Evaluación de un programa para la generación de bases de datos con secuencias del año 2017 del gen de la Hemaglutinina del virus de influenza A H1N1

Cristina Isabel Acuña Carvajal

Trabajo de Grado para Optar el título de Bióloga

Director

Carlos Jaime Barrios Hernández

Doctorado en Ciencias

Tutor

Francisco José Martínez Pérez

Doctorado en Ciencias

Universidad Industrial de Santander

Facultad de Ciencias Básicas

Escuela de Biología

Bucaramanga

2019

Agradecimientos

Mi gratitud es primeramente para Dios, por permitirme llegar al final de este largo y hermoso camino; a mis padres Alfonso Acuña y Carmenza Carvajal por todo su esfuerzo, dedicación y trabajo para darme la oportunidad de formarme como Bióloga en la Universidad Industrial de Santander, a mi nonita Ana Isabel Navarro quien ha sido un gran apoyo en mi vida, a mis tías por su preocupación, a Sneider Rodríguez quien siempre me ha dado su opinión sincera y me ha apoyado incondicionalmente en todas mis metas y sueños.

A todos los docentes que a lo largo de estos 5 años han compartido sus conocimientos conmigo, en especial a Francisco José Martínez Pérez quien ha sido más que un formador académico, un amigo, que desde los inicios de mi carrera me ha enseñado el valor de la incondicionalidad y quién ha confiado plenamente en mis capacidades.

A todos mi amigos “Los de siempre” y en especial a Liss y Vane, que desde el inicio de la carrera más que compañeras y amigas, han sido las hermanas que nunca tuve, por tantos trasnochos estudiando para los parciales o haciendo los informes, por los momentos de esparcimiento y diversión, que son momentos que guardo en mi corazón.

A todos los compañeros y amigos del Laboratorio de Genómica de Celomados, quienes con todos sus aportes han complementado mi formación académica y mi formación personal.

Mi gratitud al Grupo de Investigación de Cómputo Avanzado y a Gran Escala, quienes me acogieron para el desarrollo de mi trabajo de grado.

Tabla de Contenido

Pág.

Introducción 14

1. Objetivos 19
1.1 Objetivo General 19
1.2 Objetivos Específicos 19

2. Competencias 20
2.1 Competencias cognitivas 20
2.2 Competencias actitudinales 21

3. Materiales y métodos 21
3.1 Obtención de secuencias del gen de la Hemaglutinina del virus de Influenza A H1N1 del año 2017 21
3.2 Construcción de la base de datos del gen de la Hemaglutinina del virus de Influenza A H1N1 con el programa BioDataToolKit v1.0 22
3.3 Revisión y depuración de la base de datos 22
3.4 Verificación de la validez biológica de la base de datos y análisis bioinformáticos 23
3.5 Análisis filogenético 24

4. Resultados. 24

5. Discusión 39

6. Conclusiones 44

7. Recomendaciones 45

Referencias bibliográficas 46

Lista de Tablas

Pág.

Tabla 1. Palabras claves para la identificación del gen de la Hemaglutinina del virus de Influenza A H1N1 del año 2017 en el GenBank. 24

Lista de Figuras

Pág.

Figura 1. Ejemplo de la búsqueda del gen de la Hemaglutinina del virus de Influenza A H1N1 del año 2017 por medio de palabras claves en el GenBank 25

Figura 2. Procedimiento de la ejecución del programa BioDataToolkit v.1.0 para la obtención de la hoja de Excel con la información y las secuencias nucleotídicas del virus de Influenza A H1N1 26

Figura 3. Ejemplo del contenido de la información de cada secuencia del gen HA del virus de Influenza A H1N1del año 2017 tomada del formato GenBank full para la generación del archivo Excel 27

Figura 4. Archivo de salida de la versión 1.0 del programa BioDataToolkit 30

Figura 5. Resultado de la evaluación de la versión 2.0 del programa BioDataToolkit 31

Figura 6. Evidencia de los resultados generados por la versión 3.0 del programa para la integración de la información de genes reportados con dos CDS 32

Figura 7. Desface en la información de las secuencias del gen HA del virus de Influenza A H1N1 generado por la versión 3.0 del programa BioDataToolkit 33

Figura 8. Espacios vacíos en la información de las secuencias del gen HA del virus de Influenza A H1N1 generado por la versión 4.0 del programa BioDataToolkit 34

Figura 9. Archivo de salida de la versión 5.0 del programa BioDataToolkit con el cambio en el nombre de las secuencias de acuerdo con el código diseñado por el Grupo de Investigación 35

Figura 10. Uso de la herramienta de filtros de Excel para la depuración de la base de datos de las secuencias nucleotídicas del gen HA del virus de Influenza A H1N1 36

Figura 11. Generación de las secuencias consensos mensuales de las secuencias nucleotídicas del gen HA del virus de Influenza A H1N1 37

Figura 12. Árbol filogenético con las secuencias consenso del gen HA del virus de Influenza A H1N1 39

Lista de Apéndices

(Ver apéndices adjuntos en el CD y pueden visualizarlos en la Base de Datos de la Biblioteca UIS)

Apéndice A. Alineamiento del mes de enero

Apéndice B: Alineamiento del mes de febrero

Apéndice C: Alineamiento del mes de marzo

Apéndice D: Alineamiento del mes de abril

Apéndice E: Alineamiento del mes de mayo

Apéndice F: Alineamiento del mes de junio

Apéndice G: Alineamiento del mes de julio

Apéndice H: Alineamiento del mes de agosto

Apéndice I: Alineamiento del mes de septiembre

Apéndice J: Alineamiento del mes de octubre

Apéndice K: Alineamiento del mes de noviembre

Apéndice L: Alineamiento del mes de diciembre

Apéndice M: Alineamientos tipo BLAST

Resumen en Español

TITULO: Evaluación de un programa para la generación de bases de datos con secuencias del año 2017 del gen de la Hemaglutinina del virus de influenza A H1N1*

AUTOR: Cristina Isabel Acuña Carvajal **

PALABRAS CLAVE: Virus Influenza A H1N1, bases de datos, bioinformática, minería de datos.

DESCRIPCIÓN: En la pandemia de Influenza A H1N1 del 2009, algunos pacientes que presentaban la sintomatología de infección por este virus eran diagnosticados como falsos negativos por la RT-PCR, debido a la ausencia en la polimerización de los genes para la Hemaglutinina (HA), Nucleocápside y las Proteínas de Matriz M1 y M2. Con una base de datos que incluyó todas las secuencias genómicas hasta el año 2010 del virus, generada por 10 personas durante 18 meses, se determinó que el resultado fue debido a procesos evolutivos del genoma viral; por ello, fueron diseñados nuevos cebadores que diagnosticaron la infección en 150 pacientes. Para solucionar los tiempos de construcción de la base de datos, se generó el programa BioDataToolkit v1.0 cuyo objetivo fue obtener del GenBank: la fecha de colección, país, hospedero, organismo, segmento, serotipo, cepa, el número de acceso y el ORF de cada cepa, para ubicarlos por columnas en Excel, para manejar la información en minutos. Sin embargo, el programa no había sido determinado a nivel biológico, lo cual se realizó en esta pasantía con el gen HA del virus de Influenza A H1N1 del 2017. Se determinó que la combinación óptima para obtener la mayoría de las secuencias fue “Influenza a virus 4 segment h1n1 2017 complete CDS”; con ellas, se generó un formato GenBank full que empleó el programa para su análisis. Desde la versión 1.0 se generó la página de Excel, pero la información

no permitía análisis biológicos por tanto los programadores realizaron las modificaciones requeridas en cada validación hasta generar la versión 5.0 la cual permite obtener la información de cada secuencia en columnas y los formatos Fasta en minutos, para la generación de secuencias consenso y análisis filogenéticos. Sin embargo, es necesario la generación de la v6.0 para concluir la optimización del programa BioDataToolkit.

*Pasantía de investigación
**Facultad de Ciencias Básicas. Escuela de Biología. Tutor: Francisco José Martínez Pérez

Resumen en Inglés

Title: Evaluation of a program for the generation of databases with sequences of the year 2017 of the influenza virus A H1N1 Hemagglutinin gene*

Author: Cristina Isabel Acuña Carvajal **

Key words: Influenza A H1N1 virus, databases, bioinformatics, big data.

Description: In the pandemic Influenza A H1N1 of 2009, some patients who had the symptomatology of infection with this virus were diagnosed as negatives false by RT-PCR, due to the absence in polymerization of genes for Hemagglutinin (HA), Nucleocapsid (HA), and matrix proteins M1 and M2. With a database that included all genomic sequences up to the year 2010 of the virus, generated by 10 people for 18 months, it was determined that the result was due to evolutionary processes of the viral genome. Therefore, new primers that diagnosed the infection in 150 patients were designed. In order to solve the construction times of the database, the BioDataToolkit v1.0 program was generated whose objective was to obtain from GenBank: the collection date, country, host country, organism, segment, serotype, strain, the access number and ORF of each strain, to be placed by columns in Excel, to handle the information in minutes. However, the program had not been determined biologically, which was done in this internship with the HA gene of the Influenza A H1N1 virus of 2017. It was determined that the optimal combination to obtain most sequences was "Influenza a virus 4 Segment h1n1 2017 complete CDS"; with them, a full GenBank format was generated that used the program for analysis. Since version 1.0 the Excel page was generated but the information did not allow biological analysis. Therefore, the programmers made the modifications required in each validation until generating

version 5.0 which allows to obtain the information of each sequence in columns and Fasta formats in minutes, for the generation of consensus sequences and phylogenetic analysis. However, the generation of the V6.0 is necessary to complete the optimization of the BioDataToolkit program.

*Pasantía de investigación
**Facultad de Ciencias Básicas. Escuela de Biología. Tutor: Francisco José Martínez Pérez

Introducción

El virus de la gripe Influenza A H1N1 pertenece a la familia Orthomyxoviridae, el cual contiene 8 genes de ARN monocatenario (Christman, Kedwaii, Xu, Donis, & Lu, 2011), los subtipos del virus de Influenza están determinados por los genes codificantes para las glucoproteínas Hemaglutinina (HA) y Neuraminidasa (NA) que son proteínas de membrana (Stefańska, Dzieciatkowski, Brydak, & Romanowska, 2013); adicionalmente, estos genes presentan altas tasas mutacionales, lo que ocasiona cambios en las secuencias codificantes de HA, NA y reestructuran la composición de la membrana vírica, logrando evadir la defensa inmunitaria del hospedero (Peteranderl, Herold, & Schmoldt, 2016); estos dos genes junto con el que codifica para dos Proteínas de Matriz (M1 y M2) que por empalme alternativo forma un canal iónico (Wise et al., 2012), son los usados para realizar el diagnóstico en pacientes (Cui et al., 2016) (Ortiz, Rojo, & Sanz, 2019). Actualmente, existen 18 subtipos conocidos de HA y 11 subtipos conocidos de NA (White & Lowen, 2018) presentes en animales como aves, cerdos, humanos, entre otros (Ozawa & Kawaoka, 2013). Los virus de Influenza en especial el A H1N1 ha sido causante de pandemias como la de 1918 (Cox & Subbarao, 2000) y la más reciente en el 2009 (Christman et al., 2011).

De acuerdo a lo anterior, en el año 2009 la Organización Mundial de la Salud (OMS) declaró el inicio de la pandemia provocada por el virus de la gripe Influenza A H1N1 (Chan, 2009) y hasta agosto del año 2010 fue declarado el periodo post-pandémico (WHO, 2010). Para el

diagnóstico de esta enfermedad se implementaron diferentes métodos: entre ellos la inmunofluorescencia, cultivo viral, detección rápida de antígenos y detección molecular (McMullen, Anderson, & Burnhamfor, 2016); sin embargo, el método más usado desde ese entonces a la fecha ha sido la Reacción de Polimerización en Cadena con Transcriptasa Reversa en Tiempo Real (RT-PCR) de un paso (Cui et al., 2016). No obstante, en la pandemia del 2009 aun cuando los pacientes presentaban toda la sintomatología y eran diagnosticados clínicamente como infectados con el virus de Influenza A H1N1, al realizar la RT-PCR de un paso, el resultado era negativo para el virus de Influenza A H1N1 en algunos casos (González Barrios et al., 2016) (González Barrios et al., 2016 B).

Dado lo anterior, para determinar las causas del resultado antes mencionado dentro de la pandemia del 2009, el Consejo Nacional de Ciencia y Tecnología de México autorizó el proyecto titulado "Caracterización molecular de cepas atípicas del Virus de Influenza A H1N1", el cual fue realizado en el Laboratorio de medicina genómica del Hospital Regional 1^{ero} de Octubre del Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado (ISSSTE) y por el Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV). Posteriormente, en el año 2011 se da inicio al proyecto de colaboración internacional entre la Universidad Industrial de Santander (UIS) y las instituciones previamente mencionadas, con el proyecto titulado "Diseño de una placa de 96 pozos para la identificación de cepas de virus de Influenza A H1N1 de pacientes por RT-PCR en Tiempo Real de un solo paso". En donde, por medio de métodos bioinformáticos, evolutivos, ingeniería genética y biomedicina se determinó que la causa de los resultados falsos negativos en la RT-PCR de un paso, con muestras de pacientes que presentaban toda la sintomatología causada por el virus de Influenza A

H1N1, era debido a los procesos evolutivos en los genes virales usados para el diagnóstico (González Barrios et al., 2016) (González Barrios et al., 2016 B).

Con base en estas conclusiones y a las secuencias nucleotídicas que corroboran el patrón evolutivo de tres de los ocho genes virales, fueron diseñados nuevos cebadores y condiciones de RT-PCR de un paso; un juego de ellos fue diseñado para amplificar a partir del codón de inicio de la traducción del gen de la cápside viral, su capacidad para el diagnóstico se determinó en el Laboratorio de Medicina Genómica del Hospital Regional 1^{ero} de Octubre del ISSSTE en la ciudad de México, con 150 pacientes que presentaban el cuadro clínico antes mencionado y a diferencia del diagnóstico oficial, el nuevo juego de cebadores identificó al virus de Influenza A H1N1 lo que les salvo la vida a las personas infectadas.

