TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Aplicación de inteligencia artificial en genómica y metagenómica viral para la clasificación taxonómica y el descubrimiento de nuevas proteínas virales

Blanca Itzel Taboada Ramírez
Universidad Nacional Autónoma de México, Instituto de Biotecnología, Cuernavaca, Morelos, México.
ORCID: 0000-0003-1896-5962

Lorena Díaz-González
Universidad Autónoma del Estado de Morelos, Centro de Investigación en Ciencias,
Cuernavaca, Morelos, México.
ORCID: 0000-0003-1577-5629

Oscar Alejandro Uscanga Junco
Universidad Autónoma del Estado de Morelos, Instituto de Investigación en Ciencias Básicas Aplicadas (IICBA),
Cuernavaca, Morelos, México.
ORCID: 0000-0002-4179-6725

Alida Esmeralda Zárate Jiménez
Universidad Nacional Autónoma de México, Instituto de Biotecnología, Cuernavaca, Morelos, México.
ORCID: 0009-0006-5407-6598

Edna Cruz-Flores
Universidad Autónoma del Estado de Morelos, Instituto de Investigación en Ciencias Básicas Aplicadas (IICBA),
Cuernavaca, Morelos, México.
ORCID: 0000-0002-4187-3428

Recepción: 01 de marzo de 2026.
Aceptación: 28 de abril de 2026.

Mayo 2026 • número de revista 15 • DOI: 10.22201/dgtic.26832968e.2026.15.158

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Aplicación de inteligencia artificial en genómica y metagenómica viral para la clasificación taxonómica y el descubrimiento de nuevas proteínas virales

Resumen

La metagenómica viral permite caracterizar comunidades de virus en muestras clínicas o ambientales a partir de millones de secuencias de ADN generadas por tecnologías de próxima generación. Este trabajo describe tres aplicaciones computacionales que, apoyadas en el uso de supercómputo, buscan mejorar el análisis de datos metagenómicos: i) una metodología que elimina la redundancia de las bases de datos de secuencias de referencia de genomas de virus mediante la construcción de pangenomas, conservando secuencias específicas de cada especie y las compartidas a nivel de género, lo que permite identificar virus de manera más precisa; ii) una herramienta que usa inteligencia artificial para identificar secuencias de virus eucariontes a nivel de proteínas, facilitando la detección de virus nuevos o con baja similitud a los anotados; y iii) una herramienta, también basada en inteligencia artificial, para identificar arreglos CRISPR en genomas bacterianos, lo que favorece el estudio de las interacciones fago-bacteria en datos metagenómicos. Estas aplicaciones apoyan el análisis de datos metagenómicos, contribuyendo a comprender mejor la diversidad viral y las relaciones virus-bacteria.

Palabras Clave: metagenómica viral, clasificación taxonómica, proteínas virales, bacteriófagos, CRISPR, aprendizaje profundo, supercómputo.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Application of Artificial Intelligence in Viral Genomics and Metagenomics for Taxonomic Classification and the Discovery of Novel Viral Proteins

Abstract

Viral metagenomics enables the characterization of viral communities in clinical or environmental samples based on millions of DNA sequences generated by next-generation sequencing technologies. This work describes three computational applications that leverage high-performance computing to improve metagenomic data analysis: (i) a methodology that reduces redundancy in viral reference genome databases through the construction of pangenomes, preserving species-specific sequences as well as sequences shared at the genus level, enabling more accurate virus identification; (ii) an artificial intelligence–based tool for identifying eukaryotic viral sequences at the protein level, facilitating the detection of novel viruses or those with low similarity to annotated references; and (iii) an artificial intelligence–based tool for identifying CRISPR arrays in bacterial genomes, which supports the study of phage–bacteria interactions in metagenomic datasets. Together, these applications enhance metagenomic data analysis and contribute to a better understanding of viral diversity and virus–bacteria relationships.

