Bioinformatics

Final Project

Claudio Rota 816050
Simone Paolo Mottadelli 820786
Stefano Zuccarella 816482

Contents

1 Introduction 1
2 Description of the sequences 2
3 Description of the tools 3
4 Structure of the output format 5
5 Analysis of the alignment results 7
5.1 Results at nucleotide base level 7
5.2 Results at gene level 9
5.3 Results at amino acid level 10
6 Analysis of the phylogenetic trees 12
7 Perfect phylogeny 15
8 Conclusions 17

1 Introduction

Coronavirus disease 19, mostly known as *COVID-19*, is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (*SARS-CoV-2*), which was first identified in December 2019 in Wuhan, the capital city of Hubei Province in China, and has since then spread globally, resulting in a global pandemic. At the time of writing, the total number of confirmed cases is about 5 million people and 300,000 people died worldwide [1]. The disease is particularly fatal for the elderly and for people with past illnesses, even though many people who were healthy also died. The virus also caused major socioeconomic crises in the world.
In particular, the unknown nature of the virus has involved many researchers to study it with the aim of finding cures and answering to some research questions, such as "how did the virus move between European countries?", "how did the virus mutate its genomic structure from its first discovery?", and so on. In this regard, the virus has been sequenced many times and the results have been made available on the main websites that treat biotechnological information, such as *NCBI* or *GISAID*.
This project sets the goal of selecting some interesting genomic sequences from the aforementioned websites, aligning them with different multiple sequence alignment tools and analyzing the obtained results. In particular, since Italy has been one of the most affected country in the world and many Icelandic individuals spent their holidays in ski resorts in the Italian and Austrian Alps before resulting positive to *COVID-19* tests, this work aims to verify the hypothesis that these Icelandic individuals contracted the virus in Italy or in Austria or somewhere else. In practice, what is expected is that there are similar mutations between Icelandic sequences and the Italian and Austrian ones.
This report is organized in the following sections: Section 2 describes the sequences selected for this project and motivates the reason why they have been chosen; Section 3 provides a brief description of the multiple alignment tools chosen to conduct this study; Section 4 outlines the structure of the output report that has been designed to highlight the differences

among the sequences; Section 5 presents the results of the alignment at nucleotide base level, at gene level and at amino acid level; Section 6 shows some considerations on the phylogenetic trees derived from the alignment tools; Section 7 shows the result of the perfect phylogeny induced from the sequences and compares it with the phylogenetic trees obtained from the alignment tools and, finally, Section 8 concludes this report by providing a brief summary of the work presented in the previous sections.

2 Description of the sequences

In order to accomplish the goal of this project, different Icelandic, Austrian and Italian sequences have been selected and consequently aligned. The sequences have been obtained from the *GISAID* website because it provides useful information about the individuals, including their recent journeys. In particular, the following 19 sequences have been chosen:

- 10 Icelandic sequences, 5 of which come from individuals who recently traveled to Italy and the remaining ones come from individuals who recently traveled to Austria;
- 4 Italian sequences;
- 4 Austrian sequences;
- The reference Chinese sequence.

Further details on the gathered sequences are available in Table 1, which shows the Italian sequences in green, the Austrian sequences in red and, finally, Icelandic sequences in blue and in orange whether they come from individuals who traveled to Italy and Austria, respectively. In addition, the Italian and the Austrian sequences have a collection date that precedes the collection date of the Icelandic ones. This choice is related to the fact that an Icelandic subject got infected potentially in Italy or in Austria and, thus, it can be assumed that it contracted the virus in a period prior to the time

the swab was made. Therefore, the sequence of the considered Icelandic individual will be more similar to Italian or Austrian sequences collected in previous periods, because in the meantime they could have undergone mutations.
Another consideration concerns the variance of the age of the individuals. In fact, the sequences have been chosen with the aim of keeping the variance of the age of the individuals as low as possible, as comparing sequences coming from subjects of different ages could introduce some bias in the analysis.
The sequences have been obtained using different Next-Generation-Sequencing (NGS) technologies, which produce short reads composed of about 100-400 bp, achieving an accuracy value higher than 99% and high coverage values.
Finally, the chosen sequences are complete, that is, they are composed of more than 29.000 nucleotide bases, and they have a high coverage, i.e., the number of symbols equal to "N" (see *IUPAC code*) is less than 1%, the amino acids mutations are less than 0.05% and no insertion/deletion operations are present.

3 Description of the tools

The sequences described in Section 2 have been aligned using three multiple alignment tools:

- *Clustal Omega* [2];
- *Muscle* [3];
- *Kalign* [4].