El éxito de los proyectos antes mencionados fue gracias a una base de datos inicial de 31,835 secuencias con la cual se realizaron todos los análisis evolutivos de este virus. Inicialmente, esta fue construida con todas las secuencias virales de Influenza A H1N1 de seres humanos que tenían el marco abierto de lectura (ORF) completo; es decir, la base de datos estuvo constituida con un total de 37,708 secuencias reportadas hasta el año 2010 en la base de datos del GenBank del Centro Nacional de Información Biotecnológica (NCBI) (Agarwala et al., 2018). En esa época, su elaboración tardó aproximadamente año y medio con el trabajo de 10 personas en jornadas de búsqueda de 4 a 6 horas por día. La selección de cada secuencia de la base de datos fue de forma manual debido a que al realizar las exploraciones en el NCBI con los criterios de búsqueda, es decir, las palabras claves, independientemente de cada gen los datos obtenidos siempre mostraban secuencias: con diferente fecha de colección y reporte a las del periodo de

búsqueda seleccionado, incluían otras cepas virales como la H2N2, H3N2 entre otras, mostraban otros segmentos que no correspondían con el criterio de búsqueda, además de secuencias cuyo ORF era parcial o en su defecto se encontraban nucleótidos indeterminados.

En búsqueda de soluciones a la problemática anterior de la edición de la base de datos, el Grupo de Investigación, generó inicialmente una plantilla de Excel que ubicaba y producía el cambio de nombre a cada secuencia en formato Fasta con el código que identifica a cada una de las secuencias para su aplicación en supercómputo y otras plataformas. Sin embargo, también era necesario hacer una revisión manual del resultado, adicionalmente la plantilla de Excel no permitía la visualización o uso de cualquier otro dato de la información suministrada de cada una de las secuencias reportadas en el GenBank.

Motivo por el cual el Grupo de Investigación continuó generando una serie de programas para dar una solución y poder obtener bases de datos de genes virales en tiempos operativos menores a 8 horas para cualquier usuario. Lo anterior es fundamental independientemente del gen de estudio debido a que, la cantidad de secuencias que son depositadas en bases de datos públicas mundiales como la del NCBI han aumentado exponencialmente debido a los nuevos métodos de secuenciación de nueva generación que generan miles de secuencias en menos de una semana. Un ejemplo de esto es que al 22 de septiembre del 2018 como se indicó en la propuesta inicial se tenían reportadas en el GenBank 169.125 secuencias del virus de Influenza A H1N1, de las cuales 39.970 codifican para la HA, ahora al 5 de junio del 2019 del virus de Influenza A H1N1 se encuentran 188.611 secuencias reportadas, de las cuales 43.055 corresponden a la glucoproteína HA (Benson et al., 2017). De allí la importancia de tener sistemas de computación

que permitan obtener y actualizar cualquier base de datos incluyendo la que se obtuvo con las secuencias reportadas hasta el año 2010.

Actualmente, los avances computacionales de otros centros de investigación continúan de forma simultánea a los realizados por el Grupo de Investigación, una muestra de esto es el desarrollo de herramientas informáticas para la construcción de bases de datos como son la del NCBI del virus de Influenza denominada *Influenza Virus Resource. Information, Search and Analysis* (Bao et al., 2008) y la de *Influenza Research Database* (IRD) (National Institute of Allergy and Infectious Diseases (NIH / DHHS), 2017). Aun cuando estas herramientas permiten una selección de los virus de influenza por: cepa, organismo infectado, gen, año, mes, día y la generación de un formato Fasta, los resultados no permiten el manejo de toda la información de las secuencias requerida para análisis evolutivos o terapéuticos entre otros; por tanto, el Grupo de Investigación ha desarrollado un programa computacional llamado BioDataToolkit el cual tiene como objetivo disminuir los tiempos de elaboración de las bases de datos, para su posterior depuración y uso en análisis biológicos, sin embargo, la validez biológica de la información que genera no ha sido establecida por un experto en Biología. Es por ello, que este programa fue evaluado en este trabajo, para constatar su validez biológica empleando como criterio de funcionamiento y rendimiento las secuencias del gen de la HA de seres humanos infectados en el año 2017 reportadas en el GenBank y en caso necesario hacer las recomendaciones pertinentes para con ello contribuir a la generación de conocimiento de los procesos evolutivos del virus de Influenza A H1N1.

1. Objetivos

1.1 Objetivo General

Contribuir a la generación de conocimiento de los procesos evolutivos del virus de Influenza A H1N1 por medio de la evaluación del programa BioDataToolkit versión 1.0 diseñado dentro del proyecto UIS-ISSSTE-CINVESTAV, que permite la generación de bases de datos, con secuencias del gen de la Hemaglutinina del año 2017.

1.2 Objetivos Específicos

Evaluar el funcionamiento del algoritmo del programa para la construcción de bases de datos con secuencias del año 2017 del gen de la Hemaglutinina del virus de Influenza A H1N1 mediante la generación de formatos fasta y una hoja de cálculo con los datos de identificación de cada secuencia.

Determinar la capacidad de categorización del algoritmo, para la clasificación del gen de la Hemaglutinina del virus de Influenza A H1N1 de acuerdo con el organismo infectado, lugar, fecha y procedencia de la cepa.

Establecer el funcionamiento de base de datos del año 2017 para la generación de secuencias consenso del gen de la Hemaglutinina del virus de Influenza A H1N1 para establecer sus patrones evolutivos.

2. Competencias

2.1 Competencias cognitivas

- Diseña búsquedas en la plataforma del GenBank para la obtención de secuencias de ácidos nucleicos para la construcción de bases de datos bioinformáticas para gen de la Hemaglutinina del virus de Influenza A H1N1.
- Identifica los datos obtenidos de las secuencias del GenBank para aplicaciones bioinformáticas con fines de análisis evolutivos propuestos dentro del proyecto.
- Maneja hojas de cálculo para la clasificación de los datos de secuencias nucleotídicas del gen de la Hemaglutinina del virus de Influenza A H1N1 como base para análisis bioinformáticos.
- Construye bases de datos para el análisis bioinformático de ácidos nucleicos del gen de la Hemaglutinina del virus de Influenza A H1N1.
- Clasifica las secuencias de bases de datos nucleotídicas públicas de acuerdo con los criterios requeridos para realizar análisis bioinformáticos.

- Genera resultados de análisis evolutivos para evidenciar los cambios del gen de la Hemaglutinina del virus de Influenza A H1N1 en el año 2017 por medio del método de inferencia bayesiana.
- Desarrolla informes de análisis y discusión de resultados de la base de datos de ácidos nucleicos del gen de la Hemaglutinina para el desarrollo de los objetivos del proyecto de investigación.

2.2 Competencias actitudinales

- Participa activamente en un Grupo de Investigación para resolver preguntas enfocadas en los análisis evolutivos del virus de Influenza A H1N1.
- Integra los conocimientos para tomar decisiones lógicas de forma imparcial y razonada.
- Establece una línea de acción adecuada en la resolución de problemas.

3. Materiales y métodos

3.1 Obtención de secuencias del gen de la Hemaglutinina del virus de Influenza A H1N1 del año 2017

Las secuencias nucleotídicas del gen que codifica para la Hemaglutinina del virus de Influenza A H1N1 del año 2017 fueron descargadas de la base de datos del GenBank

(Benson et al., 2017), para ello se usó como criterio de búsqueda las siguientes palabras claves: “Influenza a virus 4 segment h1n1 2017 complete CDS”, las secuencias obtenidas fueron descargadas en formato GenBank full, el cual contiene toda la información reportada de cada secuencia en el GenBank.

3.2 Construcción de la base de datos del gen de la Hemaglutinina del virus de Influenza A H1N1 con el programa BioDataToolkit v1.0

La construcción de la base de datos se realizó mediante el uso del programa BioDataToolkit versión 1.0 y sus posteriores versiones, las cuales estuvieron en función del análisis de resultados de cada versión evaluada. A la versión inicial del programa y a las posteriores se les ingresó el archivo en formato GenBank Full previamente descargado del GenBank, lo que generó como resultado un archivo de salida que consistió en una tabla en Excel en la cual se encontraba organizada toda la información en columnas por separado, de cada una de las secuencias descargadas del gen de la Hemaglutinina del virus de influenza A H1N1 del año 2017. De la columna denominada Fasta se extrajo la información para la obtención del segundo formato que contenía todas las secuencias nucleotídicas.

3.3 Revisión y depuración de la base de datos

Se realizó la revisión de todas las secuencias del virus, presentes en el archivo Excel verificando que la información de cada una correspondiera a los datos que se encontraban en el formato reportado en el GenBank. Puesto que la versión 1.0 del programa BioDataToolkit inicialmente

estaba diseñada para obtener la información correspondiente a las columnas llamadas: *collection_date, country, host, organism, segment, serotype, strain, ACCESSION, CDS*. Sin embargo, con la evaluación de las siguientes versiones del programa, se fueron incluyendo los siguientes datos: *mol_type, strain, db_xref, note, gene, function, codon_start, producto, protein_id, translation, Fasta, Complete Fasta, Renamed Fastas*. Posteriormente, por medio de la herramienta de filtros de Excel se efectuó la depuración de las secuencias con la eliminación de las secuencias que no correspondieron a la cepa A H1N1 del gen de la Hemaglutinina del *Homo sapiens* del año 2017 y que tuvieron un CDS parcial.

3.4 Verificación de la validez biológica de la base de datos y análisis bioinformáticos

Las secuencias seleccionadas se organizaron de forma mensual. A cada mes se le realizó un alineamiento por medio del programa *Kalign* en su versión en línea (Madeira et al., 2019), con los parámetros de *Gap Open* y *Gap Extension* empleados durante la ejecución del proyecto UIS-ISSSTE-CINVESTAV. Inmediatamente a partir de cada alineamiento mensual, se obtuvo su respectiva secuencia consenso con el programa *BioEdit* (Hall, 1999) y finalmente fueron editados en formato Fasta y almacenados en un archivo de texto plano. Posteriormente, se estableció la capacidad de identificación de las secuencias mensuales consenso de la HA del virus de Influenza A H1N1 de humanos, reportadas en el año 2017 mediante una búsqueda de sus secuencias homólogas con un porcentaje de similitud del 97% al 100% y valores esperados (*e-value*) cercanos a cero, mediante el programa *Basic Local Alignment Search Tool* (BLAST) (Altschul, Gish, Miller, Myers, & Lipman, 1990).

3.5 Análisis filogenético

Se realizó un análisis filogenético mediante el método de Inferencia Bayesiana con el programa MrBayes (Huelsenbeck, JP & F. Ronquist, 2001), para el cual se emplearon las secuencias consenso mensuales del año 2017 correspondientes al gen HA del virus de Influenza A H1N1 aislado de seres humanos; se realizaron 2 carreras con 4 cadenas de Markov-Montecarlo y se corrieron 15 millones de generaciones, con muestreo de cadenas cada 1500 generaciones.

4. Resultados

Obtención de secuencias del gen de la Hemaglutinina del virus de Influenza A H1N1 del año 2017.

Las búsquedas con las palabras claves en la base de datos del GenBank, se inició mediante una palabra general hasta concluir con el criterio de búsqueda específico. Las palabras generales “Influenza” e “Influenza a” mostraron todas las secuencias nucleotídicas reportadas de este virus, las cuales fueron más de un millón (Tabla 1). Al incluir la palabra virus como siguiente criterio de selección la cantidad de secuencias se redujo al 61,43% respecto a las secuencias iniciales. Al adicionar el número “4” y la palabra “segment” que corresponde al gen de la Hemaglutinina se determinó que la búsqueda se restringió al 22% con ambos criterios. Esta disminución en la cantidad de secuencias fue más evidente al incluir la palabra H1N1, en donde únicamente el 7%

correspondió al segmento deseado y de este grupo, el 19% fueron secuencias que indicaban correspondencia al año 2017 y el 17% tenían el ORF completo (Tabla 1).

Es de resaltar, que en cualquiera de las búsquedas con las palabras de selección iniciales siempre se generaron resultados que mostraron: otras secuencias del genoma del virus de Influenza A H1N1, años distintos al de la búsqueda e inclusive otras cepas del virus de Influenza. (Fig. 1.A y 1.B).

Las palabras claves “Influenza a virus 4 segment h1n1” generaron mayor restricción en la búsqueda, ya que el resultado mostró secuencias de la cepa H1N1 del año 2017, sin embargo, también se añadieron a la búsqueda otras cepas diferentes a las del criterio de selección, además, se encontraron secuencias del año 2017 y de otros años, sin que se hubiesen incluido como criterio de búsqueda (Fig. 1.C).

Tabla 1.

Palabras claves para la identificación del gen de la Hemaglutinina del virus de Influenza A H1N1 del año 2017 en el GenBank.

