Viral Escape(SARS-CoV-2)

Kiana Seraj

March 2024

INTRODUCTION

Viruses have developed countless mechanisms over years to escape the immune system, while the biology of the host's immune system and the viruses are both complex, the governing principle underlying their interaction is simple-natural selection [5]. One common way viruses escape recognition by the immune system, including T cells, is through antigenic variation. Antigens are molecules on the surface of pathogens, such as viruses, that the immune system recognizes as foreign. T cells, a crucial component of the immune system, can identify and eliminate infected cells by recognizing specific viral antigens presented on the surface of these cells. By changing their antigenic profile, viruses can escape detection and elimination by the immune system, allowing them to persist and continue replicating within the host. The same happened for Corona virus, made the virus to spread around the world quickly and became more infectious. Most studies of genetic variation have focused on spike, but variant mutations outside spike are also key components in SARS-CoV-2's continued adaptation to human infection. Nucleocapsid mutations in SARS-CoV-2 augment replication and pathogenesis. For instance, it has been shown that the mutation of two amino acids in the Nucleocapsid protein, R203K+G204R mutation is sufficient to enhance replication, fitness, and pathogenesis of SARS-CoV-2, recreating a mutation found in the alpha and omicron variants in an early pandemic (WA-1) background[4]. Additionally, studies have shown the impact of envelope protein mutations on the pathogenicity of Omicron XBB [6]. The RdRp is critical for the replication of viral RNA, and also a promising drug target for COVID-19 treatment[3], hence, its mutation can have an important impact on covid drug design. Protein language models, trained on millions of biologically observed sequences, generates feature-rich numerical representations of protein sequences. Sequence embedding summarizes the entire sequence into a high-dimensional vector. Protein language models can be applied towards a wide variety of tasks such as secondary structure prediction, contact prediction, homology detection, and etc. A protein language model generates embedding vector of size (t, e) where each residue is represented by a token and the contextual information for each residue is encoded in (e) dimensions, and (t) represents the total number of residue tokens and special tokens. Language models learn the probability of a sequence occurring and this can be directly applied to predict the fitness of sequence mutations. Deep language models are an exciting breakthrough in protein sequence modeling, allowing us to discover aspects of structure and function from only the evolutionary relationships presented in a corpus of sequences. These models can be enriched with strong biological priors through multi-task learning [1]. In this report, with the embedded representation obtained from sequences, it was tried to find co-evolutionary patterns among SARS-CoV-2 proteins and to see how much precisely the model can predict the future mutations by giving likelihood values for each amino acids based on the context of the given sequence. As the embeddings are high dimensional, non linear methods should be applied in order to have this feature information in a lower dimension.

METHOD

0.1 Data preparation

In order to evaluate proteins' co-evolutionary effect, protein sequence datasets were divided into 3 groups based on their common strain names. 1.Spike protein and RdRP, 2.Spike protein and Nucleocapsid and 3.RdRP and Nucleocapsid. shown in the Table(1) below.

	Dataset size	Grouped with RdRP	Grouped with N	Grouped with S
Spike	2391	329	187	-
RdRP	841	-	91	329
Nucleocapsid	549	91	-	187

Table 1. Size of the used datasets

0.2 Sequence embedding

Protein bert is used for feature vectoring of the protein sequences[2].Sequence embeddings were generated using a pre-trained protein language model without fine-tuning. The dimension of the embedded vector, depending on the architecture of the model, here is 1024. Vector representation is contextualized, thanks to the self-attention mechanism. The obtained dimension ($1024 \times$sequence length), which is an amino acid-based representation of the sequence, is converted to a protein-based representation through mean pooling which gives a high-dimensional vector of length 1024 for each protein sequence.

0.3 Dimensionality reduction

A dataset with a large number of attributes, generally on the order of a hundred or more, is referred to as high-dimensional data. Some of the difficulties that come with high dimensional data manifest during analyzing or visualizing the data to identify patterns, and some manifest while training machine learning models. Therefore, dimensionality reduction is used to reduce the number of input variables in a dataset. Here, the 1024 dimensional features have been converted into 2 dimensions with UMAP. Additionally, TSNE was used for dimensionality reduction; however, UMAP grouped the clusters with clearer distances within the groups.

