UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE DOCTORADO EN CIENCIAS BIOMÉDICAS

INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

CARACTERIZACIÓN DE DIFERENTES PATRONES DE INFORMACIÓN MUTUA ENTRE VIRUS DE RNA Y DNA

TITULACIÓN POR TESIS

QUE PARA OPTAR POR EL GRADO DE DOCTOR EN CIENCIAS BIOMÉDICAS

PRESENTA:

BIÓL. VÍCTOR MANUEL SERRANO SOLÍS

TUTOR

DR. MARCO ANTONIO JOSÉ VALENZUELA

INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

COMITÉ TUTOR

DR. LUIS PADILLA NORIEGA

INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

DR. PEDRO MIRAMONTES VIDAL

FACULTAD DE CIENCIAS

MÉXICO, D. F. NOVIEMBRE 2013

RESUMEN DEL TRABAJO DE TESIS

El presente trabajo de tesis se divide en dos partes:

A. En primer lugar, se presentará el artículo de investigación ya aceptado y publicado, derivado de una sección del trabajo de investigación doctoral. Este artículo se centra en la comparación entre reconstrucciones filogenéticas y una organización jerárquica basada en la teoría de la información, teoría que permite develar patrones ocultos en series de distancia, en este caso secuencias genómicas. Una vez confirmada la hipótesis de que la teoría de información captura información evolutiva relevante, se cuantifican y determinan ganancias y pérdidas de información, medida en bits, del virus de la influenza a lo largo de 100 años de historia evolutiva pandémica en humanos. Éste es un aporte a la teoría evolutiva para el virus de la influenza A.

B. En segundo lugar se presentará trabajo experimental de caracterización viral mediante la teoría de la información, realizado de igual manera durante el período doctoral. Se aplican pruebas estadístico-matemáticas para responder la hipótesis de si existen similitudes y/o diferencias en los patrones de información entre 792 secuencias de genomas virales, bacterianos y de Archaeas, así como sus respectivos 792 controles aleatorizados. Se esperaba que una periodicidad de tres bases estuviera presente en todos los perfiles de IM de los genomas analizados dado que todos son codificantes para proteínas, sin embargo sólo los virus de DNA mantuvieron dicho patrón. Los virus de RNA presentan perfiles MIF drásticamente diferentes. De igual manera se implementa una clasificación dos-dimensional donde se muestran diferencias claras entre los genomas celulares y los genomas virales, así como diferencias entre sí, de los diversos genomas virales analizados. Actualmente este trabajo se encuentra en un avance del 70% para su pronta publicación.

PRIMERA PARTE

ARTICULO DE INVESTIGACIÓN PARA TITULACIÓN.

Flow of Information during an Evolutionary Process: The Case of Influenza A Viruses.

Víctor Serrano-Solís and Marco V. José. *Entropy* **2013**, *15*, 3065-3087; doi:10.3390/e15083065

Received: 14 March 2013 / Accepted: 23 July 2013 / Published: 29 July 2013

INTRODUCCIÓN GENERAL AL ARTICULO DE INVESTIGACIÓN

Basándonos en el objetivo de investigar cómo poder unir la teoría evolutiva con la teoría de información, obtuvimos resultados originales que nos llevaron a publicar el artículo de investigación **"Flujo de Información Durante el Proceso Evolutivo: El Caso del Virus de la Influenza A".**

A partir de un modelo experimental como lo es el virus de la influenza A [13], buscamos subtipos genómicos de relevancia especial, en este caso de importancia pandémica tanto en aves como humanos [30]. Esta colección de subtipos nos acercóla oportunidad de dilucidar la historia evolutiva de 100 años del virus de Influenza A humana. A partir de la base de datos obtenida se desarrollaron pruebas de contenido de Información Mutua, y pruebas evolutivas estándar de evolución molecular, como las reconstrucciones filogenéticas.

Se construyeron Dendrogramas a partir de los cálculos de MIF, estos pueden ser tomados como equivalentes de los análisis de similitud en bioinformática. Se generaron reconstrucciones filogenéticas con herramientas del estado del arte, en este caso tanto para árboles de Neighbor Joining, como para Máxima Verosimilitud (con acrónimo inglés ML). Un análisis comparativo entre los Dendrogramas MIF y las Filogenias calculadas con ML, nos muestra que los agrupamientos jerárquicos resultaron inesperadamente muy parecidos. Se desprende entonces la idea de que la MIF captura y refleja información evolutiva relevante de los genomas de Influenza A.

Desde el punto de vista de la información mutua se obtuvieron resultados interesantes. Se observa que los segmentos del genoma viral ganan y pierden información a lo largo del tiempo. Estos cambios se encuentran dentro de un rango constante y aparentemente cada pico es la expresión de emergencia pandémica de cada subtipo. Es importante notar que después de cada pico asociado a alguna pandemia se desarrolla un regreso a los subtipos estacionales que resulta ser la media del rango del flujo informacional para que posteriormente emerja un nuevo pico con la pandemia siguiente, tal como lo describe Wolf [15]. Este flujo de información nos muestra que el mensaje genético presenta, por lo menos, cifras distintas, en bits, a lo largo del proceso evolutivo.

PERSPECTIVAS GENERALES DEL ARTICULO DE INVESTIGACIÓN

Aun no podemos determinar si las correlaciones de Información Mutua son un resultado adaptativo, por ejemplo, debido a una presión constante de selección ejercida por el sistema inmune de los diversos hospederos.Mas sabemos que mediante la teoría de información, podemos dilucidarles, medirles y clasificarles. Son muestra inequívoca de estructura y organización genética de cualquier organismo ya sea de DNA o de RNA.

1. Surge entonces la pregunta de qué tipo de estructuras se obtendrían a partir de un mismo subtipo (A/H3N2 humano; A/H5N1 aviar; A/H1N1 humano y/o porcino) a lo largo del mayor tiempo posible de registros públicos.
2. De igual manera, se busca explorar historias evolutivas de otras familias de virus aparte de los Orthomixoviridae.
3. Más aun, se busca comprender la dinámica de ganancia y pérdida de información en Influenza A para desarrollar el respectivo modelo y predecir de forma certera, donde la bioinformática ha fallado, futuros cambios y cómo se reflejarían en cada segmento genómico del virus.
4. Conocer cómo está relacionada esta dinámica de ganancia y pérdida de información con respecto a la adaptación del virus.
5. Determinar qué variaciones hay en el sistema inmune del hospedero con respecto a estas oscilaciones de la información.

SEGUNDA PARTE

1. INTRODUCCIÓN

El objetivo de los proyectos genómicos es revelar la totalidad de Información genética de un organismo. Los procesos de secuenciación automática han aumentado en su velocidad, sensibilidad y exactitud [Dark*et al.*, 2013], lo cual ha generado una cantidad de datos enorme [Venter *et al.*, 2004]. Es evidente entonces la necesidad de metodologías *in silico* para discernir, encontrar y entender nuevos patrones y motivos en estructuras nucleotídicas tanto de DNA como de RNA [Hunter *et al.*, 2013].

Tradicionalmente los virus han sido estudiados de acuerdo a su importancia y propiedades patológicas. Los genomas virales muestran una diversidad biológica impresionante y apenas comenzamos a formar una idea de su importancia global [Breitbart *et al.,* 2002]. Los virólogos ahora buscan entender la naturaleza biológica tan especial y única de todos y cada uno de los tipos virales [Cairns *et al.*, 1966; Norrby, 2008]. Recientemente, la investigación científica ha cambiado el rumbo para entender a los virus como entidades de impacto ecológico e importancia evolutiva en la historia de la vida en la Tierra [Forterre, 2006]. Aun no existe algún tipo de clasificación taxonómica para agrupar a todos los virus según un rasgo básico común. Ponemos en la mesa la tesis de que sí existe ese rasgo. La única característica que puede incorporar a todos los virus conocidos dentro de un solo grupo es su macromolécula de información; a partir de ahí se pueden discriminar hacia dos subgrupos, los virus de genoma de DNA y los virus de genoma de RNA.

En nuestro caso, estamos interesados en conocer el contenido de información de una secuencia genómica sea de DNA o RNA. En éste trabajo, se utilizará la Función de Información Mutua, que se mencionará de ahora en adelante con su acrónimo en inglés “MIF”, como un método capaz de descubrir e identificar precursores de estos grandes grupos virales para posteriormente inferir diferencias y similitudes entre ellos. Los fundamentos de la teoría de la información fueron establecidos en un artículo de Shannon titulado en inglés, “A Mathematical Theory of Comunication” [Shannon, 1948]. Shannon estaba interesado en cuantificar qué tanta información se podría transmitir por un canal de comunicación dado. La MIF exhibe el contenido de Información en un mensaje, en este caso el mensaje genético. Los virus son únicos, perseguimos la idea de revelar patrones específicos dentro y entre virus de DNA y/o RNA.

Una característica de cualquier ser vivo es la habilidad inherente del procesamiento de información para mantener su actividad biológica. Esta actividad: a) debe ser de naturaleza física y, b) afecta a fenómenos de largo plazo como la evolución y herencia. [Román-Roldán, 1996]. Las metodologías de teoría de la información son adecuadas para el análisis de

secuencias nucleotídicas, esto es debido a su capacidad de manejar secuencias simbólicas. Una cuestión básica es obtener mediciones significativas y confiables acerca del orden, regularidad, estructura y complejidad, etc. en una secuencia dada. Esto podría permitir comparaciones con otras secuencias (o con otros segmentos de la misma secuencia), derivando en resultados de interés para estudios evolutivos (filogenias moleculares), identificación de segmentos codificantes (hallazgo de genes, exones, señales de transcripción), etc. En general, el objetivo es una medición capaz de indicar qué tan lejos está una secuencia nucleotídica de su respectiva aleatorizada. Si una secuencia no es numérica, tal como un lenguaje de texto, solamente la MIF puede ser aplicada dado que no se le asigna una cifra numérica cada símbolo del texto. [Román-Roldán, 1996; Li, 1997].

El análisis MIF fue utilizado inicialmente para buscar diferencias entre secuencias codificantes para proteínas y no codificantes para proteínas. Los perfiles MIF son especie-independiente, es decir, puede ser aplicado a cualquier genoma, sin suposiciones *ad hoc* [Grosse *et al.*, 2000b].

La MIF tiene la capacidad de distinguir entre DNA codificante para proteínas y no codificante para proteínas sin adecuaciones previas.Independientemente de su origen filogenético se sugiere que la MIF puede ser útil para identificar regiones codificantes en genomas en donde no existen curaciones o anotaciones previas. Respecto al DNA codificante, en el artículo de Grosse [Grosse *et al.,* 2000a] pudieron mostrar que la suma de los valores de información presentes en los marcos de lectura en cromosomas humanos, no es suficiente para reproducir las distribuciones de información mutua cromosómicas, es decir, para alcanzar los niveles de información mutua biológicos, se necesita la totalidad de la secuencia más otras correlaciones de larga distancia y no solamente los marcos de lectura. Este hallazgo hace tentadora la suposición de que correlaciones adicionales son vitales como ingrediente de cualquier secuencia codificante en cualquier organismo vivo [Grosse *et al.,* 2000a]. Un ejemplo es que se han hallado correlaciones a larga distancia (k=165) de repeticiones Alu en cromosomas humanos [Holste *et al.,* 2003].

De manera técnica, la MIF cuantifica la información del nucleótido X(A,C,T(U),G) respecto al nucleótido Y que es localizado a k posiciones río-abajo del nucleótido X [Grosse *et al.*, 2000], desarrollando una matriz de 16 combinaciones posibles de dinucleótidos contra el número de distancias (k) calculadas. En [Grosse *et al.* 2000b], usaron la MIF para analizar secuencias de DNA humanas tanto codificantes para proteínas como no codificantes para proteínas. La MIF reveló diferentes estructuras y valores entre el DNA codificante y el DNA no codificante[Grosse *et al.,* 2000].

Los perfiles MIF proveen una firma de especie sencilla y fácilmente computable. Esta firma genómica puede ser utilizada en aplicaciones donde relaciones evolutivas necesitan ser conocidas a partir de secuencias relativamente pequeñas, al igual que en los casos en donde se

necesita averiguar relaciones evolutivas entre organismos a partir de secuencias genómicas extensas. La estructura de los perfiles MIF es conservada, aun en sub-secuencias cortas de un genoma dado, resultando en una firma pervasiva [Bauer *et al.,* 2008].

2. HIPOTESIS

Los genomas virales son secuencias codificantes altamente compactadas. Independientemente de su tipo de macromolécula de Información, sea de DNA o RNA, todos están sujetos al código genético universal de la misma manera que sus hospederos lo están. Luego entonces, se esperan obtener de los perfiles MIF virales, patrones similares a los perfiles MIF calculados en genomas celulares.

3. RELEVANCIA E IMPACTO DEL PROYECTO EN EL CAMPO DE ESTUDIO

Dado que la diversidad viral es absolutamente apabullante y aún no ha sido descubierta en su totalidad, a la fecha no ha sido posible encontrar un rasgo común para agrupar a todos los virus en la sistemática actual. Si bien la clasificación viral de Baltimore es un sistema vigente, está solamente basado en el tipo de la semántida (DNA, RNA, cadenas sencillas y cadenas dobles) y en su sentido de la replicación (sentido o antisentido). En este esfuerzo tomamos a la secuencia genómica, sea DNA o RNA, como el único rasgo común de todos los virus, y a su contenido de Información (desde el punto de vista de la Teoría de la Información), como un rasgo evolutivo que permitirá comparaciones diversas. Según nuestro conocimiento, este es el primer esfuerzo extensivo y sistemático del uso de la MIF para analizar la diversidad de la virosfera conocida al día de hoy. En este esfuerzo, 693 secuencias virales serán analizadas, organizadas y clasificadas. La búsqueda detallada de patrones en los genomas virales nos permitirá descubrir aspectos fundamentales de sus historias evolutivas [Serrano *et al.*, 2013]. Este trabajo puede contribuir a nuevos criterios en las metodologías de clasificación taxonómica actuales debido a la característica de independencia de especie, así como la pervasividad encontrada en las respectivas firmas genómicas.

