BioHackrXiv

Determining a novel feature-space for SARS-COV2 Sequence data

Francesco Ballesio[5], Ali Haider Bangash[2, 6], Didier Barradas-Bautista[8], Justin Barton[9], Andrea Guarracino[5], Lukas Heumos[1], Aneesh Panoli[7], Marco Pietrosanto[5], Anastasios Togkousidis[3], Phillip Davis[4], and Fotis Psomopoulos[3]

1 University of Tübingen/Quantitative Biology Center, Auf der Morgenstelle 10, Tübingen, Germany **2** Shifa College of Medicine, STMU, Islamabad, Pakistan **3** Institute of Applied Biosciences, Centre for Research and Technology Hellas, 6th km Charilaou-Thermis rd, Thessaloniki, Greece **4** MRIGlobal, 425 Volker Boulevard, Kansas City, MO 64110, USA **5** University of Rome Tor Vergata, Via della Ricerca Scientifica 1, Rome, Italy **6** SYNCH, Pakistan **7** Insight Data Science, 500 3rd St, San Francisco, CA 94107, USA **8** King Abdullah University of Science and Technology, Thuwal, Saudi Arabia **9** Birkbeck, University of London, London, United Kingdom

BioHackathon series:
COVID-19 Biohackathon
Virtual conference 2020
Machine Learning group

Submitted: 08 May 2020

Published by BioHackrXiv.org

Abstract

Motivation: The end of 2019 came with the emergence of a novel virus, identified as a new strain of Coronavirus, and has since spread over the globe as a pandemic of an unprecedented scale. A global collaborative effort has lead to a number of virus samples being fully sequenced, with the data disseminated by being published in publicly accessible repositories. Given the high similarity of the sequences, both at the aminoacid as well as the nucleotide levels, a key question arises as to how to identify interesting, discriminating features across the different sequences such that the underling structure of the evolutionary story of the virus can be highlighted. In this work we present our efforts in addressing this issue, through the systematic application of machine learning methods towards meaningful feature extraction.

Results: We applied a range of methods, in order to; identify the optimal word (k-mer) size for aminoacid patterns; identify k-mers features at the nucleotide level that have predictive value; construct continuous distributed representations for protein sequences in order to create phylogenetic trees in an alignment-free manner; and predict MHC class I and II binding affinity.

Availability: All data, code and results are available under permissive licenses (CC-BY or MIT) in the team GitHub repository here

Contact: pdavis@mriglobal.org, fpsom@certh.gr

1. Introduction

In late 2019, a novel virus began spreading within the population of the Wuhan-city in the Hubei province of China (C. Huang et al., 2020). The virus, identified as a new strain of Coronavirus (Organization, 2020a), has since spread over the globe as a pandemic of an unprecedented scale (Organization, 2020b,Organization (2020c)).

A global collaborative effort has lead to a number of virus samples being fully sequenced, with the data disseminated by being published in publicly accessible repositories, such as the SARS-CoV-2 sequences GenBank and the EBI Data.

Given the high similarity of the sequences, both at the aminoacid as well as the nucleotide levels, a key question arises as to how to identify features of interest across the different sequences, such that the underling structure of the viral evolutionary story can be highlighted.

In order to address this question, the machine learning group of the COVID-19 BioHackathon, defined the following tasks:

- Identification of the potential features at the nucleotide level based on the k-mers, for various k values.
- Identification of the potential features at the aminoacid level, based on the AA frequencies, across various word sizes.
- Performance of in-silico estimates of epitopes for COVID19.
- Identification of patterns in secondary structure, compared to patterns evident in random sequences.

Each task, along with the corresponding outputs, is detailed hereinafter.

2. Approach

2.1. Data pre-Processing

Different publicly available resources were used as input for this study including the following:

- SARS-CoV-2 ORF1ab gene sequences and metadata for all *betacoronaviridae* were obtained from NCBI Virus.
- SARS-CoV-2 (COVID19) full assembly nucleotide sequences that have been identified in Humans were retrieved from GenBank

The processed version of all *betacoronaviridae* sequences used in this study can be found here (fasta, metadata). The nucleotide sequences were translated using Biopython(Cock et al., 2009) package.

The processed version of all human virus sequences can be found here and comprises of a set of 281 genome sequences of SARS-CoV-2, each one approximately 30,000 nucleotides in length. The corresponding meta-data file is here and contains information about the length of each sequence, geographical location, isolation source, collection date of the sample etc.

2.2 Scope

In order to identify features of interest across the target sequences, and investigate their utility as potential predictors, we applied a range of methods including following: 1. Identification of optimal word (k-mer) size for aminoacid patterns 2. Identification of potential k-mers features at the nucleotide level 3. Continuous distributed representations for protein sequences inorder to create phylogenetic trees in an alignment-free manner 4. Prediction of MHC class I and MHC class II binding affinity

3. Methods

Each of the aforementioned methods is described concisely in the following sections.

3.1 Determination of optimal amino acid word size for ORF1ab feature extraction.

Using the processed version of all *betacoronaviridae* SARS-CoV-2 ORF1ab sequences, the amino acid sequences were fragmented into words ranging from 1 to 9 in size. From within the available meta-data, we selected the following four fields as potential classification targets: - Species - Host - Geographical Location - Extraction Source

In order for a field to be included as a classification target, the particular label must have a representation of at least 20 within the entire dataset of 2,384 unique, by accession ID, sequences.

The produced words were embedded with `CountVectorizer` and fitted on logistic regression using the scikit-learn(Pedregosa et al., 2011) package. Model performances were evaluated using a weighted average of precision, recall, and F1-score across the test data.

3.2 Potential features at the nucleotide level based on the k-mers

This approach focused on the detection of k-mers that appear with high frequency in the data. The primary dataset used for feature extraction was the set of 281 human SARS-CoV-2 virus sequences. The analysis was conducted is an algorithmic procedure based on a pruning tree, which dynamically evaluated k-mers of different lengths, keeping only those with the highest evaluation. The evaluation parameter depended on both; the length of each k-mer as well as its frequency in the data. In this way, the most significant k-mers were isolated within a very decent time and were able to be used as features in our data.

The analysis was conducted in two different ways:

- In the first approach, the algorithm was applied to each sequence separately. In this way, the repetitiveness of k-mers within a single sequence was examined. The data that were extracted from this analysis have been joined with the meta data in a single data matrix. The elements below the k-mer columns correspond to the frequency of each k-mer within a single sequence.
- In the second approach, the algorithm was applied to the total data set and, thus, treating the genome sequences as a single set. K-mers that appeared with high frequency within all the genome sequences were successfully isolated in an output k-mers set. The next step was to remove all k-mers that appeared to every sequence from this output set, in order to reduce the dimensionality of the problem. The data that were extracted from this analysis have also been joined with the meta data set in a single data matrix. It is indicated that the elements in the k-mer columns are zeros and ones, where 1 indicates that the current k-mer appears in the corresponding sequence, while 0 indicates absence.

3.3 Continuous distributed representations for protein sequences to create phylogenetic trees in an alignment-free manner

Biological sequence comparison is a well established method in inferring the relatedness of various organisms as well as the functional role of their components. In the last years, there have been some efforts into representing biological sequences with new paradigms, especially by Natural Language Processing methods hereinafter laid down, with the aim to capture the most meaningful information of the original sequences. Although more modern solutions are present in the NLP domain universe, including but not limited to ELMo (Peters et al., 2018) & BERT (Devlin, Chang, Lee, & Toutanova, 2018), biological sequence representation still has much to explore (Kimothi, Soni, Biyani, & Hogan, 2016), especially in relation to the final task which was solved exploiting the new representation.