Combinación de palabras claves	**Número de secuencias encontradas en el NCBI**
Influenza	1.054.461
Influenza a	1.054.461
Influenza a virus	647.712
Influenza a virus 4	236.874
Influenza a virus 4 segment	225.755
Influenza a virus 4 segment h1n1	74.007
Influenza a virus 4 segment h1n1 2017	14.285
Influenza a virus 4 segment h1n1 2017 complete CDS	13.155

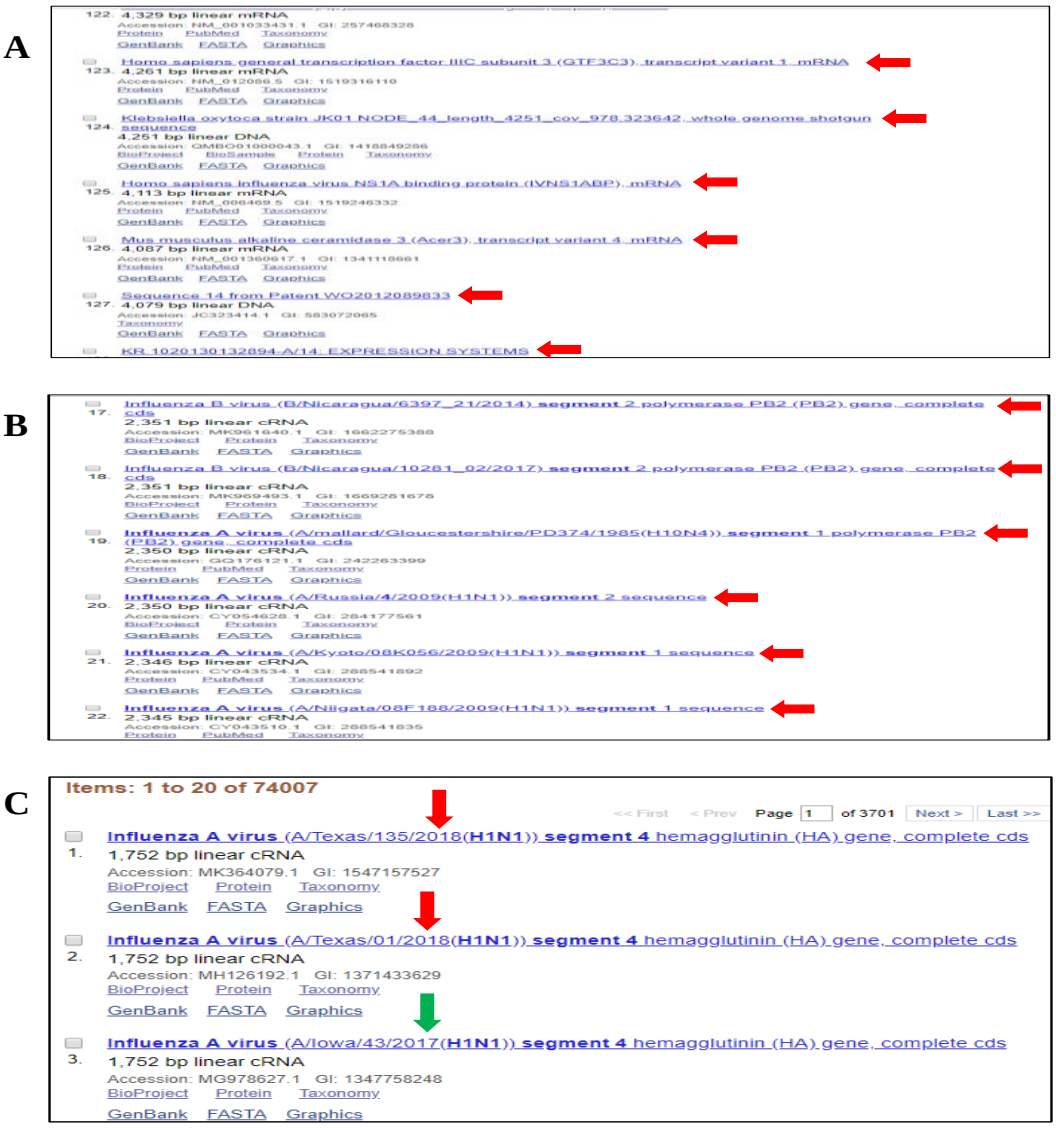


Figura 1. **Ejemplo de la búsqueda del gen de la Hemaglutinina del virus de Influenza A H1N1 del año 2017 por medio de palabras claves en el GenBank.** En "A" se muestra un ejemplo del resultado del uso de las palabras claves "Influenza a", note que la mayoría de las secuencias no corresponden al criterio de búsqueda. Como se indica en la parte "B" de la imagen, con la inclusión de las palabras "virus 4 segment" al criterio anterior se presentaron secuencias del segmento 4 del virus de Influenza, no obstante, también se integraron secuencias de otros segmentos del virus. En la imagen "C", se muestra una parte de la identificación de genes de la cepa H1N1 del virus de influenza A con las palabras claves "Influenza a virus 4 segment h1n1", con ellas se incrementó la selección de la cepa, sin embargo, esta incluyó años no solicitados como criterio de búsqueda. Las flechas rojas indican las secuencias incorrectas, mientras que la flecha verde señala el resultado correcto de la búsqueda.

Construcción de base de datos del gen de la Hemaglutinina del virus de Influenza A H1N1 con el programa BioDataToolkit v 1.0.

Una vez determinadas la mayoría de las secuencias del gen de la Hemaglutinina del virus de Influenza A H1N1 del año 2017, se generó de la página *Nucleotide* del GenBank un archivo en texto plano con toda la información de cada una de las secuencias seleccionadas con el formato GenBank full (Fig. 2.A y 2.B). Posteriormente, para la ejecución del programa BioDataToolkit v1.0 se introdujo dentro de sus comandos el nombre del archivo de texto plano con las secuencias del virus de Influenza A H1N1 (Fig. 2.C), así como el nombre del archivo de salida de Excel (Fig. 2.D), para luego ser ejecutado.

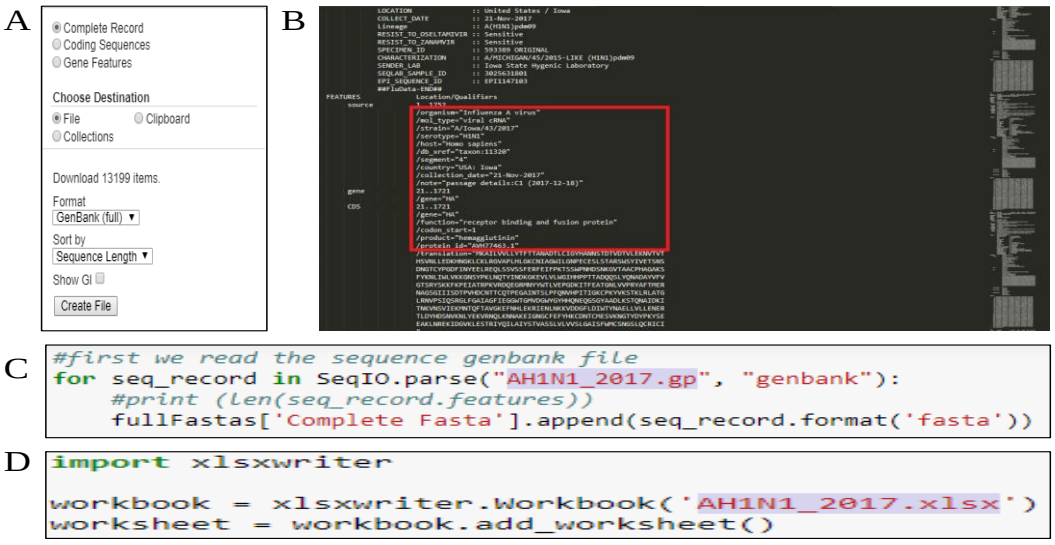


Figura 2. **Procedimiento de la ejecución del programa BioDataToolkit v.1.0 para la obtención de la hoja de Excel con la información y las secuencias nucleotídicas del virus de Influenza A H1N1.** En "A" se observa el procedimiento realizado en el GenBank para la generación del formato GenBank full el cual produjo un archivo de texto plano con toda la información de todas las secuencias seleccionadas como se muestra en "B". La ubicación en

el programa para colocar el nombre del archivo en formato GenBank full y el de salida, son indicados con letras rojas y resaltadas en color púrpura en las imágenes C y D respectivamente.

Revisión y depuración de la base de datos

En un principio el programa BioDataToolkit v.1.0 se diseñó para colocar en cada columna de la hoja de cálculo de Excel la información seleccionada de cada cepa en el GenBank, es decir, la correspondiente a: la fecha de colección, el país, el hospedero, el organismo, el segmento, el serotipo, la cepa, el número de acceso del GenBank y el ORF indicado como CDS en el GenBank (Fig. 3).

```
LOCUS       MH637478               1777 bp    cRNA    linear   VRL 25-AUG-
2018
DEFINITION  Influenza A virus (A/Baltimore/P0197/2017(H1N1)) segment 4
            hemagglutinin (HA) gene, complete cds.
ACCESSION   MH637478
VERSION     MH637478.1
DBLINK      BioProject: PRJNA297517
KEYWORDS    .
SOURCE      Influenza A virus
  ORGANISM  Influenza A virus
            Viruses; ssRNA viruses; ssRNA negative-strand viruses;
            Orthomyxoviridae; Alphainfluenzavirus.
REFERENCE   1  (bases 1 to 1777)
  CONSRTM   Centers of Excellence for Influenza Research and Surveillance
            (CEIRS)
  TITLE     Direct Submission
  JOURNAL   Submitted (17-JUL-2018) The Johns Hopkins Center for Excellence
in
            Influenza Research and Surveillance (JH-CEIRS), Baltimore, MD
            21287, USA
COMMENT     This submission was made by the CEIRS Data Processing and
            Coordinating Center (DPCC) on behalf of the The Johns Hopkins
            Center for Excellence in Influenza Research and Surveillance
            (JH-CEIRS). This work was supported by National Institute of
            Allergy and Infectious Diseases, National Institutes of Health,
            grant HHSN272201400007C.

            ##Assembly-Data-START##
            Assembly Method       :: bowtie2 v. 2.2.6
            Coverage              :: 100
            Sequencing Technology :: Illumina
```

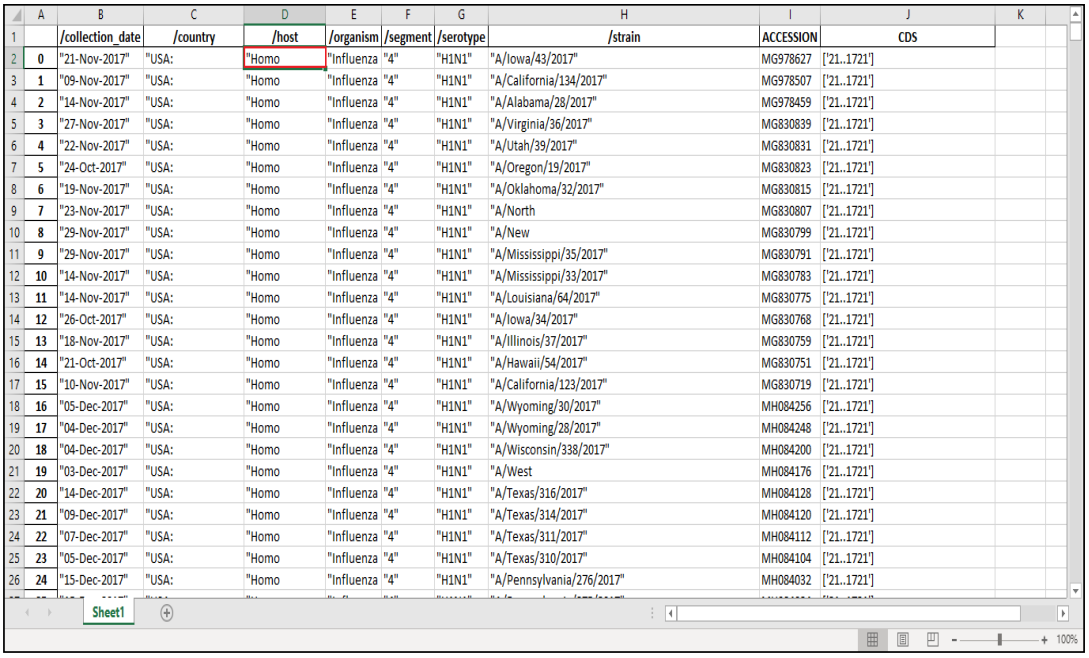
```
            ##Assembly-Data-END##
FEATURES             Location/Qualifiers
     source          1..1777
                     /organism="Influenza A virus"
                     /mol_type="viral cRNA"
                     /strain="A/Baltimore/P0197/2017"
                     /serotype="H1N1"
                     /host="Homo sapiens"
                     /db_xref="taxon:11320"
                     /segment="4"
                     /country="USA"
                     /collection_date="17-Nov-2017"
                     /PCR_primers="fwd_name: MBTuni-12, fwd_seq:
                     acgcgtgatcagcaaaagcagg, rev_name: MBTuni-13, rev_seq:
                     acgcgtgatcagtagaaacaagg"
     gene            33..1733
                     /gene="HA"
     CDS             33..1733
                     /gene="HA"
                     /function="receptor binding and fusion protein"
                     /codon_start=1
                     /product="hemagglutinin"
                     /protein_id="AXO64767.1"

/translation="MKAILVVLLYTFTTANADTLCIGYHANNSTDTVDTVLEKNVTVT

HSVNLLEDKHNGKLCKLRGVAPLHLGKCNIAGWILGNPECESLSTARSWSYIVETSNS

DNGTCYPGDFINYEELREQLSSVSSFERFEIFPKTSSWPNHDSNKGVTAACPHAGAKS

FYKNLIWLVKKGNSYPKLNQSYINDKGKEVLVLWGIHHPSTTADQQSLYQNADAYVFV

GTSRYSKKFKPEIATRPKVRDQEGRMNYYWTLVEPGDKITFEATGNLVVPRYAFTMER

NAGSGIIISDTPVHDCNTTCQTPEGAINTSLPFQNVHPITIGKCPKYVKSTKLRLATG

LRNVPSIQSRGLFGAIAGFIEGGWTGMVDGWYGYHHQNEQGSGYAADLKSTQNAIDKI

TNKVNSVIEKMNTQFTAVGKEFNHLEKRIENLNKKVDDGFLDIWTYNAELLVLLENER

TLDYHDSNVKNLYEKVRNQLKNNAKEIGNGCFEFYHKCDNTCMESVKNGTYDYPKYSE

EAKLNREKIDGVKLESTRIYQILAIYSTVASSLVLVVSLGAISFWMCSNGSLQCRICI
                     "
     sig_peptide     33..83
                     /gene="HA"
     mat_peptide     84..1064
                     /gene="HA"
                     /product="HA1"
     mat_peptide     1065..1730
                     /gene="HA"
                     /product="HA2"
ORIGIN
        1 agcaaaagca ggggaataca aaagcaacaa aaatgaaggc aatactagta gttctgctat
       61 atacatttac aaccgcaaat gcagacacat tatgtatagg ttatcatgcg aacaattcaa
```

```
      121 cagacactgt agacacagta ctagaaaaga atgtaacagt aacacactct gttaatcttc
      181 tggaagacaa gcataacgga aaactatgca aactaagagg ggtagcccca ttgcatttgg
      241 gtaaatgtaa cattgctggc tggatcctgg gaaatccaga gtgtgaatca ctctccacag
      301 caagatcatg gtcctacatt gtggaaacat ctaattcaga caatggaacg tgttacccag
      361 gagatttcat caattatgag gagctaagag agcaattgag ctcagtgtca tcatttgaaa
      421 ggtttgaaat attccccaag acaagttcat ggcccaatca tgactcgaac aaaggtgtaa
      481 cggcagcatg tcctcacgct ggagcaaaaa gcttctacaa aaacttgata tggctagtta
      541 aaaaagggaa ttcataccca aagctcaacc aatcctacat taatgataaa gggaaagaag
      601 tcctcgtgct gtggggcatt caccatccat ctactactgc tgaccaacaa agtctctatc
      661 agaatgcaga tgcatatgtc tttgtgggga catcaagata cagcaagaag ttcaagccgg
      721 aaatagcaac aagacccaaa gtgagggatc aagaagggag aatgaactat tactggacac
      781 tagtagaacc gggagacaaa ataacattcg aagcaactgg aaatctagtg gtaccgagat
      841 atgcattcac aatggaaaga aatgctggat ctggtattat catttcagat acaccagtcc
      901 acgattgcaa tacaacttgt cagacacccg agggtgctat aaacaccagc ctcccatttc
      961 agaatgtaca tccgatcaca attgggaaat gtccaaagta tgtaaaaagc acaaaattga
     1021 gactggccac aggattgagg aatgtcccgt ctattcaatc tagaggccta ttcggggcca
     1081 ttgccggctt cattgaaggg gggtggacag ggatggtaga tggatggtac ggttatcacc
     1141 atcaaaatga gcaggggtca ggatatgcag ccgacctgaa gagcacacaa aatgccattg
     1201 ataagattac taacaaagta aattctgtta ttgaaaagat gaatacacag ttcacagcag
     1261 tgggtaaaga gttcaaccac ctggaaaaga gaatagagaa tctaaataaa aaagttgatg
     1321 atggcttctt ggacatttgg acttacaatg ccgaactgtt ggttctactg gaaaatgaaa
     1381 gaactttgga ctatcacgat tcaaatgtga agaacttgta tgaaaaagta agaaaccagt
     1441 taaaaaacaa tgccaaggaa attggaaacg gctgctttga attttaccac aaatgcgata
     1501 acacatgcat ggaaagtgtc aagaatggga cttatgacta cccaaaatac tcagaggaag
     1561 caaaattaaa cagagaaaaa atagatggag taaagctgga atcaacaagg atctaccaga
     1621 ttttggcgat ctattcaact gtcgccagtt cattggtact ggtagtctcc ctgggggcaa
     1681 tcagcttctg gatgtgctct aatgggtctc tacagtgtag aatatgtatt taacattagg
     1741 atttcagaat catgagaaaa acacccttgt ttctact
//
```