0.4 K-means Clustering

To identify unique clusters of data points from a data set, k-means clustering was used. Aiming to minimize the within-cluster variance, which measures how spread out the data points are within each cluster.

0.5 Mutation predictions with amino acid likelihoods

One way to study the effect of mutation is studying protein sequences and considering the amino acid changes happening in the sequences. In here, to predict the mutations, the predicted likelihoods (ap2prob feature of the dataset) by the model for Nucleocapsid sequence amino acids, collected in 2021, has been taken from the dataset, size of 191. A list of conserved and prone-to-mutation amino acids have been prepared. The predicted conserved amino acids are the amino acids which have been predicted by the model to have the likelihood of more than 85% and they have been predicted for more than 85% of the sequences of 2021 by the model. The prone to mutation amino acids are the amino acids which have been predicted to have the likelihood of less than 65% and they have been predicted for more than 85% of the sequences of 2021 by the model. Afterwards, MSA has been applied on the sequences of Nucleocapsid in the year of 2021-2023 with using Clustal Omega and pymsaviz softwares, to see the real mutational effects happening on the sequences. subsequently, the residue likelihoods given by the model for the data of 2021 were compared with the actual mutations happened from 2021-2023, to see how many of the amino acids that they have been predicted by the model to have low likelihoods have actually changed during the selection process from the year of 2021 to 2023.

RESULT

Based on the figures 3 and 2, Nucleocapsid protein embedding shows longer distance between two groups of mutation from 2021 to 2022 compared to RdRP which can be considered that the mutational effect on Nucleocapsid has changed its biological feature more.

In the year of 2023 it seems that no significant mutations happened for the Nucleocapsid protein and RdRP protein. No separate clusters can be recognized for the year of 2023, figures 2 and 3 .

Considering Spike protein, the main reason of the virus spread and being more infectious, it has shown that in each year the protein has more than one mutation as shown in the dimensional reduced embedding in fig 4.

To evaluate the co-evolution phenomena of Spike with RdRP and Nucleocapsid proteins, having the same strain names with each other. They have been colored by the same cluster Id obtained from Spike protein and vice versa. The coloring shows co-evolutionary patterns happening within spike and these 2 proteins figs(5, 6) and figs(7, 8). Based on the figs, the sequences of Nucleocapsid and RdRP in the same neighborhoods are colored by the same cluster Id of Spike protein, demonstrating that the sequences of spike being in multiple clusters also cause the N and RdRP proteins being in the same cluster Id as spike and vice versa.

Evaluating the co-evolutionary effect between Nucleocapsid and RdRP, considering the 3 clusters in the RdRP and Nucleocapsid, hence, coloring the each protein's embeddings with another protein's cluster Id in fig 9 and fig 10, illustrates that in the near sequences are colored similarly. This shows a co-evolutionary effect between these two proteins within the year of 2021-2023

In the prediction of the mutatiton effect, the MSA applied on the sequence data of Nucleocapsid, has shown 11 of amino acids having lower frequencies in the histogram than the rest, which are amino acids that were not conserved during the year of 2021-2023 and changed, fig 1. However, the model predicted only 4 of these prone to mutation amino acids to have lower than 65% likelihood, and this likelihood is predicted for more than 85% of the sequence data of 2021 given to the model(amount of 191 sequences). Meanwhile, the false positive of the model was also high which made the model to be not precise enough (precision less than 10%). Additionally, the model predicted 70% of the conserved amino acids having the likelihood of more than 85% (precision = 95%), predicted for more than 85% of the given sequences of the year 2021. The total result has been shown in the Table(2).

	real value	Predicted	TP	FP	precision
Conserved	409	305	290	15	95%
Mutated	11	48	4	44	8%

Table 2. Number of predicted mutated and conserved Nucleocapsid's amino acids by the model, total number of amino acids = 420