One of the most successful word embedding-based models is the word2vec model (Mikolov, Chen, Corrado, & Dean, 2013) for generating distributed representations of words and phrases. Considerable advances have been made with its standard application (Asgari & Mofrad, 2015a), with the functionality being extended to modelling for DNA (Ng, 2017), RNA (Yi et al., 2020) and protein (Asgari & Mofrad, 2015b) sequences. To briefly summarize those studies, the impact of projecting sequence data on embedded spaces is likely to reduce the complexity of the algorithms needed to solve certain tasks (*e.g.* protein family classification (Asgari & Mofrad, 2015b)). Moreover, this approach is promising to represent residue-level sequence contexts for potential phosphorylation sites and demonstrate its application in both general and kinase-specific phosphorylation site predictions (Xu, Song, Wilson, & Whisstock, 2018).

Phylogenetics is the task of creating a phylogenetic tree which represents a hypothesis about the evolutionary ancestry of a set of genes, species or any other taxa. Many tree inference methods have been proposed and the current state-of-the-art approach is to perform tree inference through a two-step process of multiple sequence alignment (MSA) followed by statistical tree inference (Felsenstein, 1988). In this work we propose the use of continuous distributed representations for the protein sequences to create phylogenetic trees in an alignment-free manner, analyzing its strengths and weaknesses for this aim.

Our approach is inspired by previous works cited above, with the following characteristics:

- each protein sequence is treated as a sentence, made by overlapping words (k-mers) to incorporate some context-order information in the resulting distributed representation;
- the word size is 3, which seems to work properly to embed amino acid sequences for biological tasks (S. Cheng et al., 2019, Yi et al. (2020));
- the sequence vector is defined as the arithmetic mean of all its word vectors.
- we also explored document representation via distributed memory (doc2vec) to directly generate sequences vectors(Le & Mikolov, 2014).

By defining the sequence vector as the arithmetic mean of all its word vectors we must point out that the sequence vector loses the concept of k-mer order, (i.e. the same vector can be obtained by the same k-mers shuffled) **but** the overlapping k-mers should have processed that "order" information down to their representations. That is, if there is a k-mer "SAN" there will certainly by a k-mer "-SA" and a k-mer "AN-" (where "-" is any aminoacid), and this is, in our view, a way of loosely preserving the k-mer order information in the sequence vector. Nonetheless, for this reason, we generated also sequence vectors via doc2vec which is intrinsically supposed to preserve k-mers order. However, using this architecture, we didn't obtain good models for our task (result not shown).

As word2vec models, two architectures are available: continuous bag-of-words (CBOW) and skip gram. These models are shallow, two-layer neural networks that are trained to reconstruct semantic contexts of words. The CBOW model is trained to predict the current word by using a few surrounding context words. On the other hand, skip-gram uses the current word to predict the surrounding context words. In this work we applied the CBOW architecture, which is generally faster, therefore it is the preferred choice to have a scalable solution when a large corpus will be available for training. Importantly, the skip-gram architecture, in addition to result in a greater computational load for training the models, did not lead to significantly better models in our task (result not shown).

The data we analyzed was a collection of ORF1ab AA sequences from the NCBI, as previously mentioned NCBI Virus and metadata. We explored the hyper-parameter space trying several combinations of the following hyper-parameters: k-mers size, vector space dimension, number of epochs for the training.

- embedding size: [3, 4]
- embedding size: [10, 50, 100, 200, 300, 500, 1000]
- training epochs: [5, 10, 20, 50, 100, 200, 500, 1000]

All the experiments were performed using Gensim (Řehůřek & Sojka, 2010) and Scikit-learn

(Pedregosa et al., 2011) libraries. In particular, we focused on the following pipeline: - obtain a vectorial representations of the proteins; - build a tree by using cosine distance between sequence vectors; - compare it with the clustalOmega (Sievers & Higgins, 2013) generated tree by means of Robinson-Foulds distance; - choose the best embedding by referring to the aforementioned distance, exploring the embedded space and the resulting tree by: * analyzing the embedded space by PCA * analyzing the embedded space by tSNE * exploring the resulting tree both with the full embedded space and with the first Principal Components

The comparison between the trees built on the embeddings and the clustalOmega tree is done to have an external validation: results should not be too different from standard phylogenetic trees but should still show variations, in order to point untracked similarities between SARS-CoV-2 and other *coronaviridae*.

3.4 MHC class I and II binding affinity prediction

An integral part of the adaptive immune system is the presentation of antigen epitopes on the cell surface. The MHC is the tissue-antigen which T-cells bind to, recognize and self-tolerate. During this process the MHC molecules bind to both the T-cell receptor and glycoproteins CD4/CD8 (cluster of differentiation) on T lymphocytes. Additionally, interactions between the variable Ig-like domain of the TCR interacts with the antigen epitope located in the peptide-binding groove of the MHC molecule to trigger T cell activation. Hence, epitopes can be used to elicit specific immune response making them suitable for vaccine design (Palatnik-de-Sousa, Soares, & Rosa, 2018). To construct an epitope based vaccine it is therefore imperative to evaluate the MHC class I or II binding affinity for a given set of peptide candidates and a given set of alleles. Furthermore, determining the binding affinities for specific subgroups may aid researchers to focus on the most promising protein subunits, speeding up the vaccine development process.

To determine binding affinities of peptides to MHC molecules, time consuming experiments such as competition experiments have to be carried out. In these experiments the peptide concentration, which leads to 50% inhibition of a standard peptide is measured. This concentration is known as the IC50 value. MHC binding peptides are typically classified by resulting IC50 values of less than 500 nM (Sette et al., 1994). To allow for quick and free assessments of MHC class I and II binding affinities several machine learning based software packages have been released in recent years. All of them are based on experimentally verified databases of MHC molecule binders and non-binders, but differ in their algorithms, training datasets and accessibility (Backert & Kohlbacher, 2015).

MHCNuggets, a MHC class I and II binding affinity predictor, is based on a deep neural network, which makes use of several long-short term memory (LSTM) units to facilitate fast and peptide length-independent predictions. Moreover, the usage of transfer learning and allele clustering approaches enable the confident prediction of rare alleles. The authors demonstrated that MHCNuggets has comparable prediction performance for both classes when compared to NetMHCPan, MHCFlurry and others, while being the fastest prediction method (Shao et al., 2019). MHCNuggets v2.3 was applied within the EpitopePredict framework v0.4 (Farrell, 2019) using the predefined broad_coverage_mhc1 (26 HLA alleles providing broad coverage) and human_common_mhc2 (11 most prevalent HLA-DR alleles worldwide) allele sets for class I and II respectively on a set of 7773 subunit proteins of common corona virus including SARS-CoV-2.

The resulting epitope predictions and predicted binding affinities were used to for clustering using UMAP (McInnes, Healy, Saul, & Großberger, 2018). UMAP operates under the assumptions that the data are uniformly distributed on a Riemannian manifold, that the Riemannian metric is locally constant, and that the manifold is locally connected. This allows UMAP to find a low dimensional projection of the data, which is equivalent to a fuzzy topological structure.

UMAP has been demonstrated to retain more of the global structure than, for example, t-SNE, while having a lower run time (McInnes, Healy, & Melville, 2018).