Keywords: *viral metagenomics, taxonomic classification, viral proteins, bacteriophages, CRISPR, deep learning, high-performance computing.*

Introducción

La metagenómica es una herramienta que permite estudiar las comunidades microbianas presentes en una muestra clínica o ambiental mediante el análisis de millones de secuencias de ADN o ARN, siguiendo el flujo de trabajo descrito en la Fig.1. Esto ha permitido identificar virus en océanos, suelos, animales, plantas y humanos, entre otros, mostrando que la diversidad de virus es mucho más extensa de lo que se había pensado.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

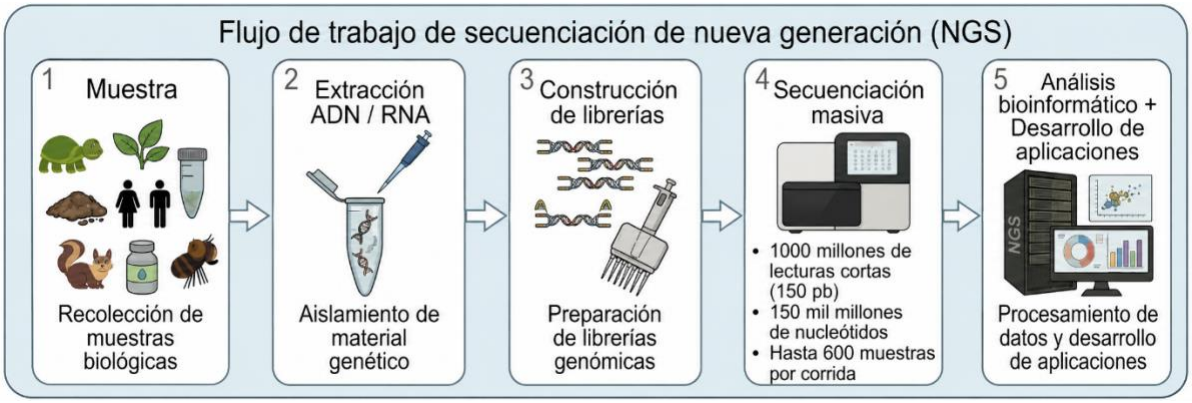

Fig. 1. Flujo de trabajo general para el análisis metagenómico mediante NGS.

Los virus se pueden dividir en dos grupos: los bacteriófagos, que infectan bacterias, y los eucariontes, que infectan a plantas, animales, hongos, insectos y humanos. Estos últimos incluyen virus de importancia médica, veterinaria y agrícola, por lo que conocer su diversidad y mejorar su detección es crucial para el desarrollo de estrategias de vigilancia, prevención y control. Por otra parte, los bacteriófagos son los virus más abundantes del planeta, con más de 10^{31} partículas en la biosfera, y desempeñan un papel importante al modular a las comunidades bacterianas. Sin embargo, la mayoría de estos virus permanece sin descubrirse [1]. Estudiar su diversidad y sus interacciones es clave para explicar su efecto en distintos ecosistemas.

Los estudios metagenómicos generan un inmenso volumen de secuencias cortas de ADN llamadas lecturas (Fig.1). Sin embargo, entre el 40 % y el 90 % de éstas no puede asignarse a ninguna especie viral, conociéndose como “materia oscura viral”. Esto es debido al uso de métodos basados en alineamientos y/o búsquedas por similitud contra bases de datos de genomas de referencia, las cuales no reflejan toda la diversidad viral. Además, la redundancia de estas bases de datos puede generar asignaciones no específicas. Todo esto afecta la vigilancia genómica, el estudio de la diversidad viral y la comprensión del papel de los bacteriófagos en diferentes ecosistemas.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

En este contexto, el desarrollo de nuevas herramientas computacionales basadas en inteligencia artificial y la mejora de las bases de datos de referencia abren nuevas posibilidades para abordar esta problemática. El objetivo de este trabajo es describir un conjunto de aplicaciones (Fig. 2), que se apoyaron en el uso de cómputo de alto desempeño para (i) reducir y enriquecer bases de datos de secuencias virales de referencia [2], (ii) clasificar taxonómicamente proteínas de virus eucariontes [3] y (iii) apoyar el estudio de interacciones fago-bacteria mediante la identificación de arreglos CRISPR.

