Genetic space exploration: Overview of low dimension projections of alignment-free characterisation methods applied to SARS-CoV-2

Roberto Cahuantzi[1] **Thomas House[1]** **Ian Hall[1]** **Katrina Lythgoe[2, 3]** **Lorenzo Pellis[1]**

[1]Department of Mathematics, University of Manchester [2]Department of Biology, University of Oxford [3]Big Data Institute, University of Oxford

Introduction

The ongoing epidemic has showed us the necessity to understand better the dynamics of viral evolution to identify variants of concern and variants of interest (VoC/VoI). The common methods of phylogenetical analysis are computationally expensive and struggle to incorporate new data, rely on multi-aligned sequences, performing a pair-wise comparison among all the sampled genomes. Alignment-free methods offer an alternative, and among them those based on word-statistics methods offer an individual representation as a single point in a $\mathbb{R}^n$ space for each genetic code, which allows for easier incorporation of new information. This project analyses approximately 250k SARS-CoV-2 sequences extracted from the GISAID database, characterising them with the aforementioned featuring methods to apply the dimensionality reduction (DR) algorithm PaCMAP[8], to visualise the structures formed within the different genetic spaces. This gives multiple perspectives to understand better the virus ecology. Furthermore, the use of clustering methods, like HDBSCAN[3] and CLASSIX[1] , shows to be a valuable tool to confirm the appearance of clusters formed by new VoC/VoI.

Methods

- Download of around 250k genetic sequences from the GISAID database procuring an appropriate representation of VoC/VoI and even distribution through time;
- Classification of sequences using Pangolin, labelling them through the Scorpio reported lineages;
- Extraction of natural vectors features (NVf). They differ in the details, but can be interpreted as an array of summary statistics of an element $\epsilon \in E$ in a sequence S of length n, generalised as:

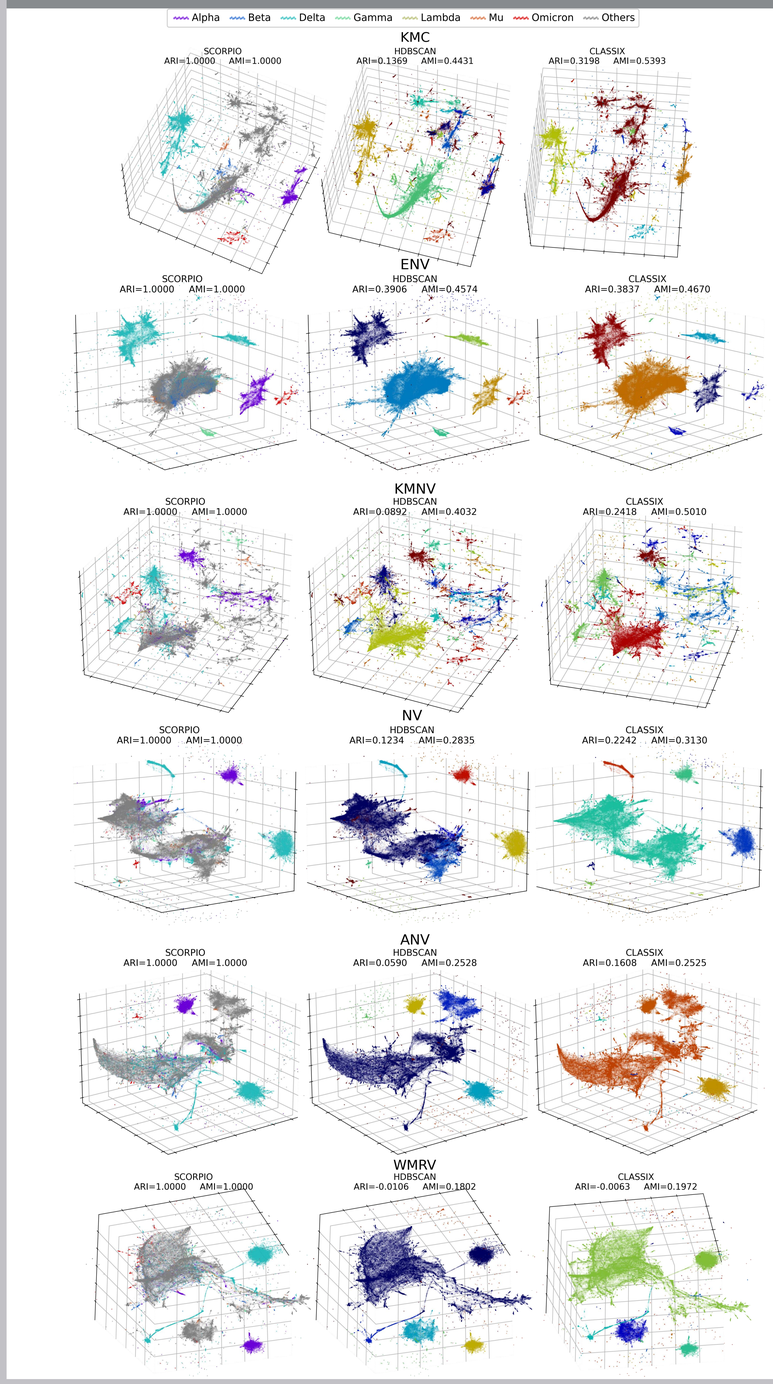
$$NVf(S) = \left(n_{\epsilon_1}, \mu_{\epsilon_1}, D_2^{\epsilon_1}, \dots n_{\epsilon_n}, \mu_{\epsilon_n}, D_2^{\epsilon_n}\right)$$