Figura 3. **Ejemplo del contenido de la información de cada secuencia del gen HA del virus de Influenza A H1N1 del año 2017 tomada del formato GenBank full para la generación del archivo Excel.** Los datos seleccionados para la primera versión del programa BioDataToolkit v1.0 y sus nuevas versiones que generaron sus respectivos archivos de Excel, con la información de las secuencias nucleotídicas del GenBank se resaltan en colores. Para la versión 1.0 los datos están en amarillo. La información incluida en la versión 2 se muestra en verde. Los respectivos datos para la versión 3 se resalta en azul y la última es resaltada en fucsia.

El resultado obtenido indicó que el programa sí generó el documento de Excel. La información la ubicó en columnas, no obstante, dentro de ellas únicamente incluyó la primera palabra de cada ítem, es decir, todos los caracteres que encontró hasta hallar un espacio. Lo anterior se evidencia en la columna host, que en lugar de ubicar el género y especie *Homo*

sapiens como está indicado desde el archivo GenBank full (/host="Homo sapiens") únicamente se mostró el género (Fig. 4). Este primer resultado indicó que sí era posible separar la información del formato del GenBank en columnas para un archivo de Excel, pero se requerían adecuaciones. Por ello, se solicitó la realización de las respectivas modificaciones las cuales originaron la versión 2.0 del programa BioDataToolkit, además de incluir las líneas: mol_type, db_xref y PCR_primers como control de la información en la generación de columnas.

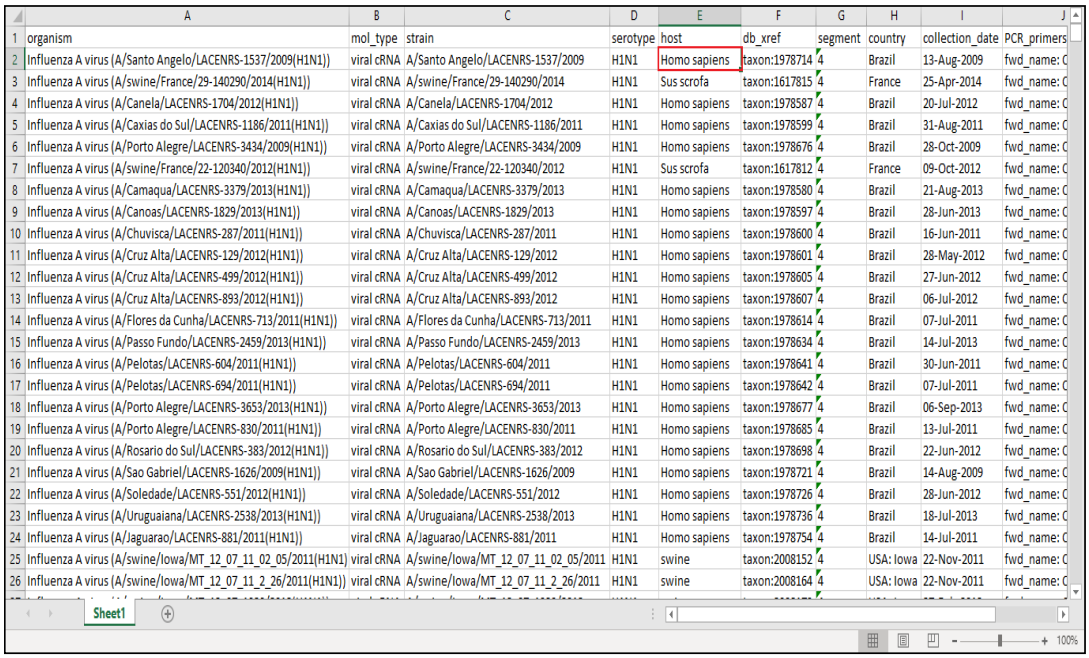


		/collection_date	/country	/host	/organism	/segment	/serotype	/strain	ACCESSION	CDS
0	0	"21-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Iowa/43/2017"	MG978627	['21..1721']
1	1	"09-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/California/134/2017"	MG978507	['21..1721']
2	2	"14-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Alabama/28/2017"	MG978459	['21..1721']
3	3	"27-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Virginia/36/2017"	MG830839	['21..1721']
4	4	"22-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Utah/39/2017"	MG830831	['21..1721']
5	5	"24-Oct-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Oregon/19/2017"	MG830823	['21..1721']
6	6	"19-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Oklahoma/32/2017"	MG830815	['21..1721']
7	7	"23-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/North	MG830807	['21..1721']
8	8	"29-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/New	MG830799	['21..1721']
9	9	"29-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Mississippi/35/2017"	MG830791	['21..1721']
10	10	"14-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Mississippi/33/2017"	MG830783	['21..1721']
11	11	"14-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Louisiana/64/2017"	MG830775	['21..1721']
12	12	"26-Oct-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Iowa/34/2017"	MG830768	['21..1721']
13	13	"18-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Illinois/37/2017"	MG830759	['21..1721']
14	14	"21-Oct-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Hawaii/54/2017"	MG830751	['21..1721']
15	15	"10-Nov-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/California/123/2017"	MG830719	['21..1721']
16	16	"05-Dec-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Wyoming/30/2017"	MH084256	['21..1721']
17	17	"04-Dec-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Wyoming/28/2017"	MH084248	['21..1721']
18	18	"04-Dec-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Wisconsin/338/2017"	MH084200	['21..1721']
19	19	"03-Dec-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/West	MH084176	['21..1721']
20	20	"14-Dec-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Texas/316/2017"	MH084128	['21..1721']
21	21	"09-Dec-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Texas/314/2017"	MH084120	['21..1721']
22	22	"07-Dec-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Texas/311/2017"	MH084112	['21..1721']
23	23	"05-Dec-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Texas/310/2017"	MH084104	['21..1721']
24	24	"15-Dec-2017"	"USA:	"Homo	"Influenza	"4"	"H1N1"	"A/Pennsylvania/276/2017"	MH084032	['21..1721']

Figura 4. **Archivo de salida de la versión 1.0 del programa BioDataToolkit.** En el recuadro rojo se evidencia que en esta versión del programa la base de datos solo muestra la primera palabra.

En la versión 2.0 del programa se corrigió la problemática determinada en la versión 1.0 y con ello, se incorporó la lectura de la respectiva línea de información de cada ítem de la sección de "Características" que se muestra en el formato GenBank full (Fig. 5). Para corroborar la exactitud en la generación de la hoja de Excel de esta nueva versión, el programa fue evaluado

con secuencias del gen de la Proteína de Matriz (Segmento 7) del virus de Influenza A H1N1, cuyo ARN tiene un procesamiento alternativo que genera un canal iónico. Por ello, la información del GenBank y concomitantemente la del formato GenBank full muestran los respectivos CDS. Sin embargo, al ejecutar la versión 2.0, el programa solo reconoció el primero de los dos CDS; es por ello, que se solicitó una nueva versión que diera solución a esta problemática, la cual dio origen a la versión 3 que permitió la integración de ambos CDS en su respectiva columna y la inclusión de las palabras: "gene, codon_start, product, protein_id y translation" como nuevo control de la información (Fig. 6).



	A	B	C	D	E	F	G	H	I	J
1	organism	mol_type	strain	serotype	host	db_xref	segment	country	collection_date	PCR_primers
2	Influenza A virus (A/Santo Angelo/LACENRS-1537/2009(H1N1))	viral cRNA	A/Santo Angelo/LACENRS-1537/2009	H1N1	Homo sapiens	taxon:1978714	4	Brazil	13-Aug-2009	fwd_name: C
3	Influenza A virus (A/swine/France/29-140290/2014(H1N1))	viral cRNA	A/swine/France/29-140290/2014	H1N1	Sus scrofa	taxon:1617815	4	France	25-Apr-2014	fwd_name: C
4	Influenza A virus (A/Canela/LACENRS-1704/2012(H1N1))	viral cRNA	A/Canela/LACENRS-1704/2012	H1N1	Homo sapiens	taxon:1978587	4	Brazil	20-Jul-2012	fwd_name: C
5	Influenza A virus (A/Caxias do Sul/LACENRS-1186/2011(H1N1))	viral cRNA	A/Caxias do Sul/LACENRS-1186/2011	H1N1	Homo sapiens	taxon:1978599	4	Brazil	31-Aug-2011	fwd_name: C
6	Influenza A virus (A/Porto Alegre/LACENRS-3434/2009(H1N1))	viral cRNA	A/Porto Alegre/LACENRS-3434/2009	H1N1	Homo sapiens	taxon:1978676	4	Brazil	28-Oct-2009	fwd_name: C
7	Influenza A virus (A/swine/France/22-120340/2012(H1N1))	viral cRNA	A/swine/France/22-120340/2012	H1N1	Sus scrofa	taxon:1617812	4	France	09-Oct-2012	fwd_name: C
8	Influenza A virus (A/Camaqua/LACENRS-3379/2013(H1N1))	viral cRNA	A/Camaqua/LACENRS-3379/2013	H1N1	Homo sapiens	taxon:1978580	4	Brazil	21-Aug-2013	fwd_name: C
9	Influenza A virus (A/Canoas/LACENRS-1829/2013(H1N1))	viral cRNA	A/Canoas/LACENRS-1829/2013	H1N1	Homo sapiens	taxon:1978597	4	Brazil	28-Jun-2013	fwd_name: C
10	Influenza A virus (A/Chuvisca/LACENRS-287/2011(H1N1))	viral cRNA	A/Chuvisca/LACENRS-287/2011	H1N1	Homo sapiens	taxon:1978600	4	Brazil	16-Jun-2011	fwd_name: C
11	Influenza A virus (A/Cruz Alta/LACENRS-129/2012(H1N1))	viral cRNA	A/Cruz Alta/LACENRS-129/2012	H1N1	Homo sapiens	taxon:1978601	4	Brazil	28-May-2012	fwd_name: C
12	Influenza A virus (A/Cruz Alta/LACENRS-499/2012(H1N1))	viral cRNA	A/Cruz Alta/LACENRS-499/2012	H1N1	Homo sapiens	taxon:1978605	4	Brazil	27-Jun-2012	fwd_name: C
13	Influenza A virus (A/Cruz Alta/LACENRS-893/2012(H1N1))	viral cRNA	A/Cruz Alta/LACENRS-893/2012	H1N1	Homo sapiens	taxon:1978607	4	Brazil	06-Jul-2012	fwd_name: C
14	Influenza A virus (A/Flores da Cunha/LACENRS-713/2011(H1N1))	viral cRNA	A/Flores da Cunha/LACENRS-713/2011	H1N1	Homo sapiens	taxon:1978614	4	Brazil	07-Jul-2011	fwd_name: C
15	Influenza A virus (A/Passo Fundo/LACENRS-2459/2013(H1N1))	viral cRNA	A/Passo Fundo/LACENRS-2459/2013	H1N1	Homo sapiens	taxon:1978634	4	Brazil	14-Jul-2013	fwd_name: C
16	Influenza A virus (A/Pelotas/LACENRS-604/2011(H1N1))	viral cRNA	A/Pelotas/LACENRS-604/2011	H1N1	Homo sapiens	taxon:1978641	4	Brazil	30-Jun-2011	fwd_name: C
17	Influenza A virus (A/Pelotas/LACENRS-694/2011(H1N1))	viral cRNA	A/Pelotas/LACENRS-694/2011	H1N1	Homo sapiens	taxon:1978642	4	Brazil	07-Jul-2011	fwd_name: C
18	Influenza A virus (A/Porto Alegre/LACENRS-3653/2013(H1N1))	viral cRNA	A/Porto Alegre/LACENRS-3653/2013	H1N1	Homo sapiens	taxon:1978677	4	Brazil	06-Sep-2013	fwd_name: C
19	Influenza A virus (A/Porto Alegre/LACENRS-830/2011(H1N1))	viral cRNA	A/Porto Alegre/LACENRS-830/2011	H1N1	Homo sapiens	taxon:1978685	4	Brazil	13-Jul-2011	fwd_name: C
20	Influenza A virus (A/Rosario do Sul/LACENRS-383/2012(H1N1))	viral cRNA	A/Rosario do Sul/LACENRS-383/2012	H1N1	Homo sapiens	taxon:1978698	4	Brazil	22-Jun-2012	fwd_name: C
21	Influenza A virus (A/Sao Gabriel/LACENRS-1626/2009(H1N1))	viral cRNA	A/Sao Gabriel/LACENRS-1626/2009	H1N1	Homo sapiens	taxon:1978721	4	Brazil	14-Aug-2009	fwd_name: C
22	Influenza A virus (A/Soledade/LACENRS-551/2012(H1N1))	viral cRNA	A/Soledade/LACENRS-551/2012	H1N1	Homo sapiens	taxon:1978726	4	Brazil	28-Jun-2012	fwd_name: C
23	Influenza A virus (A/Uruguaiana/LACENRS-2538/2013(H1N1))	viral cRNA	A/Uruguaiana/LACENRS-2538/2013	H1N1	Homo sapiens	taxon:1978736	4	Brazil	18-Jul-2013	fwd_name: C
24	Influenza A virus (A/Jaguarao/LACENRS-881/2011(H1N1))	viral cRNA	A/Jaguarao/LACENRS-881/2011	H1N1	Homo sapiens	taxon:1978754	4	Brazil	14-Jul-2011	fwd_name: C
25	Influenza A virus (A/swine/Iowa/MT_12_07_11_02_05/2011(H1N1)	viral cRNA	A/swine/Iowa/MT_12_07_11_02_05/2011	H1N1	swine	taxon:2008152	4	USA: Iowa	22-Nov-2011	fwd_name: C
26	Influenza A virus (A/swine/Iowa/MT_12_07_11_2_26/2011(H1N1))	viral cRNA	A/swine/Iowa/MT_12_07_11_2_26/2011	H1N1	swine	taxon:2008164	4	USA: Iowa	22-Nov-2011	fwd_name: C

Sheet1

Figura 5. **Resultado de la evaluación de la versión 2.0 del programa BioDataToolkit.** El recuadro rojo de la columna host muestra que el programa ya lee toda la línea de cada encabezado, es decir, ya se incorporó el género y la especie en la respectiva columna.