It follows a brief description of the functioning of the tools.
Clustal Omega performs the alignment in five steps. In the first one, the

ID	Gender	Age	Nationality	Traveled to	Collection date	Technology	Length
NC_045512 (REF)	-	-	China	-	- -/12/19	-	29903
EPI_ISL_424418	Male	42	Iceland	Austria	19/03/20	Illumina Miseq	29903
EPI_ISL_424413	Female	32	Iceland	Austria	19/03/20	Illumina Miseq	29897
EPI_ISL_424405	Female	47	Iceland	Austria	19/03/20	Illumina Miseq	29903
EPI_ISL_417875	Male	39	Iceland	Austria	10/03/20	Illumina Miseq	29903
EPI_ISL_417874	Female	40	Iceland	Austria	10/03/20	Illumina Miseq	29903
EPI_ISL_417829	Male	51	Iceland	Italy	16/03/20	Illumina Miseq	29903
EPI_ISL_417700	Female	45	Iceland	Italy	15/03/20	Illumina Miseq	29903
EPI_ISL_424451	Male	47	Iceland	Italy	20/03/20	Illumina Miseq	29903
EPI_ISL_417861	Female	21	Iceland	Italy	10/03/20	Illumina Miseq	29903
EPI_ISL_417868	Female	48	Iceland	Italy	09/03/20	Illumina Miseq	29903
EPI_ISL_419655	Male	73	Austria	-	26/02/20	Illumina NovaSeq	29859
EPI_ISL_419654	Female	27	Austria	-	03/03/20	Illumina NovaSeq	29860
EPI_ISL_419658	Male	59	Austria	-	06/03/20	Illumina NovaSeq	29866
EPI_ISL_419656	Male	41	Austria	-	26/02/20	Illumina NovaSeq	29859
EPI_ISL_417921	Male	32	Italy	-	01/03/20	Ion Torrent	29760
EPI_ISL_417922	Male	41	Italy	-	28/02/20	Ion Torrent	29760
EPI_ISL_417923	Male	53	Italy	-	04/03/20	Ion Torrent	29791
EPI_ISL_424344	Male	56	Italy	-	04/03/20	Ion Torrent	29834

Table 1: Sequences used for the study

sequences are pair-wise aligned using a heuristic method that does not guarantee to find the optimal solution, but it is still faster than the classical dynamic programming approaches. Then, the *mBed* algorithm is used to compute the distances between all the aligned sequence pairs and the *K-Means* algorithm is used to cluster them. After, a guide tree is built using either the *UPGMA* algorithm or the *Neighbor Joining* algorithm. Finally, the obtained guide tree is used to perform the multiple alignment of the sequences, also exploiting Hidden Markov Models.

Muscle works in three steps: the "Draft progressive" step, the "Improvement progressive" step and the "Refining" step. In the first step, *Muscle* produces a first draft of the multiple alignment, favoring speed over accuracy, and a guide tree is built using the *UPGMA* algorithm. In the second step, it uses the *Kimura* distance to re-estimate the guide tree and to produce a new draft, which should be more accurate than the previous one. Finally, in the last step, it applies an iterative procedure with the aim of maximizing the multiple alignment *SP score*. This procedure terminates ei-

ther when the method achieves convergence, that is, when it is not possible to increase the *SP score* anymore, or when the maximum number of iterations is reached.
In conclusion, *Kalign* exploits a strategy very similar to the standard progressive alignment. It first computes the pair-wise distances between the sequences and generates a guide tree, which is usually built with either the *UPGMA* algorithm or the *Neighbor Joining* algorithm. The *Wu-Mamber* algorithm is used to compute the distances between the pair-wise alignments. Lastly, the resulting guide tree is used to lead the multiple sequence alignment.
Since the three algorithms use different strategies to align the sequences, it is expected to obtain different alignment results.

4 Structure of the output format

The outputs of the different multiple alignment tools have been compared and summarized in a file with a JSON format, whose structure is exemplified in Listing 1, with the aim of highlighting all the differences found among the different sequences and the reference one. In particular, the example shows that a registered mutation is composed of the following fields:

- **id**: the identifying code of the mutation;
- **begin**: the initial position of the mutation;
- **end**: the final position of the mutation;
- **type**: the mutation type, which can be either “Edit”, “Delete” or “Insert”;
- **refSubstring**: the symbols in the reference sequence;
- **seqSubstring**: the symbols in the other sequences;

- **sequences**: a list containing the sequences that present the same mutation;
- **tools**: a list containing the tools that detected the mutation;
- **gene**: the gene affected by the mutation, characterized by its name, its begin position and its end position.

The entire output file is available at this *link*.
It is important to mention that the alphabet of the analyzed sequences was composed of the entire *IUPAC code* and, for this reason, all its symbols have been taken into account to determine the variations. For instance, as letter R corresponds to either letter A or G, if the reference sequence presents the letter A in a specific position and another sequence presents the letter R in the same position, then no mutation is registered.

```
[
  ...
  {
    'id': 9
    'begin': 1440,
    'end': 1440,
    'type': 'Edit',
    'refSubstring': 'G',
    'seqSubstring': 'A',
    'sequences': ['Austria_EPI_ISL_419655', 'Austria_EPI_ISL_419654'],
    'tools': ['Clustal', 'Muscle', 'Kalign'],
    'gene': {
        'gene_name': 'ORF1ab',
        'begin': 266,
        'end': 21555
     }
  },
  ...
]
```

Listing 1: Example of a difference extracted from the complete output file

5 Analysis of the alignment results

This section describes the mutations between the reference sequences and the other sequences first at nucleotide base level, then at gene level, considering that the reference sequence contains 11 genes (as also reported *here*) and, finally, at amino acid level.

