Unsupervised machine learning framework for discriminating major variants of concern during COVID-19

Rohitash Chandra

School of Mathematics and Statistics
UNSW Sydney

Seminar: Monash University

Overview

- Data
- Unsupervised machine learning methods
- Framework
- Results

COVID-19

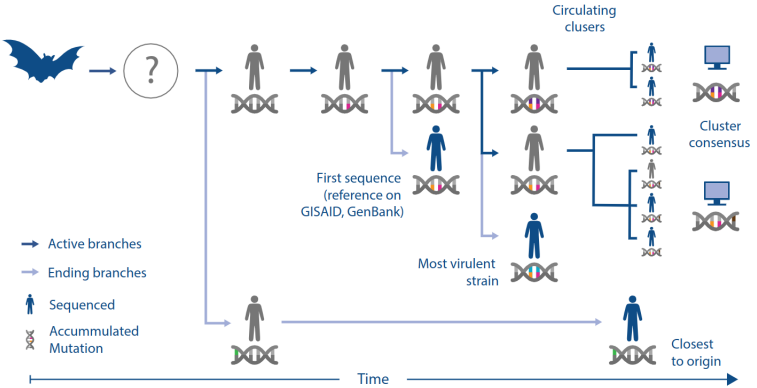


Figure: Bauer et. al (2020). Supporting pandemic response using genomics and bioinformatics: A case study on the emergent SARS-CoV-2 outbreak. Transboundary and emerging diseases, 67(4), 1453-1462.

Data

- GISAID (global initiative on sharing Avian influenza data) is recognised as a reliable portal for prompt sharing of COVID-19 data.
- Currently, GISAID is the largest publicly accessible platform, consisting of sequences and associated epidemiological data of over 12.1 million SARS-CoV-2 strains [1].
- We extracted 250 randomly selected SARS-CoV-2 isolates of complete genome sequences of human origins from GISAID on 8 July 2022.
- We note that five variants (Alpha, Beta, Gamma, Delta, and Omicron) featured 50 genome sequences each.

UNSW SYDNEY

[1] https://www.gisaid.org/hcov19-variants/

Data

Table below presents top 7 countries based on number of genome isolates for the based on the selected variants across the globe.

Country	Number of Occurrences	Number of Variants
United States	Alpha(16), Beta(5), Delta (12),Gamma(9), Omicron(18)	5
Brazil	Alpha(3), Gamma(17)	2
South Africa	Alpha(2), Beta(10) Delta(1), Omicron(1)	4
France	Beta(6), Delta(3) Gamma(2)	4
Belgium	Alpha(2), Beta(2) Delta(4), Omicron(4)	4
Canada	Alpha(4), Beta(2) Delta(2), Gamma(3)	4
Malaysia	Alpha(1), Beta(3) Delta(2)	3

K-mer Analysis

K-mers are substrings of length k contained within a biological sequence such as a DNA sequence; hence, k-mer analysis is done to calculate the frequency of fixed-length words of a sequence.
A "k-mer" refers to all of a sequence's substring of length k; for instance, the sequence "ATGG" would have four monomers (A, T, G, and G), three 2-mers (AT, TG, GG), two 3-mers (ATG and TGG), and one 4-mer (ATGG).

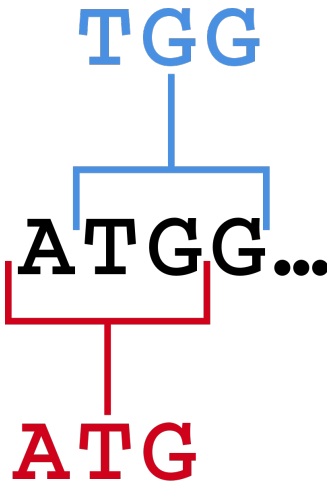


Figure: kmer example. Source: Wikipedia

Dimensional Reduction:PCA

PCA is a dimensional reduction method extensively used in various forms of data reduction, data analysis, and data visualisation which applications in computer graphics, machine learning for reducing over-fitting and model complexity.