4. METODOLOGÍA

4.1 BASE DE DATOS

Este trabajo se desarrolló en una base de datos conformada por 693 secuencias virales, 40 genomas bacterianos y 59 genomas de Archaeas.

Todas las secuencias fueron tomadas del Genebank Viral Genomes Site en el sitio web: http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=1023

Las secuencias virales y los genomas unicelulares fueron seleccionados al azar, presentando por tanto diferentes longitudes, agrupamiento taxonómico y rango de hospedero. En el caso de

Archaeas, la totalidad de los genomas existentes fueron descargados de la base de datos previamente mencionada.

4.2 ANÁLISIS MIF

La ecuación de la MIF (1) es la siguiente:

(1)

$$I(k,s) = \sum_{\alpha \in A} \sum_{\beta \in A} P_{\alpha,\beta}(k,s) \log_2 \left[\frac{P_{\alpha,\beta}(k,s)}{P_\alpha(s) P_\beta(s)} \right]$$

Para el desarrollo de la ecuación se recomienda referir al artículo [Serrano *et al.*, 2013]. El algoritmo de la MIF fue implementado en MATLAB de la siguiente manera:

A partir de una secuencia nucleotídica (*s*), se cuantifica su longitud en número de bases. Se cuantifican las apariciones de cada símbolo denotado como base nucleotídica, en la secuencia (*s*). Se establece la distancia máxima (*k*), del análisis MIF. Se cuantifica la probabilidad marginal de cada símbolo, denotado como base nucleotídica en $P_\alpha(s)$ o $P_\beta(s)$, de la secuencia (*s*). Se cuantifica la probabilidad conjunta del símbolo α seguido de *k* bases del símbolo β en la secuencia (*s*), denotado por $P_{\alpha,\beta}(k,s)$. Se sustituyen los valores previamente calculados en la ecuación (1). Se resuelve la ecuación para cada una de las 16 combinaciones de dinucleótidos; ej: AA, AC, AG, AT, CA,… (TT), para cada distancia (*k*), obteniendo los valores en bits debido a la utilización del logaritmo base dos. Se genera entonces una matriz de 16 renglones por (*k*) columnas, y se obtiene el promedio de cada columna. A éste renglón de promedios se le denota M. Se grafican entonces los valores de M (dado en bits) contra la distancia máxima (*k*) y se obtiene el perfil MIF de la secuencia analizada.

4.3 ANALISIS AMIF

El algoritmo Average Mutual Information Function (AMIF) es un cálculo reportado por Ivo Grosse en [Grosse *et al.*, 2006]. Su intención era, a partir de las frecuencias normalizadas de cada nucleótido dada una de las tres posiciones del codón calcular un dígito y utilizarlo para hacer comparaciones prácticas de los contenidos de información de varias secuencias.

A grandes rasgos, en el algoritmo de AMI se calculan las probabilidades conjuntas en términos de las 12 probabilidades posicionales de un codón. En otras palabras, se calcula la Información Mutua de un nucleótido X respecto a un nucleótido Y a K distancia, donde Y está a una distancia múltiplo de 3 y se le llama I_{in} y se pondera como 1/3 del valor de AMI; cuando se calculan las otras dos posiciones se llama I_{out} y se ponderan como 2/3 del valor de AMI.

El algoritmo AMIF fue escrito en MATLAB de acuerdo al algoritmo reportado en [Grosse *et al.*, 2006]. Con este algoritmo se analizó la base de datos previamente descrita.

Los resultados AMIF fueron graficados en MATLAB usando la Función de Densidad de Probabilidad (PDF por sus siglas en inglés) lo cual consiste en normalizar un histograma.

4.4 INFRAME Y OUTFRAME

Originalmente en [Grosse *et al.*, 2006], propusieron la ecuación AMI. En el presente trabajo, modificamos la AMI para hacer una comparación más sencilla de datos de información entre secuencias de DNA y/o RNA. A partir de los valores de información del perfil de una secuencia dada, secalcula el promedio de todas las autocorrelaciones (las 16 combinaciones nucleotídicas) a cortas distancias k=3, k=4 y k=5. Llamaremos a la distancia 3, "Inframe", que es el punto de mayor correlación observado en el patrón de periodicidad de tres bases de los perfiles MIF de genomas de DNA; se interpreta como la correlación entre codones a una distancia de tres (k=3) posiciones, es decir, la correlación del primer nucleótido de un codón respecto a los siguientes primeros nucleótidos de los codones posteriores.

Por otra parte llamamos "Outframe" al valor del promedio de las correlaciones medidas a distancias k=4 y k=5. Este valor Outframe es la correlación calculada de la 2a y 3era posición de un codón con respecto a los siguientes codones. El promedio de Inframe y Outframe resulta en el valor que llamamos AMIF. Éste algoritmo fue implementado en MATLAB y los valores resultantes son utilizados para generar (desde MATLAB) gráficas de dispersión *XY* para conocer el agrupamiento dos dimensional de los genomas virales, los genomas celulares y todas las secuencias aleatorizadas respectivas.

5. RESULTADOS

5.1 PERFILES MIF

Los análisis MIF fueron calculados en los genomas de 59 Archaeas, así como de 40 genomas bacterianos. En general el DNA no codificante para proteínas es escaso en los genomas de Archaeas y Bacterias, por tanto presenta un efecto insignificante sobre los cálculos de la MIF. Los valores de correlación de secuencias genómicas celulares fueron graficados y se observa un patrón oscilatorio cada tres posiciones a causa de la disposición codónica en tripletes del código genético [Swati, 2007]. Es importante notar que el patrón oscilatorio de tres bases es constante y fue hallado en todos los genomas celulares analizados. Por otra parte existe una disminución de valores al aumentar la distancia del análisis, debido a la longitud relativamente corta de los genomas de Archaeas y Bacterias. (Figura 1). El patrón de decaimiento oscilatorio disminuye

rápidamente en valores cercanos a cero, a cortas distancias. En cada sub-gráfica de la Figura 1 podemos apreciar una línea basal roja, correspondiente a los valores MIF de la respectiva secuencia aleatorizada. Se tomará a esta línea basal como control negativo debido a que al aleatorizar la secuencia genómica se rompe la estructura de autocorrelaciones y su cuantificación no es significativa [Berryman *et al.,* 2004]. Las distancias que existen entre la secuencia biológica original y la respectiva aleatorizada, son entonces evidentes. Tomando como control positivo a los genomas celulares analizados y como control negativo a los genomas respectivos aleatorizados, los análisis MIF en virus son el siguiente paso.

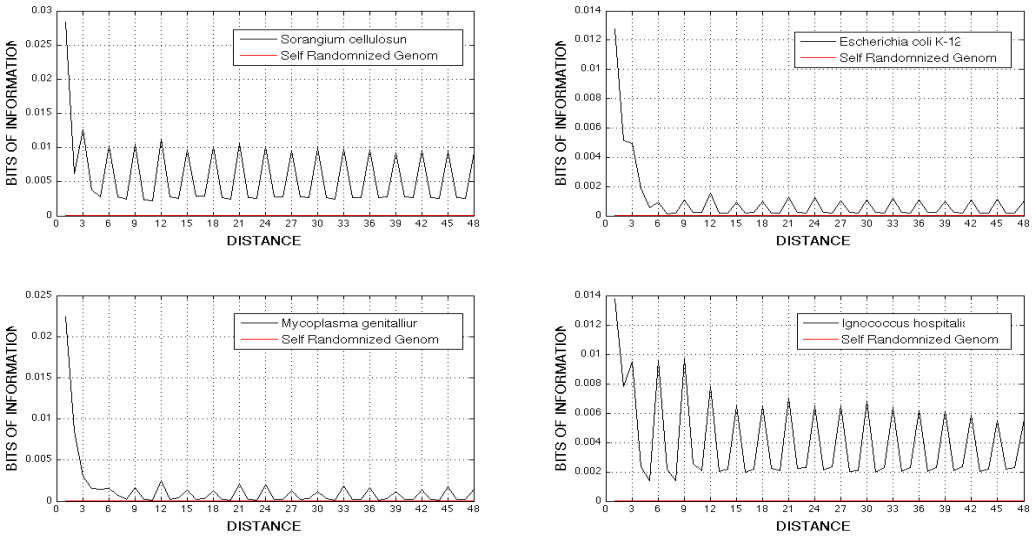


Figura 1: Análisis MIF de Genomas Completos de Archaeas y Bacterias. La línea basal roja corresponde al respectivo genoma aleatorizado. Siempre se observa claramente <u>el patrón de periodicidad de tres bases.</u>

Los primeros virus analizados con MIF (Figura 2) son los correspondientes a la Familia Adenoviridae, género Mastadenovirus. Estos virus son de doble cadena de DNA con un rango de hospedero que abarca los mamíferos. La presencia de una oscilación periódica así como el decaimiento a mayor distancia, fueron obtenidos como se esperaba. Es interesante el hecho de que en estos virus prevalezca la selección de la periodicidad de tres bases. Este tipo de comportamiento genómico viral también puede ser explicado por la presencia de Repeticiones Largas al Término del genoma, característica de los Adenovirus (Flint, 2001). Las repeticiones en las secuencias de DNA pueden contribuir a las correlaciones de larga distancia debido a su subestructura. Los LINES a menudo presentan “repeticiones terminales largas” en sus extremos

finales. Teniendo en mente los más de 20000 LINES, que llegan a abarcar extensiones de varias kilobases en los genomas, estas “repeticiones dentro de repeticiones” bien pueden inducir correlaciones a largas distancias de varias miles de bases, como ya se han encontrado empíricamente [Herzel *et al.,* 1994].

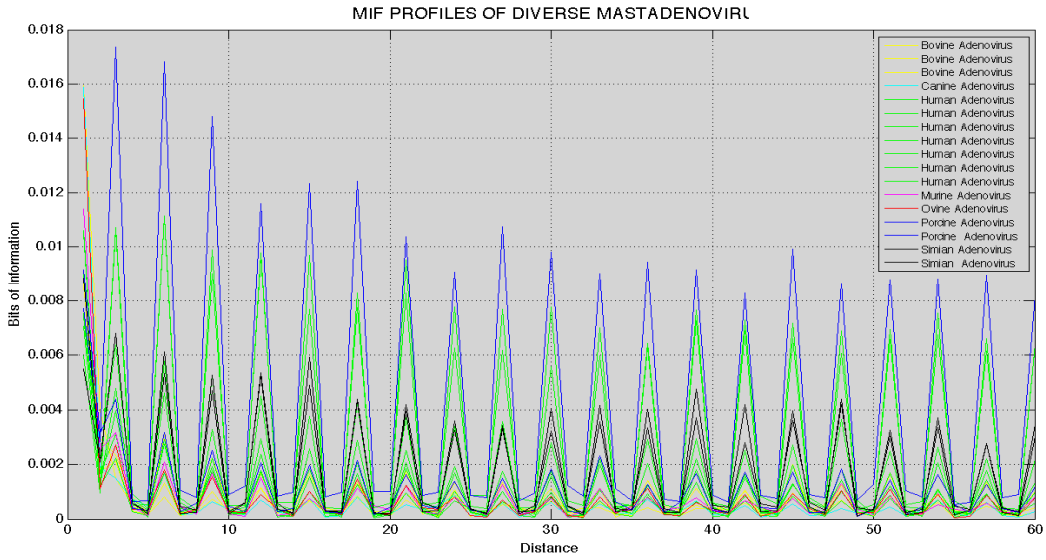


Figura 2: Análisis MIF de diversos virus de la Familia Adenoviridae, Género Mastadenovirus. La periodicidad de tres bases se manifiesta en Eukariavirus.

Analizando otro virus, como el colifago Lambda (Figura 3) podemos observar de nueva cuenta un perfil parecido al celular. Este comportamiento genómico podría denotar alguna similitud entre las correlaciones del virus y su hospedero. Al extender el análisis MIF a la diversidad de genomas virales de la base de datos, fue muy claro que la periodicidad de tres bases característica de los genomas celulares no aparece en los virus con genoma de RNA. De todos los genomas virales completos caracterizados en este estudio, solo los de genoma de DNA mostraron una periodicidad de tres bases claramente diferenciada.

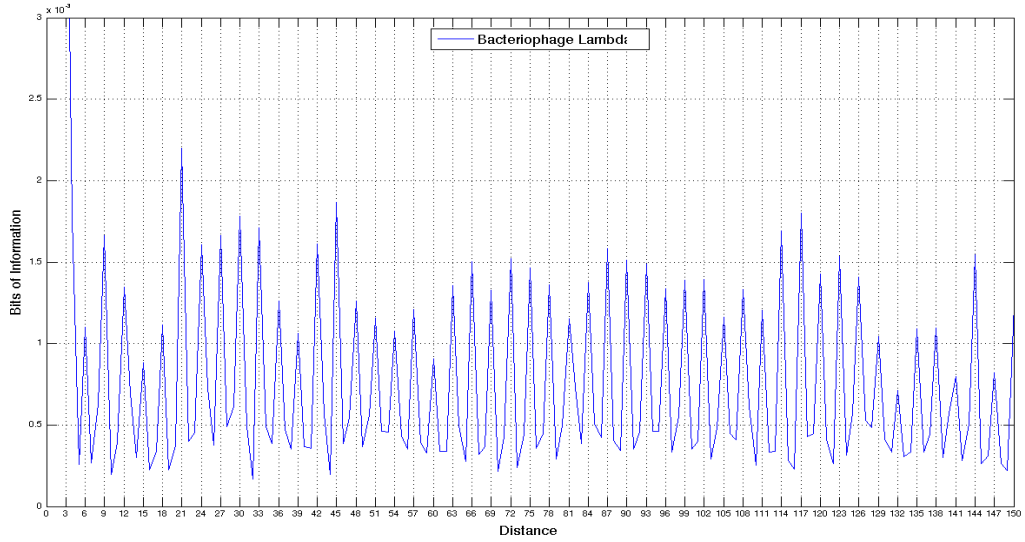


Figura 3: Perfiles MIF del Colifago Lambda. La periodicidad de tres bases se manifiesta claramente en este bacteriófago.