5.1 Results at nucleotide base level

At nucleotide base level, the main results are shown in Table 2, which does not include the "isolated cases" and the mutations found at the beginning and at the end of the sequences. The "isolated cases" are those cases in which the mutation involves only one sequence and they have not been considered so relevant for this work, since they might be caused by sequencing errors or they could be sporadic mutations. The mutations found at the beginning and at the end of the sequences have also been excluded because they are not significant.
It is worthwhile to mention that all the three tools detected the same main variations, thus confirming the real existence of mutations among the reference sequence and the other ones. Note that these main mutations are only edit operations and neither insertion nor deletion operations have been found.
The most relevant mutations are identified by id 5, 13, 18 and 23, which exactly involve 15 sequences out of 18 and the sequences are always the same. These mutations interest all the considered countries, that is, all the Icelandic and Italian sequences and just 1 out of 4 Austrian sequences. This fact shows that the Icelandic sequences may be more related to the Italian sequences than the Austrian ones. Moreover, since the Italian sequences have been collected before the Icelandic ones, this result may be interpreted as a move of the virus from Italy to Iceland and not vice versa. An additional consideration is about the affected genes, namely *ORF1ab* and *S*, where the same type of mutation occurred, i.e., from C to T in *ORF1ab* gene and from A to G in *S* gene.

Id	Begin	End	Reference substring	Sequence substring	Sequences	Gene
5	241	241	C	T	EPI_ISL_424418 EPI_ISL_424451, EPI_ISL_417874, EPI_ISL_424405, EPI_ISL_417875 EPI_ISL_417861, EPI_ISL_419656, EPI_ISL_417868, EPI_ISL_417829, EPI_ISL_417922 EPI_ISL_424413, EPI_ISL_417700, EPI_ISL_424344, EPI_ISL_417921, EPI_ISL_417923	-
8	1059	1059	C	T	EPI_ISL_424418, EPI_ISL_417874, EPI_ISL_417875, EPI_ISL_424413, EPI_ISL_417700	ORF1ab
9	1440	1440	G	A	EPI_ISL_419655, EPI_ISL_419654	ORF1ab
12	2891	2891	G	A	EPI_ISL_419655, EPI_ISL_419654	ORF1ab
13	3037	3037	C	T	EPI_ISL_424418 EPI_ISL_424451, EPI_ISL_417874, EPI_ISL_424405, EPI_ISL_417875 EPI_ISL_417861, EPI_ISL_419656, EPI_ISL_417868, EPI_ISL_417829, EPI_ISL_417922 EPI_ISL_424413, EPI_ISL_417700, EPI_ISL_424344, EPI_ISL_417921, EPI_ISL_417923	ORF1ab
18	14408	14408	C	T	EPI_ISL_424418 EPI_ISL_424451, EPI_ISL_417874, EPI_ISL_424405, EPI_ISL_417875 EPI_ISL_417861, EPI_ISL_419656, EPI_ISL_417868, EPI_ISL_417829, EPI_ISL_417922 EPI_ISL_424413, EPI_ISL_417700, EPI_ISL_424344, EPI_ISL_417921, EPI_ISL_417923	ORF1ab
23	23403	23403	A	G	EPI_ISL_424418 EPI_ISL_424451, EPI_ISL_417874, EPI_ISL_424405, EPI_ISL_417875 EPI_ISL_417861, EPI_ISL_419656, EPI_ISL_417868, EPI_ISL_417829, EPI_ISL_417922 EPI_ISL_424413, EPI_ISL_417700, EPI_ISL_424344, EPI_ISL_417921, EPI_ISL_417923	S
26	25563	25563	G	T	EPI_ISL_424418, EPI_ISL_417874, EPI_ISL_417875, EPI_ISL_424413, EPI_ISL_417700	ORF3a
29	27046	27046	C	T	EPI_ISL_417861, EPI_ISL_417868, EPI_ISL_417829	M
30	28881	28883	GGG	AAC	EPI_ISL_417861, EPI_ISL_419656, EPI_ISL_417868, EPI_ISL_417829, EPI_ISL_417922	N

Table 2: Common mutations between the reference sequence and the other sequences

The mutations with id 8, 26 and 29 involve only Icelandic sequences. This leads to hypothesize that such mutations are restricted only to the Icelandic region. It is also possible to notice that the nucleotide bases are always converted to T.
The mutations with id 9 and 12 involve the same two Austrian sequences, which have no relations with the other sequences, not even with the Austrian ones. They both happen in the same gene, namely *ORF1ab* gene, and the type of the variation is also the same, because in both the cases G is transformed into A.
Finally, the mutation with id 30 involves three Icelandic sequences of individuals coming from Italy, one Italian sequence and one Austrian sequence. This variation highlights even more the existing connection between Italy

and Iceland. It is worth noticing that the Austrian sequence with id EPI_ISL _419656 also appears in the most relevant mutations, that is, the ones with id 5, 13, 18 and 23, and this can lead to think that this Austrian sequence is strongly related to the Italian ones. In addition, this mutation is the only one characterized by 3 nucleotide base changes.

5.2 Results at gene level

At gene level, Table 3 reports the number of mutations occurred for each gene, considering both the most relevant mutations and the "isolated cases". The gene with the highest number of mutations is *ORF1ab*, in fact 16 out of 26 (61.5%) mutations manifested in this gene. This could be explained by the fact that it is the longest gene in the virus, since it starts from position 266 and it ends in position 21555 and, thus, it occupies about 71.2% of the entire virus length.