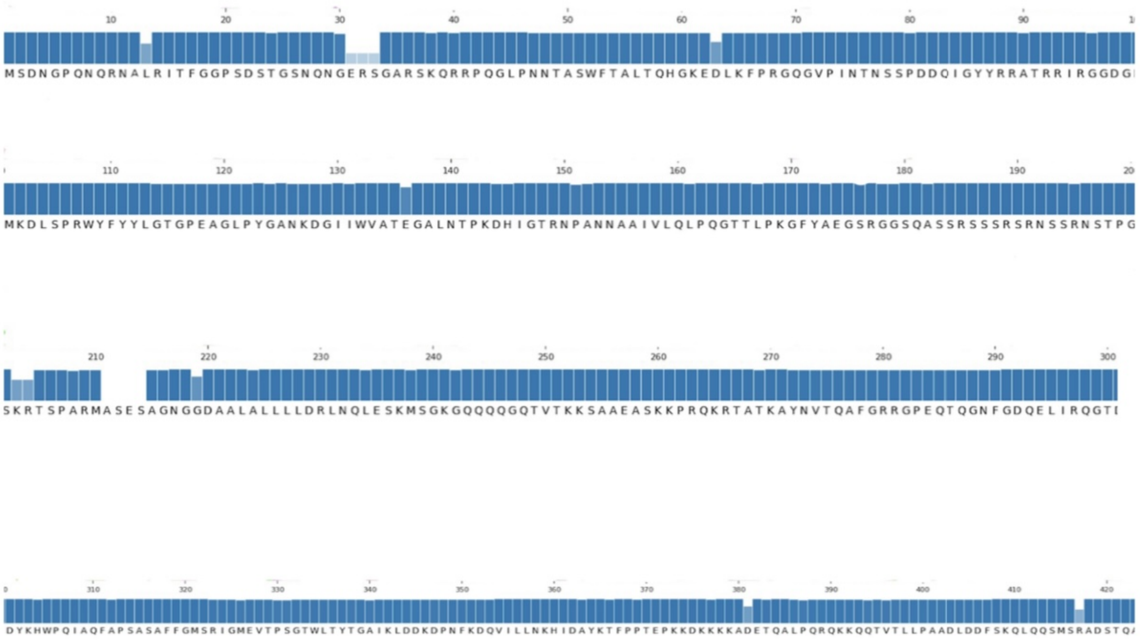

Figure 1. MSA applied on the 549 sequences of Nucleocapsid protein

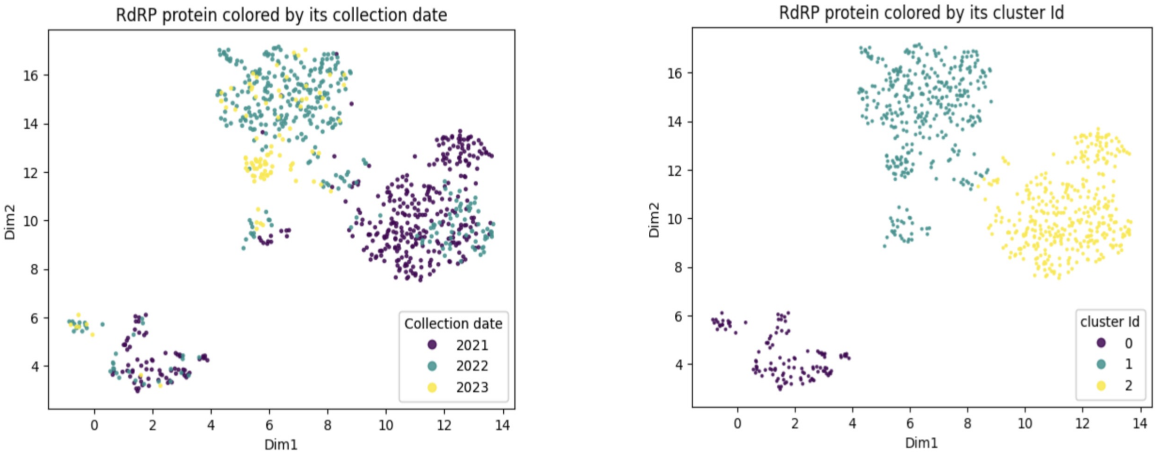

Figure 2. Clustering dimensional reduced embeddings of RdRP protein, dataset size=841