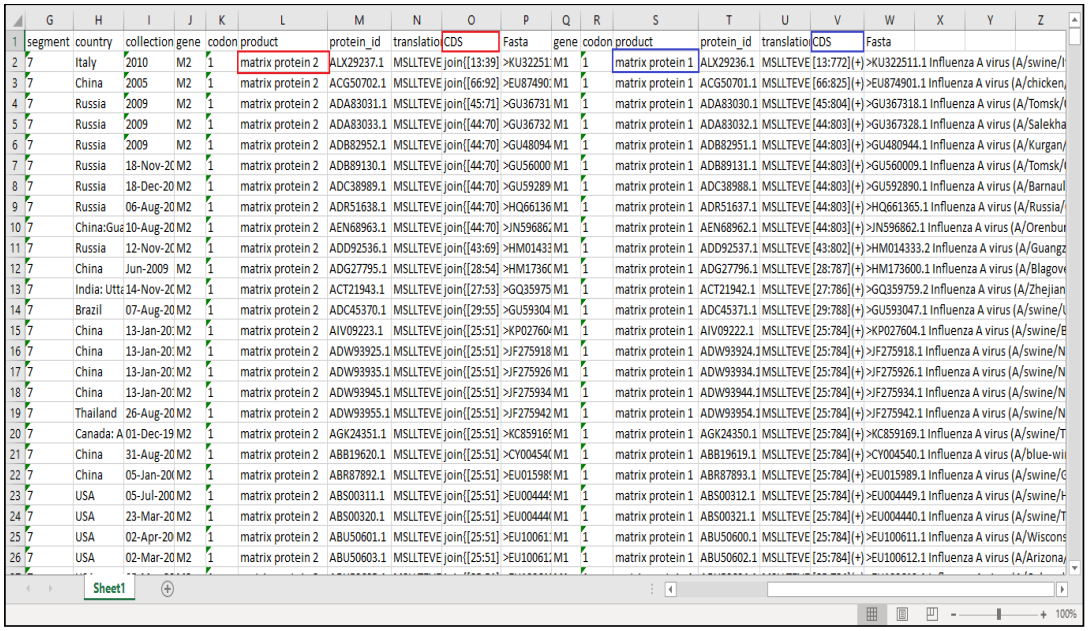
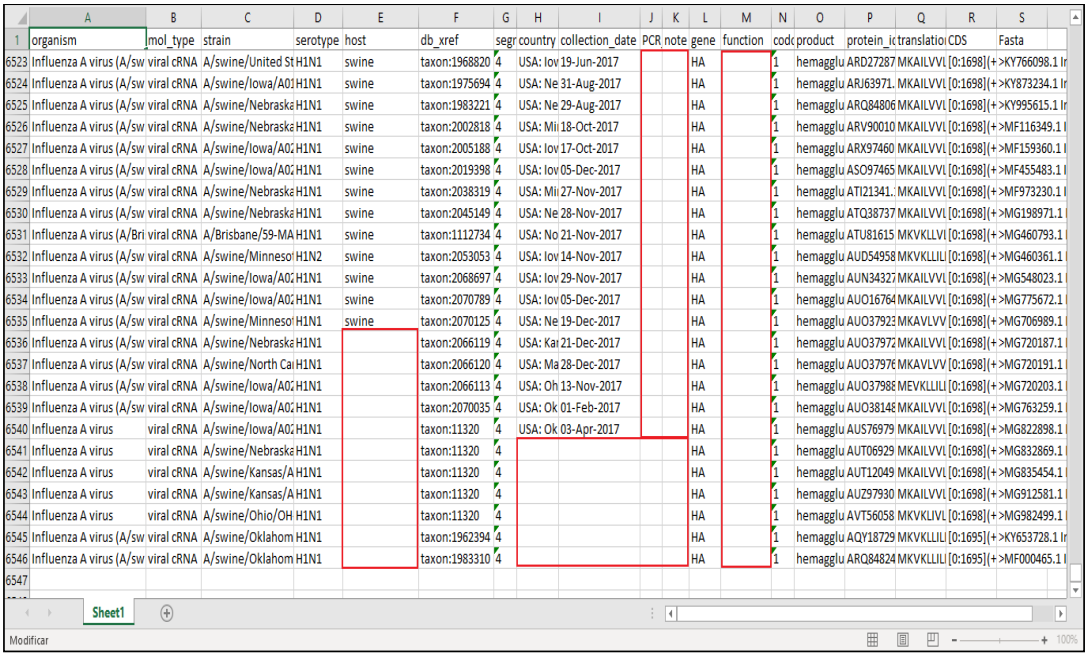


Figura 6. **Evidencia de los resultados generados por la versión 3.0 del programa para la integración de la información de genes reportados con dos CDS.** En los recuadros rojos se muestra el primer producto que es la proteína de matriz 2 con su respectivo CDS y en los recuadros azules se encuentra la proteína de matriz 1 con su CDS.

La evaluación de la versión 3.0 mostró que al final del documento Excel se presentaba un desfase de la información en las columnas de: *host*, *country*, *collection_date*, *PCR_primers*, *note* y *function*, respecto a la información de las columnas: *organism*, *mol_type*, *strain*, *serotype*, *db_xref*, *segment*, *gene*, *codon_start*, *producto*, *protein_id*, *translation*, *CDS*, *Fasta* (Fig. 7); la inconsistencia de lo anterior durante la fase de revisión del archivo Excel, se confirmó con los alineamientos de las secuencias de esta base de datos. Para determinar el origen de esta problemática, se procedió a analizar la información de las columnas de Excel con la del archivo GenBank full, encontrándose que este resultado era dado por la heterogeneidad en la información de las secuencias, ya que no todas contenían la totalidad de los datos completos de cada ítem que se reportan en el GenBank y por lo tanto el programa no pudo realizar el ordenamiento pertinente

(Fig. 7). Debido a lo antes mencionado, se solicitó una nueva versión del programa que se nombró v 4.0, sin embargo, ella continuo la inconsistencia de los espacios vacíos, ya que el programa debería poner la palabra "*empty*" en las columnas en que no encontraba el dato, las cuales fueron: *host*, *country*, *collection_date* y *note*, no obstante, en las columnas: *gen*, *fuction* y *codon_start*, continuó la ausencia de la palabra y por lo tanto se mantenía el espacio vacío (Fig. 8).



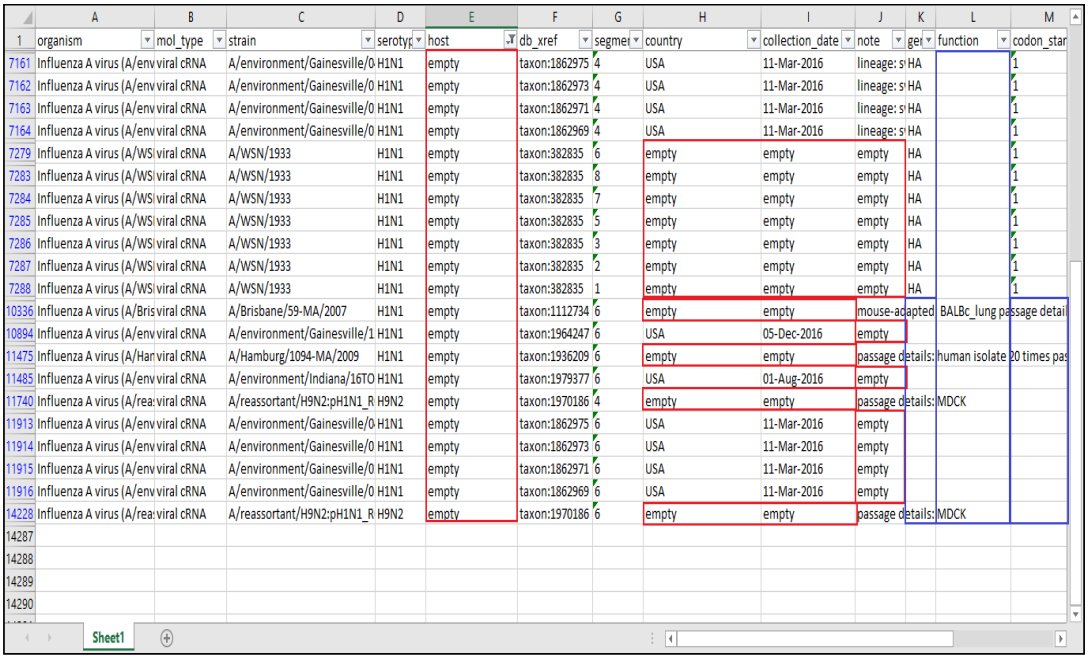
	organism	mol_type	strain	serotype	host	db_xref	segr	country	collection_date	PCR	note	gene	function	codc	product	protein_ic	translatio	CDS	Fasta
6523	Influenza A virus (A/sw	viral cRNA	A/swine/United St	H1N1	swine	taxon:1968820	4	USA: Iov	19-Jun-2017			HA		1	hemaggl	ARD27287	MKAILVVL	[0:1698](+	>KY766098.1 In
6524	Influenza A virus (A/sw	viral cRNA	A/swine/Iowa/A0	H1N1	swine	taxon:1975694	4	USA: Ne	31-Aug-2017			HA		1	hemaggl	ARJ63971.	MKAILVVL	[0:1698](+	>KY873234.1 In
6525	Influenza A virus (A/sw	viral cRNA	A/swine/Nebraska	H1N1	swine	taxon:1983221	4	USA: Ne	29-Aug-2017			HA		1	hemaggl	ARQ84806	MKAILVVL	[0:1698](+	>KY995615.1 In
6526	Influenza A virus (A/sw	viral cRNA	A/swine/Nebraska	H1N1	swine	taxon:2002818	4	USA: Mi	18-Oct-2017			HA		1	hemaggl	ARV90010	MKAILVVL	[0:1698](+	>MF116349.1 I
6527	Influenza A virus (A/sw	viral cRNA	A/swine/Iowa/A0	H1N1	swine	taxon:2005188	4	USA: Iov	17-Oct-2017			HA		1	hemaggl	ARX97460	MKAILVVL	[0:1698](+	>MF159360.1 I
6528	Influenza A virus (A/sw	viral cRNA	A/swine/Iowa/A0	H1N1	swine	taxon:2019398	4	USA: Iov	05-Dec-2017			HA		1	hemaggl	ASO97465	MKAILVVL	[0:1698](+	>MF455483.1 I
6529	Influenza A virus (A/sw	viral cRNA	A/swine/Nebraska	H1N1	swine	taxon:2038319	4	USA: Mi	27-Nov-2017			HA		1	hemaggl	ATI21341.	MKAILVVL	[0:1698](+	>MF973230.1 I
6530	Influenza A virus (A/sw	viral cRNA	A/swine/Nebraska	H1N1	swine	taxon:2045149	4	USA: Ne	28-Nov-2017			HA		1	hemaggl	ATQ38737	MKAILVVL	[0:1698](+	>MG198971.1 I
6531	Influenza A virus (A/Bri	viral cRNA	A/Brisbane/59-MA	H1N1	swine	taxon:1112734	4	USA: No	21-Nov-2017			HA		1	hemaggl	ATU81615	MKVKLLVI	[0:1698](+	>MG460793.1 I
6532	Influenza A virus (A/sw	viral cRNA	A/swine/Minneso	H1N2	swine	taxon:2053053	4	USA: Iov	14-Nov-2017			HA		1	hemaggl	AUD54958	MKVKLLIL	[0:1698](+	>MG460361.1 I
6533	Influenza A virus (A/sw	viral cRNA	A/swine/Iowa/A0	H1N1	swine	taxon:2068697	4	USA: Iov	29-Nov-2017			HA		1	hemaggl	AUN34327	MKAILVVL	[0:1698](+	>MG548023.1 I
6534	Influenza A virus (A/sw	viral cRNA	A/swine/Iowa/A0	H1N1	swine	taxon:2070789	4	USA: Iov	05-Dec-2017			HA		1	hemaggl	AUO16764	MKAILVVL	[0:1698](+	>MG775672.1 I
6535	Influenza A virus (A/sw	viral cRNA	A/swine/Minneso	H1N1	swine	taxon:2070125	4	USA: Ne	19-Dec-2017			HA		1	hemaggl	AUO37923	MKAVLVV	[0:1698](+	>MG706989.1 I
6536	Influenza A virus (A/sw	viral cRNA	A/swine/Nebraska	H1N1		taxon:2066119	4	USA: Kai	21-Dec-2017			HA		1	hemaggl	AUO37972	MKAILVVL	[0:1698](+	>MG720187.1 I
6537	Influenza A virus (A/sw	viral cRNA	A/swine/North Ca	H1N1		taxon:2066120	4	USA: Ma	28-Dec-2017			HA		1	hemaggl	AUO37976	MKAVLVV	[0:1698](+	>MG720191.1 I
6538	Influenza A virus (A/sw	viral cRNA	A/swine/Iowa/A0	H1N1		taxon:2066113	4	USA: Oh	13-Nov-2017			HA		1	hemaggl	AUO37988	MEVKLLIL	[0:1698](+	>MG720203.1 I
6539	Influenza A virus (A/sw	viral cRNA	A/swine/Iowa/A0	H1N1		taxon:2070035	4	USA: Ok	01-Feb-2017			HA		1	hemaggl	AUO38148	MKAILVVL	[0:1698](+	>MG763259.1 I
6540	Influenza A virus	viral cRNA	A/swine/Iowa/A0	H1N1		taxon:11320	4	USA: Ok	03-Apr-2017			HA		1	hemaggl	AUS76979	MKAILVVL	[0:1698](+	>MG822898.1 I
6541	Influenza A virus	viral cRNA	A/swine/Nebraska	H1N1		taxon:11320	4					HA		1	hemaggl	AUT06929	MKAILVVL	[0:1698](+	>MG832869.1 I
6542	Influenza A virus	viral cRNA	A/swine/Kansas/A	H1N1		taxon:11320	4					HA		1	hemaggl	AUT12049	MKAILVVL	[0:1698](+	>MG835454.1 I
6543	Influenza A virus	viral cRNA	A/swine/Kansas/A	H1N1		taxon:11320	4					HA		1	hemaggl	AUZ97930	MKAILVVL	[0:1698](+	>MG912581.1 I
6544	Influenza A virus	viral cRNA	A/swine/Ohio/OH	H1N1		taxon:11320	4					HA		1	hemaggl	AVT56058	MKVKLIVL	[0:1698](+	>MG982499.1 I
6545	Influenza A virus (A/sw	viral cRNA	A/swine/Oklahom	H1N1		taxon:1962394	4					HA		1	hemaggl	AQY18729	MKVKLLIL	[0:1695](+	>KY653728.1 In
6546	Influenza A virus (A/sw	viral cRNA	A/swine/Oklahom	H1N1		taxon:1983310	4					HA		1	hemaggl	ARQ84824	MKVKLLIL	[0:1695](+	>MF000465.1 I
6547																			