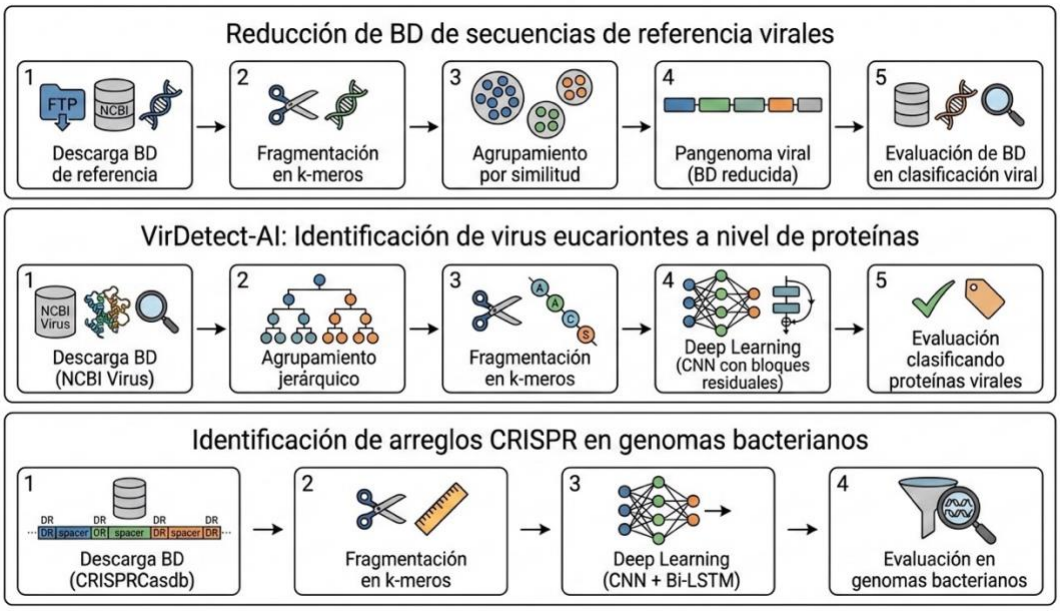

Fig. 2. Desarrollo de aplicaciones en metagenómica viral.

Este artículo tiene un enfoque de divulgación científica, en el que las aplicaciones (i) y (ii) se basan en desarrollos previamente validados, mientras que la aplicación (iii) corresponde a una línea de trabajo en desarrollo.

Marco metodológico para el análisis metagenómico de la diversidad viral

Como ya se mencionó, la metagenómica viral produce grandes volúmenes de información en forma de secuencias denominadas lecturas que son, de manera general, de una longitud

de 150 nucleótidos. A partir de estos datos, el reto consiste en encontrarles un sentido biológico, identificando virus, mejorando su clasificación y, cuando sea posible, proponiendo funciones o relaciones dentro del microbioma.

Este trabajo integra el desarrollo de aplicaciones computacionales que operan en distintos niveles. Por un lado, se emplean estrategias a nivel de nucleótido y k-meros, que trabajan directamente sobre secuencias de ADN para reducir las de bases de datos de genomas de referencia, enriquecer regiones informativas y favorecer identificaciones más específicas; dentro de esta misma línea, se ubica el método de identificación de arreglos CRISPR, útiles como evidencia para explorar interacciones fago-bacteria.

Por otro lado, cuando las secuencias presentan baja similitud con lo ya anotado, se recurre al análisis a nivel de proteínas, con el objetivo de capturar patrones más allá de la similitud inmediata y permitir la clasificación taxonómica de proteínas virales eucariontes que sean muy diferentes a lo conocido. Esto permitirá clasificar proteínas virales y descubrir nuevas, ampliando el repertorio funcional del viroma.

Dado el tamaño de los conjuntos de datos y la complejidad de los análisis, incluyendo el entrenamiento de modelos con millones de secuencias y la ejecución de procesos a gran escala, el desarrollo y evaluación de estas aplicaciones se realizaron con el apoyo de cómputo de alto rendimiento, incluyendo el uso de GPUs, en la supercomputadora Miztli (UNAM). Esta infraestructura permite entrenar modelos en tiempos razonables y ejecutar pruebas comparativas de manera sistemática, obteniendo resultados escalables.

Aplicación para reducir la base de datos de secuencias de referencia virales

De acuerdo con la International Committee on Taxonomy of Viruses (ICTV), los virus se organizan en 15 rangos taxonómicos, desde dominio hasta especie [4]. Las especies distintas de un mismo género comparten regiones genómicas conservadas, generando información redundante. En estudios metagenómicos, esto puede provocar que las lecturas obtenidas se asignen indistintamente a varias referencias, generando asignaciones poco específicas que no permiten diferenciar especies concretas. Por lo anterior, es necesario realizar análisis

adicionales que requieren un tiempo de cómputo considerable, tales como el análisis de ancestro común [5].