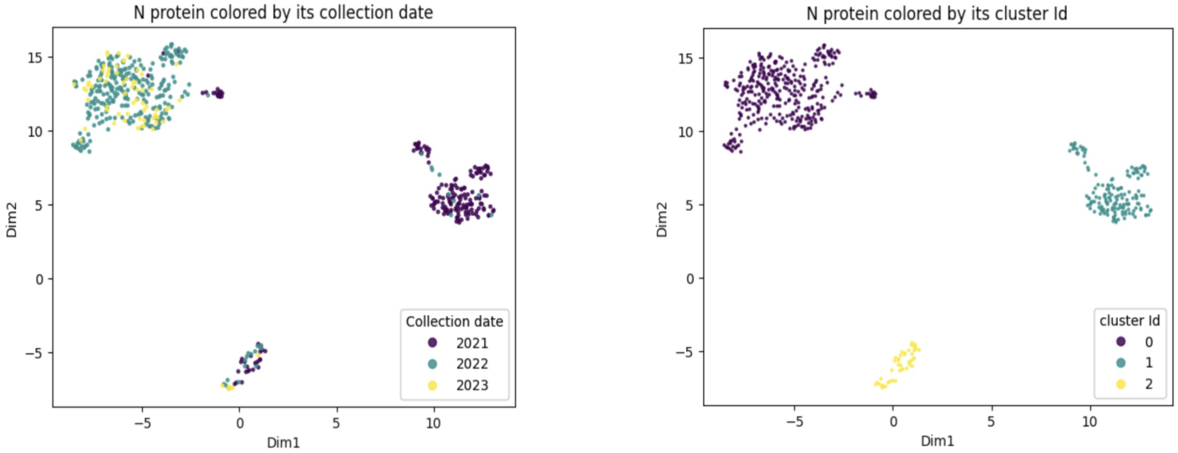

Figure 3. Clustering dimensional reduced embeddings of Nuncleocapsid protein, dataset size=549

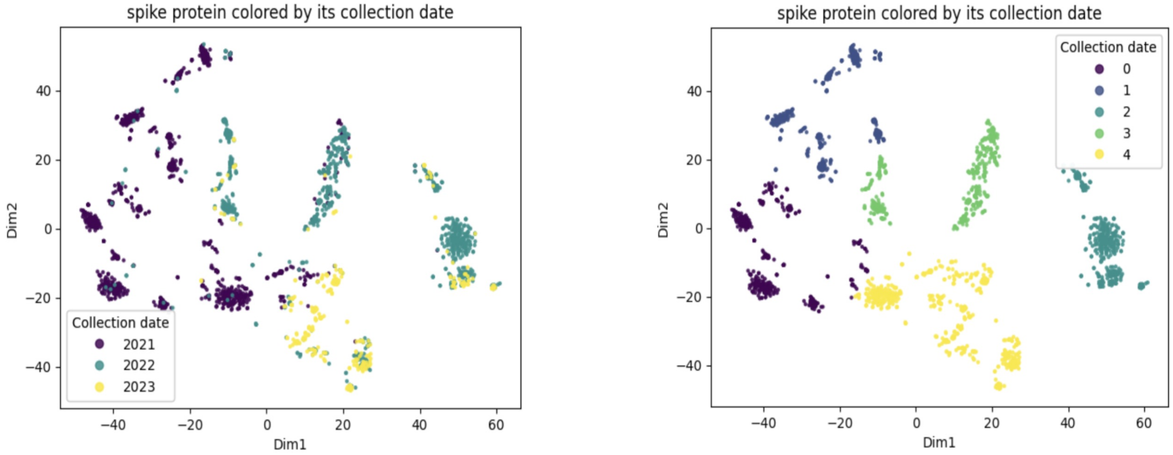

Figure 4. Clustering dimensional reduced embeddings of Spike protein, dataset size=2391

