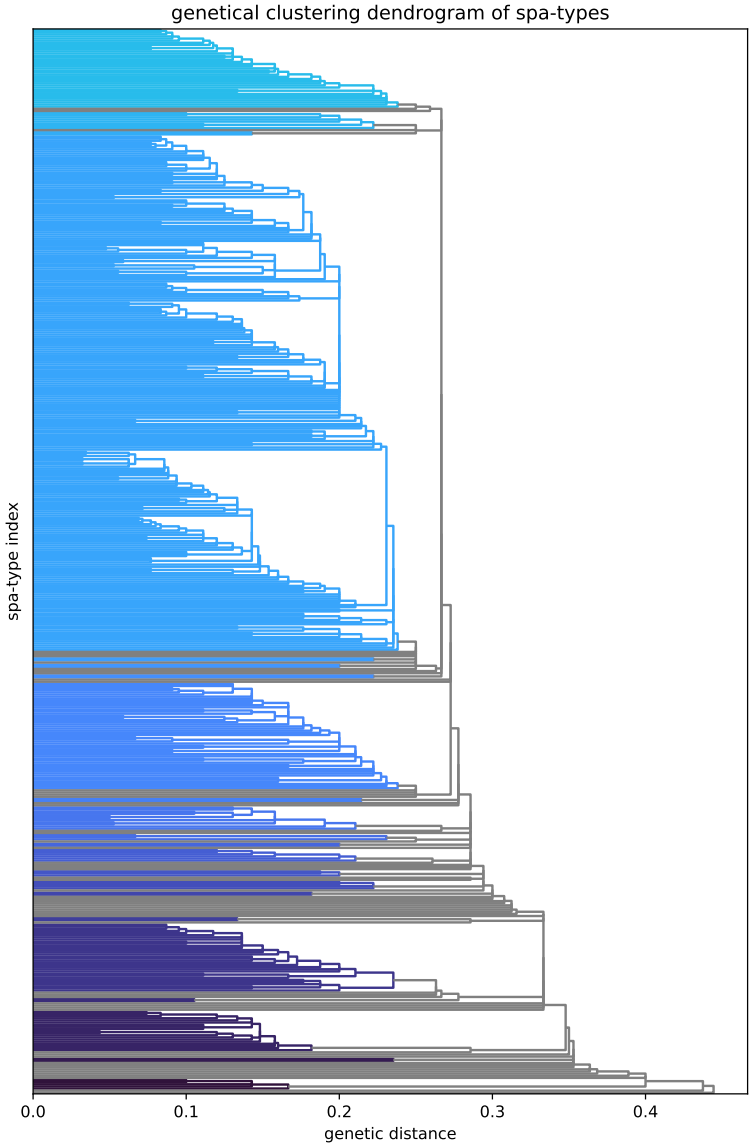

Figure 14: Dendrogram based on the selection of $n = 500$ spa-types, with HED cut-off $= 0.25$