Clasificación de los géneros de la subfamilia Coronavirinae basada en firmas genómicas

Licenciatura en Biología, Facultad de Ciencias, UNAM, México.[1]
Licenciatura en Ciencias de la Computación, Facultad de Ciencias, UNAM, México.[2]

Natalia Abasolo-Cortés[1], Héctor E. Gómez-Mora[2], y Carlos Pimentel-Ruíz[1]

Deposición de Datos: Las secuencias de los virus y las representaciones de las mismas generadas por juego del caos, así como los códigos utilizados se encuentran en el correspondiente repositorio de GitHub [3].

I. Introducción

Los coronavirus son virus envueltos que portan RNA monocatenario positivo, se caracterizan por presentar púas en su superficie y un RNA inusualmente largo aunado a una estrategia única de replicación. Estos virus pertenecen a la familia Coronaviridae (una de las cuatro familias contenidas en el orden Nidovirales) y dentro de esta a la subfamilia Coronavirinae que engloba a cuatro géneros: alfa, beta, gamma y delta (Coronavirus), división que se les fue otorgada por medio de clustering filogenético [2].

El genoma de estos virus tiene aproximadamente 30 kb, presenta una estructura de capucha 5' y una cola de poliA 3', esto permite que funcione como un mRNA para la traducción de la replicasa poliproteíca que corresponde a aproximadamente 20 kb mientras que las proteínas estructurales y accesorias son aproximadamente 10 kb de su genoma [2].

El extremo 5' contiene una secuencia líder y una región no traducida (UTR) que presenta varias horquillas necesarias para la replicación y traducción del RNA. Adicional a esto se presentan secuencias reguladoras de la transcripción (TRSs) al inicio de cada gen codificante para proteínas accesorias o estructurales, por lo tanto, la TRS son requeridas para la expresión de esos genes [2]

El extremo 3' también presenta una UTR que contiene estructuras de RNA requeridas para la replicación y síntesis del RNA viral. La organización del genoma de Coronavirus es 5' - Secuencia Líder - UTR - Replicasa - S (púa) - E (envoltura) - M (membrana) - N (nucleocápside) - 3' UTR - Cola de poliA con genes accesorios intercalados dentro de los genes estructurales del extremo 3' (Fehr y Perlman, 2015). Esos genes accesorios juegan un papel importante en la patogénesis viral [10].

Mucho se ha avanzado en el campo del reconocimiento de imágenes mediante técnicas de Machine Learning. En particular, las redes neuronales convolucionales (CNN) se han destacado como el modelo a utilizar para la resolución de dicho problema pues sus características lo posicionan como la opción más eficiente [8].

De forma somera una red neural constituye un conjunto de entidades denominadas neuronas artificiales interconectadas que emiten y reciben señales (Pattanayak S. 2017). Dichas unidades de procesamiento se encuentran agrupadas en capas para las cuáles sus inputs provienen de las neuronas pertenecientes a una capa previa y sus salidas pueden o no servir como entradas de otra capa. El esquema básico de una red neuronal comprende una capa de entrada y de salida entre las cuáles se encuentra una cantidad no fija de capas adicionales conocidas como capas ocultas [7].

Este modelo de aprendizaje automatizado se cataloga como supervisado en tanto que se requiere un entrenamiento previo del mismo a partir de un conjunto de datos específico, mismo que es denominado conjunto de entrenamiento, en el cual a cada uno de los datos se le ha asignado una etiqueta que les identifica como miembros de una clase [7].

Posterior a la etapa de entrenamiento, un conjunto de datos sin etiquetas, denominado como conjunto de prueba, es suministrado a la red neuronal para acreditar la capacidad de aprendizaje obtenido por el modelo. De esta última se espera que a partir de los patrones que logró abstraer en la anterior etapa, pueda clasificar a los datos de entrada de forma correcta [6].

De lo anterior, una red neuronal convolucional comprende las características anteriores salvo que las capas intermedias constituyen una secuencia de capas de convolución, activación y pooling, junto con capas tradicionales encargadas de identificar, cada una, patrones específicos desprendidos de la señal inicial [5].

Las firmas genómicas generadas por representación por juego del caos (CGR), proveen una visualización única de la organización de los patrones en una secuencia de nucleótidos. Una firma genómica está asociada a la longitud de una subsecuencia de nucleótidos que representa una medida de resolución en el análisis de la organización primaria del DNA [9].