Dimensional Reduction:PCA

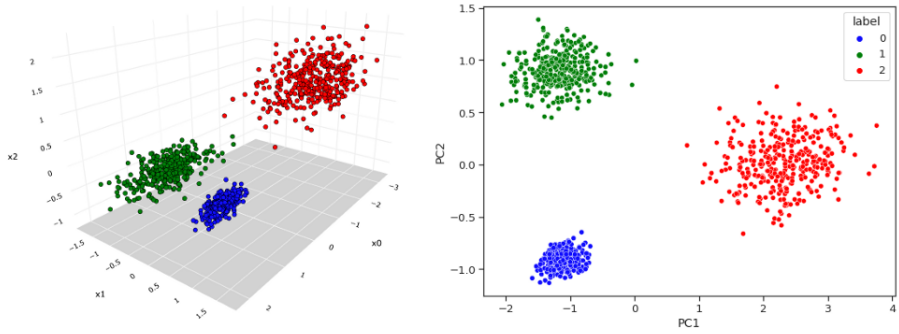


Figure: (Left) The original data and (Right) the same data reduced to 2-D with P

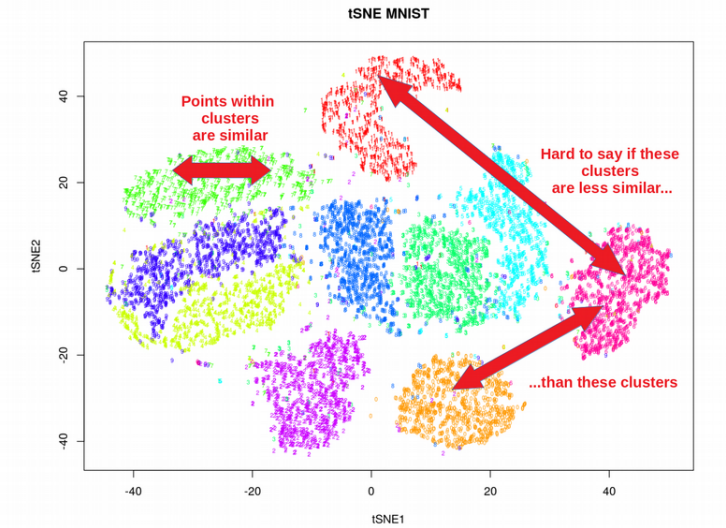


Figure: tSNE - MNIST Character recognition data.

Dimensional Reduction:tSNE

t-SNE is a nonlinear dimensionality reduction method which is also used for visualisation of high-dimensional data into a low-dimensional space of two or three dimensions.

t-SNE is an extension of stochastic neighbor embedding (SNE) with two key modifications that include a student t-distribution rather than a Gaussian and a symmetrical form of the SNE cost function with basic gradients.

t-SNE has been widely used in the domain of medicine and bioinformatics such as molecular dynamics simulations

tSNE can practically only embed into 2 or 3 dimensions, i.e. only for visualization purposes, so it is hard to use tSNE as a general dimension reduction techniques

tSNE performs a non-parametric mapping from high to low dimensions, meaning that it does not leverage features (aka PCA loadings) that drive the observed clustering.

tSNE can not work with high-dimensional data directly, Autoencoder or PCA are often used for performing a pre-dimensionality reduction before plugging it into the tSNE

Dimensional Reduction:UMAP

UMAP is a manifold learning approach for dimension reduction which employs a conceptual structure according to the Riemannian geometry and algebraic topology. UMAP has been comparable to t-SNE in terms of visualization quality, and potentially retain more global structure with less computational time. Additionally, UMAP does not have computational restriction on the dimension of embedding, enabling it to be practical as a dimension reduction approach for various problems.

