2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Unraveling the hCoV-19 Informational Architecture

S. Aguilar-Valdez[1] and J. Alejandro Morales[1] and Omar Paredes[1]

Abstract— **The hCoV-19 virus is continuously evolving to highly infectious and lethal variants. There is a latent risk that current vaccines will not be effective over these novel variants. This entails comprehending the genome-wide viral information to unveil mutagenic mechanisms of hCoV-19. To date, this virus is studied as a collection of non-related variants, making it challenging to forecast hotspots and their upcoming effects. In this work, we explore genome-wide information to disentangle informational mechanisms that lead to insights into viral mutagenicity. Towards this aim, we modeled informational compartments based on a topic-free-alignment workflow. These compartments illustrate that hCoV-19 has a complex informational architecture that addresses high-level virus phenomena, i.e., mutagenicity. This new framework represents the first step towards identifying the virus mutagenicity leading to the development of all-variants-effective vaccines.**

I. INTRODUCTION

The human coronavirus (hCoV-19) outbreak began in Wuhan, China, by the end of 2019, reaching the pandemic status on March 11, 2020 [1]. Since then, the original strain has diverged into nine clades according to the Global Initiative for Sharing All Influenza Data (GISAID): G, GR, GH, GV, L, S, O, V, and GRY. Among these clades, new variants have emerged, increasing the virus' infectivity and virulence [2], which extends the pandemic's lifespan. A strategy to contain the outbreak is the vaccine development and the vaccination campaigns as demonstrated in Israel [3]. However, there is an increasing concern about the effectiveness of current vaccines against novel variants [4].

As new variants emerge, vaccine immunity is highly prone to be compromised. Given this picture, there are two alternatives: i) a continuous development of current vaccines, thus requiring new pharmaceutical phases, and ii) prediction of forthcoming variants to develop a global coronavirus vaccine [5]. The latter seems the most feasible, which involves the outlining of hCoV-19 informational structure.

Genomes are a network of informational modules interplaying to yield phenotypic effects, where such modules consist of information quanta [6]. However, the virus research has focused on analyzing a limited number of variants without interrelating them, thereby overlooking a broader spectrum of biological features. Unraveling the information quanta and how they developed informational modules is helpful towards disentangling the hCoV-19 mutagenicity and hence forecasting hotspots.

Lou et al. [7] outlined genomic compartments with Latent Dirichlet Allocation (LDA) to unveil the informational network in cell lines. Similarly, Borrayo et al. [8] carried out a bacteria phylogenetic analysis with LDA-estimated topics. Such topics are modules of an informational network in the human genome that we refer to throughout this paper as informational compartments. In this work, we developed a Latent Dirichlet Allocation workflow to pinpoint the hCoV-19 informational compartments and their correspondence with the GISAID clades.

[1]Computer Sciences Department, Exact Sciences and Engineering University Centre, Universidad de Guadalajara, México omar.paredes@academicos.udg.mx

II. MATERIALS AND METHODS

A. Database and Data Processing

To analyze hCoV-19 virus with a free-alignment approach, we retrieved the Latin America viral genomes of all nine reported clades from the GISAID EpiCov Database (March 7th, 2021) [9]. Then, discarded genomes with undefined bases below 5% and marked as hosted by nonhuman species.

A viral genome is mapped as a long non-spaced sequence of four letters known as nucleotides which are adenine (A), thymine (T), guanine (G), and cytosine (C). Genomic sequences are oversimplified as an ensemble of individual elements and tend to be compared to others on a position-by-position basis. However, complex structures, as genomes, are built up from information units or information quanta that altogether unfold any organism's biological functions [6]. Hence, a suitable approach to cluster viral genomes is by comparing these information quanta.

Thus, we considered each sequence as a collection of overlapped k-length information quanta known as k-mers. We decomposed the viral genomes into k-mers of $k = 9$ according to [10]. These 9-mers collections that compound each genome sequence are called genome corpus. We then merged all these corpora into the hCoV-19 corpus.

B. Topic Model

Studies have modeled genome corpora as a mixture of word collections called topics that yield potential phenotypic effects [8], [11]. Towards modeling the hCoV-19 corpus into topics, we counted the 9-mer appearance of the hCoV-19 corpus for each genome, obtaining a matrix called the genome-corpus matrix. Then, we estimated the matrix topics with the LDA algorithm.