4. Results and Discussion

Preliminary results include:

4.1 A k-mer length of four is sufficient to model the distribution of ORF1ab sequences.

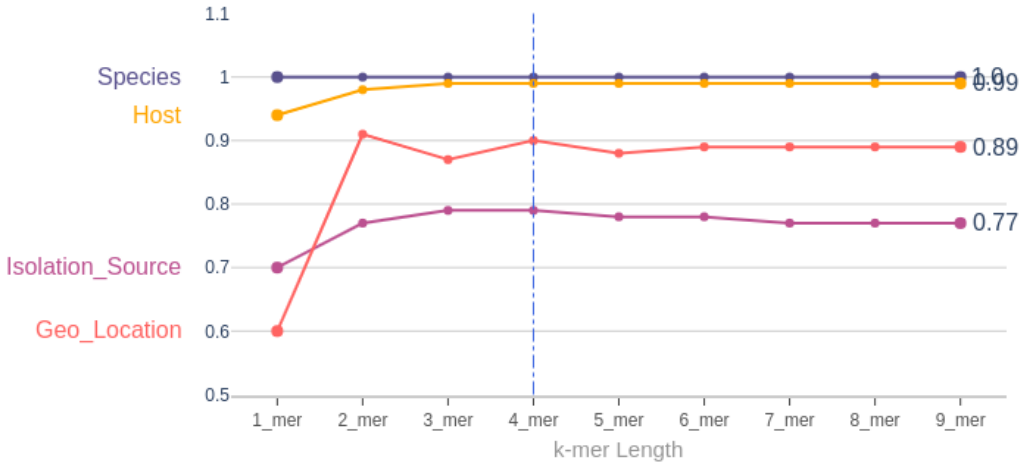

Figure 1: A line plot of weighted average F1-score for four different ORF1ab amino acid sequence classification tasks (Y-axis labels) at various kmer lengths. Y-axis denotes F1-score and X-axis denotes k-mer lengths. The vertical dotted line indicates optimal k-mer length.

We obtained the weighted averaged F1-score in different context of the data to explore the use of k-mer lengths to extract features from the ORF1ab aminoacid sequences for classification tasks. The plot shows that the species and host context are highly separable using aminoacid sequence. Geolocation has the best F1-scores at two and four k-mers. The isolation source context also shows there are differences on sequences that four k-mer length captures better. A four k-mer length presented the optimal scores to classify the test data on the different chosen context in combination.

	precision	recall	f1-score	support	kmer
20	0.803875	0.805195	0.794777	77	3_mer
27	0.803875	0.805195	0.794777	77	4_mer
34	0.791317	0.792208	0.781165	77	5_mer
41	0.791317	0.792208	0.781165	77	6_mer
55	0.77255	0.779221	0.772237	77	8_mer
62	0.77255	0.779221	0.772237	77	9_mer
48	0.776663	0.779221	0.767307	77	7_mer
13	0.773534	0.779221	0.766958	77	2_mer
6	0.701253	0.727273	0.702229	77	1_mer

Extraction Source

	precision	recall	f1-score	support	kmer
13	1	1	1	219	1_mer
27	1	1	1	219	2_mer
41	1	1	1	219	3_mer
55	1	1	1	219	4_mer
69	1	1	1	219	5_mer
83	1	1	1	219	6_mer
97	1	1	1	219	7_mer
111	1	1	1	219	8_mer
125	1	1	1	219	9_mer

Species

	precision	recall	f1-score	support	kmer
39	0.919419	0.911602	0.909453	181	2_mer
79	0.908101	0.900552	0.899728	181	4_mer
179	0.908725	0.895028	0.892738	181	9_mer
119	0.90112	0.889503	0.887768	181	6_mer
139	0.904967	0.889503	0.887735	181	7_mer
159	0.904967	0.889503	0.887735	181	8_mer
99	0.892175	0.883978	0.882799	181	5_mer
59	0.884699	0.872928	0.874994	181	3_mer
19	0.61399	0.640884	0.604679	181	1_mer

Geo_Location

	precision	recall	f1-score	support	kmer
38	0.994892	0.994624	0.994665	186	3_mer
51	0.994892	0.994624	0.994665	186	4_mer
64	0.994892	0.994624	0.994665	186	5_mer
77	0.994892	0.994624	0.994665	186	6_mer
90	0.994892	0.994624	0.994665	186	7_mer
103	0.989695	0.989247	0.989337	186	8_mer
116	0.989695	0.989247	0.989337	186	9_mer
25	0.979295	0.978495	0.978544	186	2_mer
12	0.939999	0.94086	0.940354	186	1_mer

Host

Figure 2: A table showing weighted averages scores for four different ORF1ab amino acid sequence classification tasks at various k-mer lengths.

This table shows the weighted average metrics across the test data for four classification task depending on the context.

4.2 Nucleotide k-mer features as potential predictors

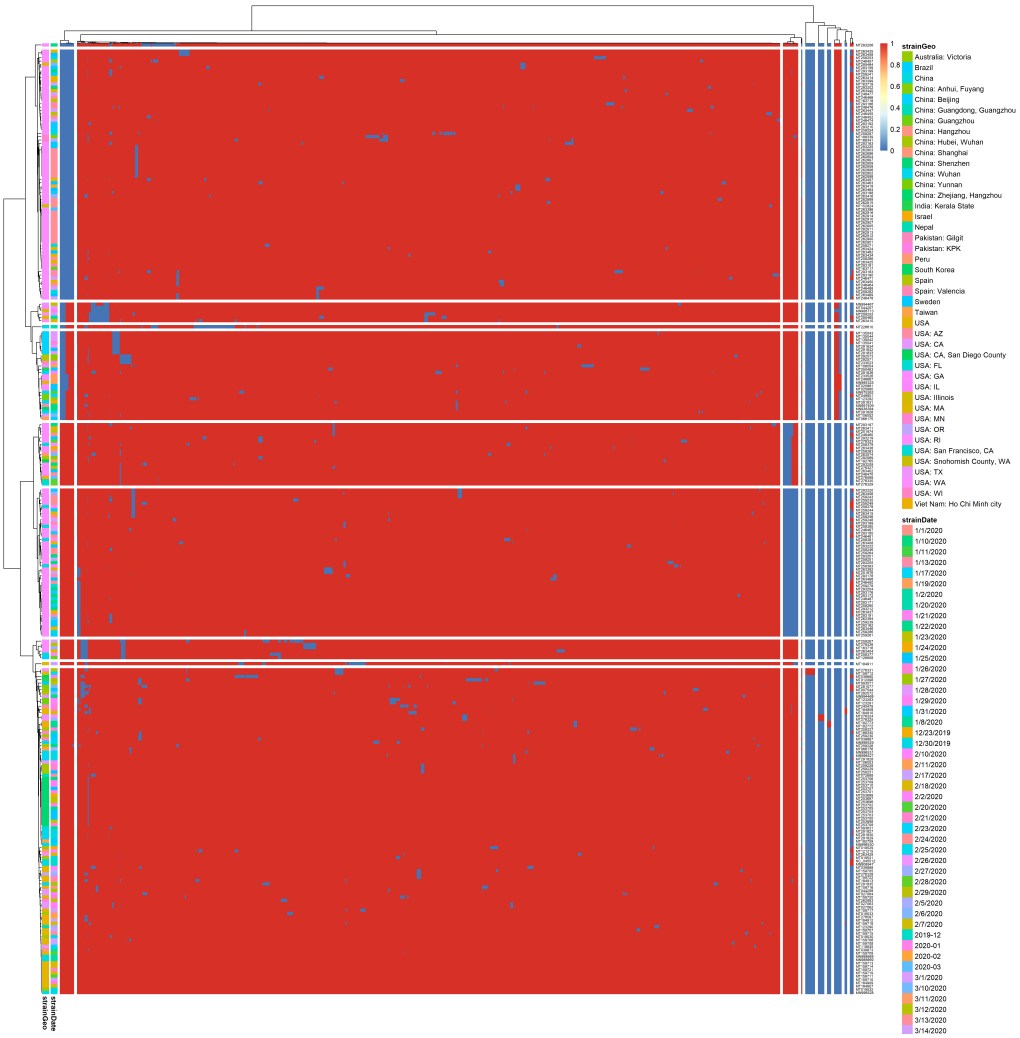

Figure 3: A heatmap of the k-mer-based features across the COVID19 sequences, annotated both by GEO and Date. Rows and columns were clustered using ward.D2. Sequence clusters (y-axis) where produced by applying a tree cut-off at 10 clusters.