Un ejemplo de los perfiles MIF de las secuencias genómicas virales analizadas se ilustra en la Figura 4, donde un grupo de perfiles MIF de diversos virus de Genoma de DNA son comparados.

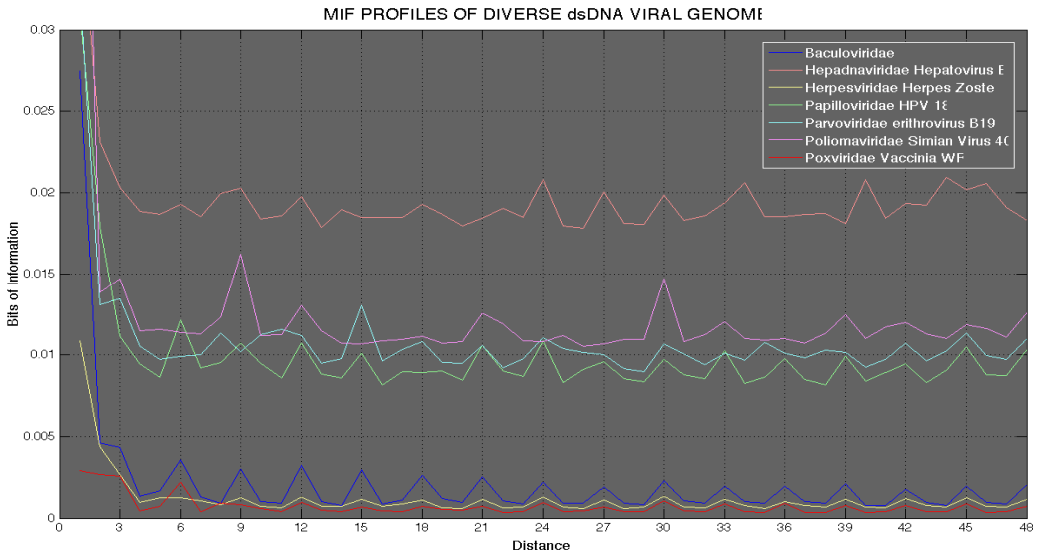


Figura 4: Perfiles MIF de diversos virus de Genoma de DNA. La periodicidad de tres bases se presenta en los perfiles de virus de genoma de DNA.

A primera vista, los perfiles MIF se distribuyen en tres subgrupos. Esto puede darse debido al resultado de la composición genética o al uso de codones de cada secuencia analizada, dado que los subgrupos formados no reflejan un criterio de agrupamiento taxonómico vigente. El virus Vaccinia (importante por su papel en la erradicación de la viruela), el Herpes zoster (mismo que el de la varicela) y el Baculovirus (virus de invertebrados) presentan los menores valores de correlación de todos los virus de la gráfica, pero muestran una clara periodicidad de tres bases. El perfil MIF del Virus 40 de simio presenta la mayor correlación a distancias 9 y 30, interesantemente ambos múltiplos de 3.

En la Figura 4 se ilustra la independencia entre los valores de correlación y la longitud genómica, siendo que el genoma del virus Vaccinia mide en promedio 200 kb mientras que el genoma de Parvovirus (ubicuo en mamíferos) presenta apenas un promedio de 5 kb de longitud.

Cuando los análisis MIF abarcaron a los virus de genoma de RNA, una extraña característica fue descubierta. De manera sorprendente estas secuencias de RNA no presentaban la periodicidad de tres bases. El primer genoma de RNA viral analizado fue un miembro de la familia de los Reoviridae, un Rotavirus serotipo C. Este virus tiene un genoma segmentado de RNA de doble cadena de once secuencias codificantes. Sus perfiles MIF se ilustran en la Figura 5 (líneas S1 a S11). En ésta misma Figura podemos ver los perfiles MIF del virus de la Enfermedad Bursal y el perfil del virus de la Necrosis Pancreática B1, ambos con genomas de RNA de doble cadena integrantes de la familia Birnaviridae. Sorpresivamente los once perfiles respectivos del Rotavirus C, presentan diferentes valores MIF entre ellos. Algunos segmentos del Rotavirus C comparten los rangos de valores de correlación con los Birnavirus, en lugar de consigo mismo como se esperaría por coherencia genética y uso de nucleótidos y codones. El segmento 1 del genoma de Rotavirus C codifica para la RNA polimerasa dependiente de RNA y presenta los menores valores de correlación. Otros perfiles de segmentos codificantes para proteínas estructurales del Rotavirus C están cerca de los valores del perfil del Segmento 1, al igual que de los perfiles de la Enfermedad Bursal y el virus de la Necrosis Pancreática B1. Interesantemente, mientras que los perfiles respectivos a segmentos codificantes para proteínas de replicación muestran los menores valores de correlación, los perfiles de los segmentos que codifican para proteínas no estructurales del Rotavirus C (Segmento 10 y Segmento 11), presentan los mayores valores de correlación en la gráfica.

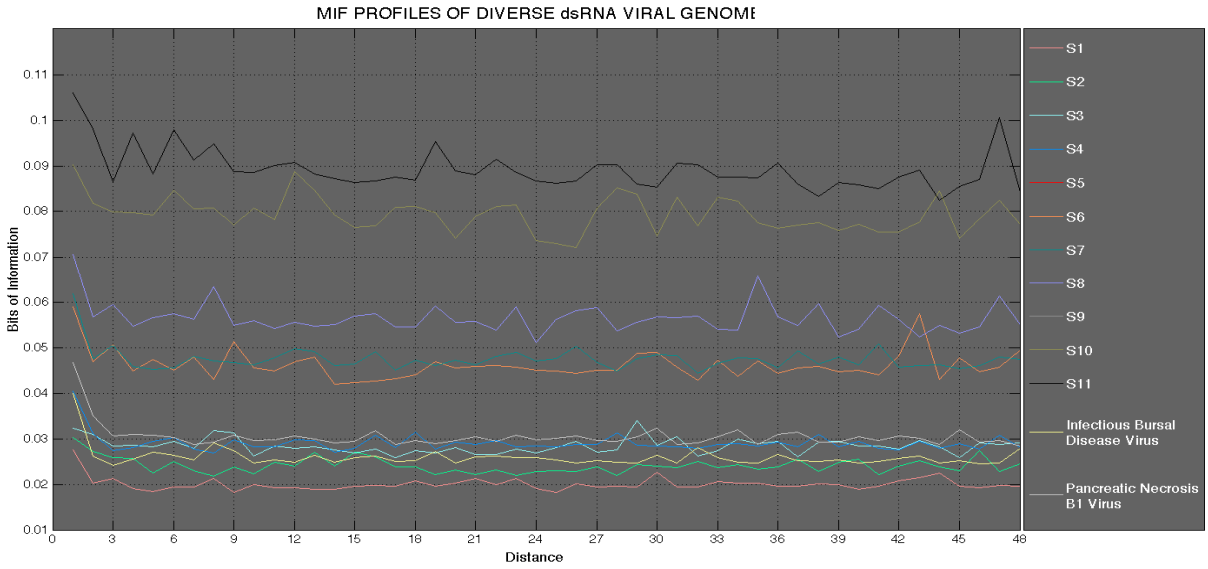


Figura 5: Análisis MIF de virus de RNA de doble cadena. <u>La periodicidad de tres bases deja de ser discernible en los virus de RNA de doble cadena.</u>

Hasta ahora solo una pequeña fracción (7%) del total de secuencias virales analizadas se han graficado. La Figura 6 presenta una gráfica de genomas segmentados virales de RNA de la familia Nanoviridae. Se observa un comportamiento similar al de la mayoría de perfiles MIF de virus de genoma de RNA. Es por esto que el siguiente paso fue determinar y utilizar un algoritmo para aplicar una métrica a cada perfil y generar comparaciones cuantitativas y sistemáticas de todos y cada uno de los perfiles MIF de las secuencias biológicas originales así como de sus respectivos controles negativos (genoma aleatorizado) de genomas virales, de Bacterias y de Archaeas de nuestra base de datos.

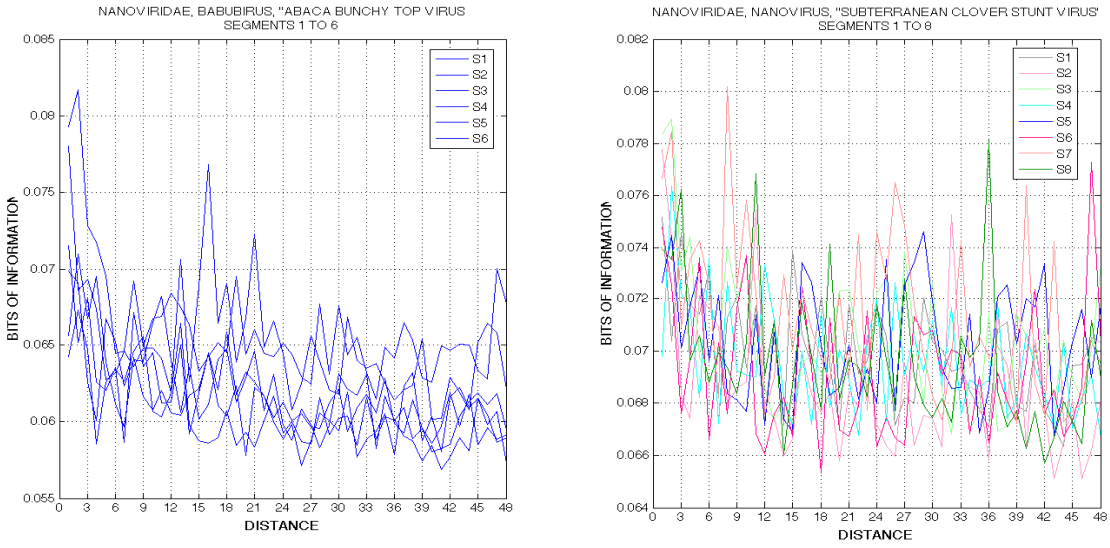


Figura 6: Análisis MIF de virus de la Familia Nanoviridae, Género Babivirus; ambos virus son de genoma segmentado de RNA de cadena sencilla. La periodicidad de tres bases desaparece.

5.2 AMIF

Calculando el valor de AMIF de cada secuencia genómica viral y celular obtuvimos un único valor que nos permite una comparación más sencilla. La función de densidad de probabilidad (de acrónimo inglés PDF) de los 59 genomas de Archaeas y la distribución de los 40 genomas bacterianos analizados con AMIF, muestran bimodalidad (Figura 7). En el caso de la distribución de Archaeas, una revisión de los componentes de los bines nos dice que existe una gran cantidad de Archaeas Metanogénicas en la moda de más alta Información, mientras que en la moda de baja información se encuentran todas las Archaeas del género *Sulfolobus*.

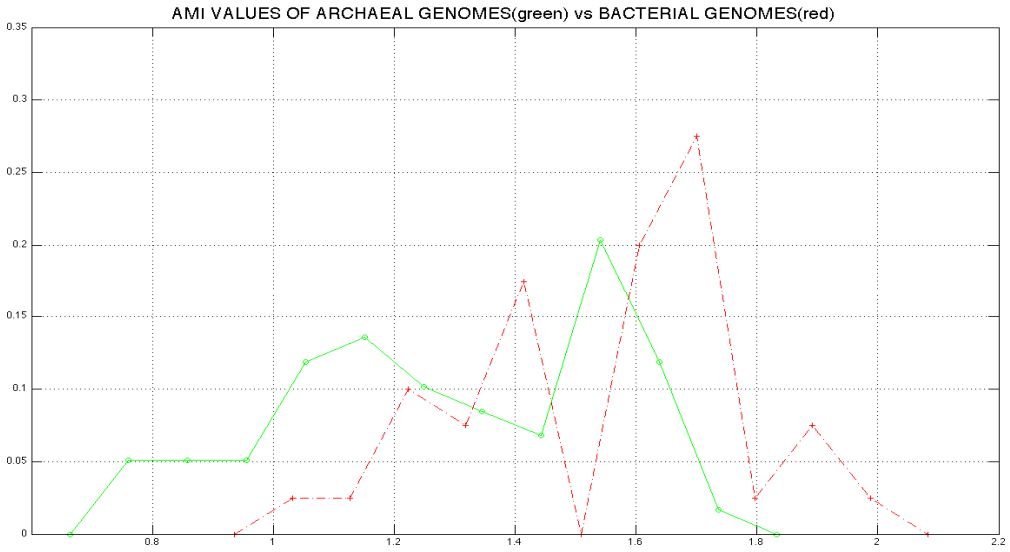


Figura 7: Gráfica de Frecuencia Normalizada de valores AMIF de Genomas Celulares, nuestra modificación a la fórmula de AMI de Grosse, para implementación en genomas completos.

El siguiente paso es comparar las distribuciones de secuencias virales y celulares de los tres Dominios de la vida. La primera comparación será Bacterias contra Bacteriófagos. La distribución de los Bacteriófagos muestra bimodalidad (Figura 8). La moda de mayor información contiene una mezcla de genomas de Bacteriófagos de RNA y DNA, mientras que la moda de menor información corresponde a una distribución conformada únicamente por Bacteriófagos de genoma de DNA.