Deschavanne et al., [1] mostraron que la firma genómica elaborada con al menos 1kb resultaba ser muy similar a la que se generaba con el genoma completo, siendo esto la base del concepto firma genómica.

Estos antecedentes nos llevaron a la idea de generar firmas genómicas de secuencias conocidas y reportadas por medio de la representación por juego del caos de virus de la subfamilia Coronavirinae con la finalidad de entrenar una red neuronal convolucional que funcione como clasificador de secuencias que aún no se han clasificado en uno de los 4 géneros que esta subfamilia comprende.

II. Metodología

La generación de imágenes se realizó a partir de genomas completos correspondientes a cada uno de los géneros de la subfamilia Coronavirinae. Estos últimos se obtuvieron como lecturas en formato FASTA de los registros dispuestos por el NCBI en su base de datos específica de virus.

Para la parte de la red neuronal se tuvieron que conformar los conjuntos de entrenamiento y validación a partir de un criterio que pudiese particionar la información disponible en dos conjuntos disjuntos. Para el conjunto de entrenamiento se utilizaron los genomas completos y para el conjunto de validación utilizaron genomas parcialmente completos y para ello se utilizaron los siguientes filtros:

Entrenamiento

- El taxid (identificador de taxón) correspondiente
- db (base de datos) = genbank
- Nucleotide Completeness (Completud del genoma) = complete

Validación

- El taxid correspondiente
- db = genbank
- Nucleotide Completeness = partial

Debido a que la descarga masiva de genomas se realiza depositando todas las lecturas en un sólo archivo, fue necesario realizar un pre-procesamiento con la finalidad de separar cada genoma y a partir de estos generar su firma genómica mediante representación por juego del caos.

Se obtuvieron 724 imágenes (firmas genómicas) de entrenamiento, 200 correspondientes a cada uno de los siguientes géneros alphacoronavirus, betacoronavirus y gammacoronavirus respectivamente, y las 124 restantes para deltacoronavirus.

Para el conjunto de validación fueron generadas 200 imágenes en proporciones iguales para cada uno de los géneros.

Por su parte, la construcción de la red neuronal convolucional se realizó con ayuda de la biblioteca Keras, emulando la arquitectura MiniVGG de 3 bloques disponible en el repositorio de nuestro proyecto final [3].

A partir de esta se realizó un primer entrenamiento utilizando el conjunto de imágenes sin ruido (algunas imágenes generadas para virus representativos se encuentran en el anexo 1). Los parámetros utilizados se encuentran disponibles de forma detallada en el código fuente [3], sin embargo se destacan `cross_entropy` como función de costo durante 200 épocas de entrenamiento.

Posteriormente se construyó una nueva colección de imágenes a partir de la anterior con la intención de aumentar su tamaño mediante la creación de instancias con ruido.

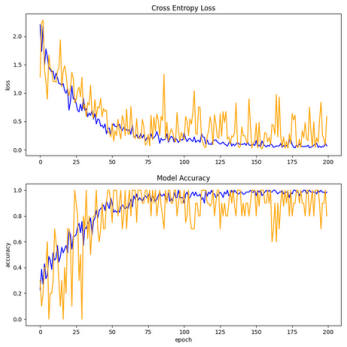
Para ello se utilizaron los filtros mean filter y Gaussian filter, dispuestos ambos por la bilbioteca Open-CV, con la cual se espera reducir los efectos del sesgo derivado de la diferencia en el número de muestras entre las clases alpha, beta y gamma con respecto a la clase delta. Estas clases hacen referencia a los géneros comprendidos en la subfamilia Coronavirinae.

Aunado a ello, la inserción de ruido ayuda a minimizar el fenómeno de overfitting [4] que es causado por la alta similitud que comparten las imágenes.

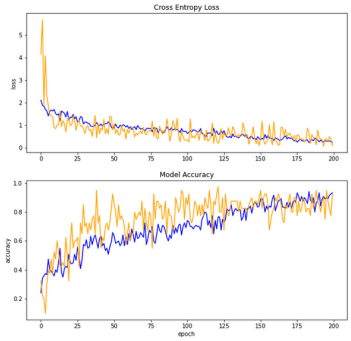
Utilizando el conjunto con ruido se realizó un segundo entrenamiento de la arquitectura VGG bajo las mismas condiciones descritas anteriormente.