LDA is a topic modeling technique that estimates the latent topics from high-dimensional data. This approach defines each genome as a probability distribution of these latent topics, and these topics represent a 9-mers probability distribution. Both topic and 9-mers probability distributions provide an explicit representation of hCoV-19 genome [7].

LDA assumes the following generative process for each genome w in a corpus D:

1) Choose $N \sim$Poisson(ξ).

2) Choose $\Theta \sim \text{Dir}(\alpha)$.
3) For each of the N 9-mers w_n:
    a) Choose a topic $z_n \sim \text{Multinomial}(\Theta)$.
    b) Choose a 9-mer w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

To set up the topic number suitable to describe the hCoV-19 virus, we calculated the following metrics [7]:

1) The data posterior likelihood given by multiple LDA models.
2) The Kullback–Leibler (KL) divergence of the genome-corpus matrix.
3) The average cosine distance γ among latent topics.

We set the topic number where both the maximum of Griffiths' metric and the minimum of KL and cosine distance converge. Then, we computed the topic distribution of each viral genome and the 9-mer distribution associated with each topic.

C. Hierarchical Clustering

To uncover hCoV-19 virus clustering in Latin America based on its topic distribution, we carried out an agglomerative hierarchical clustering. This clustering method indexes each genome as a single cluster and then joins clusters by ranking inter-cluster distances through a linkage function. In this work, we implemented a Ward criterion to cluster the Euclidean distance of the topics.

Lastly, to project hCoV-19 virus topic distribution and its clustering into a low-dimensional representation, we embedded the genome-corpus matrix by the Uniform Manifold Approximation and Projection (UMAP) technique. UMAP is a dimension reduction technique that preserves local topological features among topic distribution by assuming these distributions build a Riemann manifold and estimating such manifold. Code is available in Google Drive.

III. RESULTS AND DISCUSSION

The GISAID database consists of worldwide hCov-19 viral genomes grouped in 9 phylogenetic clades, in this work we focused on those viral genomes corresponding to the Latin America region, to outline a picture of the least explored countries during the hCov-19 pandemic. The total retrieved genomes are 9 471 ranging 29 014 to 30 008 nucleotides and distributed as follows: 1 512 in clade G, 6 024 in clade GR, 1 402 in clade GH, 34 in clade GV, 21 in clade L, 270 in clade S, in for clade O, 35 in clade V, and 52 in clade GRY.

To explore the clade robustness of the current hCov-19 clustering, we developed a topic-free-alignment approach that clusters viral genomes modeling k-mers collections called topics and their occurrence in hCov-19 genomes. Such approach is grounded on the idea that complex information encoded in genomes is built of information quanta, that interact in steady pools to unfold phenotypic features [6].

We defined information quanta as overlapped fixed-length subsequences within genomes called k-mers. We fixed the k-mer length at 9 nucleotides according to Zhang et. al. [10]. For each viral genome, we obtained $L-8$ 9-mers where L is the genome size, this 9-mer set was known as genome corpus. Next, we merged all genome corpora into a new corpus called the hCov-19 corpus. The latter consists of 87 520 unique 9-mers corresponding to 33.38% of all available 9-mer permutations ($4^9 = 262144$). This hints that the viral corpus biases towards a singular ordered structure that potentially differentiates hCov-19 from other coronavirus and viruses in general. Such orderedness seems to occur for energetic or informational constraints, as in natural language vocabularies [12].

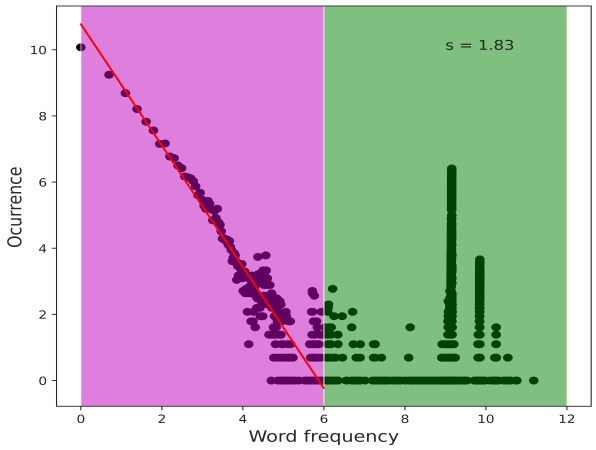

Fig. 1. Zipf's Law distribution of hCoV-19 corpus. Two regions are shown: Region I corresponding to viral genome regularities associated with hCoV-19 identity, and Region II, describing viral clades divergence.