En un análisis más fino de la distribución de Bacteriófagos, una clara diferencia fue encontrada. La moda de mayor información está compuesta por los genomas de Bacteriófagos de cadena simple, mientras que la moda de menor información contiene los Bacteriófagos de genoma de doble cadena. Esta bimodalidad puede ser explicada por la biología intrínseca de la macromolécula informacional; el DNA tiende a organizarse en doble cadena, lo cual resulta en una macromolécula muy estable y eso a su vez lleva a aumentar la diversidad de los genomas virales de DNA [Shackelton *et al.*, 2004]. Por otra parte en el caso de la moda de cadenas sencillas, podemos encontrar la mayoría de los Bacteriófagos de genoma de RNA. Esto se puede interpretar como que una cadena sencilla de RNA permite una tasa mayor de evolución que a su vez resulta en una mayor tasa de adaptabilidad al ambiente, pero que limita la diversidad de estos virus de RNA [Forterre *et al.*, 2009]. El bajo número total de Bacteriófagos de genoma de RNA de cadena sencilla disponibles en el Genebank al momento del muestreo puede explicar la bimodalidad.

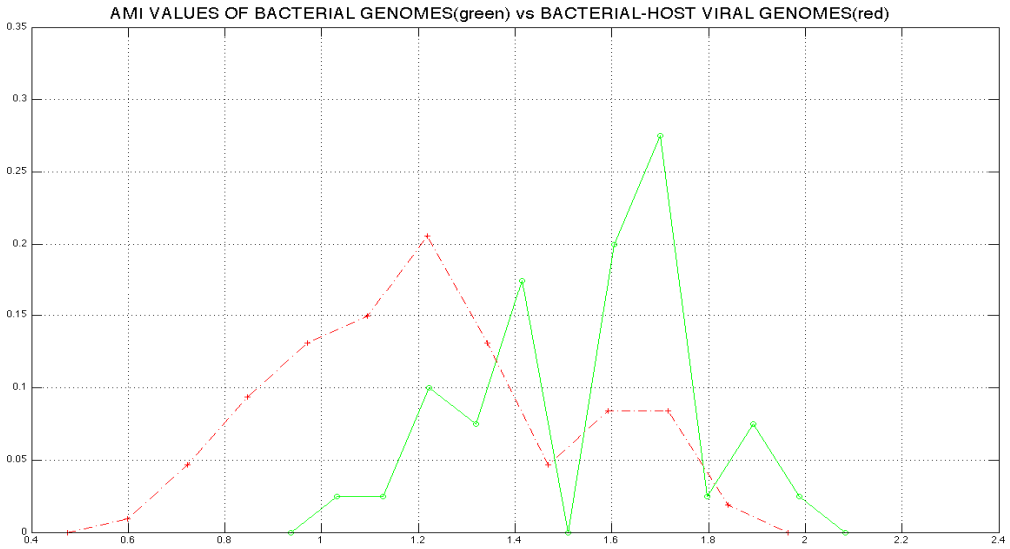


Figura 8: Distribuciones de valores AMIF de Bacteriófagos y Genomas Completos de Bacterias. <u>Comparación General entre Bacteriófagos y algunas Bacterias respectivas</u>.

La Figura 9 ilustra las distribuciones de valores AMIF de genomas de Archaeas respecto a los genomas de Archaeafagos. Estas distribuciones se encuentran casi traslapadasy la moda de mayor información de las Archaeas está cerca del promedio de los Archaeafagos. Ambas distribuciones muestran mayores niveles de Información respecto a la Figura 8.

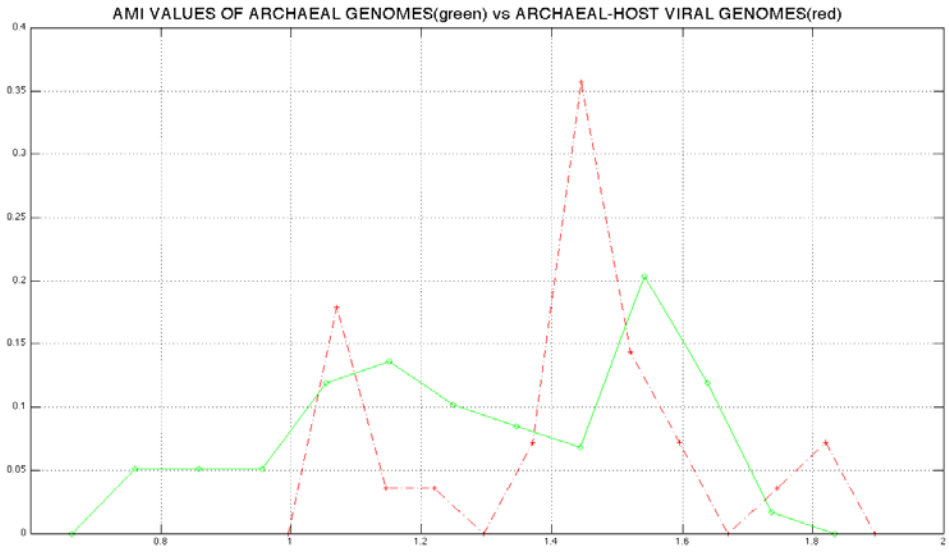


Figura 9: Distribuciones de valores AMIF de Archaeafagos y Genomas Completos de Archaeas.

Las distribuciones AMIF de las 213 secuencias virales de DNA y las 480 secuencias virales de RNA analizados en este trabajo, se muestran en la Figura 10. Como en el caso de las distribuciones de los Bacteriófagos, las secuencias de RNA (rojas) muestran mayoresvalores de información comparadas con las secuencias de DNA (verdes), y ambos tipos virales muestran mayores niveles de información respecto a los genomas celulares de Archaeas analizados. Es importante notar una característica al parecer intrínseca de los virus: muchos de los genomas de RNA virales son de cadena sencilla, mientras que la mayoría de los genomas virales de DNA son de doble cadena.

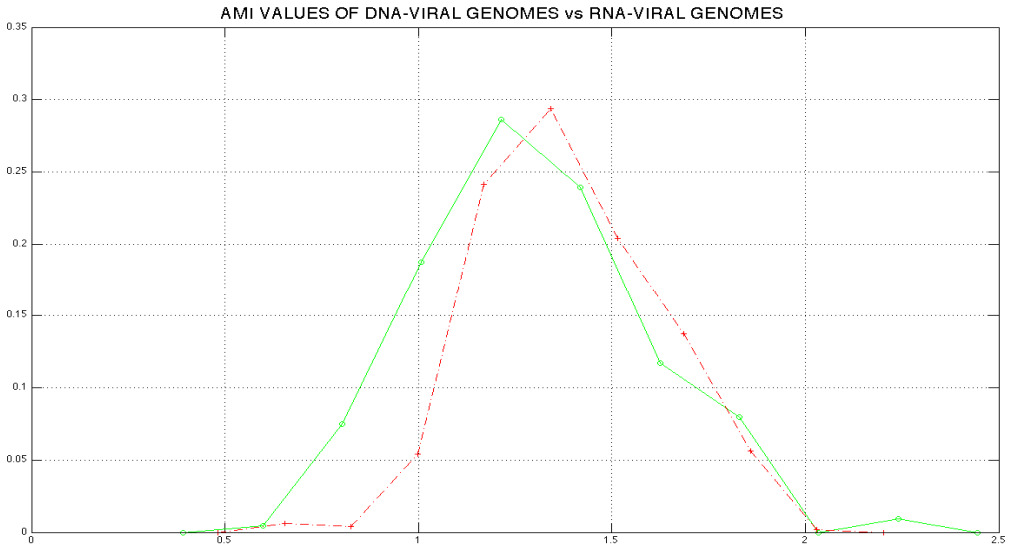


Figura 10: Distribuciones de valores AMIF de virus de Genoma de DNA y virus de Genoma de RNA. La vasta mayoría de los virus de DNA son de doble cadena, en contraste al grupo de los geomas de RNA que presentan cadena sencilla.

A lo largo de este trabajo, numerosas gráficas han sido mostradas con un solo objetivo: ilustrar las diferencias entre grupos de genomas virales utilizando los perfiles MIF y las PDFs de AMIF. Ahora bien, cuando todas las secuencias virales son agrupadas en arreglos de acuerdo al dominio de la vida del respectivo hospedero (ya sea Archaea, Bacteria o Eukarya), su distribución de PDF establece diferencias en los niveles de información, siendo la distribución de los Eucariavirus la de mayor Información, seguida por los Bacteriófagos y en último término la distribución de los Archaeafagos con casi 2 órdenes de magnitud menos en sus valores de Información (Figura 11).

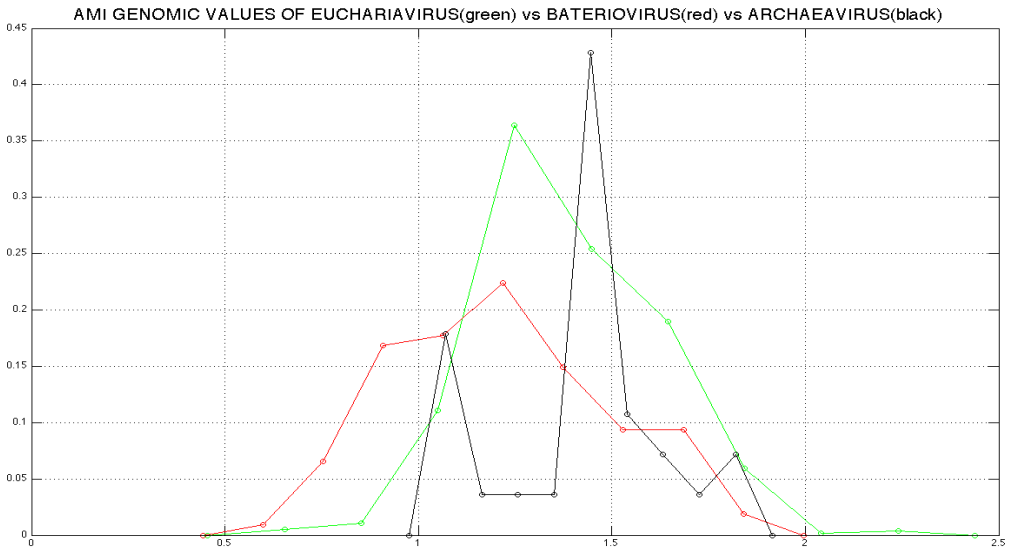


Figura 11: Distribución de valores AMIF en Eucariofagos, Bacteriofagos y Archaeafagos. Esta Figura es el resultado del cálculo de toda la base de datos de genomas virales. Los tres Dominios (Eukarya, Archaea, Bacteria) de la virósfera son mostrados.

La mayor cantidad de Información de los Eucariavirus puede ser explicada por la mayor complejidad de los Eucariontes, ej: metazoarios ya sean plantas, animales u hongos. Otro argumento a tomar en mente, es la descendencia viral masiva que rinde cada célula infectada del hospedero [Coffin *et al.*, 2013] y que por tanto aumenta la capacidad máxima de información mutua de la molécula de información.

Los mayores valores de AMIF fueron encontrados particularmente en dos virus de genoma de RNA:

a) el Rotavirus C que es segmentado y de doble cadena de RNA, específicamente en los segmentos 10 y 11,

b) el virus de cadena sencilla de RNA, influenza A(H5N1) en sus segmentos 7 y 8.

Los segmentos 10 y 11 de Rotavirus codifican para proteínas no estructurales; el segmento 8 del genoma segmentado de influenza A también codifica para una proteína no estructural.

En contraste, los valores más bajos de Información fueron los del Enterofago T5 y el Fago de Bacillus SPO1, ambos virus de cadena doble de DNA. También, el virus Vaccinia, de la familia Poxviridae, es de los de menor valor de AMIF. Los virus Vaccinia presentan un genoma lineal de DNA con longitudes que van de los 130 a los 338 kb. Este genoma viral gigante codifica para cerca de 300 proteínas. Este virus infecta desde artrópodos hasta cordados. Es interesante que estos valores menores de información lleguen a traslapar al promedio de los valores AMIF de las Bacterias correspondiendo también con la moda de menor Información de los genomas de Archaeas. No genera sorpresa que los análisis MIF del virus Vaccinia muestren entonces la periodicidad de tres bases (Figura 4).

5.3 INFRAME Y OUTFRAME

Como se adelanta en la sección de Métodos, otro algoritmo desarrollado en este trabajo, fue la intersección resultante de los valores de Inframe contra los valores Outframe.

En la Figura 12, un gráfico de dispersión log-log es mostrado. Ambos ejes X y Y representan valores de Información medidos en bits. Se presentan los intersectos (ilustrados como puntos) del conjunto de virus de genoma de DNA y del grupo de virus de genoma de RNA. Esto nos da una idea de las dinámicas a corta distancia de los genomas medidos con la MIF. En general, estos resultados sugieren que existe un rango constante de información en el cual habitan los genomas virales.

Dos comportamientos se pueden observar:

a) Hacia los valores más altos de información, existe una relación directamente proporcional entre el incremento de los valores de Inframe y Outframe; en otras palabras, si el pico de mayor información del análisis MIF (Inframe) aumenta, los otros dos puntos de la periodicidad de tres bases (Outframe) serán proporcionalmente mayores en sus valores de MIF. Esto nos permite predecir que al expandir este análisis a distancias mayores de k=3, un comportamiento continuo similar al de la periodicidad de tres bases característico de genomas celulares estará presente en los virus de RNA, probablemente descubriendo una oscilación de baja frecuencia característica en este tipo de virus.

b) Hacia menores valores de Información, los intersectos no son directamente proporcionales.