Para comprobar el desempeño alcanzado por ambos modelos se utilizaron genomas de virus clasificados en los géneros de la subfamilia Coronavirinae.

III. Resultados



(a) Modelo VGG sin ruido.



(b) Modelo VGG con ruido.

Figura 1: *Desempeño alcanzado por el modelo tras 200 épocas. La línea azul representa el desempeño alcanzado por el conjunto de entrenamiento mientras que la línea naranja denota el desempeño alcanzado por el conjunto de validación. La gráfica superior contrasta la minimización de la función de costo a lo largo de 200 épocas, mientras que la gráfica inferior representa la precisión para clasificar alcanzada por el modelo en 200 épocas.*

Del desempeño al optimizar la función de costo en el conjunto de imágenes sin ruido, puede notarse que, aunque la tendencia del conjunto de entrenamiento es decreciente, el desempeño del conjunto de validación es inestable y tiende a errar en tanto que, a medida que las predicciones para el conjunto de entrenamiento son más acertadas la pérdida del conjunto de prueba aumenta.

Esto último puede ser un indicio de overfitting pues el modelo logró optimizar de buena manera la función de costo

ID	**HCoV-229E**	ID	**SARS-CoV2**
Alpha	0.99971253	Alpha	0.99954563
Beta	0.00009756162	Beta	0.00010292028
Delta	0.0000712035	Delta	0.00014498606
Gamma	0.00011876108	Gamma	0.00020643209

ID	**Porcine Diahrrea Cov**	ID	**Beluga Whale Cov**
Alpha	0.99967992	Alpha	0.99952006
Beta	0.00010776492	Beta	0.0001995612
Delta	0.00011975131	Delta	0.00006335269
Gamma	0.00009254125	Gamma	0.00021700891

Cuadro I: *Predicciones bajo el modelo entrenado con el conjunto de imágenes sin ruido.*

con respecto al conjunto de entrenamiento aumentando así su capacidad de predicción para el mismo mientras que las predicciones sobre el conjunto de validación no mostraron mejoras significativas.

En el caso del conjunto de imágenes con ruido, la tendencia de optimización y pérdida de la función de costo es decreciente y por tanto la capacidad de predecir para ambos conjuntos mejoró a lo largo de las épocas. Puede notarse incluso cómo la suma del error adquiere un comportamiento lineal a partir de 50 épocas de entrenamiento.

Por su parte, las gráficas de precisión tienen como objetivo mostrar el porcentaje de instancias de entrenamiento o prueba que fueron clasificadas correctamente al transcurrir las épocas.

De ello, puede notarse que a partir de 25 épocas la precisión en la colección de imágenes sin ruido tiende a aumentar para su conjunto de entrenamiento mientras que el conjunto de validación oscila entre el 60 y 90 % y nunca alcanza la estabilidad.

Con base en los resultados anteriores, y a pesar de el entrenamiento con la inserción de ruido, pudimos observar

ID	**HCoV-229E**	ID	**SARS-CoV2**
Alpha	0.9089777	Alpha	0.89561015
Beta	0.00195871	Beta	0.00243213
Delta	0.00656711	Delta	0.01359664
Gamma	0.08249648	Gamma	0.08836111

ID	**Porcine Diahrrea Cov**	ID	**Beluga Whale Cov**
Alpha	0.8996626	Alpha	0.91656125
Beta	0.00152908	Beta	0.00301626
Delta	0.0211022	Delta	0.00623145
Gamma	0.07770605	Gamma	0.07419103

Cuadro II: *Predicciones bajo el modelo entrenado con el conjunto de imágenes con ruido.*

resultados similares pero con una variación temporal, debido a que el grupo con ruido (Figura 1,b.) alcanza una precisión cercana a 1.

En el Cuadro 1 se presenta la certeza de la predicción de clasificación a uno de los 4 géneros de la subfamilia Coronavirinae siendo el valor más cercano a 1 la clasificación más probable.

Todos los genomas presentados en el Cuadro 1 tienen un género ya establecido por el ICTV (International Committee on Taxonomy of Viruses) sin embargo, nuestra red neuronal no ha sido capaz de clasificarlos en su género correspondiente y todos los ha clasificado como Alphacoronavirus.