Para resolver esta problemática, diseñamos K-FluDB, una metodología computacional basada en el agrupamiento de subsecuencias de genomas virales [2]. Esto permite identificar subsecuencias que son específicas de cada especie y otras que son comunes entre varias especies. Estas últimas se denominan "piezas genómicas dispensables", pues son secuencias que permiten clasificar a nivel de género. Al conjunto total de secuencias específicas y dispensables, se le conoce como pangenoma viral.

Resumen de la metodología

La base de datos de virus de referencia se descargó de NCBI [6], obteniendo 14,331,967 secuencias de nucleótidos de referencia correspondientes a 31,573 especies y 3,328 géneros.

La metodología de reducción, ilustrada en la Fig. 3, consiste en obtener, para cada conjunto de secuencias de una especie s, todas las subsecuencias o k-meros de tamaño k = 200 con un sobrelape(o) de 150 nucleótidos. Esta configuración de k y de o garantiza que cualquier subsecuencia de 150 nucleótidos, como las lecturas generadas por las tecnologías de secuenciación masiva en estudios genómicos o metagenómicos, se encuentre dentro de, al menos, un k-mero.

Los k-meros obtenidos de cada especie se agrupan por similitud al 94%, utilizando la herramienta CD-HIT [7], y se selecciona una secuencia representativa para cada grupo. A partir de esto, se definen dos grupos:

a) $CUEs$ (Clústeres de un elemento): grupos que tienen una sola secuencia específica de la especie, sin similitud con secuencias de otras especies.
b) $CMEs$ (Clústeres de múltiples elementos): grupos de múltiples secuencias compartidas entre distintas especies, por lo que representan información redundante que debería estar relacionada a un nivel taxonómico superior, en este caso género.

Las secuencias representativas de los $CUEs$ y los $CMEs$ se concatenan si son consecutivas y provienen de la misma secuencia. El conjunto final de secuencias representa el pangenoma viral de la especie s.

Este procedimiento se repite utilizando todos los pangenomas virales generados de todas las especies asociadas al género G. De este modo, se obtienen secuencias específicas que permiten identificar especies virales con precisión, derivadas de los $CUEs$, y secuencias compartidas a nivel de género que son derivadas de los $CMEs$. Estas últimas secuencias compartidas representan la información que se hubiera obtenido de un análisis LCA. En el pangenoma a nivel de género, los $CUEs$ y los $CMEs$ se denominan CGs y CSs, respectivamente, por las secuencias representativas de géneros y especies.

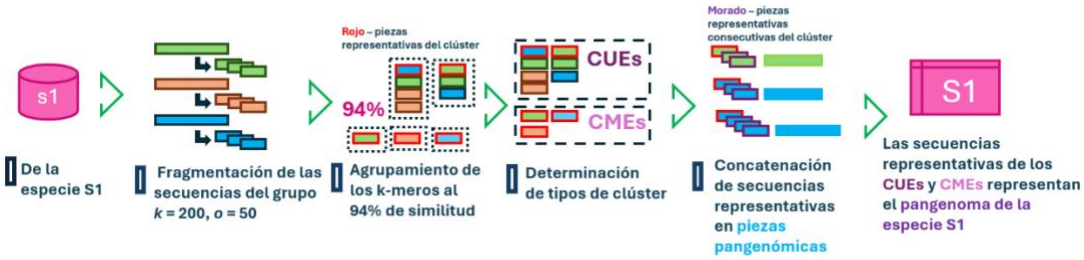

Fig. 3. Esquema de metodología para obtención de pangenomas virales.

Resultados de la compresión y validación del pangenoma

Para evaluar la eficacia de la metodología, se construyó un índice de búsqueda rápida del pangenoma viral. La validación se realizó mediante la simulación de más de 105 millones de lecturas de 150 nucleótidos, pertenecientes a las 31,573 especies virales, las cuales se mapearon contra este índice para encontrar su anotación a nivel de especie o género.