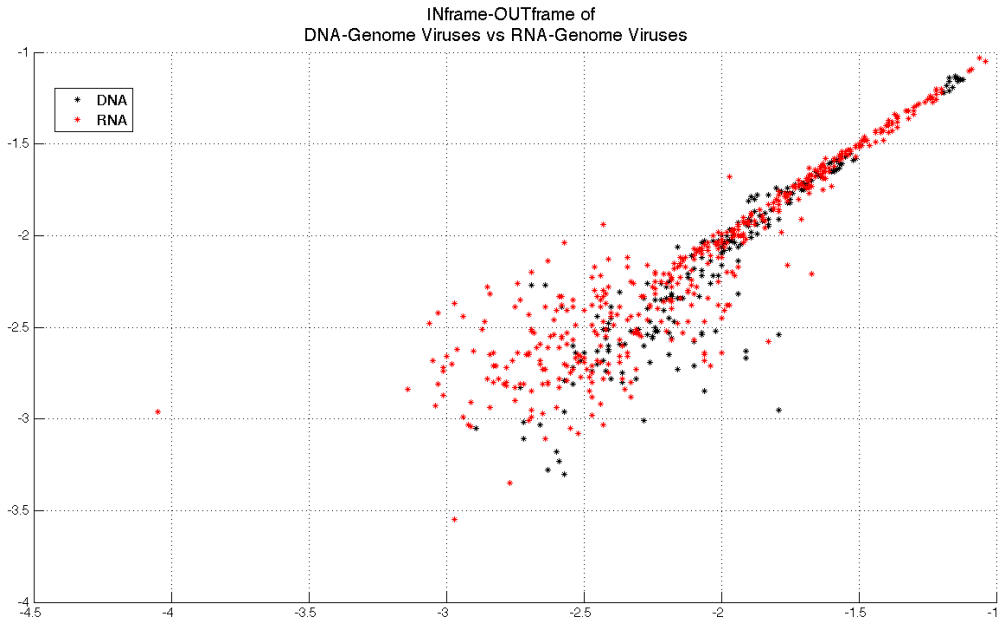


Figura 12: Intersectos de los valores Inframe y Outframe de secuencias virales de DNA y RNA. Gráfica de dispersión de los valores Inframe (valor genómico de la tercera posición en el cálculo MIF), contra Outframe (valor genómico del promedio de la 4ª y 5ª posiciones en el cálculo MIF) de la base de datos completa. Para efectos ilustrativos, dos colores son utilizados, negro para genomas virales de DNA y rojo para genomas virales de RNA.

Reorganizando la base de datos para separar conforme a los tres Dominios de la vida, otro punto de vista puede ser explorado (Figura 13). La mayoría de los virus de Eucariontes presentan una relación proporcional entre sus valores de Inframe y Outframe, mas existe otro grupo menor con un patrón diferente en su relación Inframe-Outframe donde al decaer los valores de información, la intersección entre Inframe y Outframe se convierte en un grupo disperso.

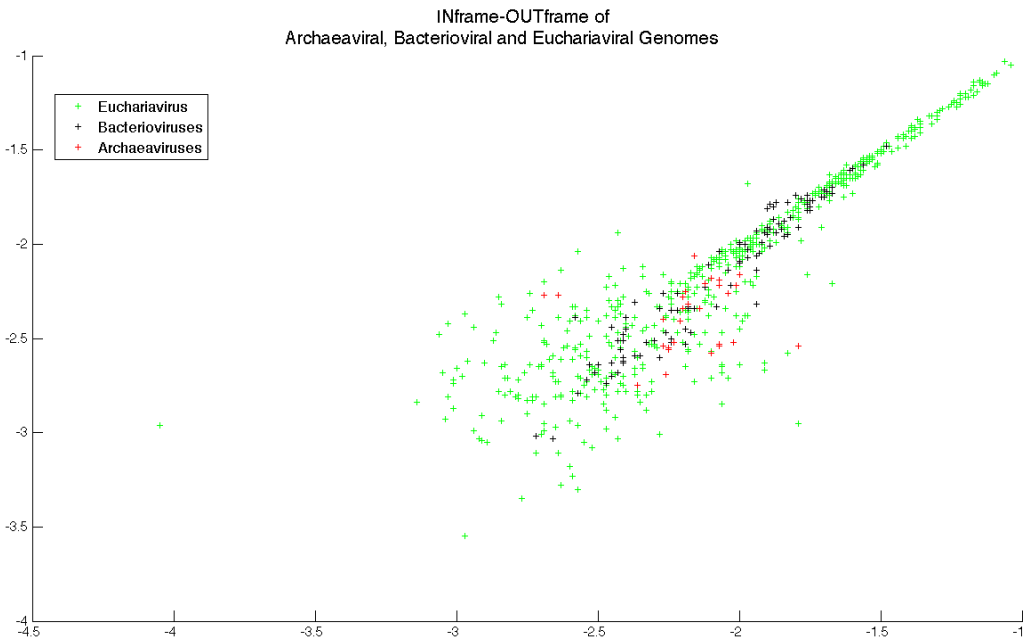


Figura 13: Intersectos de los valores Inframe y Outframe de las secuencias virales infectivas a cada dominio de la vida. Como en la Figura 10, la base de datos completa es utilizada para apreciar la virósfera infectiva a los tres Dominios de la vida (Eukarya, Archaea y Bacteria).

El conjunto de datos de Archaeavirus se comporta de manera similar al subgrupo de baja información Eucariavirus en el sentido de que cohabita en la zona de dispersión. En contraste, los Bacteriófagos son los virus que mejor se acomodan en la línea de 45° debido a la constante relación lineal entre los valores de Inframe y Outframe.

Ya por último pero no menos importante, se reajustó el conjunto de datos hacia la conformación de "categorías complejas" tales como:

a) secuencias virales de genoma de DNA de hospedero animal,

b) secuencias virales de genoma de RNA de hospedero animal,

c) secuencias virales de DNA de hospedero planta,

d) secuencias virales de RNA de hospedero planta,

e) los 40 genomas completos bacterianos,

f) los 59 genomas completos de Archaeas y

g) todo el conjunto de datos de las respectivas secuencias aleatorizadas.

Luego entonces se graficaron estas nuevas categorías lo cual resultó en un poco más de 1584 valores independientes representados como puntos correspondientes al espectro completo de análisis de Inframe y Outframe realizados en este trabajo (Figura 14). Los controles negativos inmediatamente confirman la distancia entre una secuencia biológica original y su versión aleatorizada tal como los análisis MIF y AMIF demostraron anteriormente. Es posible apreciar que las secuencias biológicas celulares son las más sensibles a la pérdida de información por el proceso de aleatorización. Es interesante que posterior a la aleatorización, las secuencias que pierden más información sean aquellas que presentaron el orden periódico y constante de la periodicidad de tres bases.

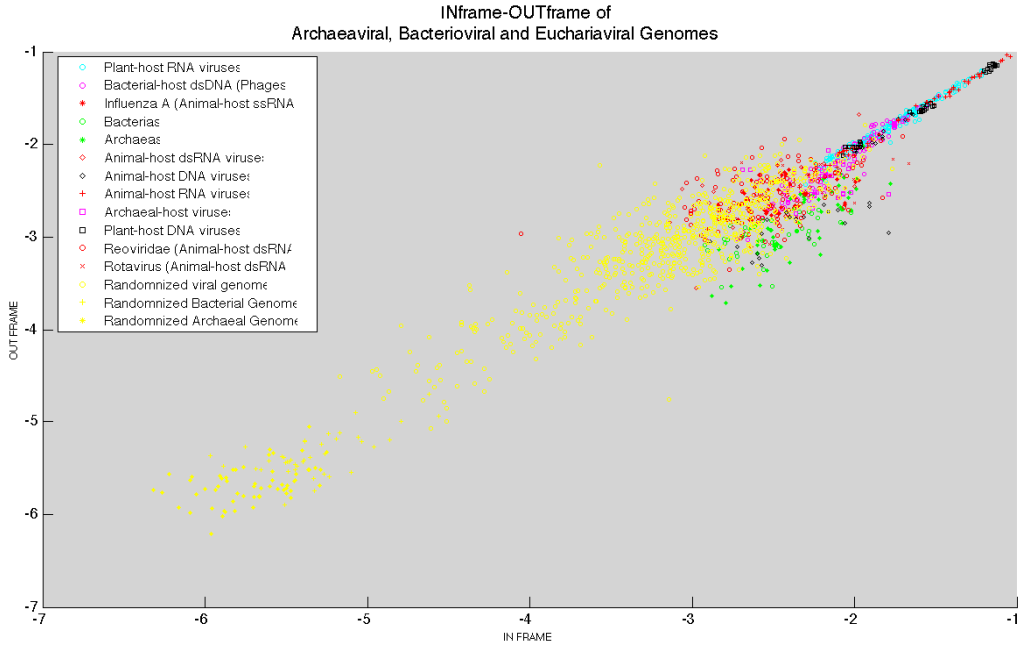


Figura 14: Intersectos de valores de Inframe y Outframe de 1584 secuencias analizadas. Los genomas de Bacterias y Archaeas son incluidos en la gráfica de dispersión. Los genomas celulares interesantemente se sitúan por debajo de la línea de 45° formada por los genomas virales. Igualmente se aprecian los genomas virales aleatorizados y los genomas celulares aleatorizados, los cuales en el proceso de aleatorización perdieron los mayores niveles de Información en comparación con los genomas virales aleatorizados.

Existe otra particularidad acerca de los genomas celulares: presentan en su totalidad mayores valores de Inframe que casi cualquier secuencia analizada; esto puede ser explicado por la presión selectiva ejercida sobre el mantenimiento invariante del código genético en la primera posición de codón. Muchas secuencias virales están localizadas a mayores valores de MIF en comparación con los genomas celulares completos.

Una gráfica más prístina presentando solamente las secuencias biológicas originales analizadas en este grupo resulta en la Figura 15. Ésta aborda diferencias entre las categorías mixtas, sobre todo notables entre los virus de genoma de DNA de hospedero plantas, que se agrupan en tres pequeñas regiones a lo largo de la línea de 45° mientras que las secuencias restantes se depositan como un continuo.

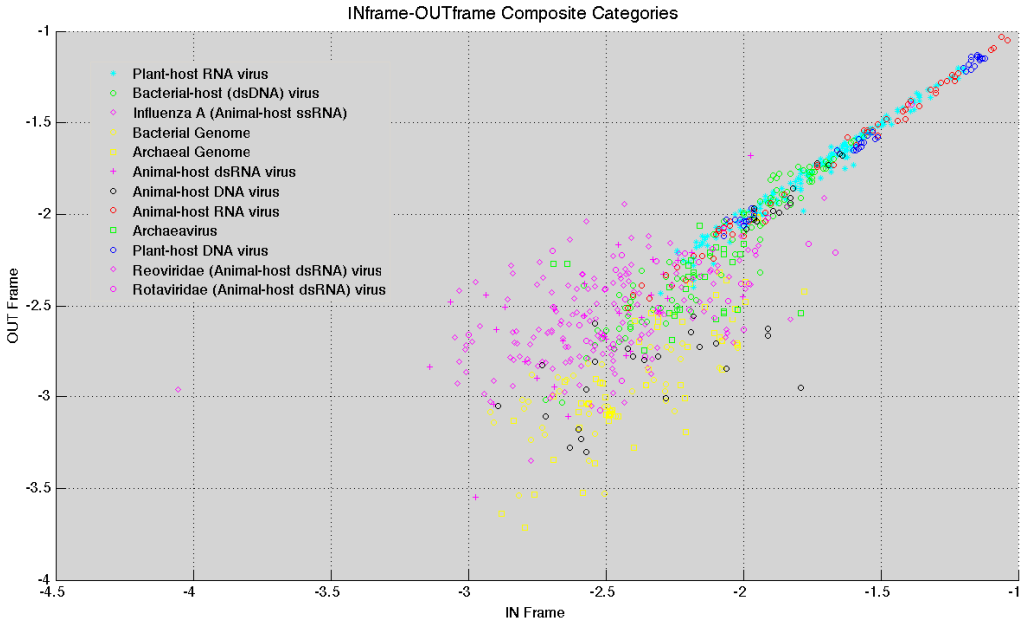


Figura 15: Intersectos de valores de Inframe y Outframe de las categorías compuestas de secuencias virales.

6. CONCLUSIONES PRELIMINARES.

En este trabajo hemos guiado en el uso de la Información Mutua en series simbólicas denotadas por genomas virales, bacterianos y de archaeas, para encontrar y comparar patrones ocultos de información. El uso de la teoría de la información en biología no es algo nuevo. Durante el texto se han abordado diferentes trabajos seminales en los cuales se aplicaban pruebas estándar de la teoría de la información para revelar los patrones de información de secuencias de DNA codificante y no codificante. Es importante contextualizar que dichos trabajos fueron publicados en los días en que el monumental proyecto de secuenciación del Genoma Humano, y el gran esfuerzo de Identificación de Elementos Funcionales del DNA Humano ENCODE por sus siglas en inglés (ENCODE, 2003; Guigó*et al.*, 2006; Raney*et al.*, 2011), aun no era lo ampliamente conocido y extendido como lo es ahora, de tal manera que a lo que los autores llamaban DNA codificante, se refería a segmentos de DNA presentes en las bases de datos, que tenían importancia biológica por codificar para una proteína conocida; en contraste, llamaban DNA no codificante a los segmentos que no codificaban para una proteína conocida.

En este trabajo hemos demostrado que existen diferencias significativas, cualitativa y cuantitativamente, entre los perfiles del contenido de información de las secuencias de RNA-virus y todas las secuencias de DNA, no importando su origen filogenético.