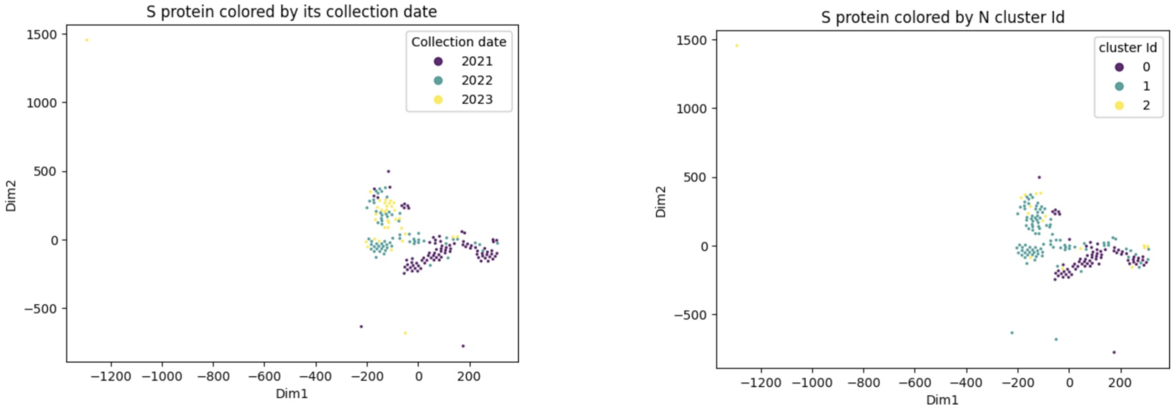

Figure 5. Plotting dimensioncal reduced embeddings of S Protein after being grouped by N protein based on their similar strain names, and being colored by N cluster Id, dataset size=187

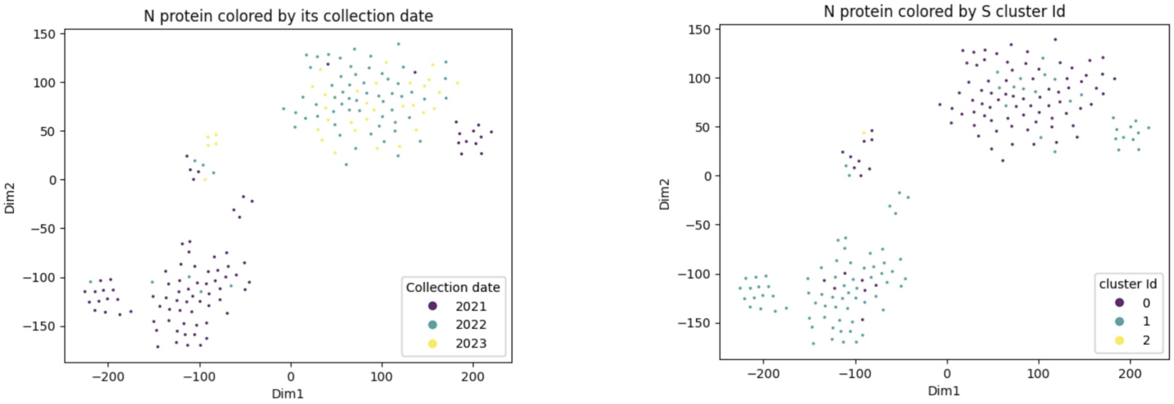

Figure 6. Plotting dimensioncal reduced embeddings of N Protein after being grouped by S protein based on their similar strain names, and being colored by S cluster Id, dataset size=187

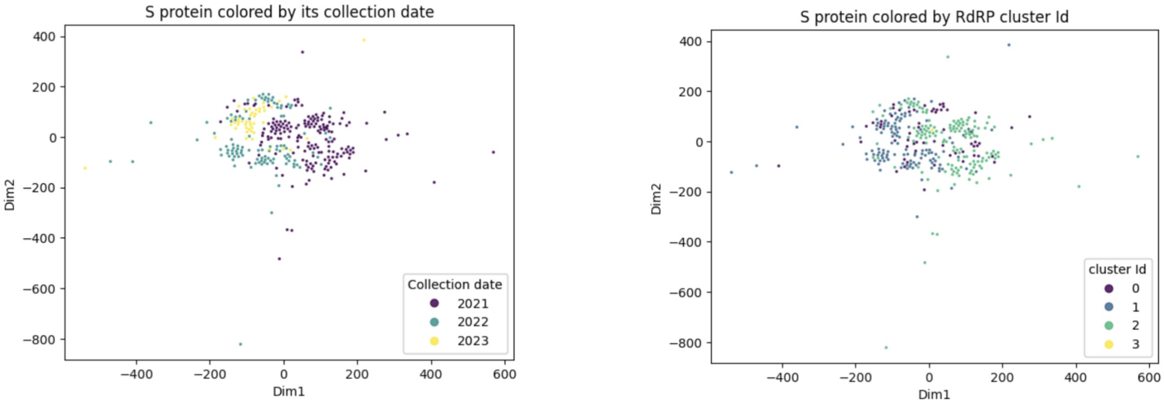

Figure 7. Plotting dimensioncal reduced embeddings of S Protein after being grouped by RdRP protein based on their similar strain names, and being colored by RdRP cluster Id, dataset size=329