Los resultados mostraron un desempeño sobresaliente, con una recuperación promedio del 99.86% de las secuencias simuladas, lo que indica una alta precisión y sensibilidad. Estos valores confirman que la reducción de la base de datos a pangenomas no compromete la capacidad de identificación, permitiendo una clasificación eficiente.

En cuanto a la eficiencia de la compresión, se observa que, de más de 9 mil millones de nucleótidos iniciales, el pangenoma final representa solo el 12.8% del total. Esto permite realizar búsquedas eficientes, utilizando poco más de 1.1 mil millones de nucleótidos, sin sacrificar resolución a nivel de especie y género.

VirDetect-AI: una nueva forma de ver la diversidad viral eucarionte

En la última década, se han desarrollado herramientas computacionales para identificar virus en datos metagenómicos, como VirFinder, DeepVirFinder y VirSorter2, entre otras [8], [9], [10]. Sin embargo, la mayoría analiza secuencias a nivel de ADN, lo que limita estos métodos en identificar virus con baja homología con secuencias de referencia. Además, muchas funcionan como clasificadores binarios, distinguiendo sólo entre secuencias virales y no virales, sin aportar información sobre la familia, género o especies virales identificadas, la función de las proteínas virales o el posible hospedero. Además, estas herramientas suelen ser específicas en un ámbito, por ejemplo, la identificación sólo de bacteriófagos o de muestras humanas, lo que limita su desempeño en la detección de virus eucariontes y deja una fracción considerable de la diversidad viral sin explorar.

Para enfrentar esta problemática, desarrollamos VirDetect-AI [3], cuyo proceso se muestra en la Fig. 4. VirDetect-AI es una herramienta innovadora, basada en inteligencia artificial, diseñada específicamente para la identificación de virus eucariontes en muestras metagenómicas.

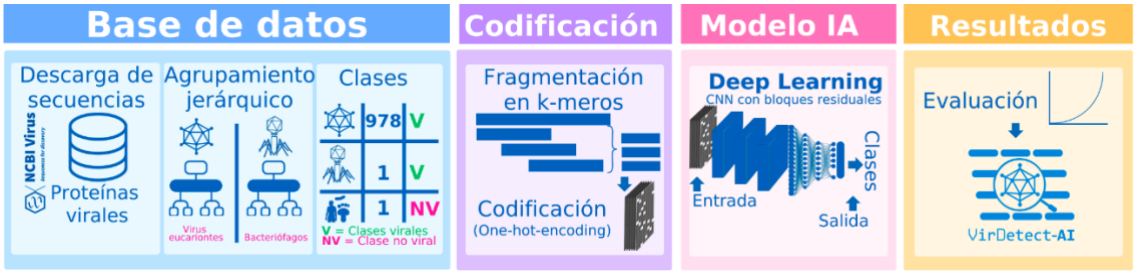

Fig. 4. Metodología detallada para el desarrollo de la herramienta VirDetect-AI para la identificación de secuencias de proteínas virales eucariontes en datos metagenómicos.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

VirDetect-AI trabaja a nivel de aminoácidos (aa) para identificar proteínas virales o fragmentos de éstas. Este cambio es clave, ya que las proteínas suelen conservar motivos funcionales, incluso cuando sus secuencias de nucleótidos cambian considerablemente. De este modo, se puede identificar relaciones evolutivas aun cuando la similitud entre las secuencias sea baja. Esto permite identificar virus poco parecidos a los ya descritos en las bases de datos actuales [11].

Para poder construir esta herramienta, se descargaron secuencias de proteínas virales de NCBI [6]. A continuación, estas secuencias fueron agrupadas de manera jerárquica usando umbrales de similitud desde el 80% hasta el 30%. Posteriormente, a partir de cada secuencia, se generaron k-meros de 300 aa de longitud con saltos de 20 aa.

VirDetect-AI está basado en un modelo de deep learning que combina arquitecturas de redes neuronales convolucionales (CNN) [12] con bloques residuales (ResNet) [13]. Éste clasifica secuencias en 978 clases de proteínas virales eucariontes, abarcando decenas de familias virales y géneros que representan diversas funciones y tipos de hospederos. De esta forma, no sólo identifica si una secuencia es viral, sino que ofrece una clasificación mucho más rica y detallada, lo que permite interpretar mejor los resultados biológicos.