Name	**Begin**	**End**	**Count**
ORF1ab	266	21555	16
S	21563	25384	4
ORF3a	25393	26220	2
E	26245	26472	0
M	26523	27191	2
ORF6	27202	27387	0
ORF7a	27394	27759	0
ORF7b	27756	27887	0
ORF8	27894	28259	0
N	28274	29533	2
ORF10	29558	29674	0

Table 3: Number of modifications for each gene of the reference sequence

5.3 Results at amino acid level

At amino acid level, the results of this analysis are reported in Table 4.

Id	Sequence Ids	Gene Id	Begin	End	Reference Codon	Sequence Codon	Begin position from CDS	Reference amino acid	Sequence amino acid
1	EPI_ISL_424451	ORF1ab	266	21555	GGT	GAT	379	G	D
2	EPI_ISL_417861	ORF1ab	266	21555	CCT	TCT	577	P	S
3	EPI_ISL_424418, EPI_ISL_417874, EPI_ISL_417875, EPI_ISL_424413, EPI_ISL_417700	ORF1ab	266	21555	ACC	ATC	793	T	I
4	EPI_ISL_419655, EPI_ISL_419654	ORF1ab	266	21555	GGC	GAC	1174	G	D
5	EPI_ISL_424344	ORF1ab	266	21555	GCC	GCT	1795	A	A
6	EPI_ISL_424413	ORF1ab	266	21555	GTA	GTG	1915	V	V
7	EPI_ISL_419655, EPI_ISL_419654	ORF1ab	266	21555	GCA	ACA	2626	A	T
8	EPI_ISL_424418, EPI_ISL_424451, EPI_ISL_417874, EPI_ISL_424405, EPI_ISL_417875, EPI_ISL_417861, EPI_ISL_419656, EPI_ISL_417868, EPI_ISL_417829, EPI_ISL_417922, EPI_ISL_424413, EPI_ISL_417700, EPI_ISL_424344, EPI_ISL_417921, EPI_ISL_417923	ORF1ab	266	21555	TTC	TTT	2770	F	F
9	EPI_ISL_419658	ORF1ab	266	21555	TTG	TTT	10816	L	F
10	EPI_ISL_424413	ORF1ab	266	21555	ATG	A–	11005	M	?
11	EPI_ISL_424413	ORF1ab	266	21555	GTT	—	11008	V	?
12	EPI_ISL_424413	ORF1ab	266	21555	GAT	-AT	11011	D	?
13	EPI_ISL_419656	ORF1ab	266	21555	CAG	CAA	12565	Q	Q
14	EPI_ISL_417700	ORF1ab	266	21555	TCG	TTG	13192	S	L
15	EPI_ISL_424418, EPI_ISL_424451, EPI_ISL_417874, EPI_ISL_424405, EPI_ISL_417875, EPI_ISL_417861, EPI_ISL_419656, EPI_ISL_417868, EPI_ISL_417829, EPI_ISL_417922, EPI_ISL_424413, EPI_ISL_417700, EPI_ISL_424344, EPI_ISL_417921, EPI_ISL_417923	ORF1ab	266	21555	CCT	CTT	13142	P	L
16	EPI_ISL_419658	ORF1ab	266	21555	TAC	TAT	14538	Y	Y
17	EPI_ISL_419658	ORF1ab	266	21555	CGT	CGC	16980	R	R
18	EPI_ISL_417923	ORF1ab	266	21555	TTA	TTG	20001	L	L
19	EPI_ISL_424405	S	21563	25384	GGT	TGT	1777	G	C
20	EPI_ISL_424418, EPI_ISL_424451, EPI_ISL_417874, EPI_ISL_424405, EPI_ISL_417875, EPI_ISL_417861, EPI_ISL_419656, EPI_ISL_417868, EPI_ISL_417829, EPI_ISL_417922, EPI_ISL_424413, EPI_ISL_417700, EPI_ISL_424344, EPI_ISL_417921, EPI_ISL_417923	S	21563	25384	GAT	GGT	1840	D	G
21	EPI_ISL_419654	S	21563	25384	TCA	TCT	2203	S	S
22	EPI_ISL_424405	S	21563	25384	ACA	ACG	3298	T	T
23	EPI_ISL_424418, EPI_ISL_417874, EPI_ISL_417875, EPI_ISL_424413, EPI_ISL_417700	ORF3a	25393	26220	CAG	CAT	169	Q	H
24	EPI_ISL_419658	ORF3a	25393	26220	GGT	GTT	751	G	V
25	EPI_ISL_424451	M	26523	27191	GAT	GGT	7	D	G
26	EPI_ISL_417861, EPI_ISL_417868, EPI_ISL_417829	M	26523	27191	ACG	ATG	523	T	M
27	EPI_ISL_417861, EPI_ISL_419656, EPI_ISL_417868, EPI_ISL_417829, EPI_ISL_417922	N	28274	29533	AGG	AAA	607	R	K
28	EPI_ISL_417861, EPI_ISL_419656, EPI_ISL_417868, EPI_ISL_417829, EPI_ISL_417922	N	28274	29533	GGA	CGA	610	G	R
29	EPI_ISL_417875	N	28274	29533	GAA	CAA	1132	E	Q