Dimensional Reduction:UMAP

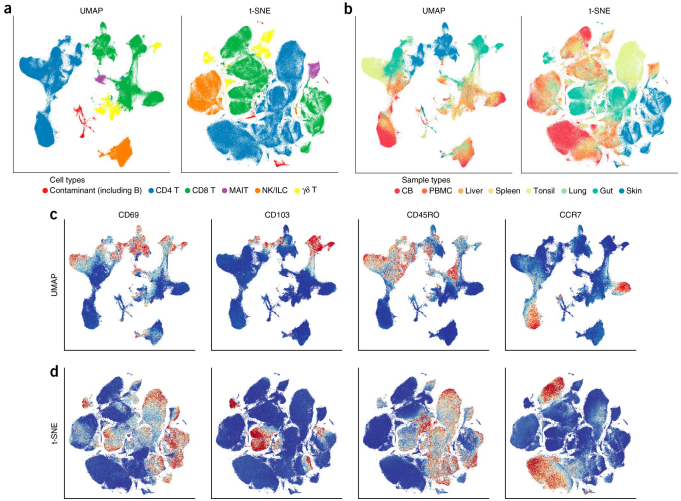
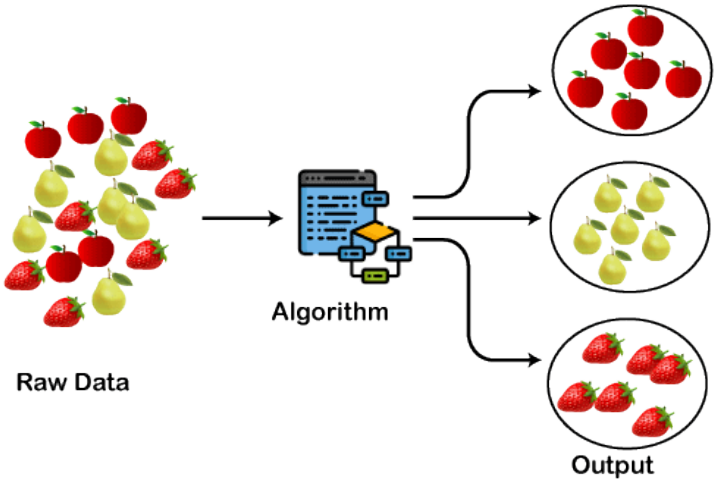


Figure: Visualizing single-cell data using UMAP, Nature Biotechnology, volume 37, 38–44 (2019)

Clustering



Clustering:AGNES

Hierarchical agglomerative clustering, also known as agglomerative nesting (AGNES) provides a better approach by addressing the problem of k-means clustering where k needs to be manually tuned.
In an agglomerative clustering model, the clustering initiates with individual collections of every data point.
AGNES has been extensively used various medical domains, such as categorizing patients with severe aortic stenosis, and mapping molecular substructures.
However, AGNES has been ineffective in some problems since finding the nearest pair of clusters can be challenging when data is sparse and noisy.

