Optimal balance between computational cost and classification accuracy for SARS-CoV-2 lineages using Natural-Vectors-based methods

Roberto Cahuantzi[1] **Thomas House**[1] **Ian Hall**[1] **Katrina Lythgoe**[2] **Lorenzo Pellis**[1]

[1]Department of Mathematics, University of Manchester [2]Big Data Institute, University of Oxford

Introduction

The SARS-CoV-2 variants of concern and interest (VOC/VOI) can have major impacts on the global epidemiological situation. Therefore, we need to efficiently identify and monitor emerging variants. State-of-the-art phylogenetic methods are the gold standard, but are computationally expensive. One alternative is to use of Natural vectors (NV) [1], which characterise genetic sequences as lower dimensional arrays. We explore the extraction of NV-based features (NVf) to augment more traditional methods, by applying them to machine-learning (ML) classification and dimension reduction (DR) algorithms.

Methods

- Download of around 150,000 genetic sequences from the GISAID database procuring an appropriate representation of VOC/VOI and even distribution through time;
- Application of method of aligning sequences and nucleotide extraction from specific locations (loci) to be used as one-hot features on traditional algorithms: random forest (rf) and decision trees (dt);
- Classification of sequences using Pangolin [6], to extract labels from the Scorpio reported lineages.
- Extraction of NVf. They differ in the details, but can be interpreted as an array describing the distribution of an element $\epsilon \in \Gamma$ in a sequence S of length n, generalised as:

$$NV(S) = (n_{\epsilon_1}, \mu_{\epsilon_1}, D_2^{\epsilon_1}, ...n_{\epsilon_n}, \mu_{\epsilon_n}, D_2^{\epsilon_n})$$

The count of ϵ (n_ϵ), mean distance of ϵ to the origin (μ_ϵ) and variance of said distance (D_2^ϵ), characterise its distribution within S, these three magnitudes can be defined as:

$$n_\epsilon = \sum_{i=0}^{n} w_\epsilon(s_i), \quad \mu_\epsilon = \sum_{i=0}^{n} i\frac{w_\epsilon(s_i)}{n_\epsilon}, \quad D_2^\epsilon = \sum_{i=0}^{n} \frac{(i-\mu_\epsilon)^2 w_\epsilon(s_i)}{n_\epsilon n}$$

Where $s_i \in S$, and w_ϵ is a weight of 1 if $\epsilon = s_i$. NV is a 12-D array, since $\epsilon \in \{A, C, G, T\}$. Accumulated NV (ANV) [2] adds the covariance of the accumulation of nucleotides through S, degenerate bases NV (WMRV) [3] maps S by pairs of degenerate bases, namely $\epsilon \in \{W, S, M, K, R, Y\}$, k-mers NV (KMNV) [8] describes k-mer distribution, and extended NV (ENV) [5] characterise the intensities of the pixels in a 2-dimensional frequency chaos representation of S, in which $\epsilon \in \{0, ...255\}$. Additionally, the method of k-mer counts (kmc) was also explored as a benchmark. The value k was set to 3 for all k-dependant algorithms;

- Exploration of ML algorithms to evaluate trade-offs between accuracy and processing costs, these were: rf, dt, support vector machines (svm) and k-nearest neighbors (knn);
- DR projection for the different NVf were produced to observe structures in these genetic spaces. From these the novel approach PaCMAP [7] showed to be robust for replication.

Results

The use of dt and rf using loci demonstrated high accuracy at low processing cost in comparison to other relatively successful features-algorithm pairs like KMNV-dt or kmc-rf. The algorithm with highest accuracy was svm, but was computationally expensive. The pair kmc-knn showed less costs and reasonable accuracy, see figure 1.

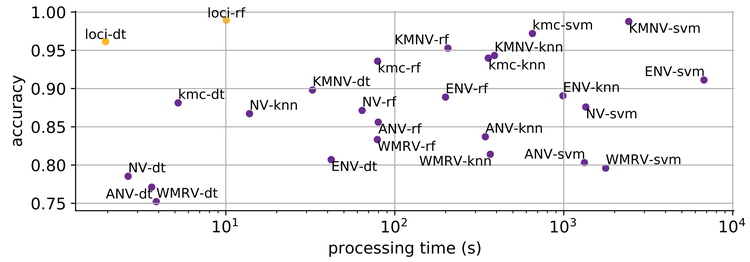

Figure 1: Accuracy vs processing costs for different features and ML algorithms.

Analysing the DR projection, two NVf were particularly interesting. One of these was ENV, see figure 2, here it is possible to see the formation of clusters among significant VOC such as Alpha (B.1.1.7) and Delta (B.1.617.2 and AY.4.x) and the so-called Delta+ (AY.4.2), whereas most of the least remarkable variants are clustered together.