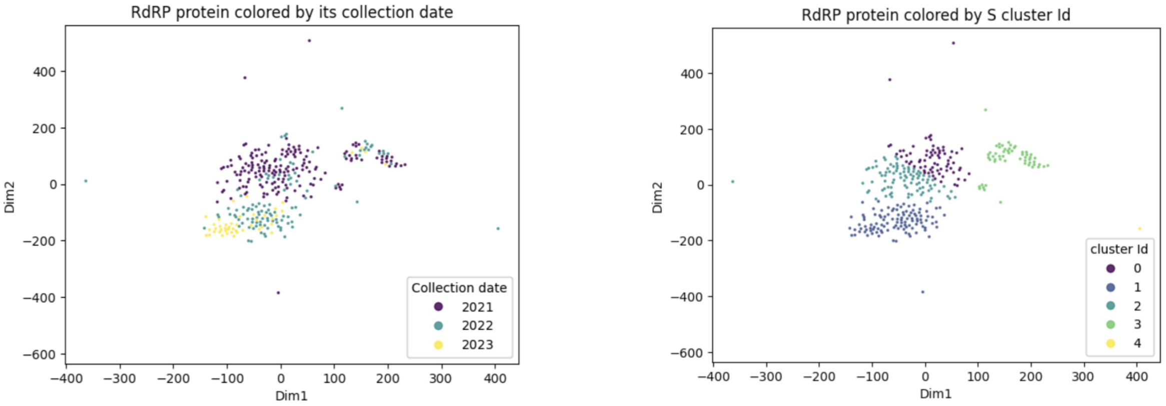

Figure 8. Plotting dimensioncal reduced embeddings of RdRP Protein after being grouped by S protein based on their similar strain names, and being colored by S cluster Id, dataset size=329

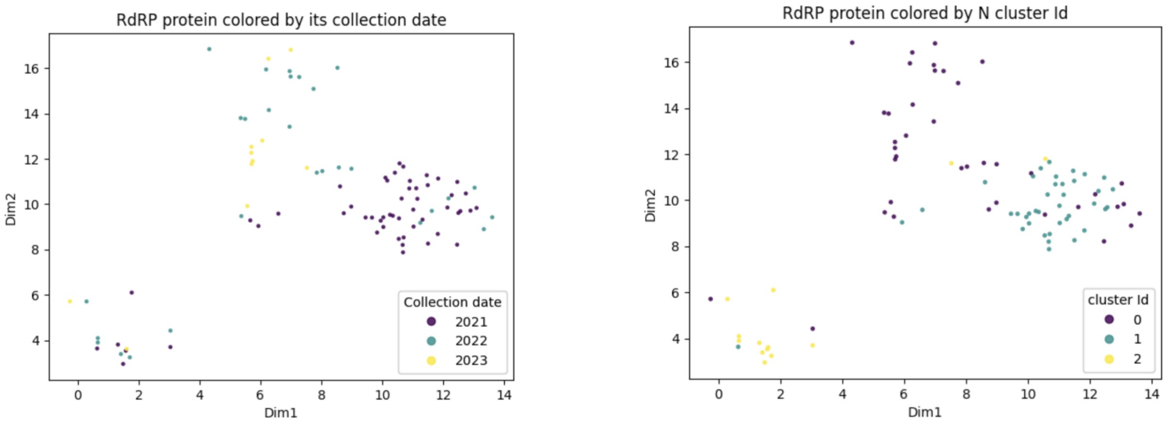

Figure 9. Plotting dimensioncal reduced embeddings of RdRP Protein after being grouped by N protein based on their similar strain names, and being colored by N cluster Id, dataset size=91