De acuerdo a nuestra prueba de AMIF, obtenemos que los valores de Inframe y Outframe van correlacionados a lo largo de la distancia, en este caso en autocorrelaciones a corta distancia (k=3, k=4, k=5) de tal manera que coincidimos y confirmamos con los resultados similaresde [Grosse *et al.*, 2000b] en el sentidoque los niveles de Inframe y Outframe van ligados en correlaciones de largas distancias, lo cual nos permite predecir que existirán oscilaciones de baja frecuencia las cuales pueden ser fundamentales para una clasificación sistemática más completa y amplia de los virus analizados en este esfuerzo.

Una particularidad revelada por los genomas virales segmentados, es que cada segmento del genoma viralpresenta un perfil único de información, lo cual puede ser explicado por un origen polifilético [Shackelton *et al.,* 2004]. Este resultado contrasta con los resultados de pervasividad reportados en [Bauer *et al.*, 2008]. Para encontrar una pervasividad representativa de cada virus segmentado, probablemente habría que generar un promedio de firmas genómicas o alguna secuencia compartida por todos los segmentos del genoma viral. En paralelo, utilizando una aproximaciónbioinformática amplia de escrutinio respecto a las bases de datos disponibles, podría arrojar resultados de sus relaciones evolutivas y por tanto de sus orígenes.

Este es el primer esfuerzo de clasificación viral mediante MIF tanto en PDFs como en dos dimensional.

En concordancia con [Grosse *et al.*, 2000a; Herzel *et al.*, 1994; Swati, 2007] se mostró que las secuencias de DNA codificante para proteínas llevan un claro sesgo hacia la periodicidad de tres bases; el DNA no codificante para proteínas y ahora también los genomas de RNA, no presentan ese sesgo y puede ser por tener otras estructuras adicionales y diferentes o carecer de ellos. Será necesario profundizar a este respecto para hallar diversos patrones que nos revelen estructuras genómicas aun no reportadas en la literatura, que inevitablemente tendrán repercusiones evolutivas como la que se demuestra con las repeticiones Alu de [Holste *et al.*, 2003] y los agrupamientos de subtipos de VIH de [Bauer *et al.*, 2008].

Desarrollamos un aporte metodológico para diferenciar el origen evolutivo de secuencias, sin importar su longitud, excluyendo la necesidad de cualquier alineamiento bioinformático previo. Mediante el algoritmo de Inframe-Outframe, podemos identificar en un espacio dos dimensional el tipo de secuencia analizado, ya sea viral, bacteriano o de Archaea. Adicionalmente, a partir de este análisis se puede distinguir la zona de la vida celular, y la zona de la vida no celular que abarca a la virosfera.

La posición de los genomas en el espacio dos dimensional puede ser explicado por los eventos mutacionales en regiones no conservadas que pueden dar origen a correlaciones de larga distancia [Berryman *et al.,* 2004]aumentando sus valores de MIF, dada la potencialidad de mutación de la secuencia [Forterre *et al.*, 2009].

Reconocimientos.

Agradecemos a TzipeGovezensky por el apoyo en las primeras etapas de este trabajo, y a Juan R. Bobadilla por el apoyo computacional.

REFERENCIAS

- Bauer M, Schuster SM, Sayood K.: The average mutual information profile as a genomic signature. BMC Bioinformatics. **2008** Jan 25;9:48.

- Berryman M.J., Allison A., Abbott D.: Mutual information for examining correlations in DNA. Fluctuation and Noise Letters **2004**, 4(2):L237-246.

- Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. Genomic analysis of uncultured marine viral communities. ProcNatlAcadSci U S A. **2002** Oct 29;99(22):14250-5.

- Cairns J, Stent G.S. and Watson J.: (**1966**) Phage and the Origins of Molecular Biology. Plainview, NY: Cold Spring Harbor Laboratory Press.

- Coffin J, Swanstrom R. HIV Pathogenesis: Dynamics and Genetics of Viral Populations and Infected Cells. Cold Spring HarbPerspect Med. January **2013**;3:a012526doi:10.1101/cshperspect.a012526.

- Dark MJ. Whole-genome sequencing in bacteriology: state of the art. InfectDrug Resist. **2013** Oct 8;6:115-123. Review.

- Flint JS.: Adenoviruses. Encyclopedia of Life Sciences. 2002.

- Forterre P. The origin of viruses and their possible roles in major evolutionary transitions. Virus Res. **2006** Apr;117(1):5-16.

- Forterre P, Prangishvili D. The origin of viruses. Res Microbiol. **2009** Sep;160(7):466-72.

- Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, Castelo R, Eyras E, Ucla C, Gingeras TR, Harrow J, Hubbard T, Lewis SE, Reese MG (**2006**). "EGASP: The human ENCODE Genome Annotation Assessment Project". Genome Biology 7: S2.

- Grosse I., Herzel H., Buldyrev S.V., Stanley H.E.: Species independence of mutual information in coding and noncoding DNA. Phys Rev E Stat Phys Plasmas Fluids RelatInterdiscip Topics. **2000a** May; 61:5624-9.

- Grosse I., Buldyrev S.V., Stanley H.E., Holste D., Herzel H.: Average mutual information of coding and noncoding DNA. Pac SympBiocomput. **2000b**:614-23.

- Herzel H, Ebeling W, Schmitt AO. Entropies of biosequences: The role of repeats. Phys Rev E Stat Phys Plasmas Fluids RelatInterdiscip Topics. **1994** Dec;50(6):5061-5071.

- Holste D, Grosse I, Beirer S, Schieg P, Herzel H. Repeats and correlations in human DNA sequences. Phys Rev E Stat Nonlin Soft Matter Phys. **2003** Jun;67(6 Pt 1):061913.

- Hunter S, Corbett M, Denise H, Fraser M, Gonzalez-Beltran A, Hunter C, Jones P, Leinonen R, McAnulla C, Maguire E, Maslen J, Mitchell A, Nuka G, Oisel A, Pesseat S, Radhakrishnan R, Rocca-Serra P, Scheremetjew M, Sterk P, Vaughan D, Cochrane G, Field D, Sansone SA. EBI metagenomics--a new resource for theanalysis and archiving of metagenomic data. Nucleic Acids Res. **2013** Oct 27.

- Li W. The study of correlation structures of DNA sequences: a critical review. Comput Chem. **1997**;21(4):257-71. Review.

- Norrby, E. Nobel Prizes and the emerging virus concept. Arch. Virol .2008, 153, 1109-1123. Res. **2006**, 117, 5-16.

- Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, Meyer LR, Sloan CA, Malladi VS, Roskin KM, Suh BB, Hinrichs AS, Clawson H, Zweig AS, Kirkup V, Fujita PA, Rhead B, Smith KE, Pohl A, Kuhn RM, Karolchik D, Haussler D, Kent, WJ (**January 2011**). "ENCODE whole-genome data in the UCSC genome browser (**2011 update**)". Nucleic Acids Res. 39.

- Román-Roldán R., Bernaola-Galván P., Olivier J.L.: Application of Information Theory to DNA Sequence Análisis: A Review. Pattern Recognition **1996**, 29(7):1187-1194.

- Shackelton LA, Holmes EC. The evolution of large DNA viruses: combining genomic information of viruses and their hosts. Trends Microbiol. **2004**. Oct;12(10):458-65. Review.

- Shannon, C.E.A. Mathematical theory of communication. Bell Syst. Tech. J. **1948**, 27, 379–423, 623–656.

- Swati D. In silico comparison of bacterial strains using mutual information. JBiosci. **2007**. Sep;32(6):1169-84.

- Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO. Environmental genome shotgun sequencing of the Sargasso Sea. Science. **2004** Apr 2;304(5667):66-74.

Entropy **2013**, *15*, 3065-3087; doi:10.3390/e15083065

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Flow of Information during an Evolutionary Process: The Case of Influenza A Viruses

Víctor Serrano-Solís and Marco V. José *

Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Apdo. Postal 70228, México D.F. 04510, Mexico;
E-Mail: vserrano@biomedicas.unam.mx

***** Author to whom correspondence should be addressed; E-Mail: marcojose@biomedicas.unam.mx; Tel.: +52-55-5622-3894; Fax: +52-55-5622-3894.

Received: 14 March 2013; in revised form: 18 July 2013 / Accepted: 23 July 2013 / Published: 29 July 2013

Abstract: The hypothesis that Mutual Information (MI) dendrograms of influenza A viruses reflect informational groups generated during viral evolutionary processes is put forward. Phylogenetic reconstructions are used for guidance and validation of MI dendrograms. It is found that MI profiles display an oscillatory behavior for each of the eight RNA segments of influenza A. It is shown that dendrograms of MI values of geographically and historically different segments coming from strains of RNA virus influenza A turned out to be unexpectedly similar to the clusters, but not with the topology of the phylogenetic trees. No matter how diverse the RNA sequences are, MI dendrograms crisply discern actual viral subtypes together with gain and/or losses of information that occur during viral evolution. The amount of information during a century of evolution of RNA segments of influenza A is measured in terms of bits of information for both human and avian strains. Overall the amount of information of segments of pandemic strains oscillates during viral evolution. To our knowledge this is the first description of clades of information of the viral subtypes and the estimation of the flow content of information, measured in bits, during an evolutionary process of a virus.

Keywords: mutual information function; mutual information dendrograms; viral evolution; flow of information; Influenza A; phylogenetic reconstruction

---

1. Introduction

Viruses have been considered as important players in the history of life on this planet [1]. Living beings cannot be conceived without viruses. They impinge on a wide variety of biological processes ranging from the abiotic and fundamental global carbon cycles [2] to such an exquisite process as placentation [3]. As a rule, they act as information transmitting vectors in processes such as horizontal gene transfer, or as infectious agents. Viral evolution has been studied thoroughly in many research fields [4]. Viruses are ubiquitous in the biosphere and they show an astonishing genetic diversity so that viral genomes do not resemble to their hosts' genomes [5,6], and consequently they have not been considered, so far, as part of any tree of life. RNA viruses evolve extremely rapidly, often with mutation rates one million times greater than those of vertebrate species [7]. This rate of mutation allows viral populations to rapidly cope with selective pressures. Viral pathogens, such as influenza virus, HIV, hepatitis C virus, and dengue virus, place a substantial burden on global human health. Despite the fact RNA viruses display high mutation and recombination rates, they remain as discrete recognizable evolutionary units. In fact, the error rates of RNA viruses usually approach the Eigen's error threshold [8,9]. Current methodologies for determining the type of evolutionary paths of a taxon consist in the comparison of sequences in order to infer kinship relationships as displayed by a phylogenetic reconstruction [10]. Those *in silico* methods have been shown to be useful both for DNA and RNA viruses [11]. Viruses are suitable models for evolutionary studies, e.g., the HIV-1 virus [12], influenza A virus [13] and for studying even the origin of genetic information [8]. Influenza A virus is the result of several evolutionary processes such as genetic drift, random mutations, purifying and positive selection, plus genetic recombination and the generation of genomic reassortants [14–16]. There are two major evolutionary approaches for describing viral evolution: the quasispecies theory [17–19] and standard population genetics theory [20,21]. However, none of these theories contemplate the flow of information through the viral evolution.

The extensive flu databases offer a great opportunity to uncover and to examine relevant mutational information associated to influenza A pandemics for almost a century of registries. Based on the antigenic specificities of 16 hemagglutinin (HA) (H1-H16) and nine neuraminidase (NA) (N1-N9) subtypes, there are 144 possible subtypes of influenza A viruses [22]. Unlike most pathogens, where persistent immunity arises after recovery of illness, influenza A virus presents a moving antigenic target, evading any specific immunity triggered by previous infections. The virological basis for annual recurrent epidemics is a continual process of small changes in influenza surface antigens to escape host immunity. This process, called antigenic drift, is the result of the selective fixation of mutations in the gene encoding the HA protein, the major target for the host immune response [23]. HA variants that best escape the host immune response are thought to have a significant reproductive advantage [23]. Influenza A viruses host range comprise avian as well as mammal species, and their genomic segments are subject to occasional reassortment (shift) giving rise to new viral strains consisting of novel HA and/or NA genes [24]. Although less common than antigenic drift, antigenic shift is considered another major force in the evolution of influenza viruses [23–25]. By definition antigenic shift occurs when the virus acquires an HA and/or NA of a different influenza subtype via reassortment of one or more gene segments and is thought to be the basis for the more devastating influenza pandemics that occurred several times in the last century [26]. Inter-pandemic evolution of influenza A virus involves a

complex interplay between neutral evolution during periods of antigenic stasis, positive selection during relatively short intervals of rapid change in fitness, and multiple effects of reassortment [15]. A living entity can be described as a complex adaptive system which differs from any, however complex, chemical structure by its capability of functional self-organization based on the processing of information [8]. Evolutionary processes not only entail loss and gain of relevant information for survival, but its regulation is certainly critical. The mutual information (MI) measures the differences between the average uncertainty in the input of an information channel before and after the output are received [27].

In this work, we propose an approach that considers the MI from Shannon's theory of information perspective [27] in order to measure the amount of flow of information throughout ~100 years of evolution of influenza A virus. The use of the MI has already been applied to study the genetic drift of influenza A/H3N2 virus [28]. The MI concept is useful when analyzing symbolic RNA or DNA sequences. The main goal of this work was to determine if the MI could capture relevant viral evolutionary information associated to antigenic shifts. To this end, we hypothesized that similarity hierarchizations (dendrograms) of MI as derived from actual RNA sequences of influenza A subtypes could display biologically informational coherent groups. Therefore, these dendrograms should also be, in principle, in agreement with the hierarchical clusterization and topology of phylogenetic trees as obtained from *in silico* methodologies of molecular evolution bioinformatics. The article is organized as follows. First, we calculated the MI profiles for each of the eight influenza A genomic segments from 39 antigenic shift subtypes (20 human and 19 avian) that have been clearly identified during the period of 1918 to 2012 for human subtypes and from 1902 to 2011 for avian subtypes. Secondly, dendrograms were constructed from the MI values for each of the eight influenza A segments. Next, phylogenetic reconstructions were obtained and, remarkably, they resemble the hierarchical clusters of the MI dendrograms. In particular, we illustrate results for the MI dendrograms of segments 1, 4, 6, and 8, here denoted by S1, S4, S6, and S8, which encode a highly conserved RNA polymerase (PB2), HA, NA, and a bi-cistronic sequence coding for both the only Non-Structural protein of influenza A (NS1) and the Nuclear Export Protein (NEP), respectively. Finally, a discussion of the present results in terms of viral evolutionary theory and practical implications is presented.