Using the single data matrix representation of the k-mer-based feature, we applied hierarchical clustering across both sequences and features. As shown in the figure above, there is a distinct clustering of sequences - notably, the reference sequence (*AccID: NC_045512*) is clustered together with several other sequences, but at the same time, there are a few singletons as outliers, that should be investigated further.

Additionally, it is equally important to note that in this analysis all derived features were utilized. However, it is evident that the feature variance (column) is very limited, which implies that the corresponding set of predictors will be significantly smaller.

4.3 Continuous distributed representations results

Initial results indicate that higher dimensional embeddings are better at capturing the complexity of the aminoacidic sequences in terms of the resulting tree. The best results against the

clustalOmega tree are in fact obtained for the word2vec model for a k-mer length of 3, a vector size of 1000, trained for 100 epochs. All subsequent analyses are related to this model. In addition, we briefly investigated the robustness of the method by analyzing the second-best model and all the following results were maintained (data not shown):

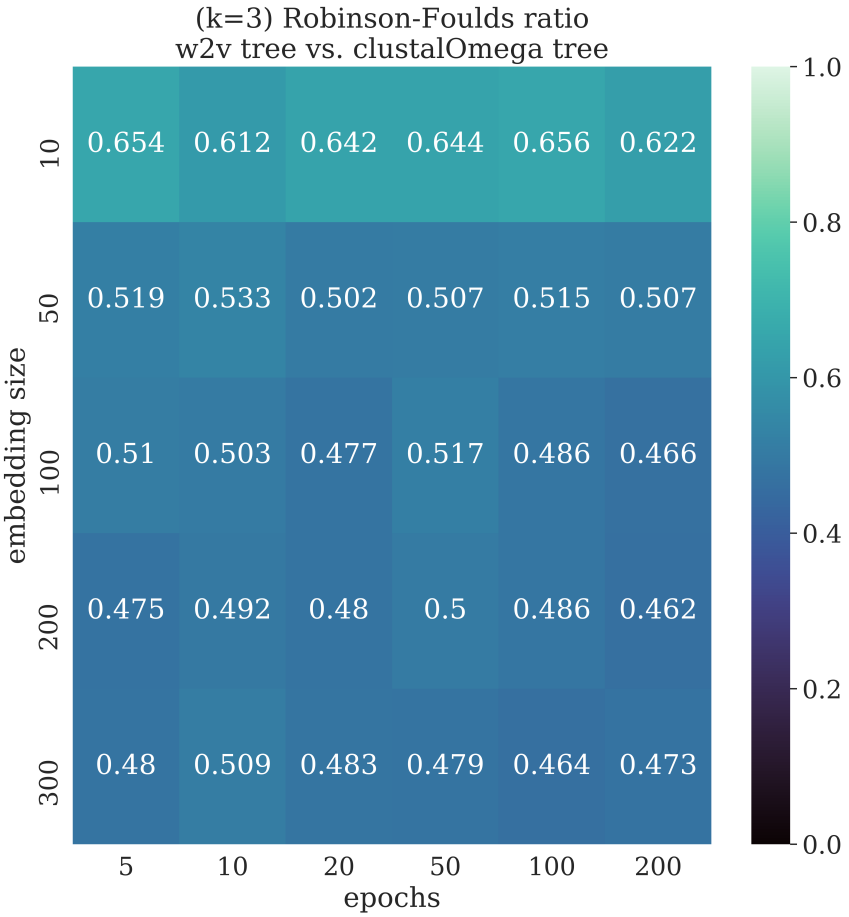

Figure 4: Heatmap reporting for all the hyper-parameter combinations performed the Robinson-Foulds distance between the trees build on the embeddings and the clustalOmega tree for all the hyper-parameter combinations performed.

To understand how the underlying space is distributing its variability we performed a PCA up until 90% explained variance, and even if the best embedding required high dimensions (1000), the majority of the variance can be found in 10 Principal components.

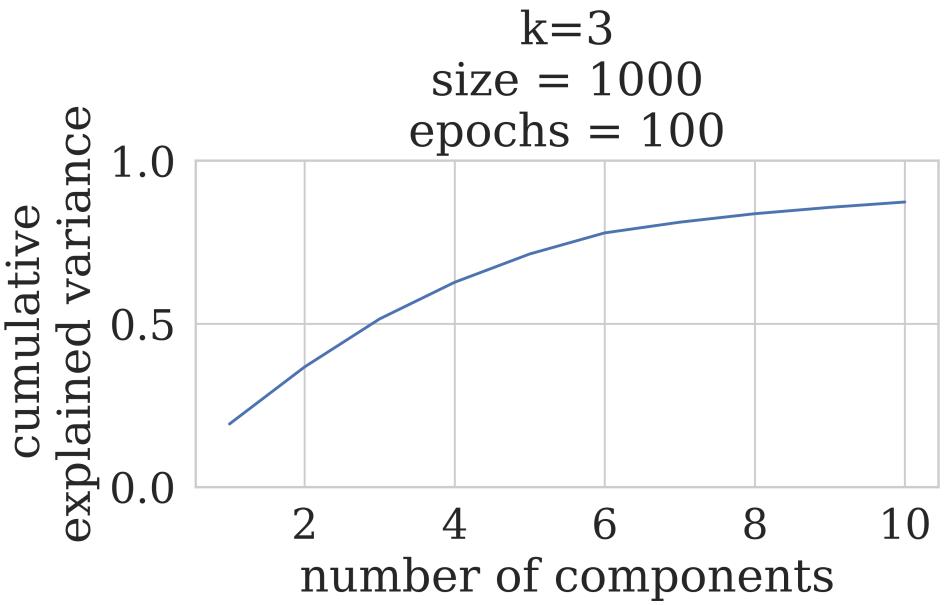

Figure 5: Principal Component Analysis of the vector-space of the best model shows that the majority of the variance lies on few principal components.

In parallel we performed a tSNE in 2-dimensions to have an indication on how the groups of different virus species were clustered and if any confounding effect was present (e.g. clustering for country). By plotting only those species that were present no less than 5 times we can see that SARS-CoV-2 clusters near the bat coronavirus, as expected.