Table 4: Variations between the sequences with respect to the reference sequence in terms of amino acids

The tools have returned the same results, except for *Kalign*, which detected several deletion events at the end of the sequences. Since these variations did not refer to a particular gene, they have not been further analyzed.
In particular, the table shows 29 variations, despite Table 3 reports just 26 mutations, because the variations with id 10, 11 and 12 refer to the same mutation and the same happens for the ones with id 27 and 28. This different representation is due to the fact that Table 4 reports the mutations in triplets to highlight amino acid variations. From the results it is possible to notice that 9 out of 29 variations did not lead to changes in terms of amino acids.
It is worthwhile to mention that the CDS of *ORF1ab* is obtained by concate-

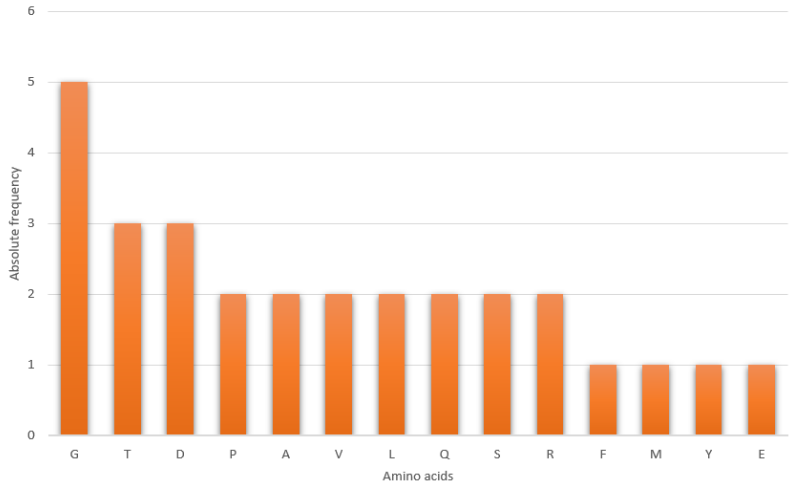

Figure 1: Frequency of the amino acids involved in the mutations considering the reference sequence

nating two CDS: the first starts from 266 up to 13468 while the second one starts from 13468 up to 21555. Note that the nucleotide base in position 13468 belongs to both the CDS and, thus, it is considered twice. For this reason, *ORF1ab* has been considered as a single CDS starting from 266 up to 21555.
In addition, Figure 1 shows the absolute frequency of the amino acids in-

volved in the mutations of the reference sequence (*Reference amino acid* column in Table 4), while Figure 2 shows the absolute frequency of the amino acids involved in the mutations considering the other sequences ((*Sequence amino acid* column in Table 4)). It is possible to notice that the most impacted amino acid in the reference sequence is Glycine (G), which has been involved 5 times in the mutations, while the most frequent amino acid within the other sequences is Leucine (L). Finally, 3 amino acids in the aligned sequences have not been encoded because of the presence of deletions, which did not allow to identify the exact codon.

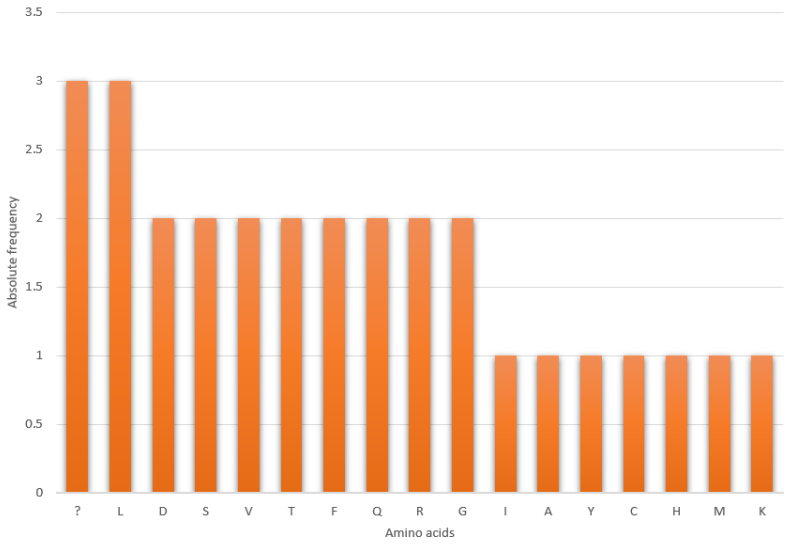

Figure 2: Frequency of the new amino acids involved in the mutations considering the analyzed sequences

6 Analysis of the phylogenetic trees

This section presents the phylogenetic trees generated by the alignment tools, which have been reported in Figures 3, 4 and 5. Although the tools

implement different strategies to perform the multiple alignment, the resulting phylogenetic trees are very similar.
Analyzing these trees, different considerations can be done. First of all, it is worth noticing that all the trees place the reference Chinese sequence in an isolated subtree with three Austrian sequences. This means that the reference sequence is more similar to Austrian sequences than to the others and leads to think that the virus reached Austria before spreading itself in the other two countries.
The trees place sequences with the same mutations in the same subtree. For example, the sequences that present the mutation with id 30 (referred to Table 2) are located in the same substree. Since the Austrian sequence with id EPI_ISL_419656 is closer to Italian sequences than Austrian ones, it is possible to hypothesize that the Austrian individual was infected in Italy.