Sheet1

Modificar

Figura 7. **Desfase en la información de las secuencias del gen HA del virus de influenza A H1N1 generado por la versión 3.0 del programa BioDataToolkit.** En los recuadros rojos se muestra la información faltante de las columnas *host, country, collection_date, PCR_primers, note* y *function,* respecto de las demás.

Nuevamente se solicitó la modificación del programa y la incorporación del código de identificación de las secuencias del virus de Influenza A H1N1 diseñado por el Grupo de Investigación dentro del proyecto inicial para su futura aplicación, lo que produjo la versión 5.0 (Fig. 9).



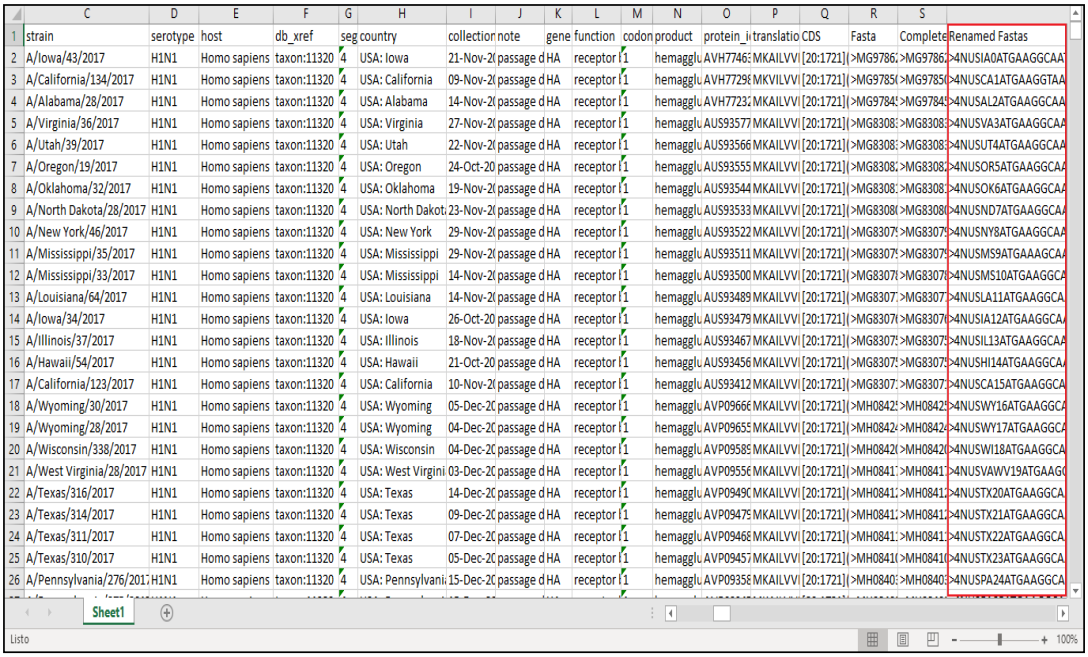
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	organism	mol_type	strain	serotyp	host	db_xref	segmer	country	collection_date	note	ge	function	codon_star
7161	Influenza A virus (A/env	viral cRNA	A/environment/Gainesville/0	H1N1	empty	taxon:1862975	4	USA	11-Mar-2016	lineage: s	HA		1
7162	Influenza A virus (A/env	viral cRNA	A/environment/Gainesville/0	H1N1	empty	taxon:1862973	4	USA	11-Mar-2016	lineage: s	HA		1
7163	Influenza A virus (A/env	viral cRNA	A/environment/Gainesville/0	H1N1	empty	taxon:1862971	4	USA	11-Mar-2016	lineage: s	HA		1
7164	Influenza A virus (A/env	viral cRNA	A/environment/Gainesville/0	H1N1	empty	taxon:1862969	4	USA	11-Mar-2016	lineage: s	HA		1
7279	Influenza A virus (A/WS	viral cRNA	A/WSN/1933	H1N1	empty	taxon:382835	6	empty	empty	empty	HA		1
7283	Influenza A virus (A/WS	viral cRNA	A/WSN/1933	H1N1	empty	taxon:382835	8	empty	empty	empty	HA		1
7284	Influenza A virus (A/WS	viral cRNA	A/WSN/1933	H1N1	empty	taxon:382835	7	empty	empty	empty	HA		1
7285	Influenza A virus (A/WS	viral cRNA	A/WSN/1933	H1N1	empty	taxon:382835	5	empty	empty	empty	HA		1
7286	Influenza A virus (A/WS	viral cRNA	A/WSN/1933	H1N1	empty	taxon:382835	3	empty	empty	empty	HA		1
7287	Influenza A virus (A/WS	viral cRNA	A/WSN/1933	H1N1	empty	taxon:382835	2	empty	empty	empty	HA		1
7288	Influenza A virus (A/WS	viral cRNA	A/WSN/1933	H1N1	empty	taxon:382835	1	empty	empty	empty	HA		1
10336	Influenza A virus (A/Bris	viral cRNA	A/Brisbane/59-MA/2007	H1N1	empty	taxon:1112734	6	empty	empty	mouse-adapted		BALBc_lung passage detail	
10894	Influenza A virus (A/env	viral cRNA	A/environment/Gainesville/1	H1N1	empty	taxon:1964247	6	USA	05-Dec-2016	empty			
11475	Influenza A virus (A/Har	viral cRNA	A/Hamburg/1094-MA/2009	H1N1	empty	taxon:1936209	6	empty	empty	passage details:		human isolate 20 times pas	
11485	Influenza A virus (A/env	viral cRNA	A/environment/Indiana/16TO	H1N1	empty	taxon:1979377	6	USA	01-Aug-2016	empty			
11740	Influenza A virus (A/rea	viral cRNA	A/reassortant/H9N2:pH1N1_R	H9N2	empty	taxon:1970186	4	empty	empty	passage details:		MDCK	
11913	Influenza A virus (A/env	viral cRNA	A/environment/Gainesville/0	H1N1	empty	taxon:1862975	6	USA	11-Mar-2016	empty			
11914	Influenza A virus (A/env	viral cRNA	A/environment/Gainesville/0	H1N1	empty	taxon:1862973	6	USA	11-Mar-2016	empty			
11915	Influenza A virus (A/env	viral cRNA	A/environment/Gainesville/0	H1N1	empty	taxon:1862971	6	USA	11-Mar-2016	empty			
11916	Influenza A virus (A/env	viral cRNA	A/environment/Gainesville/0	H1N1	empty	taxon:1862969	6	USA	11-Mar-2016	empty			
14228	Influenza A virus (A/rea	viral cRNA	A/reassortant/H9N2:pH1N1_R	H9N2	empty	taxon:1970186	6	empty	empty	passage details:		MDCK	
14287													
14288													
14289													
14290													

Sheet1

Figura 8. **Espacios vacíos en la información de las secuencias del gen HA del virus de influenza A H1N1 generado por la versión 4.0 del programa BioDataToolkit.** En los recuadros rojos se muestran las columnas en las que el programa introduce la palabra *empty* y en los recuadros azules las que no se colocó la palabra.

En la versión 5.0 del programa ya no se encontraron los espacios vacíos en la hoja Excel y generó el cambio de la nomenclatura del GenBank de las secuencias en formato Fasta, al establecido por el Grupo de Investigación. Para ello, el programa contuvo un algoritmo que identificó el país del cual fue colectada la cepa y con ello generó el código del grupo, siendo este: número del segmento + código del continente + código del país + un número que se genera consecutivamente con todas las secuencias descargadas, que inicia desde 0 hasta la cantidad de secuencias descargadas menos 1. Lo anterior, se encuentra ubicado en una columna adicional que el programa genera con nombre *Renamed Fastas.* (Fig. 9); para el uso de la información de los formatos Fasta presentes en dicha columna se tuvo que copiar y pegar en un documento tipo Word para la eliminación de un par de comillas que genera el programa de forma automática.

Finalmente, otro punto a resaltar de la versión 5.0 del programa es que el archivo de salida de Excel que contiene la información de la secuencias del virus de Influenza A H1N1 se obtuvieron en menos de un minuto.



	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	strain	serotype	host	db_xref	seg	country	collection	note	gene	function	codon	product	protein_i	translatio	CDS	Fasta	Complete	Renamed Fastas
2	A/Iowa/43/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Iowa	21-Nov-2(	passage d	HA	receptor	1	hemaggl	AVH7746	MKAILVVI	[20:1721](	>MG9786	>MG9786	>4NUSIA0ATGAAGGCAA
3	A/California/134/2017	H1N1	Homo sapiens	taxon:11320	4	USA: California	09-Nov-2(	passage d	HA	receptor	1	hemaggl	AVH7729	MKVILVVI	[20:1721](	>MG9785	>MG9785	>4NUSCA1ATGAAGGTAA
4	A/Alabama/28/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Alabama	14-Nov-2(	passage d	HA	receptor	1	hemaggl	AVH7723	MKAILVVI	[20:1721](	>MG9784	>MG9784	>4NUSAL2ATGAAGGCAA
5	A/Virginia/36/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Virginia	27-Nov-2(	passage d	HA	receptor	1	hemaggl	AUS93577	MKAILVVI	[20:1721](	>MG8308	>MG8308	>4NUSVA3ATGAAGGCAA
6	A/Utah/39/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Utah	22-Nov-2(	passage d	HA	receptor	1	hemaggl	AUS93566	MKAILVVI	[20:1721](	>MG8308	>MG8308	>4NUSUT4ATGAAGGCAA
7	A/Oregon/19/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Oregon	24-Oct-20	passage d	HA	receptor	1	hemaggl	AUS93555	MKAILVVI	[20:1721](	>MG8308	>MG8308	>4NUSOR5ATGAAGGCAA
8	A/Oklahoma/32/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Oklahoma	19-Nov-2(	passage d	HA	receptor	1	hemaggl	AUS93544	MKAILVVI	[20:1721](	>MG8308	>MG8308	>4NUSOK6ATGAAGGCAA
9	A/North Dakota/28/2017	H1N1	Homo sapiens	taxon:11320	4	USA: North Dakot	23-Nov-2(	passage d	HA	receptor	1	hemaggl	AUS93533	MKAILVVI	[20:1721](	>MG8308	>MG8308	>4NUSND7ATGAAGGCAA
10	A/New York/46/2017	H1N1	Homo sapiens	taxon:11320	4	USA: New York	29-Nov-2(	passage d	HA	receptor	1	hemaggl	AUS93522	MKAILVVI	[20:1721](	>MG8307	>MG8307	>4NUSNY8ATGAAGGCAA
11	A/Mississippi/35/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Mississippi	29-Nov-2(	passage d	HA	receptor	1	hemaggl	AUS93511	MKAILVVI	[20:1721](	>MG8307	>MG8307	>4NUSMS9ATGAAAGCAA
12	A/Mississippi/33/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Mississippi	14-Nov-2(	passage d	HA	receptor	1	hemaggl	AUS93500	MKAILVVI	[20:1721](	>MG8307	>MG8307	>4NUSMS10ATGAAGGCA
13	A/Louisiana/64/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Louisiana	14-Nov-2(	passage d	HA	receptor	1	hemaggl	AUS93489	MKAILVVI	[20:1721](	>MG8307	>MG8307	>4NUSLA11ATGAAGGCA
14	A/Iowa/34/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Iowa	26-Oct-20	passage d	HA	receptor	1	hemaggl	AUS93479	MKAILVVI	[20:1721](	>MG8307	>MG8307	>4NUSIA12ATGAAGGCA
15	A/Illinois/37/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Illinois	18-Nov-2(	passage d	HA	receptor	1	hemaggl	AUS93467	MKAILVVI	[20:1721](	>MG8307	>MG8307	>4NUSIL13ATGAAGGCAA
16	A/Hawaii/54/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Hawaii	21-Oct-20	passage d	HA	receptor	1	hemaggl	AUS93456	MKAILVVI	[20:1721](	>MG8307	>MG8307	>4NUSHI14ATGAAGGCA
17	A/California/123/2017	H1N1	Homo sapiens	taxon:11320	4	USA: California	10-Nov-2(	passage d	HA	receptor	1	hemaggl	AUS93412	MKAILVVI	[20:1721](	>MG8307	>MG8307	>4NUSCA15ATGAAGGCA
18	A/Wyoming/30/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Wyoming	05-Dec-2(	passage d	HA	receptor	1	hemaggl	AVP0966	MKAILVVI	[20:1721](	>MH0842	>MH0842	>4NUSWY16ATGAAGGCA
19	A/Wyoming/28/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Wyoming	04-Dec-2(	passage d	HA	receptor	1	hemaggl	AVP0965	MKAILVVI	[20:1721](	>MH0842	>MH0842	>4NUSWY17ATGAAGGCA
20	A/Wisconsin/338/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Wisconsin	04-Dec-2(	passage d	HA	receptor	1	hemaggl	AVP0958	MKAILVVI	[20:1721](	>MH0842	>MH0842	>4NUSWI18ATGAAGGCA
21	A/West Virginia/28/2017	H1N1	Homo sapiens	taxon:11320	4	USA: West Virgini	03-Dec-2(	passage d	HA	receptor	1	hemaggl	AVP0955	MKAILVVI	[20:1721](	>MH0841	>MH0841	>4NUSVAWV19ATGAAG
22	A/Texas/316/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Texas	14-Dec-2(	passage d	HA	receptor	1	hemaggl	AVP0949	MKAILVVI	[20:1721](	>MH0841	>MH0841	>4NUSTX20ATGAAGGCA
23	A/Texas/314/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Texas	09-Dec-2(	passage d	HA	receptor	1	hemaggl	AVP0947	MKAILVVI	[20:1721](	>MH0841	>MH0841	>4NUSTX21ATGAAGGCA
24	A/Texas/311/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Texas	07-Dec-2(	passage d	HA	receptor	1	hemaggl	AVP0946	MKAILVVI	[20:1721](	>MH0841	>MH0841	>4NUSTX22ATGAAGGCA
25	A/Texas/310/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Texas	05-Dec-2(	passage d	HA	receptor	1	hemaggl	AVP0945	MKAILVVI	[20:1721](	>MH0841	>MH0841	>4NUSTX23ATGAAGGCA
26	A/Pennsylvania/276/2017	H1N1	Homo sapiens	taxon:11320	4	USA: Pennsylvani	15-Dec-2(	passage d	HA	receptor	1	hemaggl	AVP0935	MKAILVVI	[20:1721](	>MH0840	>MH0840	>4NUSPA24ATGAAGGCA