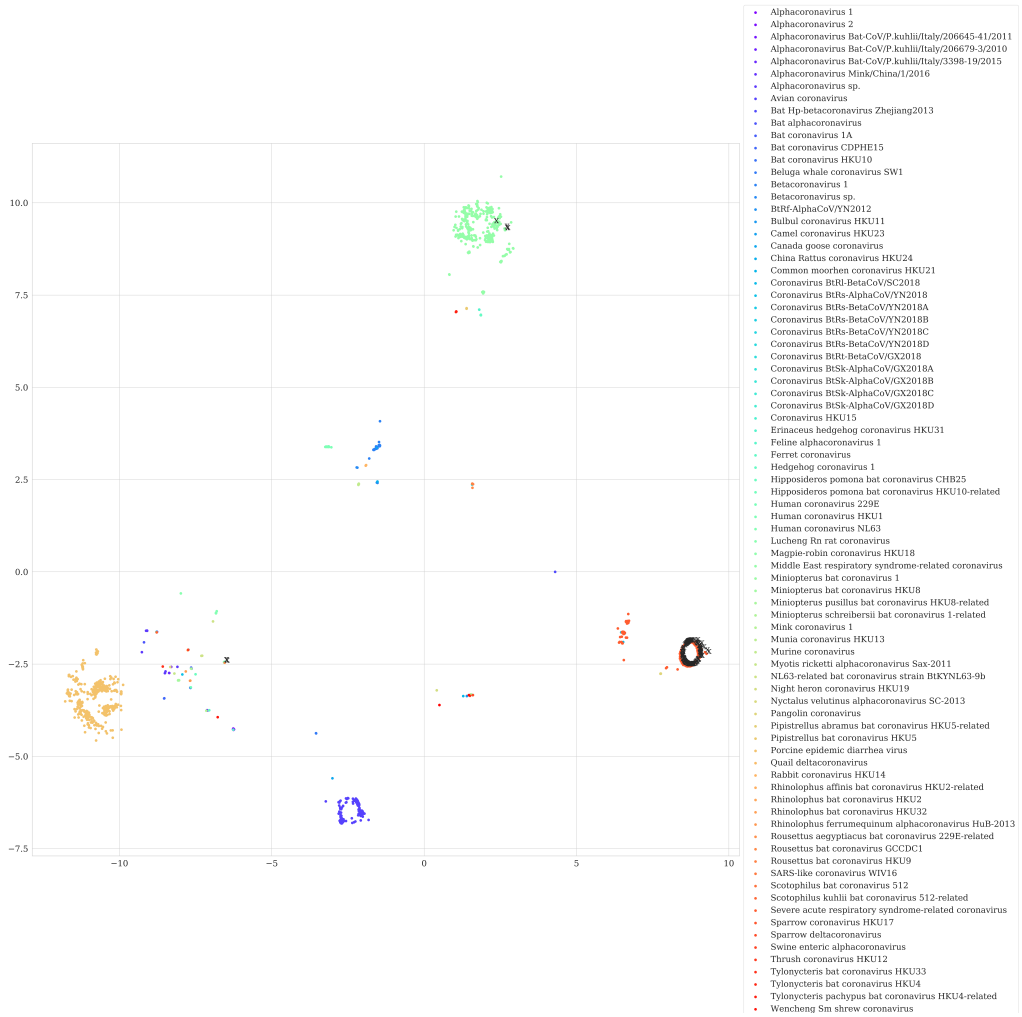

Figure 6: t-distributed stochastic neighbor embedding space in 2-dimension shows expected clusters.

No country-related clustering was evident.

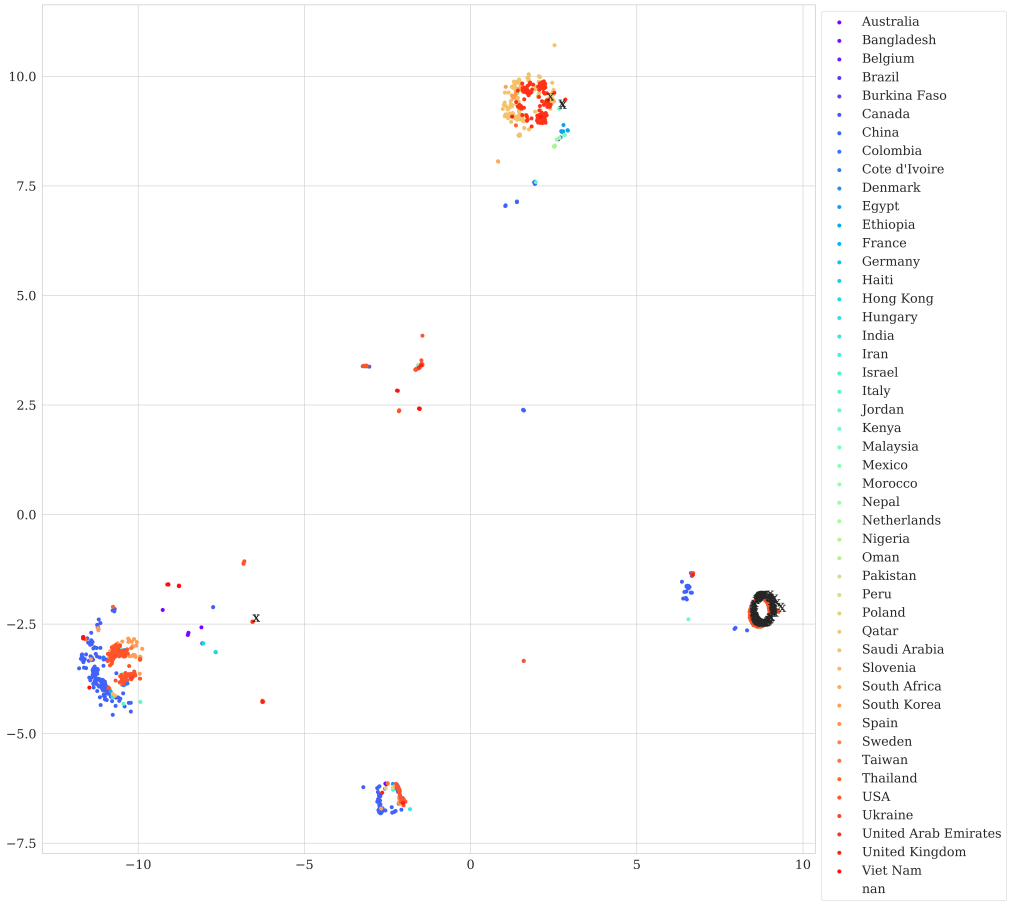

Figure 7: t-distributed stochastic neighbor embedding space in 2-dimension confirms the absence of country as a confounding effect.

Those analyses were necessary to ensure that the embedding space was reflecting the underlying phylogeny that is usually caught by multiple alignment methods.

Finally, by using the cosine distance we built a distance tree and inspected the resulting clusters formed around SARS-CoV-2, visualized using Interactive Tree Of Life (iTOL) (Letunic & Bork, 2019).