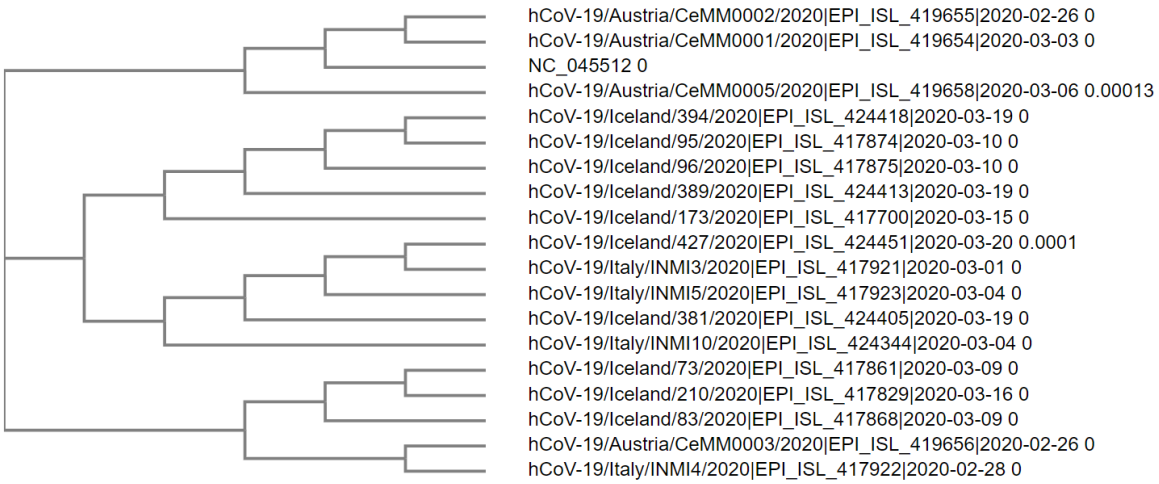

Figure 3: Clustal Omega phylogenetic tree

Observing the phylogenetic trees at a finer grain level, it is possible to notice that there exist two subtrees in which the sequence with id EPI_ISL_424418 is very similar to the one with id EPI_ISL_417874 and, in the same way, the sequence with id EPI_ISL_417861 is very close to the one with id EPI_ISL_41