To evaluate if the hCov-19 corpus behaves as a vocabulary, we fit the 9-mers' probability distribution to a Zipf's Law distribution. The Zipf's Law is a distribution that measures language regularities, where the frequency n of the m-th most frequent word (in our case 9-mer) of a text (hCov-19 corpus) follows the next equation:

$$n(m) \sim m^{-s} \tag{1}$$

where s is the score characterizing the Zipf's Law distribution [13].

Figure 1 shows the Zipf's distribution fit of the hCov-19 corpus. There are two differentiable regions, the region I (in magenta) corresponding to those words (60 894) with a frequency below $ln = 6$ that follows accurately the Zipf's Law, and region II (in green) with the words (26 626) above the latter threshold showing a discrete behavior. Region I hints that the virus has a blueprint built by an informational regular structure (around 70% of the corpus) that yields the virus identity, the analysis if such identity differs from other viruses should be explored but is beyond the scope of this work. However, the hCov-19 diverges into well-studied clades, thereby unveiling that there are hotspots lying in the viral blueprint potentially corresponding to 9-mers in region II.

COVID-19 displays complex symptomatology, hinting towards information compartmentalization within its genome

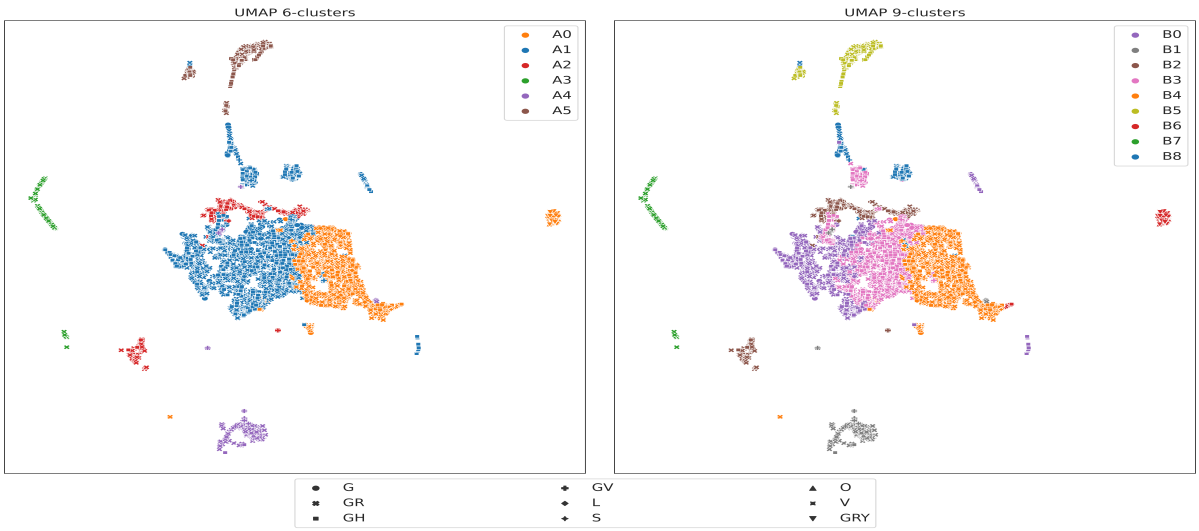

Fig. 2. Clustering UMAP embedding of hCoV-19 informational compartments.

[14], [15]. These informational compartments are information quanta interacting together that potentially yield functional pathways that cause such a myriad of symptoms. Borrayo et. al. [8] clustered bacteria with different phenotypes by modeling such compartments through topics. Hence, we modeled the viral genome topics using LDA to compare hCov-19 clades based on their informational compartments.

To determine the suitable amount of informational compartments, we estimated the optimum topic number by the metrics explained in the methodology. The optimal performance for the three metrics was at 6 topics, therefore we modeled the 6 informational compartments (T1 to T6) for all the hCov-19 genomes with LDA.

Next, we estimated the compartment probability distribution for each viral genome. Then, clustered the genomes based on their topic probability distribution with hierarchical clustering. We performed two clusterings, one with 6 clusters, and the second with 9. The former considering that viral genomes tend to mainly be yielded by a single informational compartment, while the latter is based on the approach that clades are a mixture of the informational compartments.