2. Viral Sequences

A set of 312 sequences comprising 39 antigenic shift influenza A whole genomes, (20 human and 19 avian), were selected from the Influenza Virus Resource [29]. Accession numbers for each sequence as well as relevant information are provided in Table 1. The strains were selected according to its relevance in antigenic shift history [30] associated to different pandemics. An important criterion for selection was that all of them were complete sequences. We remark a peculiarity in the analysis of the alignment of S4: for the subtype A/Brevig_Mission/1918 genome [31], the HA sequence (S4) at the 3' side, lacks a little more than ¼ of the average length. However, this is the only record of a subtype of 1918, which we could not afford to rule it out from our genomic study. All sequences were ordered by year of report, geographical location, and host species (Table 1). Note that we selected 1, 2, or more human shift subtypes per decade.

Table 1. List of influenza A strains associated to different pandemics.

Acc. Numb.	Year	City	Host	Strain
GU186783	1902	Brescia	Chicken	H7N7
CY014998	1927	Dobson	Fowl	H7N7
CY077417	1934	Rostock	Chicken	H7N1
CY014678	1949	Germany	Chicken	H10N7
CY015088	1959	Scotland	Chicken	H5N1
CY045334	1956	Czech_Republic	Duck	H4N6
CY014991	1961	South Africa	Tern	H5N3
CY015071	1963	England	Turkey	H7N3
DQ376870	1972	Taiwan	Duck	H6N1
CY024792	1976	Victoria	Chicken	H7N7
CY015043	1979	Leipzig	Goose	H7N7
CY015080	1983	Pennsylvania	Chicken	H5N2
CY117372	1988	Alberta	Mallard	H2N3
CY025084	1992	Victoria	Chicken	H7N3
CY005836	1994	Hidalgo	Chicken	H5N2
DQ997136	1997	Hubei	Chicken	H5N1
DQ997416	2002	Zhejiang	Duck	H5N1
CY103466	2008	Delaware Bay	Shorebird	H3N2
JX175257	2011	Guangdong	Duck	H3N2

Acc. Numb.	Year	City	Host	Strain
DQ208309	1918	Brevig_Mission	Human	H1N1
CY009611	1933	Wilson-Smith	Human	H1N1
CY020447	1936	Henry	Human	H1N1
CY013275	1940	Hickox	Human	H1N1
CY045779	1946	Melbourne	Human	H1N1
CY009347	1954	Malaysia	Human	H1N1
CY087804	1957	Japan	Human	H2N2
CY032268	1964	Cottbus	Human	H2N2
CY080530	1968	Hong_Kong	Human	H3N2
CY021964	1976	New_Jersey	Human	H1N1
DQ508894	1977	USSR	Human	H1N1
CY021044	1982	Christs_Hospital_UK	Human	H1N1
CY113369	1987	Shanghai	Human	H3N2
CY112844	1997	Hong_Kong	Human	H3N2
AJ278649	1999	Hong_Kong	Human	H9N2
CY006674	2003	New_York	Human	H1N1
CY118918	2006	Malaysia	Human	H3N2
GQ132145	2009	Mexico_InDRE4114	Human	H1N1
CY049984	2009	Mexico_City_WR1087T	Human	H1N1
JX046923	2012	Moscow	Human	H1N1

3. Methods

3.1. An Overview of the Mutual Information Function

Our goal is to quantify the degree of covariation of mutations of the 8 RNA segments of influenza A by using mutual information, a concept from information theory [32,33]. The identification of covarying sites (particular pairings with high mutual information values) is likely to confer a selective advantage in terms of either structure or function that facilitates the propagation of the virus. A formal measure of variability [34] at position i is the Shannon entropy, $H(i)$. $H(i)$ is defined in terms of the probabilities, $P(sym_i)$, of the different symbols, sym, that can appear at sequence position i (e.g., $sym = \mathrm{U, A, G, C, ..}$ for the four nucleotide bases of RNA). $H(i)$ is defined as:

$$H(i) = -\sum_{sym=A,U,G,C,...} P(sym_i)\log_2 P(sym_i). \quad (1)$$

Mutual information (MI) is defined in terms of entropies involving the joint probability distribution, $P(sym_i, sym'_j)$, of occurrence of symbol sym at position i, and sym' at position j. The probability, $P(sym_i)$, of a symbol appearing at position i regardless of what symbol appears at position j is defined by $P(sym_i) = \sum_{sym'_j} P(sym_i, sym'_j)$ and similarly, $P(sym'_j) = \sum_{sym_i} P(sym_i, sym'_j)$. Given the above probability distributions, one can form the associated entropies:

$$H(i) = -\sum_{sym_i} P(sym_i)\log_2 P(sym_i),$$
$$H(j) = -\sum_{sym_j} P(sym'_j)\log_2 P(sym'_j),$$

and:

$$H(i, j) = -\sum_{sym'} \sum_{sym'_j} P(sym_i, sym'_j)\log_2 P(sym_i, sym^i_j).$$

The mutual information, $I(i, j)$, is defined as:

$$I(i, j) = H(i) + H(j) - H(i, j). \quad (2)$$

The MI profile is a general measure of correlation between discrete variables, analogous to the Pearson's product-moment correlation coefficient for continuous variables. For RNA (or DNA) symbolic sequences, a MI profile between two symbols separated by a distance k is a function of k, called the mutual information function (MIF) [34]. The MI is particularly useful for analyzing correlation properties of symbolic sequences [34]. Lets denote by $A = \{\mathrm{A, U, G, C}\}$ an alphabet and $s = (\ldots, a_0, a_1, \ldots)$ an infinite string with $a_i \in A$, $i \in \mathbb{Z}$, where $\mathbb{Z}$ represents the set of all integer numbers and the values of a_i can be repeated. The MI of the string s and an identical string shifted k positions upstream is defined as:

$$I(k,s) = \sum_{\alpha \in A}\sum_{\beta \in A} P_{\alpha,\beta}(k,s)\log_2\left[\frac{P_{\alpha,\beta}(k,s)}{P_\alpha(s)P_\beta(s)}\right] \quad (3)$$

where $P_{\alpha,\beta}(k,s)$ is the joint probability of having the symbol α followed k sites away by the symbol β on the string s, and $P_\alpha(s)$ and $P_\beta(s)$ are the marginal probabilities of finding α or β in the string s. By

choosing the logarithm in base 2, $I(k,s)$ is measured in bits. Both the joint probability and the marginal probabilities are estimated throughout the sequence as a global property. The function $I(k,s)$ can be interpreted as the average information over all positions that one can obtain about the actual value of a certain position in the string, given that one knows the actual value of the position $k-$characters away. The mutual information vanishes if, and only if, the events are statistical independent, *i.e.*, if all 16 joint probabilities $P_{\alpha,\beta}(k,s)$ factorize. Thus, the MI is a function capable of detecting any deviation from statistical independence. It must be noted from Equation (2) $I(i,j) = I(j,s),$ or from Equation (3) that $I(k,s) = I(s,k),$ and that $I(k,s) \geq 0.$ The computation of the MI for a given sequence using different shifts of magnitude k provides an autocorrelation profile. Mutual Information values were obtained over the entire sequences of each of the eight segments and the plots illustrate the first 100 positions with the software MATLAB R2010b.

3.2. MI Dendrograms

Using complete sequences of the influenza A virus genome, similarity pairwise tests were performed with the criterion of Euclidian distance. The Euclidean distance between points p and q is the length of the line segment $\overrightarrow{pq}$. In Cartesian coordinates, if $p=(p_1,p_2,...,p_n)$ and $q=(q_1,q_2,...,q_n)$ are two points in Euclidean $n-$space, then the distance from p to q is given by:

$$d = \|p-q\| = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \tag{4}$$

The Euclidean distance performs an unbiased pairwise distance between pairs of objects. In MATLAB pdist(X) computes by default the Euclidean distance between pairs of objects in the $m\times n$ data matrix X. Columns of X correspond to observations (here the MI distances calculated), and rows correspond to variables (the genomic sequences). Given a $m\times n$ data matrix X, which is treated as m $(1\times n)$ row vectors $x_1,x_2,...x_m,$ the Euclidean distance between the vector x_s and x_t is defined as: $d_{st}^2=(x_s-x_t)(x_s-x_t)^{transpose}$. The next step was the generation of the dendrogram, a hierarchization method obtained from the linkage of the pairwise distances. The distance matrix was linked with an Un-Weighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm [35], which in evolutionary theory assumes a constant substitution rate for all sites on the given sequence. The UPGMA algorithm is recurrently used in phenetics, but it does not integrate an actual evolutionary model.

3.3. Phylogenetic Reconstructions

Sequences were aligned using MUSCLE v3.7 [36] under default parameters, taking advantage of the refine option in a second iteration. Complete Multiple Sequence Alignments (MSA) were visually inspected with Seaview 4 [37]. The Generalized Time-Reversible (GTR) evolution model [38] was selected relying on jModelTest analyses [39]. The evolutionary history was inferred using the Maximum Likelihood (ML) method [40]. The ML tree inferred from FastTree [41] under default parameters, is taken to represent the evolutionary history of the taxa analyzed. The ML tree is drawn as a cladogram as depicted by FigTree [42] for the respective comparison to MI dendrograms. A cladogram represents only the tree topology and its branch lengths do not represent time or relative amount of character change used to infer the phylogenetic tree. Each analysis involved 20 nucleotide sequences of human Influenza A (Table 1). All positions containing gaps and missing data were

eliminated. The total positions for each final set were: for S1, 2343 positions; for S4, 1837 positions; for S6, 1499 positions; for S8, 892 positions. Original ML trees as well as the alignments can be found in Supplementary Information.

4. Results

4.1. MI Profiles

For illustrative purposes, here we show a set of MI profiles that were obtained by using complete sequences of segments S1, S4, and S6. We selected the recently reconstructed 1918 Spanish flu virus [43] and compared it with the MI profile of isolates of virus in Mexico from the 2009 influenza pandemics [44,45] (Figure 1). In each plot only the first 100 positions are displayed for the sake of clarity. Note that both profiles of S1 A (H1N1) from 1918 and 2009 subtypes, which contains the gene for PB2, show an irregular oscillatory behavior along their whole sequences (Figure 1A). It is noteworthy to mention that there are regions (from positions 1 to 70) in which the oscillations are practically similar; but there are other regions (positions 75–90) in which there are clearly departures from each other. In contrast, the MI profiles for S4 (Figure 1B), which contains the gene for HA, are in general different from each other, being the 1918 strain the one that displays higher peaks than the 2009 strain. In both segments the behavior of the MI shows an irregular oscillatory behavior although they are roughly parallel. The discrepancies between the peaks of each segment correspond to relevant changes in information that presumably correspond to mutations that have occurred close to a century of evolution of A (H1N1). When comparing the MI's of S6 (Figure 1C), which contains the gene for NA, it is notorious that the MI of the subtype 2009 is ~tenfold higher in extent of information than the MI of flu virus of 1918, and it shows more peaks at intermediate and long distances. Despite the difference in the level of information, they are essentially parallel. This finding was unexpected but it is coherent with both the dendrogram and the phylogenetic analyses (See Section 4.2), since they are not closely similar in terms of MI values, and they are not closely related along sampling times. A MI profile of zero would indicate that mutations would be independent events and this would correspond to the case of neutral mutations which can be used as a control for neutral or nearly neutral evolution.

4.2. MI Dendrograms and Phylogenetic Reconstructions

Using all the values of MI profiles, dendrograms for each of the eight segments of the 39 influenza A subtypes, were calculated. Here we illustrate the dendrograms for S1, S4, S6, and S8 together with their respective Maximum Likelihood Phylogeny visualized as cladograms (Figure 2). Original ML trees as well as their corresponding alignments are provided in Supplementary Information. Dendrograms are rotated 90 degrees counterclockwise so that the abscissa is at the bottom and the units are in bits. Hence the number of bits associated for each split can be observed (Figure 2). The MI dendrogram of S1, that code for PB2 (a basic RNA polymerase) is presented in Figure 2A1. PB2 is a protein of the replication-transcription complex of influenza virus. PB2 is considered to be an evolutionary conserved molecule, meaning that small amount of changes should be expected.

Figure 1. Mutual Information of A(H1N1) from 1918 (red curves) and from 2009 (black curves): (**A**) S1; (**B**) S4; (**C**) S6.