Clustering

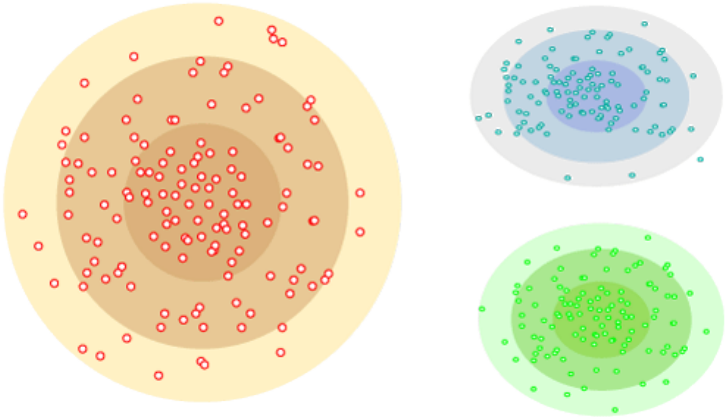


Figure: AGNES clustering

Clustering

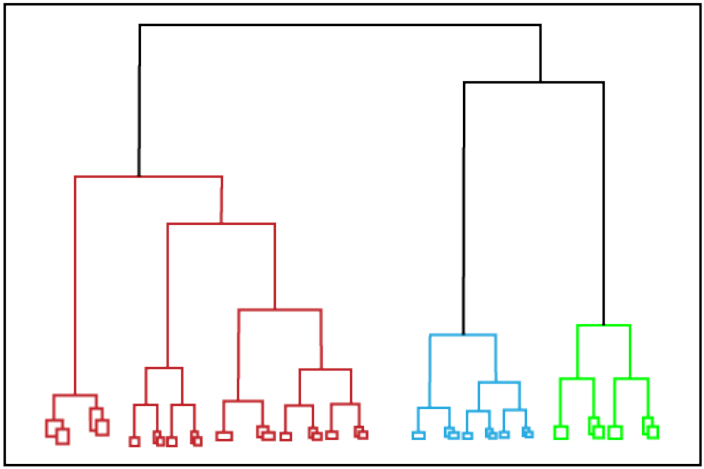
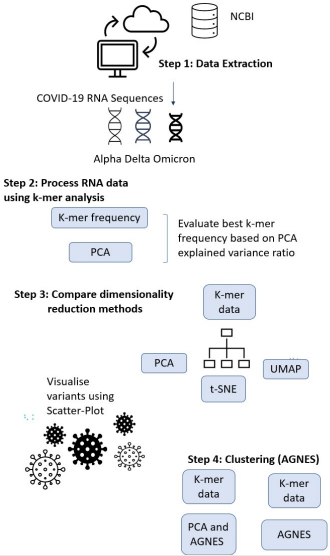


Figure: Dendogram

Framework


NCBI
Step 1: Data Extraction
COVID-19 RNA Sequences
Alpha Delta Omicron
Step 2: Process RNA data using k-mer analysis
K-mer frequency
PCA
Evaluate best k-mer frequency based on PCA explained variance ratio
Step 3: Compare dimensionality reduction methods
K-mer data
PCA
UMAP
t-SNE
Visualise variants using Scatter-Plot
Step 4: Clustering (AGNES)
K-mer data
K-mer data
PCA and AGNES
AGNES

Framework

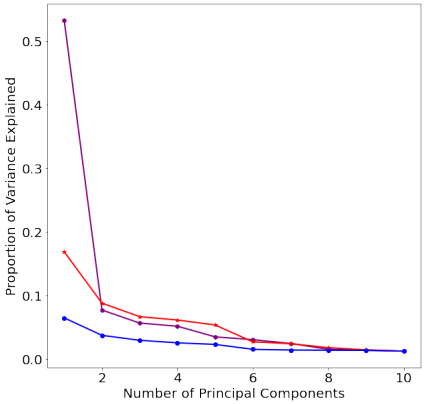


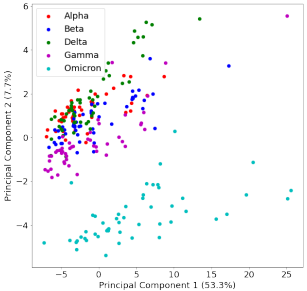
Figure: Scree-plot

Results

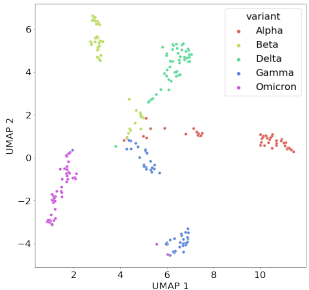
	PC1	PC2	PC3	PC4	PC5	Total
k = 3	0.5330	0.0774	0.0569	0.0519	0.0352	0.7544
k = 5	0.1690	0.0881	0.0670	0.0617	0.0538	0.4396
k = 7	0.06498	0.03754	0.0298	0.0258	0.0233	0.1814

Table: Explained variance ratio of top 5 principal components (PC) for selected values of k in k-mer analysis.