En trabajos previos se han reportado el buen funcionamiento de estas redes neuronales pero se utilizaron niveles taxonómicos superiores al género, además de que se utilizaron secuencias de ADN de organismos bacterias, hongos, plantas y animales principalmente. Sin embargo, los virus son clasificados con base en criterios diferentes a los grupos mencionados anteriormente.

Creemos que uno de los factores por los que las secuencias fueron clasificadas en el mismo grupo es debido a la similitud de estas mismas, ya que en la conformación del genoma se encuentran diferencias demasiado pequeñas como para que el funcionamiento del clasificador sea lo suficientemente meticuloso.

Aunado a esto, una de las grandes similitudes que presentan estos géneros son los marcos de lectura abierta que ocupan del genoma (aproximadamente 20 kb) y que en el tercio restante, que contiene información para proteínas estructurales, presentan las diferencias mínimas que los dividen.

IV. Conclusiones

Como una de las conclusiones podemos decir que la constitución de los genomas no fue suficientemente representativa para generar firmas genómicas que se pudieran clasificar en los distintos géneros.

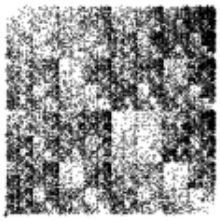
También resaltamos que podría ser necesario el uso de niveles taxonómicos superiores (i.e. familia, clases, etc.) al género para lograr una correcta clasificación, pues se esperaría una mayor diferencia entre los genomas.

Referencias

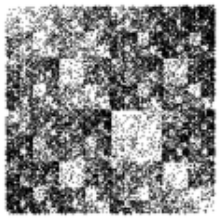
[1] Deschavanne, P., Giron, A., Vilain, J., Vaury, A., Fertil, B., 2000. *Genomic signature is preserved in short DNA fragments. IEEE International Symposium on Bioinformatics and Biomedical Engineering (BIBE'00)*, pp. 161–167.

[2] Fehr, A. and Perlman, S. 2015. Coronaviruses: An Overview of Their Replication and Pathogenesis. Coronaviruses, pp.1-23
[3] Gómez-Mora, H. E., Abasolo-Cortés, N. y Pimentel-Ruíz, C. 2020. https://github.com/cpimentelGH/genomica_computacional_proyecto_final.git
[4] Goodfellow, et. al. 2016. Deep Learning. 7: Regularization For Deep Learning. MIT Press, Disponible en: https://www.deeplearningbook.org/
[5] Habibi, A. H. y Jahani, H. E. 2017. Convolutional Neural Networks. In: Guide to Convolutional Neural Networks. Springer, Cham.
[6] Pattanayak, S. 2017. Introduction to Deep-Learning Concepts and TensorFlow. In: Pro Deep Learning with TensorFlow. Apress, Berkeley, CA.
[7] Russel, S. y Norvig, P. 2010. Artificial Intelligence: A Modern Aproach. Prentice Hall, pp. 729-737.
[8] Sun, Y., Xue, B., Zhang, M. y Yen, G.G. 2020. Evolving Deep Convolutional Neural Networks for Image Classification, in IEEE Transactions on Evolutionary Computation, vol. 24, no. 2, pp. 394-407.
[9] Wang, Y., Hill, K., Singh, S., & Kari, L. 2005. The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene, 346, 173–185.
[10] Zhao, L., Jha, B., Wu, A., Elliott, R., Ziebuhr, J., Gorbalenya, A., Silverman, R. and Weiss, S., 2012. Antagonism of the Interferon-Induced OAS-RNase L Pathway by Murine Coronavirus ns2 Protein Is Required for Virus Replication and Liver Pathology. Cell Host & Microbe, 11(6), pp.607-616.

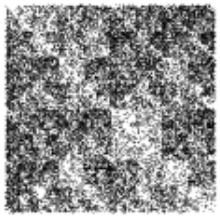
APÉNDICE A
INDIVIDUOS USADOS PARA LA PREDICCIÓN DE AMBOS MODELOS



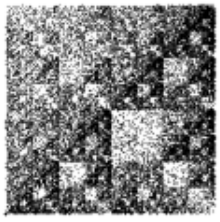
(a) Coronavirus Humano 229E (HCov-229E).



(b) Coronavirus tipo 2 del síndrome respiratorio agudo grave (SARS-CoV-2)



(c) Coronavirus de diarrea porcina (PDCoV)



(d) Coronavirus de beluga SW1 (beluga whale coronavirus SW1)