Asimismo, VirDetect-AI incluyó una clase viral negativa que identifica bacteriófagos y una clase negativa para identificar secuencias de humano, bacterias, hongos y arqueas. El total de datos se dividió en 80% para entrenamiento, 10% para validación y 10% como conjunto de prueba.

Identificación de arreglos CRISPR en genomas bacterianos

Como ya mencionamos, los bacteriófagos o fagos son impulsores clave de la evolución bacteriana, moldeando la composición, diversidad y dinámica de sus comunidades. Comprender estas interacciones fago-bacteria es fundamental para desentrañar la dinámica ecológica microbiana [14].

Para contrarrestar los ataques virales, las bacterias han desarrollado diversos sistemas de defensa, entre ellos CRISPR-Cas (*Clustered Regularly Interspaced Short Palindromic Repeats*

- CRISPR-*associated protein*), el cual funciona como un sistema inmune adaptativo, ya que almacena fragmentos del ADN fago invasor como memoria molecular. Cuando una bacteria sobrevive a una infección, integra fragmentos del genoma del fago invasor (conocidos como espaciadores) en un arreglo CRISPR. Esto le permite reconocer y neutralizar al mismo virus en infecciones posteriores [15].

Los arreglos CRISPR-Cas tienen una estructura característica, secuencias de repetidores CRISPR (*Direct Repeats* o DR) separadas por espaciadores únicos. Estas repeticiones suelen tener entre 23 y 55 nucleótidos de longitud y están separadas por espaciadores o distancias específicas que van de los 16 a 109 nucleótidos.

Actualmente, no existe un método computacional que permita predecir con alta confiabilidad las interacciones fago-bacteria. No obstante, la identificación precisa de sistemas CRISPR, junto con la extracción y el análisis comparativo de sus espaciadores, representan una aproximación prometedora, al ser un registro molecular de encuentros históricos entre bacterias y fagos [16].

A diferencia de los métodos tradicionales, basados en conservación de secuencias, nuestro enfoque, ilustrado en la Fig. 5, utiliza aprendizaje profundo capaz de identificar arreglos CRISPR poco conservados. El conjunto positivo se construyó con los sistemas CRISPR-Cas validados experimentalmente de la base de datos CRISPRCasdb [17]. El conjunto negativo combinó secuencias de nucleótidos de baja complejidad y secuencias sintéticas del trabajo de Wang & Liang en 2017 [18]. El conjunto de datos final incluyó 84,065 ejemplos positivos (DR) y 85,949 negativos (no-DR), dividido en 80% para entrenamiento y 20% para prueba/validación. Todas las secuencias se fragmentaron en k-meros de 30 nucleótidos, usados como entrada del modelo.

La arquitectura del modelo integra una capa de *embedding* para codificar numéricamente los nucleótidos de las secuencias, capas convolucionales (Conv1D) para detectar patrones locales, capas Bi-LSTM para modelar dependencias de mayor alcance en ambas direcciones de la secuencia y una capa densa para generar una salida binaria (DR/no-DR). Se evaluaron diferentes combinaciones de hiperparámetros incluyendo capas Conv1D, número de filtros y tamaño del *kernel* en cada Conv1D; número de unidades y profundidad en Bi-LSTM; y valores de regularización *dropout*.