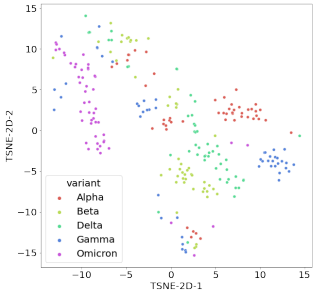
Results



(a) PCA



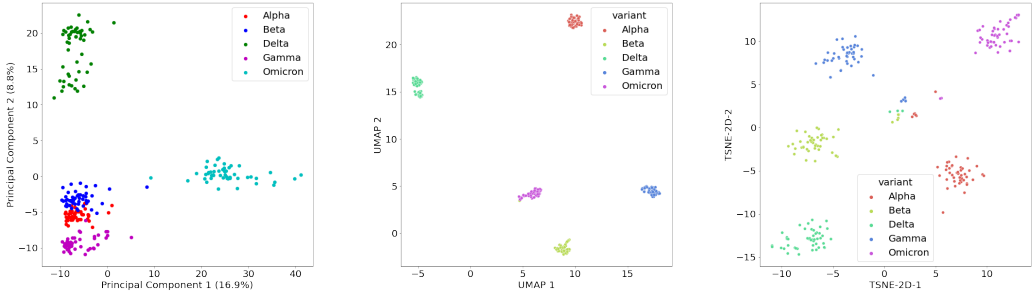
(b) UMAP



(c) tSNE

Figure: Comparison of different methods with k-mer analysis - k is 3

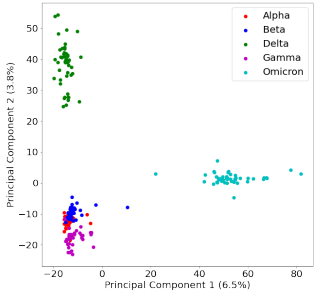
Results



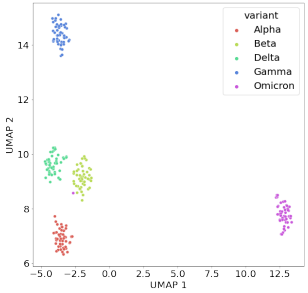
(a) PCA (b) UMAP (c) tSNE

Figure: Comparison of different methods with k-mer analysis - k is 5

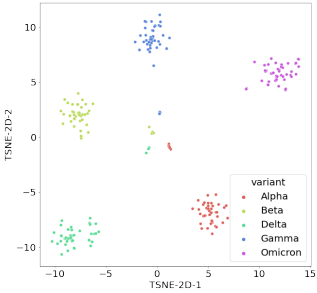
Results



(a) PCA



(b) UMAP



(c) tSNE

Figure: Comparison of different methods with k-mer analysis - k is 7

Results

	PCA	t-SNE	UMAP	Num. features
k = 3	0.0215	3.7273	0.2905	64
k = 5	0.0241	1.2987	0.3190	1024
k = 7	0.2475	1.5757	0.3269	16384

Table: Execution time (seconds) for selected values in k with different number of features in data via k-mer analysis.

Results

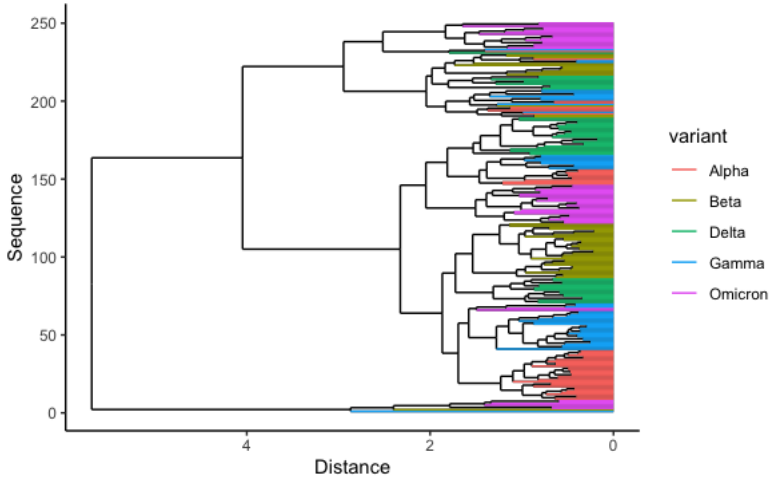


Figure: Dendrogram obtained from hierarchical clustering.