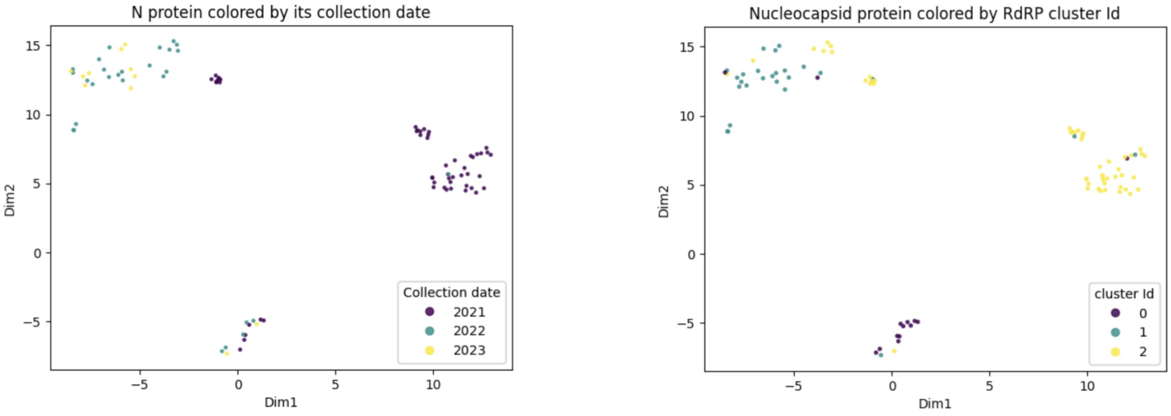

Figure 10. Plotting dimensioncal reduced embeddings of N Protein after being grouped by RdRP protein based on their similar strain names, and being colored by RdRP cluster Id, dataset size=91

CONCLUSION

we have learned that clustering of high-dimensional data is challenging due to the Curse of Dimensionality and limitations of the clustering methods. Frustratingly, most of clustering algorithms require the number of clusters to be specified a-priori[7]. The obtained embeddings have shown a co-evolutionary pattern between spike and RdRP and Nucleocapsid proteins by showing near sequences of each protein being colored by same cluster Ids of another protein. Additionally, to see how the ap2prob values are precise, Nucleocapsid dataset has been considered, and around 70% of Nucleocapsid's amino acids were predicted to be conserved by the model, by providing likelihood of more than 85% for each of the amino acid, collected in 2021. For predicting the prone to mutation amino acids, however, the model did not give a good result and its precision was so low. Nucleocapsid and RdRP has shown a mutational pattern happening once from 2021 and no specific group has been found from 2022 to 2023, whereas spike protein has shown various mutations within each year. Also, Nucleocapsid sequences have shown a longer distance between the clusters than RdRP which can show the possibility of more different biological effect happening between two groups in Nucleocapsid protein. Finally, embeddings can capture biological charactersitics of proteins and be beneficial in predicting multiple phenomena that conventional bioninformatic tools are not able to do.

REFERENCES

[1] Bepler, T. and Berger, B. (2021). Learning the protein language: Evolution, structure, and function. *Cell systems*, 12(6):654–669.

[2] Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

[3] Jiang, Y., Yin, W., and Xu, H. E. (2021). Rna-dependent rna polymerase: Structure, mechanism, and drug discovery for covid-19. *Biochemical and biophysical research communications*, 538:47–53.

[4] Johnson, B. A., Zhou, Y., Lokugamage, K. G., Vu, M. N., Bopp, N., Crocquet-Valdes, P. A., Kalveram, B., Schindewolf, C., Liu, Y., Scharton, D., et al. (2022). Nucleocapsid mutations in sars-cov-2 augment replication and pathogenesis. *PLoS pathogens*, 18(6):e1010627.

[5] Lucas, M., Karrer, U., Lucas, A., and Klenerman, P. (2001). Viral escape mechanisms–escapology taught by viruses. *International journal of experimental pathology*, 82(5):269–286.

[6] Wang, Y., Pan, X., Ji, H., Zuo, X., Xiao, G.-F., Li, J., Zhang, L.-K., Xia, B., and Gao, Z. (2023). Impact of sars-cov-2 envelope protein mutations on the pathogenicity of omicron xbb. *Cell Discovery*, 9(1):80.

[7] Zhao, N., Zhou, N., Ding, J., and He, M. (2022). Mutations and phylogenetic analyses of sars-cov-2 among imported covid-19 from abroad in nanjing, china. *Frontiers in Microbiology*, 13:851323.