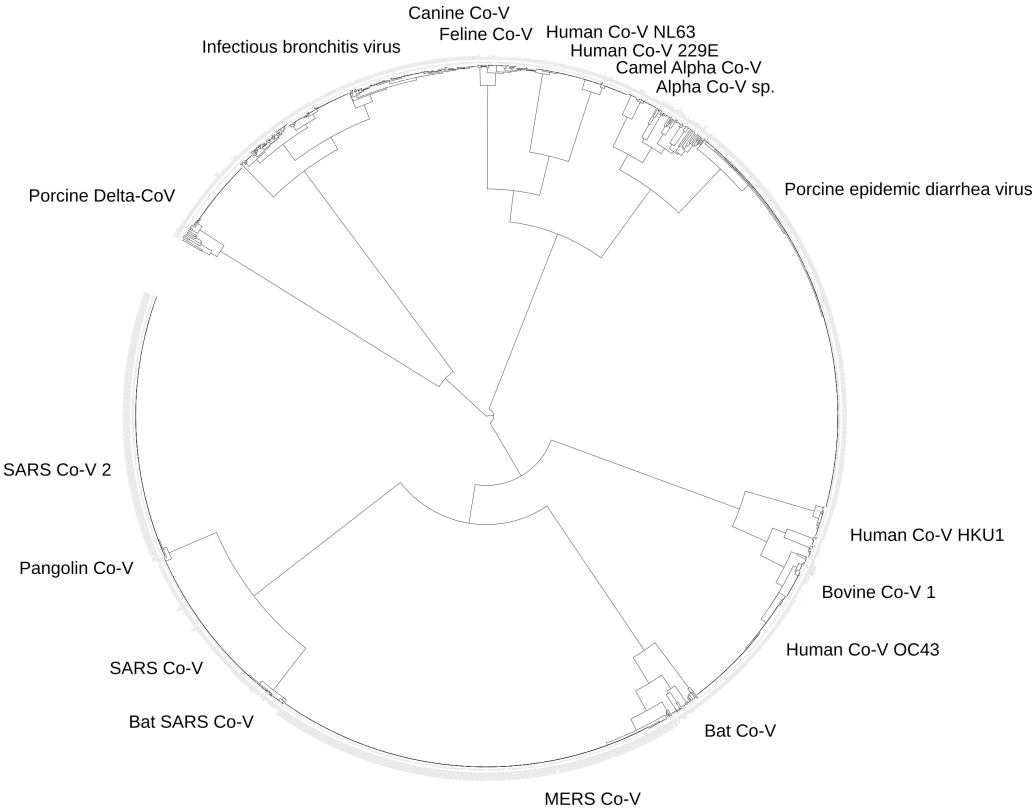

Figure 8: Distance tree from the best model.

As expected SARS-CoV-2 has as nearest neighbours: Pangolin coronavirus (Lam et al., 2020), SARS-Co-V, and Bat coronavirus. There are not apparent unexpected neighbours, and the most distant species from SARS-CoV-2 is the porcine *Deltacoronavirus*, which actually has been seen as related to SARS-Co-V in a recent study (Boley et al., 2020). A possible explanation for this discrepancy could be attributed to the distance metric used in the evaluation of the tree, which does not incorporate the "importance" of each node in the tree. More studies are needed to explore a more sensible distance metric, and the resulting best phylogenetic trees.

4.4 Epitope predictions suggest important variation by HLA allele and viral subunit protein

Out of 3730 sequences of the common corona virus sequences dataset, with an average AA sequence length of 414, MHCnuggets identified 867,231 epitopes as binding (using a theshold of IC50<500).

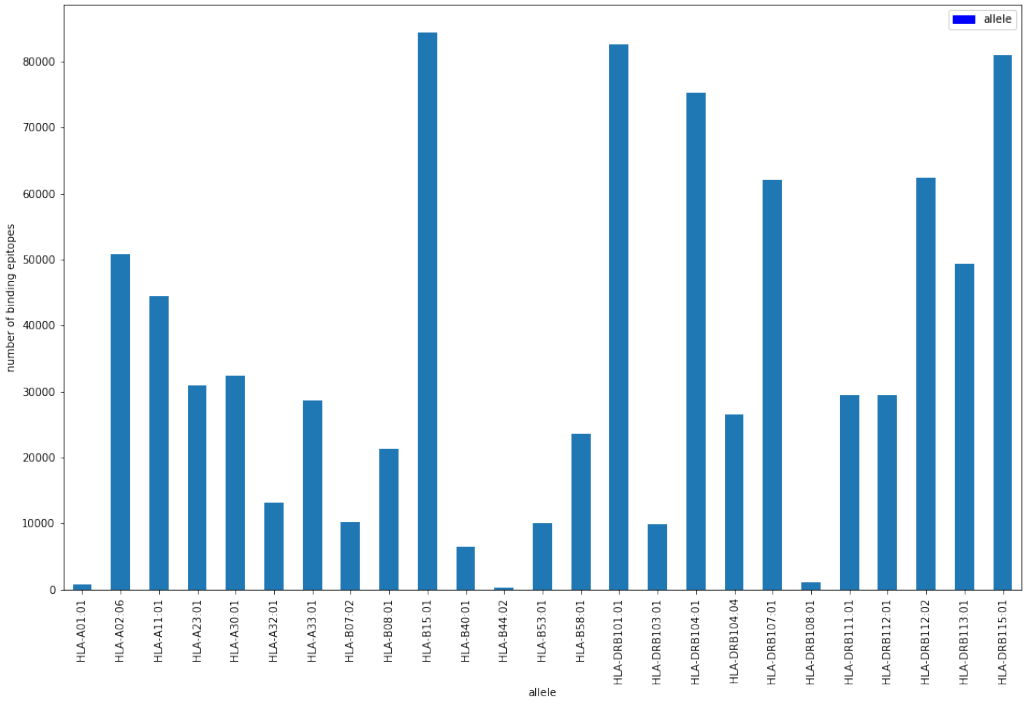

Figure 9: Number of putative binding epitopes per allele.

For the alleles, which have predicted binders, the number of strong and weak binders is not evenly distributed. HLA-B15:01, HLA-A11:01, HLA-A23:01, HLA-A30:01, and HLA-A33:01 seem to contain many strong binders, indicating that populations which cover these specific alleles well may respond strongly to vaccines based on the respective epitopes.

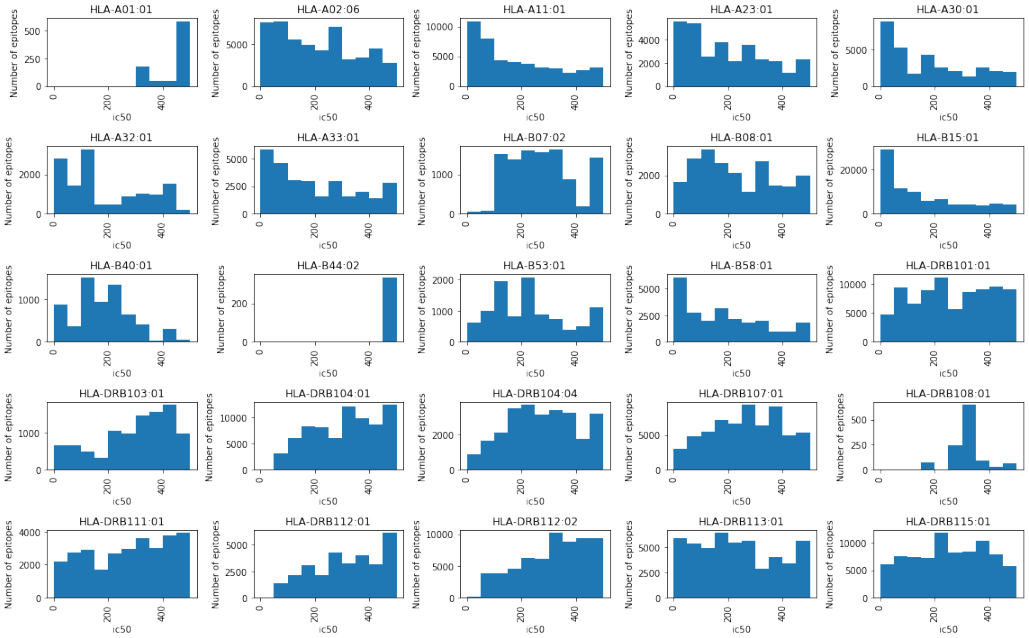

Figure 10: ic50 distribution of putative binding epitopes per allele.