Results

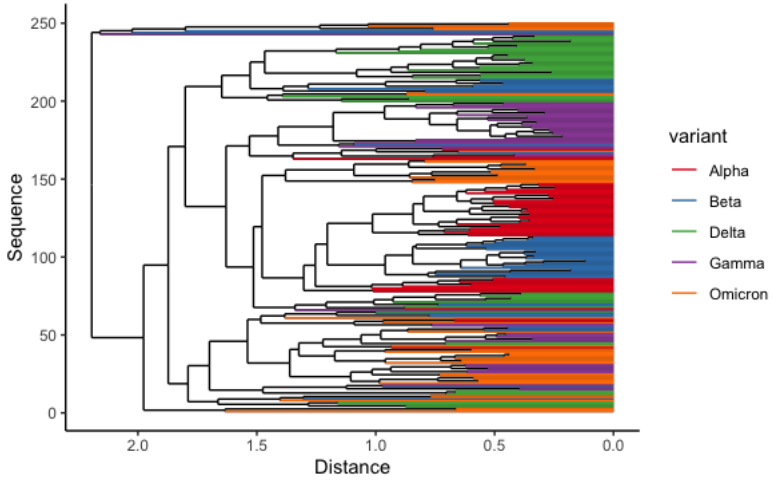


Figure: Dendrogram represents PCA-based 95% explained variance.

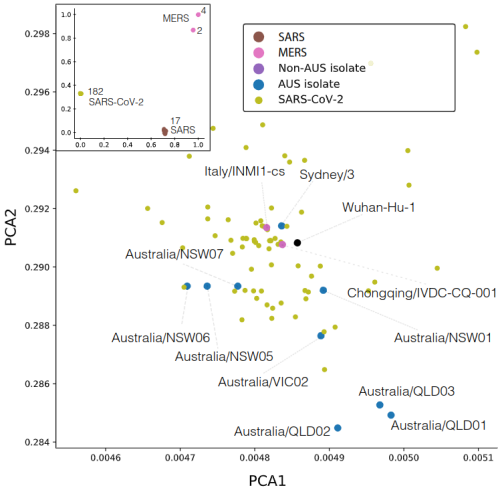


Figure: Bauer et. al (2020). Supporting pandemic response using genomics and bioinformatics: A case study on the emergent SARS-CoV-2 outbreak. Transboundary and emerging diseases, 67(4), 1453-1462.

Conclusions

- In our study, we evaluated different components of the framework with different parameter settings and found that UMAP provides the best dimensionalality reduction and visualisation tool for the genome sequences since it not only scales well given different variations in k-mer analysis, but also provides a visual representation with good computational time when compared to PCA and t-SNE.
- PCA on the other hand, provides further insights using explained variance ratio which in addition with UMAP gives a good overview of the data. We also note that it is reasonable not to go further than $k = 7$, which can take further computational time and storage during genome sequence pre-processing.

Software and Data Availability

- Please note all projects have open software and data published via Github.
- Kang, M., Vasan, S., Wilson, L. O., Chandra, R. (2022). Unsupervised machine learning framework for discriminating major variants of concern during COVID-19. arXiv preprint arXiv:2208.01439.
- https://github.com/sydney-machine-learning

UNSW SYDNEY

Acknowledgements

(a) Seshadri Vasan (Uni. of York)

ming.jpeg

(b) Mingyue Kang (UNSW)

(c) Laurence Wilson (CSIRO)