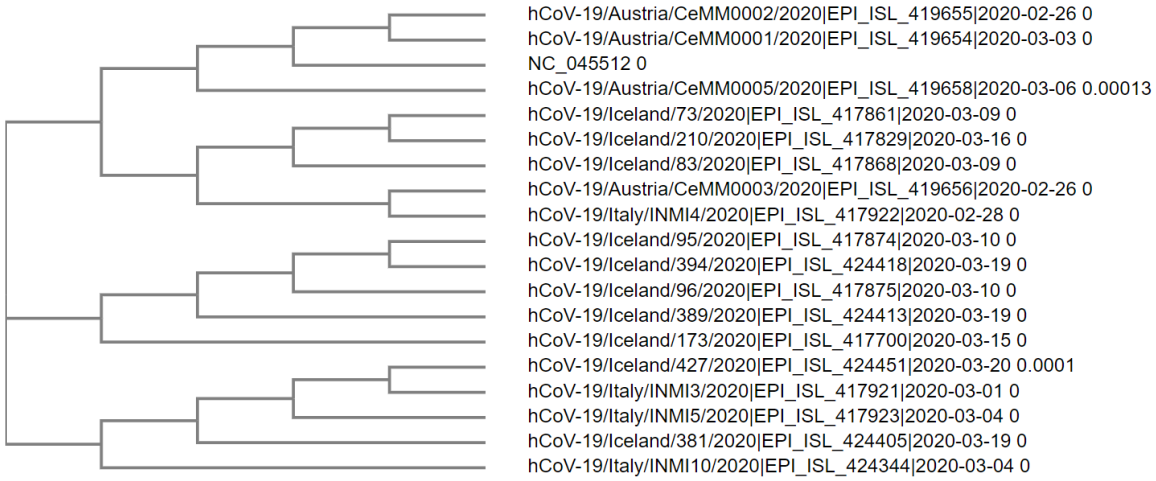

Figure 4: Muscle phylogenetic tree

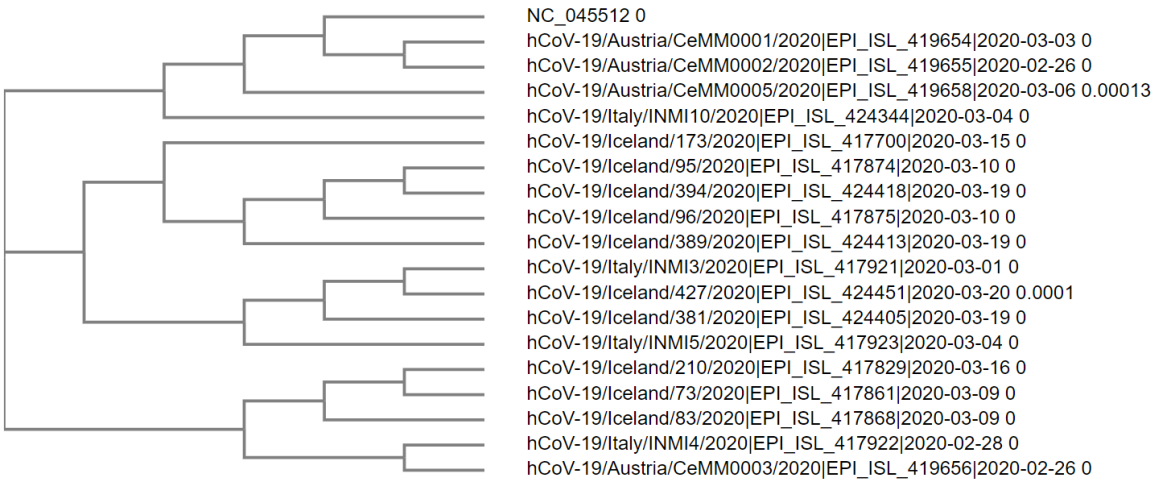

Figure 5: Kalign phylogenetic tree

7829. In particular, the first pair of sequences belongs to Icelandic subjects who traveled to Italy, while the second one to subjects who traveled to Austria. With this consideration, it can be assumed that subjects that have contracted the virus in a specific country are more likely to have the same mutations.

7 Perfect phylogeny

This section shows the perfect phylogenetic tree and compares it with the ones obtained by the three alignment tools described in Section 6.
In order to build the perfect phylogenetic tree, the edit mutations have been considered as the characteristics for the binary matrix represented in Table 5, where its rows represent the sequences and its columns represent the mutations identified by their beginning position with respect to the reference sequence. If a sequence i manifests a mutation j, then the entry of the matrix in position (i, j) is set to 1, otherwise it is set to 0. In total, it consists of 18 rows and 29 columns.
The binary matrix resulted laminar and thus it has been entirely used to

	151	241	645	842	1059	1440	2062	2182	2891	3037	11083	12832	13458	14408	14805	17247	20268	23339	23403	23767	24862	25563	26144	26530	27046	28881	28882	28883	29405
EPI_ISL_424418	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
EPI_ISL_424413	0	1	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
EPI_ISL_424405	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0
EPI_ISL_417875	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1
EPI_ISL_417874	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
EPI_ISL_417829	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	1	0
EPI_ISL_417700	0	1	0	0	1	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
EPI_ISL_424451	1	1	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
EPI_ISL_417861	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	1	0
EPI_ISL_417868	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	1	0
EPI_ISL_419655	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EPI_ISL_419654	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
EPI_ISL_419658	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0
EPI_ISL_419656	0	1	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0
EPI_ISL_417921	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
EPI_ISL_417922	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0
EPI_ISL_417923	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
EPI_ISL_424344	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

Table 5: Binary matrix used to create the perfect phylogenetic tree

create the perfect phylogenetic tree illustrated in Figure 6. The obtained tree is very similar to the ones generated by the three alignment tools. For instance, it is possible to notice that the sequences with id EPI_ISL_417861, EPI_ISL_417829 and EPI_ISL_417868 are very close to each other both in the perfect phylogenetic tree and in the phylogenetic trees obtained with the tools. In addition, this consideration can also be done for the sequences with id EPI_ISL_419655 and EPI_ISL_419654, which are the most similar sequences with respect to the reference sequence in the phylogenetic trees obtained with the tools and the ones that are closest to the root in the perfect phylogenetic tree. This can be explained by the fact that they have the least number of variations in comparison to the other sequences.