The space of predicted epitopes and IC50 values were explored with UMAP clustering. For this

purpose, an approach analogous to one hot encoding was used. For each of the original 3730 sequences a vector was generated with length equal to the number of distinct (predicted) allele-epitope pairs. As with one hot encoding, a 0 was used to indicate that a given allele-epitope was not predicted for the sequence. However, instead of using a 1 to indicate the prediction of an allele-epitope pair for a given sequence, the inverse of the IC50 value was used (1/IC50). This was done to include some information about predicted binding affinities. Future work could expand on this by using more direct representations of predicted immunogenicity and screening out epitopes resembling self-peptides.

The UMAP clustering of the predicted epitope space yielded some interesting results that warrant further investigation. In particular, it suggested very little within-species variation of the immune-presentation of envelope and membrane proteins.

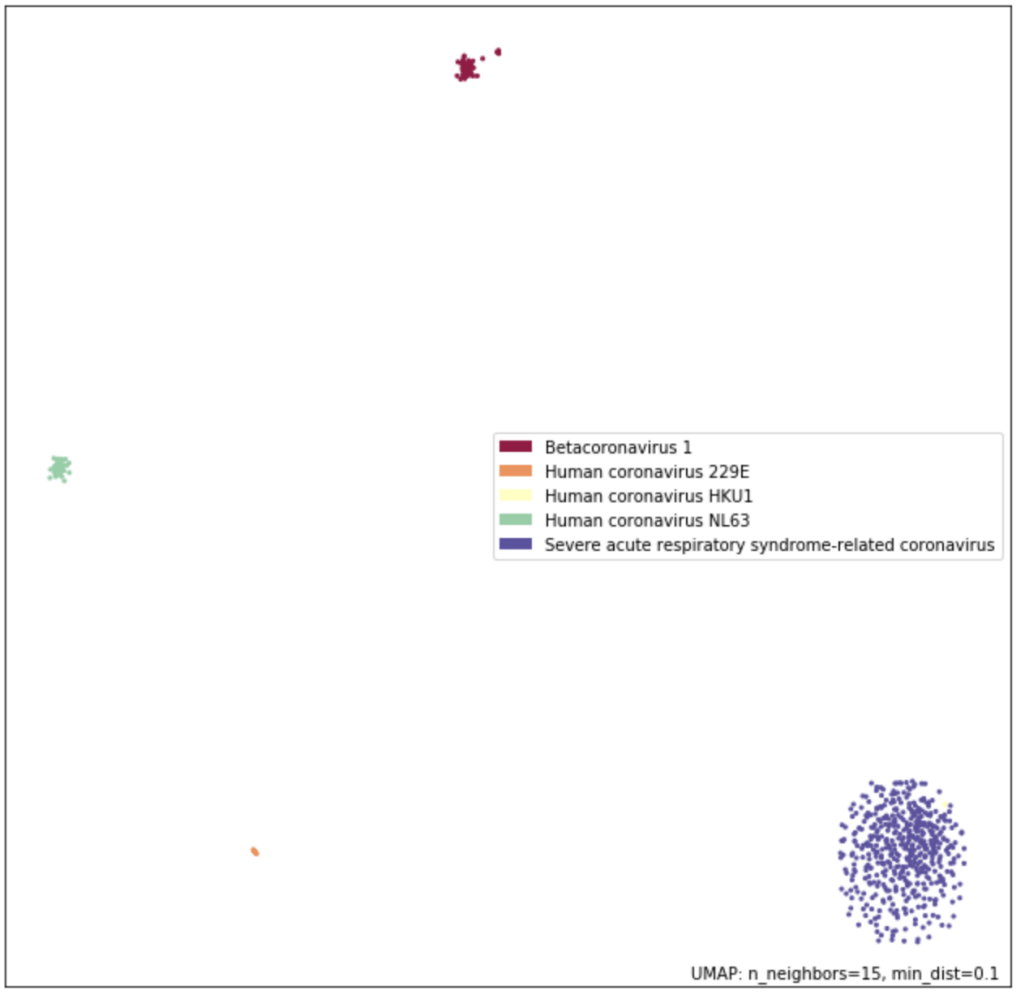

Figure 11: UMAP clustering of envelope proteins.

On the other hand, it suggests a great deal of variation in the immune-presentation of the spike proteins, even within species.

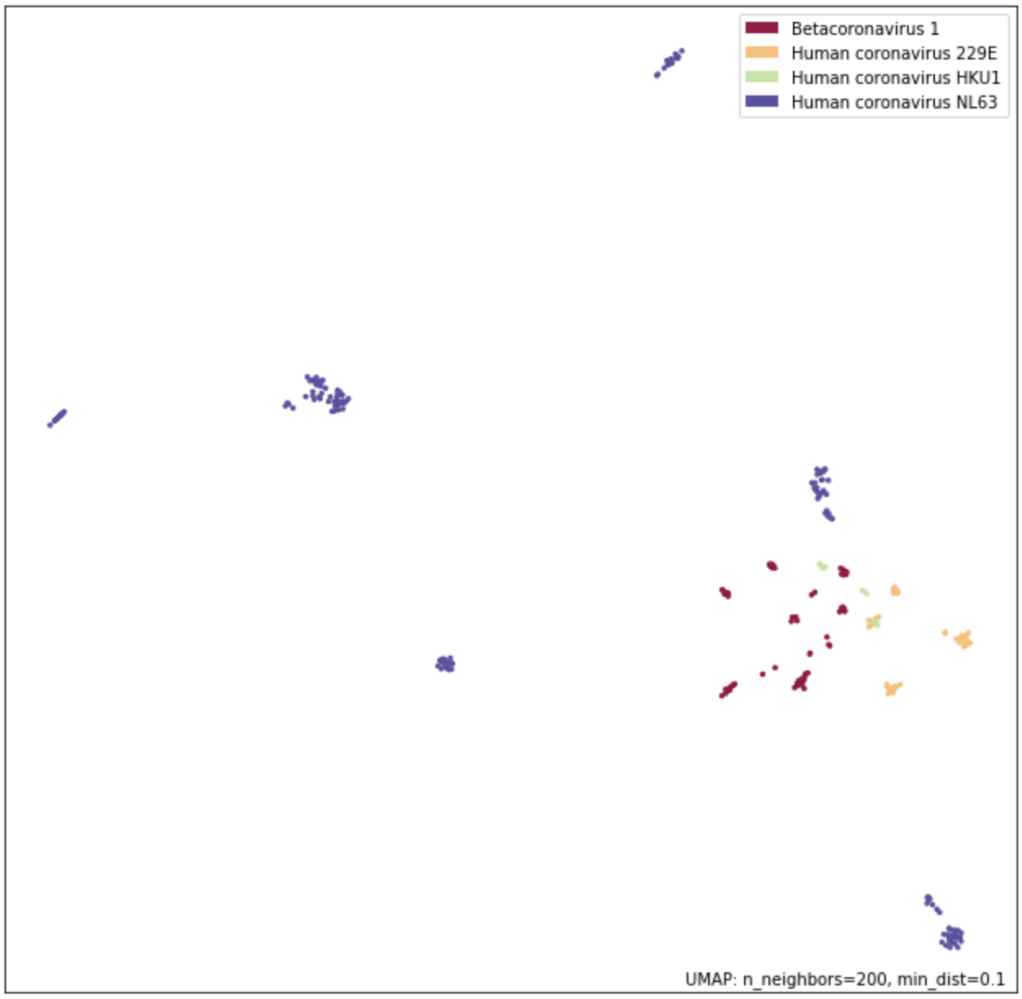

Figure 12: UMAP clustering of spike proteins.