Since the topic probability distributions are high-dimensional data, clustering visualization is complex. Thus, we embedded the data into a 2D space that was estimated with UMAP. Figure 2 shows the 2D-space clustering for the two approaches. In both clusterings, the peripheral components are labeled as isolated clusters; for the 6-cluster approach, these are clusters A3, A4, and A5, whereas, for the 9-cluster approach, these correspond to clusters B1, B2, B5, B6, and B7. These scattered from the core component because their composition consists predominantly of a single informational compartment, i.e., A3 mainly built by T2, and B5 by T5.

The core component gathers the remaining clusters that are a quasi-homogeneous mixture of the 6 informational compartments. The majority of these sequences are from clades G, GR, GH, GV, the most widely spread throughout Latin America. This homogenization is likely due to the loss of hotspots as the variant spreads, resulting in a virus built with regular 9-mers from region I.

Figure 3 (Interactive Version) shows the intersection between hCoV-19 clades and clusters for the two approaches. It is observed that the S, O, GRY and V clades group into a single cluster for both approaches. This indicates that they are informationally similar albeit they differ due to some genomic variations. Otherwise, the overrepresented clades (G, GH, and GR) are distributed into multiple clusters, revealing their heterogeneous informational composition.

This indicates that the variant-based classification is narrow since it ignores a complex informational structure in the hCoV-19 genome. To test this, further analysis is necessary, including all hCoV-19 genomes sequenced worldwide. This approach may help to unveil informational mechanisms underlying viral mutations and thus facilitate the vaccine development that anticipates future hCoV-19 mutations.

IV. CONCLUSIONS

Modeling hCoV-19 mutagenicity is a key challenge for the current pandemic for developing a global coronavirus vaccine. This requires unraveling the informational structure of the virus to predict hotspots. In this paper, we introduced an LDA-based workflow to identify informational compartments that build up the virus. We pointed out that these compartments consist of information quanta, that in the case of hCoV-19 split into two sets, a regular one that yields the identity of the virus and a quasi-random set containing the potential virus hotspots.

Our results showed that by analyzing the virus informational compartments, instead of focusing on single emerging variants, the clades reorganized into new clusters. These novel clusters are likely to feature high-level virus phenomena i.e. mortality, infectivity, and mutagenicity. This new framework represents the first step towards identifying the

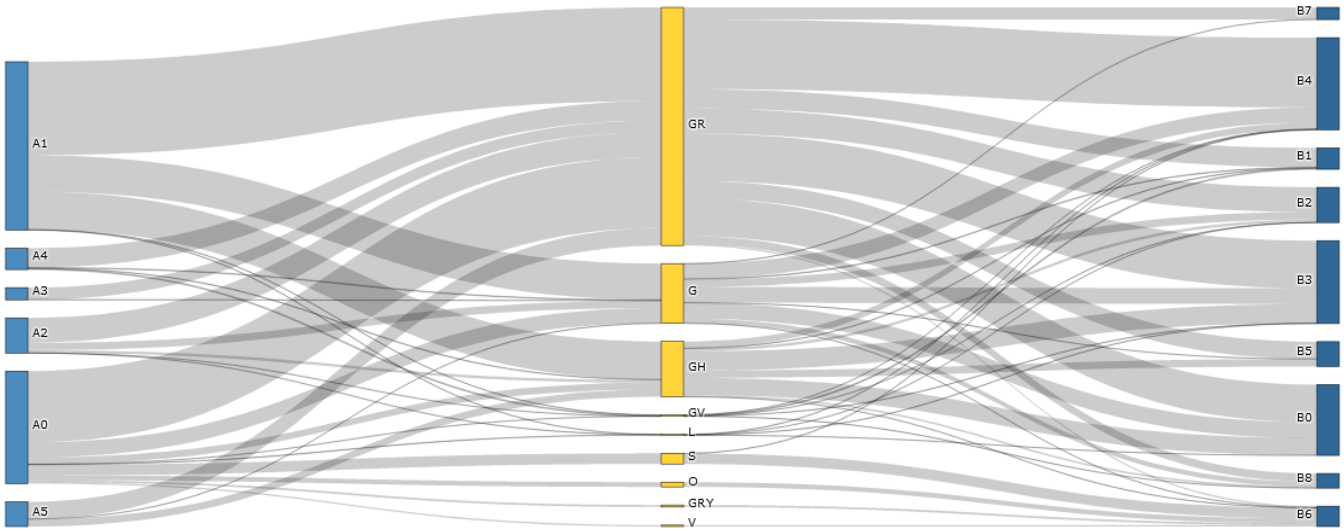

Fig. 3. Correspondence of informational compartments clustering with hCoV-19 clades. On the left for 6 clusters, and on the right for 9 clusters.

virus mutagenicity leading to the development of all-variants-effective vaccines.