The count of (n_ϵ), mean distance of 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:

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

Where $s_i \in S$, and w_ϵ is a weight of 1 if $\epsilon = s_i$. Among the explored NVf we can mention: natural vector (NV)[5], which is a 12-D array, since $\epsilon \in \{A,C,G,T\}$; Accumulated NV (ANV)[4] adds the covariance of the accumulation of nucleotides through S; degenerate bases NV ($WMRV$)[7] maps S by pairs of degenerate bases, namely $\epsilon \in \{W,S,M,K,R,Y\}$; k-mers NV ($KMNV$)[2] describes k-mer distribution; and extended NV (ENV)[6] characterise the intensities of the pixels in a 2-d frequency chaos representation of S, in which $\epsilon \in \{0, \dots 255\}$. Additionally, the method of k-mer counts (KMC) was also included among the NVf. The value k was set to 3 for all k-dependant algorithms;
- DR projection for the different NVf were produced to observe structures in these genetic spaces. From these the PaCMAP[8] method proved to be robust for replication, while reliable on representing global and local structures.
- Two clustering algorithms were applied and evaluated, HDBSCAN[3] and CLASSIX[1], both yielded similar and consistent results according feature projected.

Results



Results (cont.)

- The figure (left) shows the formation of structures and clusters on the genetic spaces that in most of them correspond to VoC/VoI,
- The unsupervised algorithms, HDBSCAN and CLASSIX, are able to detect the formation of these clusters,
- The best NVf, by ARI and AMI metrics, appear to be KMC and ENV, while the worst $WMRV$, results are summarised on right table.

Feature	HDBSCAN		CLASSIX	
	ARI	AMI	ARI	AMI
KMC	0.1369	0.4431	0.3198	0.5393
ENV	0.3906	0.4574	0.3837	0.4670
$KMNV$	0.0892	0.4032	0.2418	0.5010
NV	0.1234	0.2835	0.2242	0.3130
ANV	0.0590	0.2528	0.1608	0.2525
$WMRV$	-0.0106	0.1802	-0.0063	0.1972

Conclusions and Future work

- The low dimension projection of most of NVf seem to rescue important information to identify new variants through clustering methods,
- Monitoring of emergence of clusters could make possible to police the emergence of variants of concern,
- Simple KMC was the one of the bests NVf for clustering formation followed by ENV, modification on this methods could improve their accuracy and sensibility
- The application of different featuring methods, such as: protein NV[9], graphical representation[10], of Fourier power spectrum should be investigated[11],
- It would be also of interest to assess different metrics of distance like: Kull-Leiber, Yaus-Hausdorff[10], among others; to explore possibilities.

References

[1] X. Chen and S. Guettel, 'Fast and explainable clustering based on sorting', 2022, doi: 10.48550/ARXIV.2202.01456.
[2] J. Wen, et al., 'K-mer natural vector and its application to the phylogenetic analysis of genetic sequences', Gene, vol. 546, 2014.
[3] L. McInnes and J. Healy, 'Accelerated hierarchical density based clustering', Nov. 2017.
[4] R. Dong et al., 'A novel approach to clustering genome sequences using inter-nucleotide covariance', Frontiers in Genetics, vol. 10, 2019.
[5] M. Deng, et al., 'A novel method of characterizing genetic sequences: Genome space with biological distance and applications', PLoS ONE, vol. 6, 2011.
[6] S. Pei, et al., 'Fast and accurate genome comparison using genome images: The Extended Natural Vector Method', Molecular Phylogenetics and Evolution, vol. 141, 2019.
[7] Y. Li, et al., 'A novel fast vector method for genetic sequence comparison', Scientific Reports, vol. 7, 2017.
[8] Y. Wang, et al., 'Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization', Journal of Machine Learning Research, vol. 22, 2021.
[9] C. Yu, et al., 'Protein space: A natural method for realizing the nature of protein universe', Journal of Theoretical Biology, vol. 318, 2013
[10] K Tian, et al., 'Two Dimensional Yau-Hausdorff Distance with Applications on Comparison of 417 DNA and Protein Sequences', PLoS ONE, vol. 10, 2015.
[11] C. Yin, et al., 'A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering', Journal of Theoretical Biology, vol. 359, 2014.

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