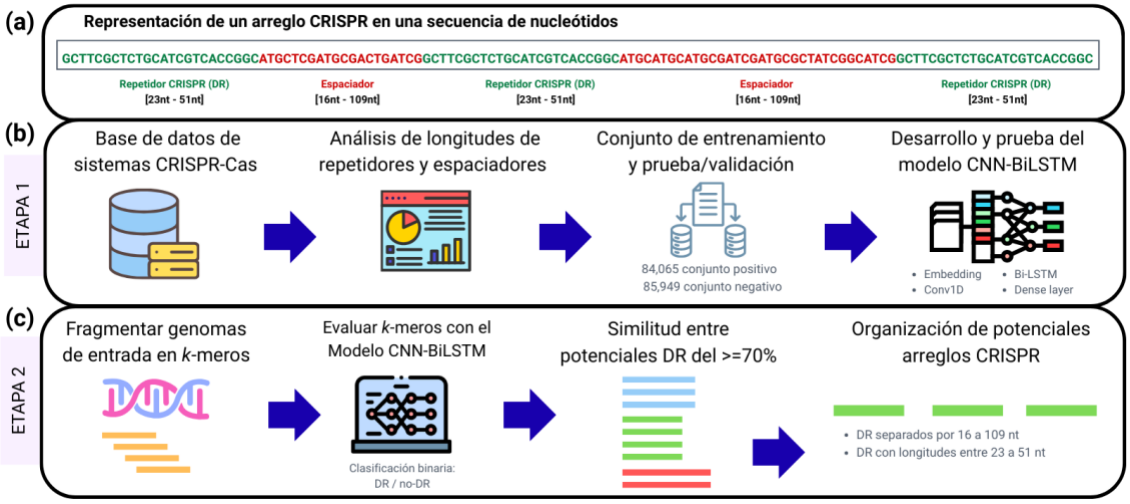

Fig. 5. (a) Representación de un arreglo CRISPR en una secuencia de nucleótidos. (b) Etapa 1 del desarrollo del modelo CNN-BiLSTM. (C) Etapa 2 del análisis bioinformático para validar los candidatos a arreglos CRISPR.

La arquitectura del modelo con mejor desempeño consistió en una capa de *embedding* seguida de dos capas convolucionales Conv1D de 64 y 32 filtros de tamaño 7, normalización por lotes (*batch normalization*), reducción de dimensionalidad mediante *max pooling* y *dropout* de 0.2. Posteriormente, se incorporó una capa Bi-LSTM de 64 unidades, así como una capa densa de 128 neuronas y, finalmente, la capa de salida con una única unidad.

Este modelo alcanzó una exactitud y precisión del 98.4% sobre el conjunto de prueba, lo cual muestra la capacidad del enfoque para discriminar de manera robusta entre repetidores directos (DR) y secuencias no-DR, incluso en escenarios con variación en los patrones de secuencia.

En la segunda etapa, implementamos un proceso bioinformático para validar los DR identificados por el modelo y pertenecientes al mismo sistema CRISPR, usando tres criterios: i) distancia espacial entre los DR, que es de 16 a 109 nucleótidos (rango típico de espaciadores); ii) similitud entre DR, que debía ser ≥70%; y iii) organización característica, que consiste en el patrón repetitivo de arreglos CRISPR auténticos.

En conjunto, este enfoque combina un modelo de aprendizaje profundo para la detección de repeticiones directas y un proceso bioinformático para delimitar DRs y extraer

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

los espaciadores, ofreciendo dos ventajas: mayor flexibilidad para identificar arreglos atípicos, que no se detectan por homología, y la aplicabilidad a secuencias cortas típicas de metagenomas fragmentados, donde los métodos convencionales muestran sensibilidad limitada [19].

Conclusión

En este trabajo, se presentan tres aplicaciones basadas en la inteligencia artificial y el supercómputo que contribuyen a los estudios de metagenómica viral: (i) K-FluDB, una herramienta de optimización de la base de datos de secuencias de referencia virales, mediante la construcción de pangenomas virales a nivel especie y género, para mejorar la clasificación taxonómica de virus; (ii) VirDetect-AI, una nueva herramienta para clasificar proteínas virales y descubrir nuevas, ampliando el repertorio funcional del viroma; y (iii) la identificación de arreglos CRISPR en genomas bacterianos para ampliar el estudio de las interacciones fago-bacteria en datos metagenómicos.

Estas nuevas herramientas contribuyen al análisis de datos metagenómicos virales, teniendo un impacto en la vigilancia genómica, el estudio de la diversidad viral y la comprensión del papel de los bacteriófagos en diferentes ecosistemas.

Financiamiento

Este trabajo fue financiado por los proyectos PAPIIT-DGAPA-IN230523 y PAPIIT-DGAPA-IN225126 de la DGAPA-UNAM (a B.T.), así como por el proyecto CBF-2025-I-1026 de SECIHTI (a B.T.).

Agradecimientos

Agradecemos a Jérôme Verleyen, Roberto Bahena y Juan Manuel Hurtado por su invaluable apoyo en el ámbito computacional.

Referencias

[1] A. D. Rowan-Nash, B. J. Korry, E. Mylonakis, and P. Belenky, "Cross-Domain and Viral Interactions in the Microbiome," *Microbiology and Molecular Biology Reviews*, vol. 83, no. 1, Feb. 2019, doi: 10.1128/MMBR.00044-18.