If confirmed these results would have implications on potential immunotherapies as well as the effectiveness of subunit vaccines against within-species variants of the virus.

5. Conclusion

Based on the analysis so far, some broad conclusions can be derived.

Using constructed sequence features as a predictor set

All utilized methods produced promising results; species and host context are highly separable using aminoacid sequence, there are unique nucleotide-based k-mers that can be potentially used as additional predictors in a classification scheme, and the application of a "bag-of-words" model across RNA secondary structures leads to a small number of word structures that have non-zero coefficients.

Continuous distributed representations

The alignment-free approach shows promising features, including the ability to mirror the standard alignment methods in recognizing the nearest neighbours of a long sequence. The ideal behaviour was to be halfway between the classic phylogenetic trees and new information, and the tree distance used to assess the best model is crucial at this step. The Robinson-Foulds distance may have been too generic to grasp the details needed to be used as an objective function (the best model in this work is the one which minimizes the RF distance), and while *easier* features are present, like the nearness to SARS-Co-V, bat and pangolin, more subtle similarities are still not caught. The reason for this, in our opinion, should be searched in the human made choices (*e.g.* objective functions and hyperparameter search strategies), not in the method itself, which has yielded promising results, mirroring classical results with an alignment-free approach.

Predicted epitope feature space

The cursory analysis of the feature space of predicted epitopes suggests two salient results. First, it suggests that there is a great deal of variation in immunopresentation by HLA allele. Secondly, it suggests that viral protein subunits have very different degrees of variation in how they are presented, with respect to epitopes and predicted binding affinity. Both of these results, if confirmed, would have profound implications for vaccine design. Further exploration is warranted. Avenues for improvement would include using HLA allele-specific thresholds for binding prediction and removing epitopes with high similarity to known self-peptides.

Future work

This report reflect the work initiated within the COVID-19 BioHackathon of April 2020 and mostly performed during the event itself. However, there are several interesting outcomes, and the authors are committed in further pursuing them. Specifically: - the report will be considered for a submission to a conference - the code produced in the context of this effort will continue to be developed, aiming for a full automated toolkit for feature extraction.

GitHub repository, Jupyter notebooks, tools and data repositories

GitHub repository

All the work presented here is available in our GitHub repository under the MIT license.

Analysis of ORF1ab dataset

- A dashboard for exploring ORF1ab dataset.
  - data
  - metadata
  - Dashboard notebook
- K-mer feature extraction at the aminoacid level, based on AA frequencies. Each dataset consists of 1 - 9-mers, and each K-mer has a corresponding `class, feature, weight` table, a `prediction` table and a classification report containing `F1, Precision, Rec all` and averaged metrics for that specific classification task.
  - Species level classification and feature extraction results.

- Host level classification and feature extraction results.
- Geographic location level classification and feature extraction results.
- Extraction source level classification and feature extraction results.

Acknowledgements

This work was done within the COVID-19 BioHackathon of April 2020.

References

Asgari, E., & Mofrad, M. R. K. (2015a). Continuous distributed representation of biological sequences for deep proteomics and genomics. (F. H. Kobeissy, Ed.)*PLOS ONE*, *10*, e0141287.

Asgari, E., & Mofrad, M. R. K. (2015b). ProtVec: A continuous distributed representation of biological sequences. *PLOS ONE*. Retrieved from https://arxiv.org/abs/1503.05140

Backert, L., & Kohlbacher, O. (2015). Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Med*, *7*, 119.

Boley, P. A., Alhamo, M. A., Lossie, G., Yadav, K. K., Vasquez-Lee, M., Saif, L. J., & Kenney, S. P. (2020). Porcine deltacoronavirus infection and transmission in poultry, united states1. *Emerging Infectious Diseases*, *26*, 255–265.

Cheng, S., Zhang, L., Tan, J., Gong, W., Li, C., & Zhang, X. (2019). DM-rpis: Predicting ncRNA-protein interactions using stacked ensembling strategy. *Computational Biology and Chemistry*, *83*, 107088. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/31330489

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. doi:10.1093/bioinformatics/btp163

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. Retrieved from https://arxiv.org/abs/1810.04805

Farrell, D. (2019). Dmnfarrell/epitopepredict: V0.4.0. doi:10.5281/zenodo.2562235

Felsenstein, J. (1988). Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics*, *22*, 521–565. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/3071258

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., & al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*. doi:10.1016/S0140-6736(20)30183-5

Kimothi, D., Soni, A., Biyani, P., & Hogan, J. M. (2016). Distributed representations for biological sequence analysis. Retrieved from https://arxiv.org/abs/1608.05949

Lam, T. T.-Y., Shum, M. H.-H., Zhu, H.-C., Tong, Y.-G., Ni, X.-B., Liao, Y.-S., Wei, W., et al. (2020). Identifying sars-cov-2 related coronaviruses in malayan pangolins. *Nature*.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *CoRR*, *abs/1405.4053*. Retrieved from http://arxiv.org/abs/1405.4053

Letunic, I., & Bork, P. (2019). Interactive tree of life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*, *47*, W256–W259.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and

projection for dimension reduction.

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, *3*(29), 861. doi:10.21105/joss.00861

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from https://arxiv.org/abs/1301.3781

Ng, P. (2017). Dna2vec: Consistent vector representations of variable-length k-mers. Retrieved from https://arxiv.org/abs/1701.06279

Organization, W. H. (2020a). Coronavirus disease 2019 (COVID-19): Situation Report – 1. Retrieved from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10_4

Organization, W. H. (2020b). WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. Retrieved from https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

Organization, W. H. (2020c). Coronavirus disease 2019 (COVID-19): Situation Report – 78. Retrieved from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200407-sitrep-78-covid-19.pdf?sfvrsn=bc43e1b_2

Palatnik-de-Sousa, C. B., Soares, I. da S., & Rosa, D. S. (2018). Editorial: Epitope discovery and synthetic vaccine design. *Frontiers in Immunology*, *9*, 826. doi:10.3389/fimmu.2018.00826

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. Retrieved from https://arxiv.org/abs/1802.05365

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. Retrieved from https://is.muni.cz/publication/884893/en

Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W. M., Melief, C. J., et al. (1994). The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J. Immunol.*, *153*(12), 5586–5592.

Shao, X. M., Bhattacharya, R., Huang, J., Sivakumar, I. A., Tokheim, C., Zheng, L., Hirsch, D., et al. (2019). High-throughput prediction of mhc class i and ii neoantigens with mhcnuggets. *Cancer Immunology Research*. doi:10.1158/2326-6066.CIR-19-0464

Sievers, F., & Higgins, D. G. (2013). Clustal omega, accurate alignment of very large numbers of sequences. *Methods in Molecular Biology*, 105–116.

Xu, Y., Song, J., Wilson, C., & Whisstock, J. C. (2018). PhosContext2vec: A distributed representation of residue-level sequence contexts and its application to general and kinase-specific phosphorylation site prediction. *Scientific Reports*, *8*.

Yi, H.-C., You, Z.-H., Cheng, L., Zhou, X., Jiang, T.-H., Li, X., & Wang, Y.-B. (2020). Learning distributed representations of rna and protein sequences and its application for predicting lncRNA-protein interactions. *Computational and Structural Biotechnology Journal*, *18*, 20–26. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6926125/