It is also possible to notice that sequences coming from the same countries

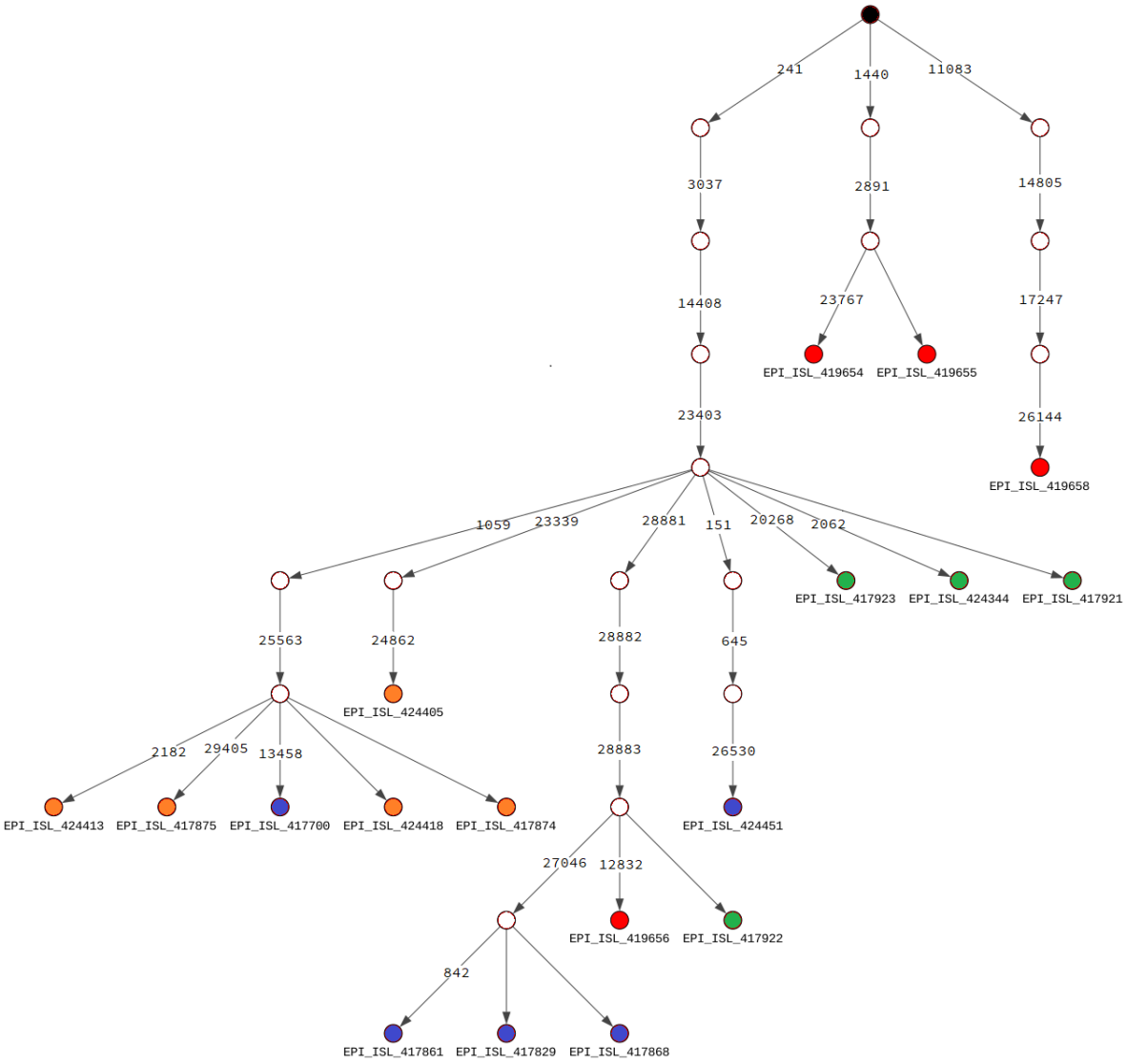

Figure 6: Perfect phylogenetic tree

are closer to each other than sequences coming from other countries. The only exception is the Austrian sequence with id EPI_ISL_419656, which had already been detected in the previous sections, and the Italian sequence with id EPI_ISL_417922.

Finally, from the perfect phylogenetic tree it is possible to see that Austrian sequences have an average number of mutations less than the other sequences, followed by the Italian sequences and the Icelandic ones. In particular, all the Italian and Icelandic sequences share 4 mutations, that is, the ones in position 241, 3037, 14408 and 23403. This consideration further confirms that the Italian and Icelandic sequences, regardless of the country they traveled to, are strongly connected. Moreover, as the Austrian sequences do not share any common mutations with the other sequences (except for the Austrian sequence with id EPI_ISL_419656), it seems that there are no relations between the Austrian virus strain and the Italian and Icelandic ones.
It is worth mentioning that, since the tools available online did not work properly, the algorithm for checking the laminarity of the binary matrix, building the perfect phylogenetic tree and generating the graphical representation of the tree has been implemented from scratch.

8 Conclusions

This work had the purpose of studying the new *Sars-Cov2* virus, focusing on the relations among Icelandic, Italian and Austrian sequences, as it is known that many Icelandic subjects have traveled to Italy and Austria during their holidays before the lockdown.
In this regards, 10 Icelandic sequences belonging to individuals who traveled to Austria and Italy have been compared with other 4 Italian sequences and 4 Austrian sequences.
To perform the multiple alignment, 3 different tools have been exploited, namely *Clustal Omega*, *Kalign* and *Muscle*. The results of the alignment have been gathered together in a JSON file able to highlight the different variations between the reference sequence and the other sequences.
Analyzing the results obtained both at nucleotide base level and at gene level, it emerged a strong relation between Icelandic and Italian sequences, since they presented many common mutations. Furthermore, analyzing

the variations at the amino acid level, it is possible to say that the most involved amino acid in the reference sequence is the Glycine (G) and the one that appeared the most within the aligned sequences is the Leucine (L).
From the phylogenetic trees derived by the tools it is possible to hypothesize that the virus has first reached Austria and then it has spread to the other two countries. In fact, the reference Chinese sequence is closer to the Austrian ones than to the others.
In conclusion, from the comparison between the perfect phylogenetic tree and the trees obtained with the three alignment tools, it emerged that these trees are very similar. In addition, the Icelandic sequences seem more correlated to Italian sequences than Austrian sequences because they share more mutations in common, regardless of the country they traveled to.

Bibliography

[1] Worldometers.info. Covid-19 coronavirus pandemic. `https://www.worldometers.info/coronavirus/`. Accessed: 2020-05-18.

[2] Fabian Sievers and Desmond Higgins. Clustal omega, accurate alignment of very large numbers of sequences. *Methods in molecular biology (Clifton, N.J.)*, 2014.

[3] Robert Edgar. Muscle: Multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 2004.

[4] Timo Lassmann and Erik Sonnhammer. Kalign: An accurate and fast multiple sequence alignment algorithm. *BMC bioinformatics*, 2005.