Sheet1

Listo

100%

Figura 9. **Archivo de salida de la versión 5.0 del programa BioDataToolkit con el cambio en el nombre de las secuencias de acuerdo con el código diseñado por el Grupo de Investigación.** En el recuadro rojo se muestra la columna en la que fueron ubicados los formatos Fastas de las secuencias con los nombres ya modificados de acuerdo con el código establecido por el Grupo de Investigación.

Construcción mensual del año 2017de las secuencias consenso del gen HA del virus de Influenza A H1N1.

A parir del formato GenBank full se realizó la verificación de los datos obtenidos del documento Excel, lo que determinó que la mayoría de las secuencias correspondieron al gen HA del virus de

Influenza A H1N1 del año 2017, sin embargo, al realizar el alineamiento de todas ellas se observó la presencia de *Gaps* y secuencias que no correspondieron a los criterios de búsqueda iniciales. No obstante, el uso de la herramienta de filtros en la hoja de Excel demostró la presencia de secuencias incompletas del gen HA y completas del gen de la Neuraminidasa (Segmento 6) dentro de la base de datos (Fig. 10). Este problema se solucionó por la selección específica de las secuencias del gen HA del año 2017 correspondientes a seres humanos con la herramienta de filtros y así se logró obtener los formatos Fasta para la generación de los alineamientos mensuales (Apéndices A – L) para posteriormente generar las secuencias consenso (Fig. 11.A) a las cuales se nombraron con la siguiente nomenclatura: signo de mayor que + año de colecta de la muestra + las primeras 3 letras del nombre del mes + la cantidad de secuencias usadas para realizar el consenso, (Fig. 11.B)

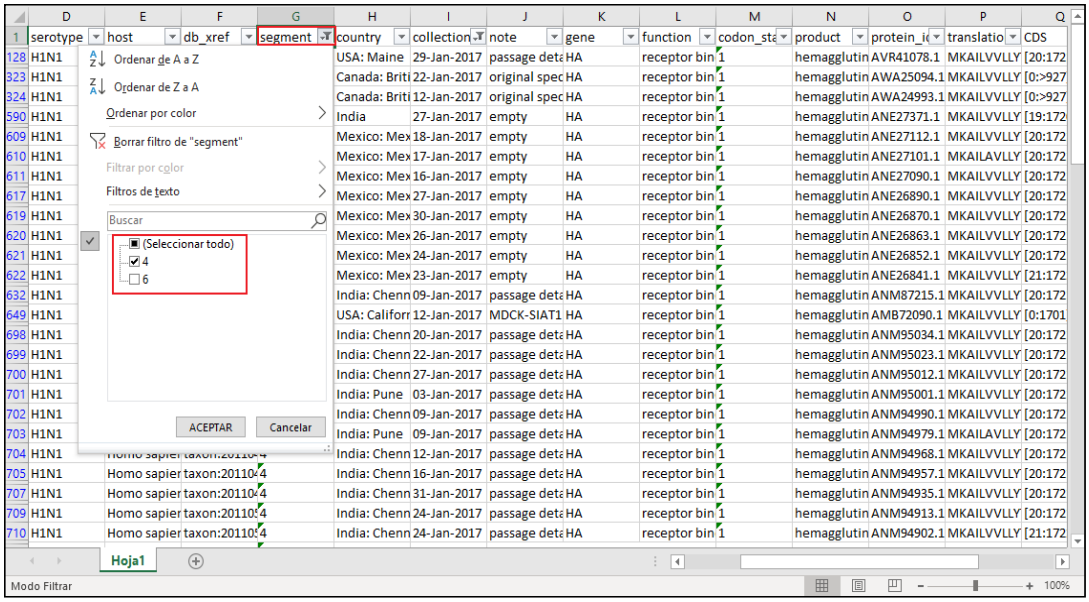


Figura 10. **Uso de la herramienta de filtros de Excel para la depuración de la base de datos de las secuencias nucleotídicas del gen HA del virus de Influenza A H1N1.** En los recuadros rojos se muestra la depuración específica de los datos que se puede realizar mediante la herramienta de filtros.

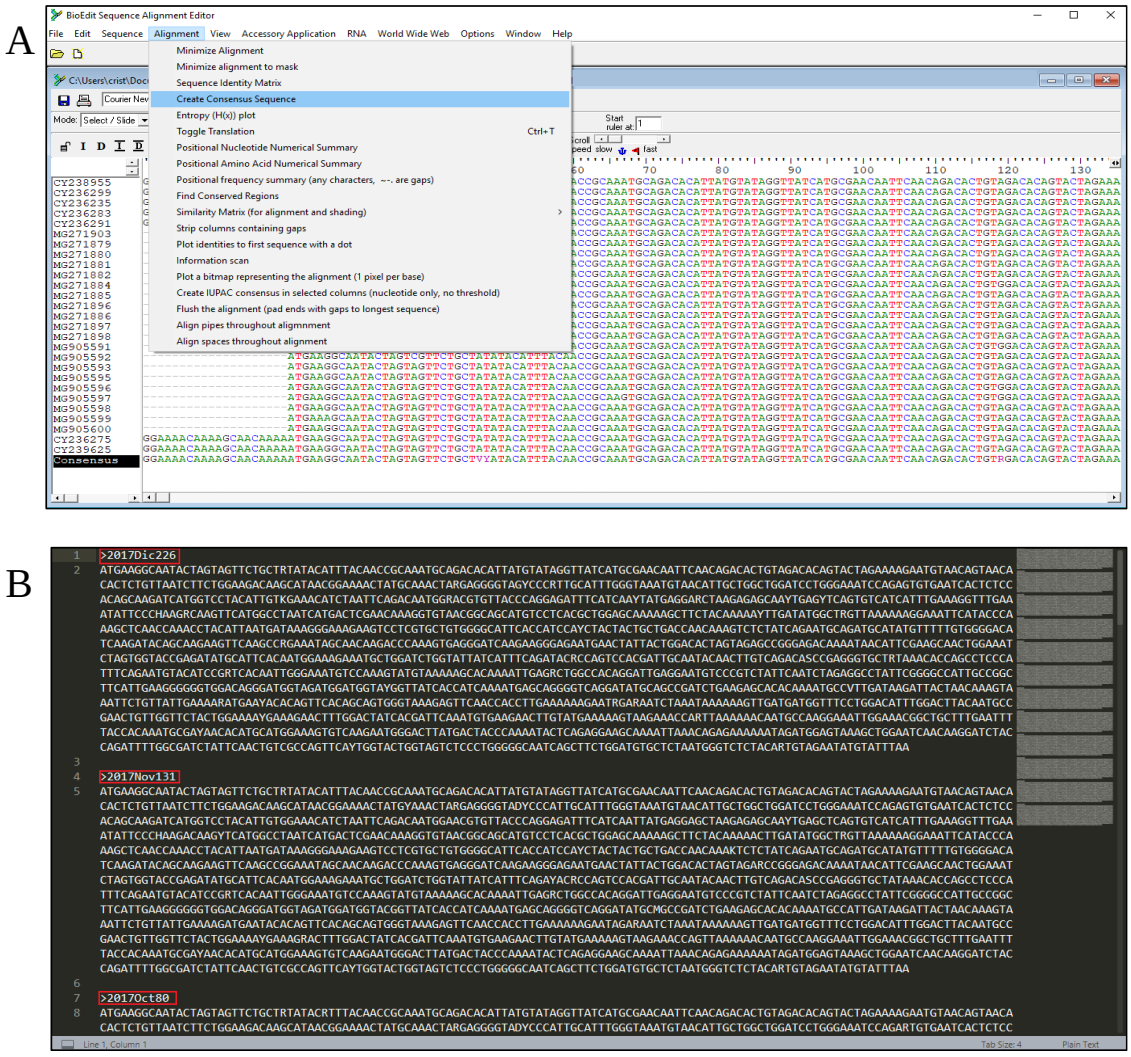


Figura 11. **Generación de las secuencias consensos mensuales de las secuencias nucleotídicas del gen HA del virus de Influenza A H1N1.** En "A" se observa la forma en que se generó el consenso mensual a partir de cada alineamiento. En "B", los recuadros en rojo muestran el nombre asignado a cada consenso mensual, de acuerdo con el código creado.

Verificación de la validez biológica de la base de datos y análisis bioinformáticos

Posteriormente, la capacidad de depuración de las secuencias consenso mensuales por medio de la selección de la hoja de Exel, fue corroborada con los resultados obtenidos por su comparación

con la base de datos del GenBank por medio de alineammientos tipo BLAST. En todos ellos se determinaron homologias del 96.47 al 98.59 % y un valor de *e* de cero para todas las secuencias consenso (Tabla 2 en Apéndice M). Estos valores de porcentaje de identidad se confirmaron con del árbol obtenido del análisis filigenético con el método de Inferencia Bayesiana, en donde se observó que las secuencias del virus de Influenza A H1N1 durante los meses de enero a diciembre del año 2017 no presentan cambios relevantes. (Fig. 12).

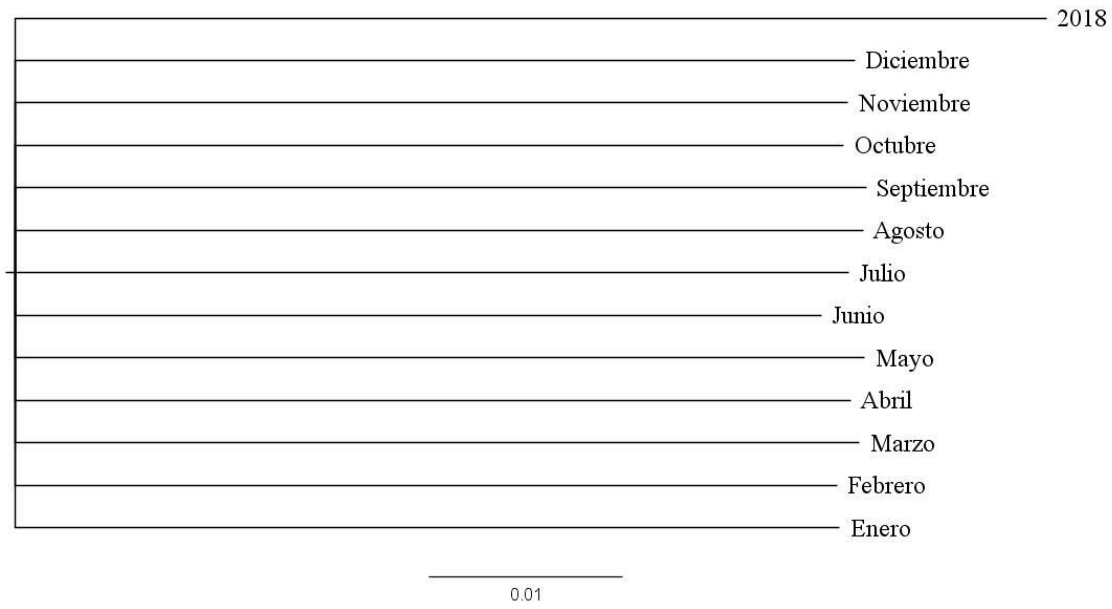


Figura 12 **Árbol filogenético con las secuencias consenso del gen HA del virus de Influenza A H1N1.** Se puede apreciar una gran similitud entre los meses de enero a diciembre del año 2017.