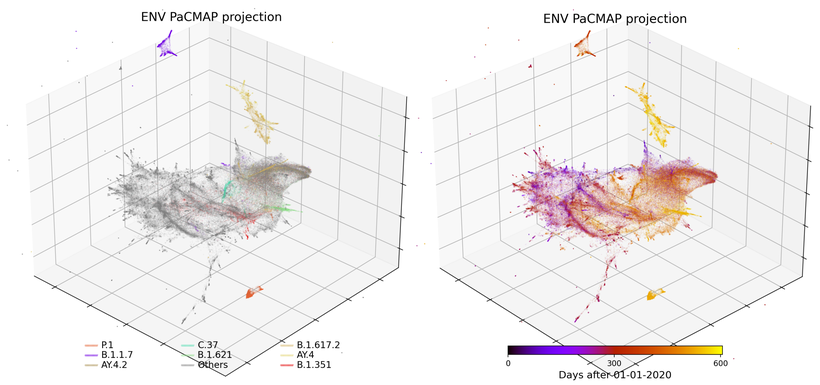

Figure 2: PaCMAP projection of ENV feature, (left) colorised by VOC/VOI, (right) colorised by day of sequencing.

A pipeline assembling this feature, PaCMAP and an innovative clustering algorithm like HDBSCAN [4], could a generate a tool to guard this genetic space and recognise the appearances of new clusters, warning us of emerging VOC/VOI. Another feature that showed interesting projections was the kmc, see figure 3. This projection showed a pattern similar of a branching tree, which could hint at the phylogenic relationships among the viruses. Additionally, it also forms remarkable clusters of VOC/VOI. The combination of these NVf could yield a more reliable clustering and classification.

Results (continuation)

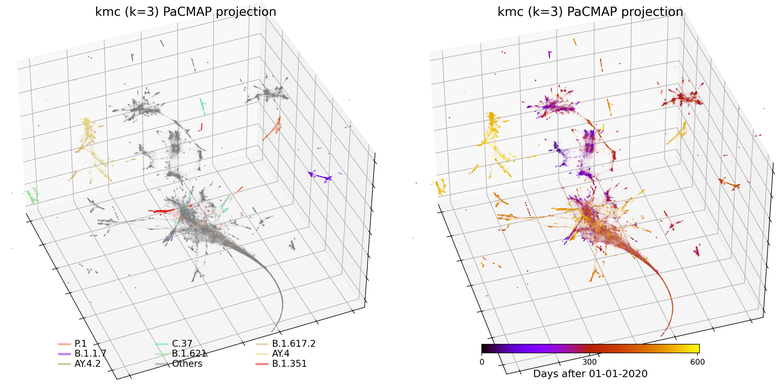

Figure 3: PaCMAP projection of kmc feature, (left) colorised by VOC/VOI, (right) colorised by day of sequencing

Conclusions and Future work

- **Acceleration of identification of new VOC/VOI** by applying NVf, further exploration of hyperparameters, specific sequence range application, DR preprocessing, and mixing of NVf is required;
- **DR projections might show phylogenic relations** a deeper phylogenic analysis of these sequences could clarify the origin of the appearing structures;
- **Guarding of genetic spaces** by the application of pipelines combining NVf, PaCMAP and HDBSCAN (or alike) we could become quickly aware of the emergence of new VOC/VOI.

References

[1] Mo Deng et al. "A novel method of characterizing genetic sequences: genome space with biological distance and applications". In: *PloS one* 6 (2011), e17293.

[2] Rui Dong et al. "A Novel Approach to Clustering Genome Sequences Using Inter-nucleotide Covariance". In: *Frontiers in Genetics* 10 (2019), p. 234.

[3] Yongkun Li et al. "A novel fast vector method for genetic sequence comparison". In: *Scientific Reports* 7 (2017), p. 12226.

[4] Leland McInnes and John Healy. "Accelerated Hierarchical Density Based Clustering". In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 2017, pp. 33–42.

[5] Shaojun Pei et al. "Fast and accurate genome comparison using genome images: The Extended Natural Vector Method". In: *Molecular Phylogenetics and Evolution* 141 (2019), p. 106633.

[6] Andrew Rambaut et al. "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology". In: *Nature Microbiology* 5.11 (2020), pp. 1403–1407.

[7] Yingfan Wang et al. "Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization". In: *Journal of Machine Learning Research* 22.201 (2021), pp. 1–73.

[8] Jia Wen et al. "K-mer natural vector and its application to the phylogenetic analysis of genetic sequences". In: *Gene* 546.1 (2014), pp. 25–34.

Contact email:
roberto.cahuantzi@manchester.ac.uk