ACKNOWLEDGMENT

This work was supported by the Consejo Nacional de Ciencia y Tecnología — CONACyT [Scholarship to CVU 713526]

REFERENCES

[1] S. M. Hamed, W. F. Elkhatib, A. S. Khairalla, and A. M. Noreddin, "Global dynamics of sars-cov-2 clades and their relation to covid-19 epidemiology," *Scientific reports*, vol. 11, no. 1, pp. 1–8, 2021.

[2] M. S. Graham, C. H. Sudre, A. May, M. Antonelli, B. Murray, T. Varsavsky, K. Kläser, L. S. Canas, E. Molteni, M. Modat, *et al.*, "Changes in symptomatology, reinfection, and transmissibility associated with the sars-cov-2 variant b. 1.1. 7: An ecological study," *The Lancet Public Health*, 2021.

[3] H. Rossman, S. Shilo, T. Meir, M. Gorfine, U. Shalit, and E. Segal, "Covid-19 dynamics after a national immunization program in israel," *Nature Medicine*, pp. 1–7, 2021.

[4] D. M. Altmann, R. J. Boyton, and R. Beale, "Immunity to SARS-CoV-2 variants of concern," *Science*, vol. 371, no. 6534, pp. 1103–1104, 2021.

[5] J. Cohen, "Vaccines that can protect against many coronaviruses could prevent another pandemic," *Science*, 2021.

[6] V. Y. Tsvetkov, "Information units as the elements of complex models," *Nanotechnology Research and Practice*, no. 1, pp. 57–64, 2014.

[7] S. Lou, T. Li, X. Kong, J. Zhang, J. Liu, D. Lee, and M. Gerstein, "Topicnet: A framework for measuring transcriptional regulatory network change," *Bioinformatics*, vol. 36, no. Supplement_1, pp. i474–i481, 2020.

[8] E. Borrayo, I. May-Canche, O. Paredes, J. A. Morales, R. Romo-Vázquez, and H. Vélez-Pérez, "Whole-genome k-mer topic modeling associates bacterial families," *Genes*, vol. 11, no. 2, p. 197, 2020.

[9] S. Elbe and G. Buckland-Merrett, "Data, disease and diplomacy: GISAID's innovative contribution to global health," *Global Challenges*, vol. 1, no. 1, pp. 33–46, 2017.

[10] Q. Zhang, S.-R. Jun, M. Leuze, D. Ussery, and I. Nookaew, "Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer," *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017.

[11] L. Juan, Y. Wang, J. Jiang, Q. Yang, G. Wang, and Y. Wang, "Evaluating individual genome similarity with a topic model," *Bioinformatics*, vol. 36, no. 18, pp. 4757–4764, 2020.

[12] A. Mazzolini, J. Grilli, E. De Lazzari, M. Osella, M. C. Lagomarsino, and M. Gherardi, "Zipf and heaps laws from dependency structures in component systems," *Physical review E*, vol. 98, no. 1, p. 012 315, 2018.

[13] Á. Corral, G. Boleda, and R. Ferrer-i-Cancho, "Zipf's law for word frequencies: Word forms versus lemmas in long texts," *PloS one*, vol. 10, no. 7, e0129031, 2015.

[14] J. Y. Dutheil, G. Mannhaupt, G. Schweizer, C. MK Sieber, M. Münsterkötter, U. Güldener, J. Schirawski, and R. Kahmann, "A tale of genome compartmentalization: The evolution of virulence clusters in smut fungi," *Genome biology and evolution*, vol. 8, no. 3, pp. 681–704, 2016.

[15] H. Wong, J. Soh, P. M. Gordon, T. Yu, C. W. Sensen, E. Parr, and R. N. Johnston, "Genomic compartmentalization of gene families encoding core components of metazoan signaling systems," *Genome*, vol. 56, no. 4, pp. 215–225, 2013.