5. Discusión

La importancia del estudio del virus de Influenza A H1N1 radica en el que ha sido uno de los agentes infecciosos que ha generado diferentes pandemias debido a que su transmisión es por vía respiratoria (Fineberg, 2014). Otra característica de este virus es que tiene la capacidad de

infectar, además de los seres humanos, a organismos utilizados para la sobrevivencia de la especie humana como son aves, cerdos, entre otros (Ozawa & Kawaoka, 2013). Sin embargo, para poder identificar la presencia del virus de Influenza A H1N1 en un organismo infectado en especial en el ser humano, se ha establecido que el mejor sistema de diagnóstico es la amplificación de algunas regiones de sus genes por medio de la RT-PCR (Ellis et al., 2009); no obstante, este sistema no fue eficiente para algunos pacientes durante la pandemia del año 2009 debido a que los cebadores indispensables para la reacción no presentaron hibridación con su región complementaria, lo que conllevó a una ausencia de polimerización en la región génica empleada para el diagnóstico (González Barrios et al., 2016) (González Barrios et al., 2016 B).

La conclusión anterior y las mejoras que se realizaron para los estudios evolutivos y sistemas de diagnóstico en pacientes de ese entonces fueron gracias a la construcción de una base de datos de todos los genomas virales de Influenza A H1N1 obtenidos del GenBank hasta el año 2010; sin embargo, su construcción implicó tiempos mayores a la velocidad de las secuencias reportadas en cualquier base de datos pública. Este factor fue fundamental para la construcción de programas computacionales que permitieran el manejo: rápido, eficiente y veraz de la información disponible en las bases de datos nucleotídicas internacionales.

Con esta directriz, para la elaboración de una correcta base de datos de cualquier secuencia nucleotídica o proteica se debe partir, desde el diseño de las búsquedas en las bases de datos ya existentes bien sea privadas o públicas como la del GenBank, ya que dependiendo de la especificidad de esta, los datos encontrados son más exactos de acuerdo con los criterios de búsqueda, el no considerar lo anterior, conlleva a que al momento de la depuración de los datos, el proceso sea más complejo, ya que habrían demasiadas secuencias mezcladas que no corresponderían a las deseadas, lo que dificulta en el tiempo y en la cantidad de datos la depuración de la misma. Además, de la inclusión de datos parciales o que no concuerden con el

criterio de búsqueda, produciendo resultados que no corresponden a los conceptos biológicos del estudio. Todo ello que puede generar conclusiones incompletas o en muchos casos erróneas, conllevando a un gasto inadecuado de los recursos destinados para responder la pregunta de investigación, validar o rechazar la hipótesis y el cumplimiento de los objetivos.

Dado lo anterior, se el Grupo de Investigación desarrollo el programa computacional BioDataToolkit v1.0, que permite organizar los datos de una secuencia nucleotídica obtenidos directamente desde el GenBank, en una tabla de Excel, lo que garantiza un mejor manejo de estos y tiempos operacionales viables para un investigador; no obstante, al ser un nuevo programa, nunca había sido determinada su utilidad biológica, lo cual se realizó en esta pasantía.

Aun cuando la primera versión permitió un mejor manejo de la información al generar la hoja de Excel, la utilidad biológica de esta fue nula, debido a que en la columna correspondiente a la identificación del hospedero viral solo mostró el género y no la especie; por lo tanto, no era posible realizar un análisis más detallado de otros animales como las aves, las cuales también se obtuvieron en la búsqueda y al tener únicamente su género no podía analizar las especies presentes en el formato GenBank full, por lo que en el análisis de la información únicamente se podía analizar el género.

En la segunda versión del programa se resolvió el problema anterior, haciendo que los análisis bioinformáticos fuesen aplicables a todos los hospederos; empero, al aplicar un control de la información nucleotídica con un gen que presenta variantes de procesamiento alternativo en su ARN como es el de la Proteína de Matriz (Wise et al., 2012), se observó que el programa no

obtenía toda la información proveniente del GenBank, lo que hasta este punto hacía que el programa fuese específico solo para secuencias de un solo ORF como es el caso del gen HA del virus de Influenza A H1N1; a pesar de ello, el objetivo del programa siempre fue que sea útil para todo tipo de secuencias, sin embargo, hasta ese momento con esta versión no se daba cumplimiento a lo estipulado.

Con la solución de la problemática anterior, el programa se tornó más versátil y aplicable no solo a secuencias del gen HA del virus de Influenza A H1N1, sino a secuencias que contengan más de un ORF. No obstante, se determinó que esta nueva versión produjo un desfase de los datos, lo que volvía al programa ineficiente, puesto que al realizar la depuración con las herramientas de Excel se determinó se determinaron resultados que no correspondían a los criterios biológicos establecidos por medio de los alineamientos con los formatos Fasta generados con el programa y con ello, los investigadores tendrían que realizar nuevamente la depuración manual, ya que no es fiable la información proporcionada por el programa, lo que nuevamente a llevaría tiempos operacionales mayores. Es de resaltar que lo antes mencionado ocurrió debido a la heterogeneidad de los datos obtenidos de la plataforma del GenBank, pues no todas las secuencias tenían todos los datos virales completos y por lo tanto esta versión del programa no tuvo consideró dicho detalle.

En la versión 4.0 del programa BioDataToolkit, se soluciona esta carencia de información al implementar la palabra *empty* cuando no se encontraban los datos, lo que hizo que nuevamente la depuración volviese a llevar tiempos operacionales cortos y que al manejar gran cantidad de

datos programa fuese eficiente, pero aún se continuó el error de los espacios vacíos en la hoja de Excel.

En la última versión del programa (la 5.0) se soluciona la problemática anterior, además se logra el renombramiento de las secuencias lo que permite una identificación más sencilla, se disminuyen los tiempos operacionales ya que la lectura del formato GenBank full y la generación del archivo Excel ocurre en minutos, lo que genera una ventaja para el su uso con una cantidad de datos nucleotídicos.

Los alineamientos con las secuencias obtenidas de acuerdo a los parámetros de alineamiento usados dentro del proyecto permitieron determinar los cambios en las secuencias nucleotídicas del virus de Influenza A H1N1 colectado de seres humanos durante el año 2017; se encontró que estas no presentaron cambios mutacionales relevantes como ocurrieron en la pandemia del 2009. Todo ello fue corroborado con los alineamientos tipo BLAST, que confirmaron que la depuración realizada con el programa fue efectiva al obtener de forma mayoritaria secuencias del gen HA correspondiente al mes de búsqueda.

El análisis filogenético evidenció de manera más contundente que la selección de las secuencias del gen HA de virus de Influenza A H1N1 con las secuencias analizadas del año 2017 no presentaron cambios relevantes, lo que corrobora la utilidad de la información nucleotídica de la hoja de Excel para su aplicación en estudios biológicos. Aun cuando se pudiese considerar que el programa es óptimo para la obtención de bases de datos de secuencias nucleotídicas, es necesario la generación de la versión 6.0 la cual debe realizar la eliminación de los procesos

manuales que se requieren para la obtención de los datos biológicos y con ello lograr la automatización de todo el proceso de la generación de la base de datos.

6. Conclusiones

El uso adecuado de palabras claves para la correcta selección de secuencias de bases de datos públicas es fundamental e indispensable para la generación de una base de datos de secuencias nucleotídicas.

El desarrollo de programas computacionales con aplicaciones en bioinformática requiere de la permanente comunicación entre los diseñadores del programa y los biólogos, para obtener herramientas que permitan el correcto uso de la información genómica y proteica de cualquier virus u organismo, para la correcta y exacta generación de conocimiento, ejemplo de ello es el desarrollo del programa BioDataToolkit.

El gen de la Hemaglutinina del virus de Influenza A H1N1 del año 2017 es un buen modelo nucleotídico para la validación de herramientas computacionales para estudios biológicos, prueba de ello es que se logó la generación de alineamientos, secuencias consenso, alineamientos tipo BLAST y su filogenia por el método de Inferencia Bayesiana a partir de los datos generados con el programa BioDataToolkit versión 5.0 al determinar que no presentó cambios relevantes en sus secuencias nucleotídicas.

Aun cuando el programa BioDataToolkit versión 5.0 tiene la capacidad de generar la depuración de los datos nucleotídicos reportados en bases de datos públicas, es indispensable que

se concluyan los procesos de automatización para eliminar la selección manual de secuencias nucleotídicas que no están incluidas en los criterios de selección en la búsqueda del GenBank, para la obtención del formato GenBank full.

7. Recomendaciones

Se recomienda realizar la nueva versión del programa BioDataToolkit 6.0 la cual debe eliminar las comillas de la columna llamada Fasta para suprimir la generación del documento Word en el que actualmente se deben copiar primero las secuencias para luego copiarlas al texto plano, ya que esto le resta practicidad el programa. Por otra parte, se recomienda que cuando el programa lea el archivo en formato GenBank full y no encuentre alguno de los datos, realice la búsqueda en el ítem denominado "*strain"*, ya que este ítem en ocasiones contiene la información que no se encuentra en otras partes del documento. Adicionalmente es recomendable que el programa genere una columna complementaria en donde se indique si el CDS de la secuencia nucleotídica es completo o parcial, esto lo puede hacer realizando la búsqueda en la información del ítem llamado Fasta.

Referencias Bibliográficas

Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., … Zbicz, K. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *46*(D1), D8–D13. https://doi.org/10.1093/nar/gkx1095

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., … Lipman, D. (2008). The Influenza Virus Resource at the National Center for Biotechnology Information. *Journal of Virology*, *82*(2), 596–601. https://doi.org/10.1128/JVI.02005-07

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2017). GenBank. *Nucleic Acids Research*, *45*(D1), D37–D42. https://doi.org/10.1093/nar/gkw1070

Chan, M. (2009). *WHO | World now at the start of 2009 influenza pandemic*. *WHO*. World Health Organization. Retrieved from https://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/

Christman, M. C., Kedwaii, A., Xu, J., Donis, R. O., & Lu, G. (2011). Pandemic (H1N1) 2009 virus revisited: An evolutionary retrospective. *Infection, Genetics and Evolution*, *11*(5), 803–811. https://doi.org/10.1016/j.meegid.2011.02.021

Cox, N. J., & Subbarao, K. (2000). GLOBAL EPIDEMIOLOGY OF INFLUENZA: Past and Present *. *Annual Review of Medicine*, *51*, 407–421. https://doi.org/10.1146/annurev.med.51.1.407

Cui, D., Zhao, D., Xie, G., Yang, X., Huo, Z., Zheng, S., … Chen, Y. (2016). Simultaneous detection of influenza A subtypes of H3N2 virus, pandemic (H1N1) 2009 virus and reassortant avian H7N9 virus in humans by multiplex one-step real-time RT-PCR assay. *SpringerPlus*, *5*(1), 4–11. https://doi.org/10.1186/s40064-016-3733-9

Ellis, J., Iturriza, M., Allan, R., Bermingham, A., Brown, K., Gray, J., & Brown, D. (2009). Evaluation of four real-time PCR assays for detection of influenza a (H1n1)v viruses. *Eurosurveillance*, *14*(22), 20–22. https://doi.org/https://doi.org/10.2807/ese.14.22.19230-en

Fineberg, H. V. (2014). Pandemic Preparedness and Response — Lessons from the H1N1 Influenza of 2009. *New England Journal of Medicine*, *370*(14), 1335–1342. https://doi.org/10.1056/nejmra1208802

González Barrios, J. A., Thompson Bonilla, M. D. R., Martínez Pérez, F. J., Barrios Hernández, C. J., Bautista Rozo, L. X., Rodríguez Vázquez, R., & Martínez Fong, D. (2016). *WO 2017/195063 A1*.

González Barrios, J. A., Thompson Bonilla, M. D. R., Martínez Pérez, F. J., Barrios Hernández, C. J., Rodríguez Vázquez, R., Martínez Fong, D., … Madero, G. A. (2016). OLIGONUCLEOTIDOS Y PROCESO PARA DETECTAR EL VIRUS DE LA INFLUENZA A H1N1. México. Retrieved from http://siga.impi.gob.mx/newSIGA/content/common/ficha.jsf?idFicha=7268103

Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95-98.

Huelsenbeck, JP, y F. Ronquist. 2001. MRBAYES: Inferencia bayesiana de la filogenia. Bioinformática 17: 754-755.

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., … Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz268

McMullen, A. R., Anderson, N. W., & Burnhamfor, C. A. D. (2016). Pathology Consultation on Influenza Diagnostics. *American Journal of Clinical Pathology*, *145*(4), 440–448. https://doi.org/10.1093/AJCP/AQW039

National Institute of Allergy and Infectious Diseases (NIH / DHHS). (2017). Influenza Research Database (IRD). Retrieved June 5, 2019, from https://www.fludb.org/brc/influenza_sequence_search_segment_display.spg?method=ShowCleanSearch&decorator=influenza

Ortiz, R., Rojo, S., & Sanz, I. (2019). Retos diagnósticos de la gripe. *Enfermedades Infecciosas y Microbiología Clínica*, *37*(Supl 1), 47–55. https://doi.org/10.1016/S0213-005X(19)30182-X

Ozawa, M., & Kawaoka, Y. (2013). Cross talk between animal and human influenza viruses. *Annual Review of Animal Biosciences*, *1*(1), 21–42. https://doi.org/10.1146/annurev-animal-031412-103733

Peteranderl, C., Herold, S., & Schmoldt, C. (2016). Human Influenza Virus Infections. *Semin Respir Crit Care Med*, *37*(4), 487–500. https://doi.org/DOI: 10.1055/s-0036-1584801

Stefańska, I., Dzieciatkowski, T., Brydak, L. B., & Romanowska, M. (2013). Application of three duplex real-time PCR assays for simultaneous detection of human seasonal and avian influenza viruses. *Archives of Virology*, *158*(8), 1743–1753. https://doi.org/10.1007/s00705-013-1648-0

White, M. C., & Lowen, A. C. (2018). Implications of segment mismatch for influenza A virus evolution. *Journal of General Virology*, *99*(1), 3–16. https://doi.org/10.1099/jgv.0.000989

WHO. (2010). *WHO | H1N1 in post-pandemic period*. *WHO*. World Health Organization. Retrieved from https://www.who.int/mediacentre/news/statements/2010/h1n1_vpc_20100810/en/

Wise, H. M., Hutchinson, E. C., Jagger, B. W., Stuart, A. D., Kang, Z. H., Robb, N., … Digard, P. (2012). Identification of a Novel Splice Variant Form of the Influenza A Virus M2 Ion Channel with an Antigenically Distinct Ectodomain. *PLOS Pathogens*, *8*(11), 1–14. https://doi.org/10.1371/journal.ppat.1002998