[2] A. Uscanga Junco, L. Díaz-González, and B. Taboada, "K-FluDB: A Novel K-Mer Based Database for Enhanced Genomic Surveillance of Influenza A Viruses," *Bioinformatics Advances*, Oct. 2025, doi: 10.1093/bioadv/vbaf254.

[3] A. Zárate, L. Díaz-González, and B. Taboada, "VirDetect-AI: a residual and convolutional neural network–based metagenomic tool for eukaryotic viral protein identification," *Brief. Bioinform.*, vol. 26, no. 1, Nov. 2024, doi: 10.1093/bib/bbaf001.

[4] E. J. Black, C. S. Powell, D. M. Dempsey, R. C. Hendrickson, L. R. Mims, and E. J. Lefkowitz, "Virus taxonomy: the database of the International Committee on Taxonomy of Viruses," *Nucleic Acids Res.*, vol. 54, no. D1, pp. D776–D789, Jan. 2026, doi: 10.1093/nar/gkaf1159.

[5] Y. Wang, T. S. Korneliussen, L. E. Holman, A. Manica, and M. W. Pedersen, "ngs LCA - A toolkit for fast and flexible lowest common ancestor inference and taxonomic profiling of metagenomic data," Methods Ecol. Evol., vol. 13, no. 12, pp. 2699–2708, Dec. 2022, doi: 10.1111/2041-210X.14006.

[6] National Center for Biotechnology Information (NCBI), "Virus genomes – All nucleotide sequences," NCBI FTP Server. [Online]. Available: https://ftp.ncbi.nlm.nih.gov/genomes/Viruses/AllNucleotide/. [Accessed: Aug. 24, 2025].

[7] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012, doi: 10.1093/bioinformatics/bts565.

[8] J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, and F. Sun, "VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data," Microbiome, vol. 5, no. 1, p. 69, Dec. 2017, doi: 10.1186/s40168-017-0283-5.

[9] J. Ren, K. Song, C. Deng, N. A. Ahlgren, J. A. Fuhrman, Y. Li, X. Xie, R. Poplin, and F. Sun, "Identifying viruses from metagenomic data using deep learning," *Quantitative Biology*, vol. 8, no. 1, p. 64, 2020, doi: 10.1007/s40484-019-0187-4.

[10] J. Guo *et al.*, "VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses," *Microbiome*, vol. 9, no. 1, p. 37, Dec. 2021, doi: 10.1186/s40168-020-00990-y.

[11] C. Chothia and A. M. Lesk, "The relation between the divergence of sequence and structure in proteins.," *EMBO J.*, vol. 5, no. 4, pp. 823–826, Apr. 1986, doi: 10.1002/j.1460-2075.1986.tb04288.x.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[14] L. Beller and J. Matthijnssens, "What is (not) known about the dynamics of the human gut virome in health and disease," *Curr. Opin. Virol.*, vol. 37, pp. 52–57, Aug. 2019, doi: 10.1016/j.coviro.2019.05.013.

[15] E. V. Koonin and K. S. Makarova, "Origins and evolution of CRISPR-Cas systems," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 374, no. 1772, p. 20180087, May 2019, doi: 10.1098/rstb.2018.0087.

[16] D. J. Nasko, B. D. Ferrell, R. M. Moore, J. D. Bhavsar, S. W. Polson, and K. E. Wommack, "CRISPR Spacers Indicate Preferential Matching of Specific Virioplankton Genes," *mBio*, vol. 10, no. 2, Apr. 2019, doi: 10.1128/mBio.02651-18.

[17] C. Pourcel *et al.*, "CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers," *Nucleic Acids Res.*, Oct. 2019, doi: 10.1093/nar/gkz915.

[18] K. Wang and C. Liang, "CRF: detection of CRISPR arrays using random forest," *PeerJ*, vol. 5, p. e3219, Apr. 2017, doi: 10.7717/peerj.3219.

[19] C. Coclet and S. Roux, "Global overview and major challenges of host prediction methods for uncultivated phages," *Curr. Opin. Virol.*, vol. 49, pp. 117–126, Aug. 2021, doi: 10.1016/j.coviro.2021.05.003.