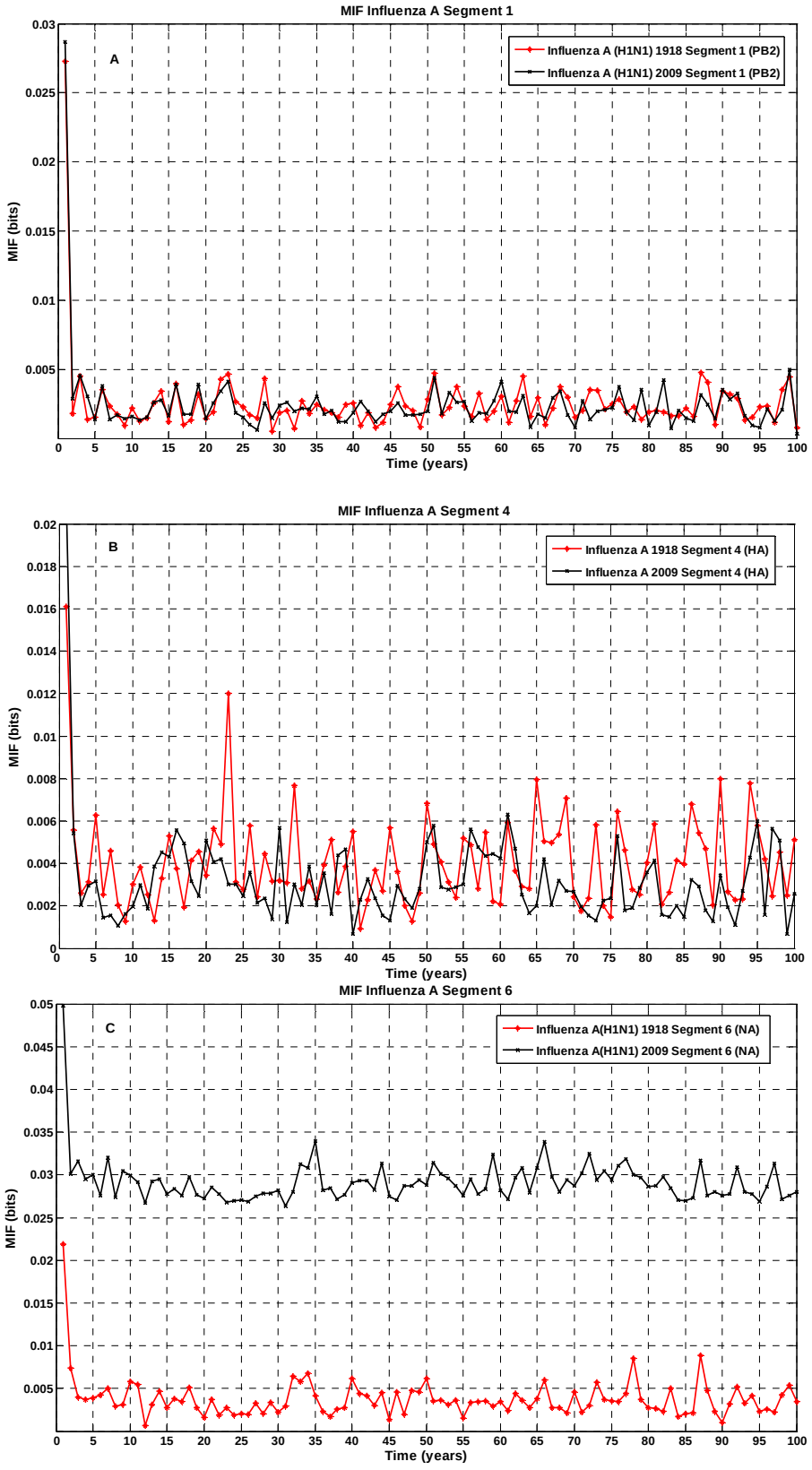
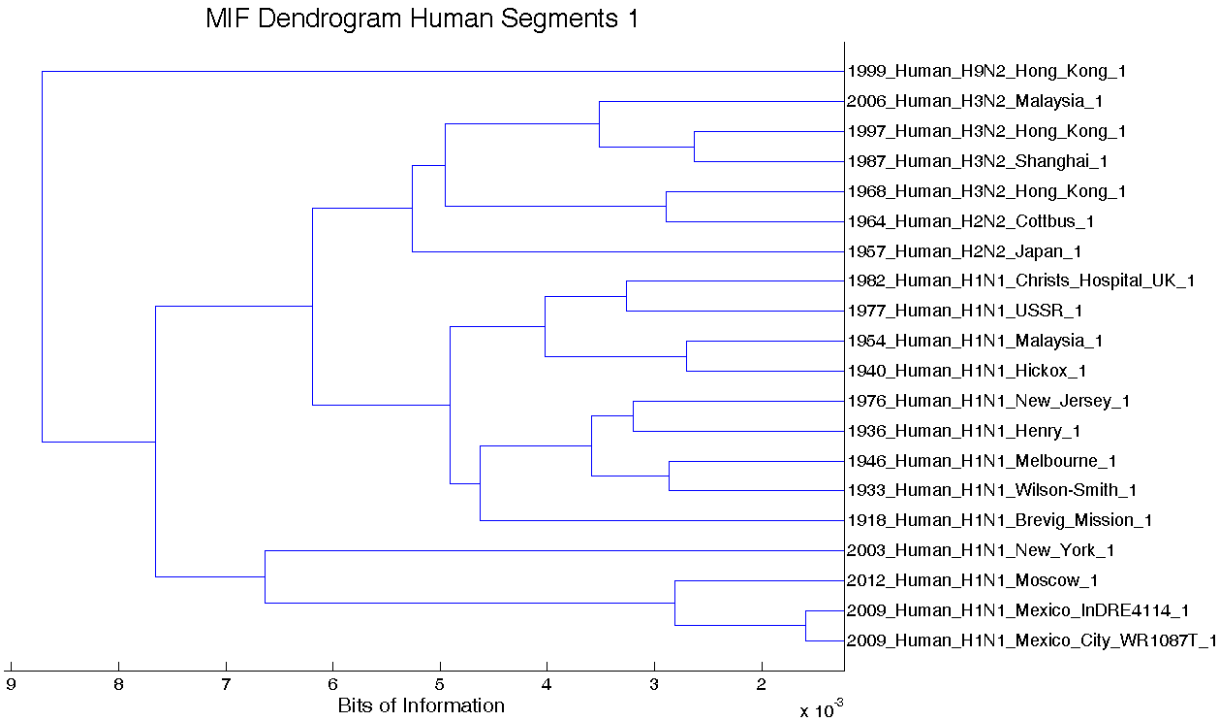
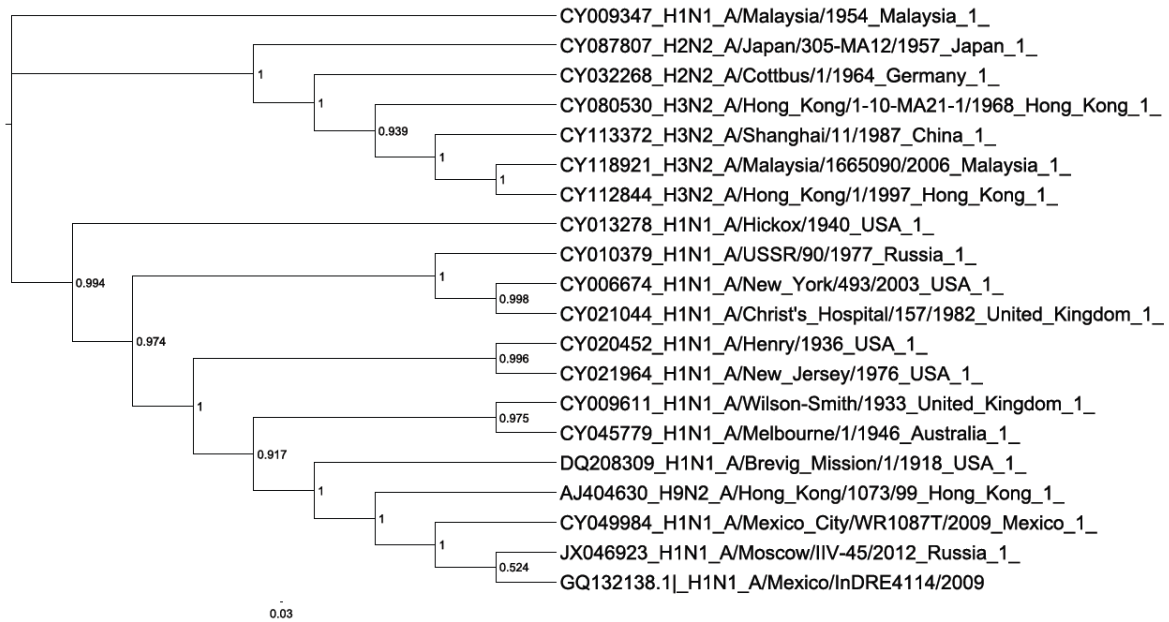


Figure 2. Dendrograms of the Mutual Information for: S1 (**A1**); S4 (**B1**); S6 (**C1**); and S8 (**D1**) segments of the RNA genome of influenza A. Maximum Likelihood Phylogenetic Reconstructions visualized as cladograms for: S1 (**A2**) LogLk (Log-likelihood) = −9150.268; S4 (**B2**) LogLk = −12865.908; S6 (**C2**) LogLk = −8565.225; and S8 (**D2**) LogLk = −3181.513. The numbers at each split in ML trees correspond to the Shimodaira-Hasegawa reliability estimate test which is part of the default parameter values of FastTree.

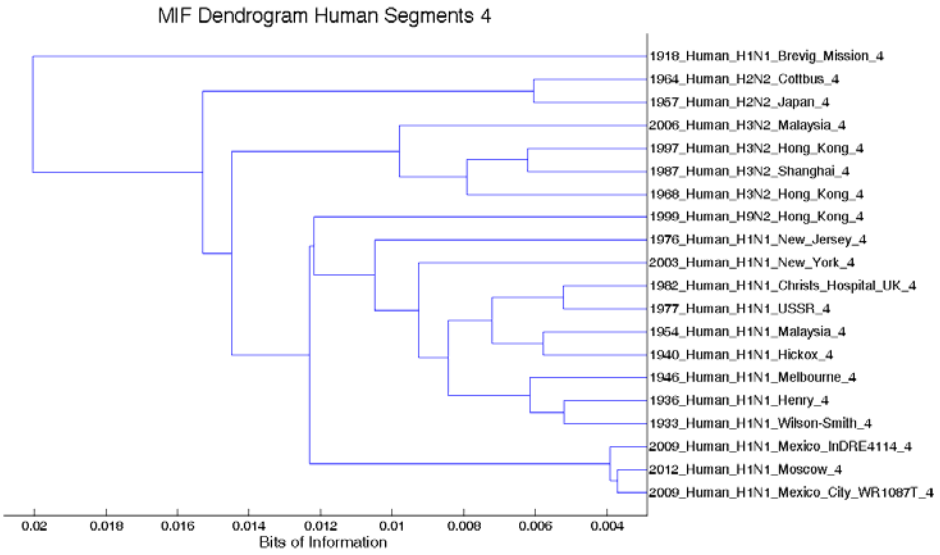


(**A1**)

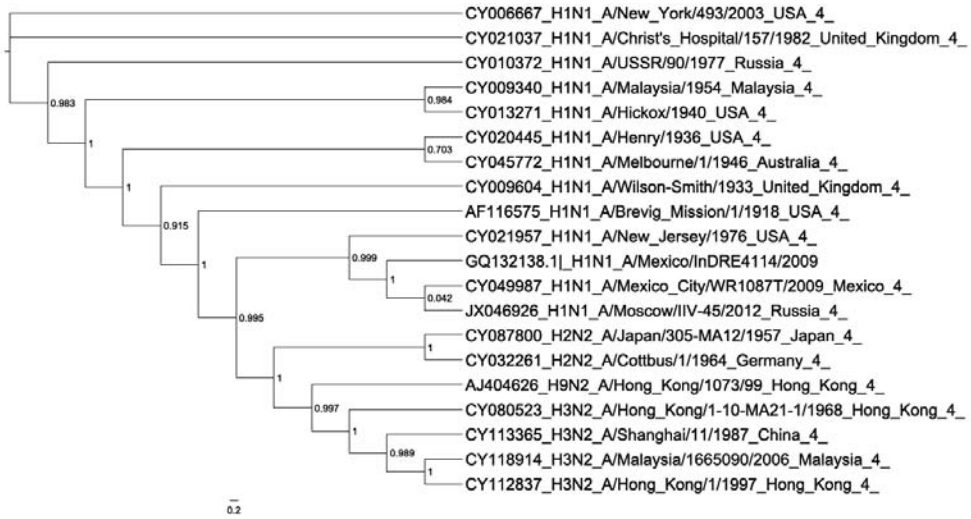


(**A2**)

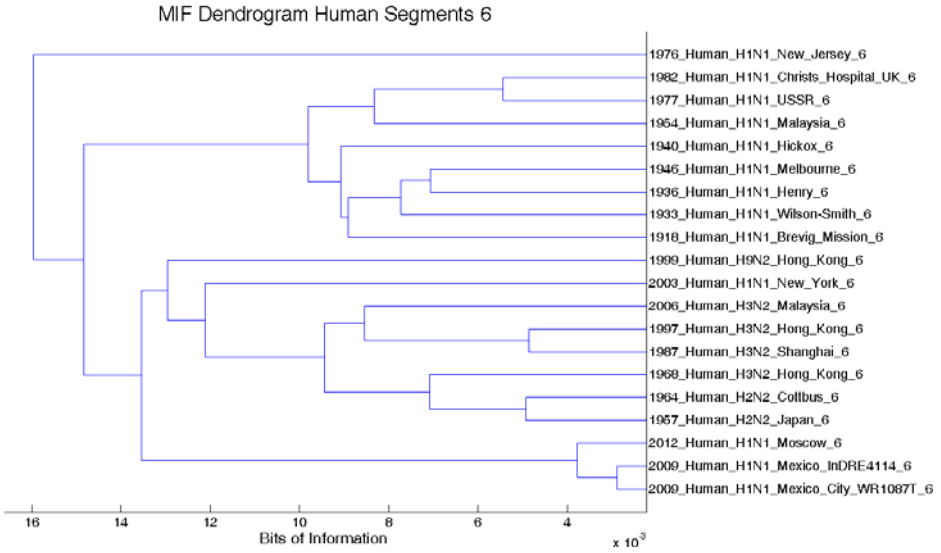
Figure 2. *Cont.*



(**B1**)

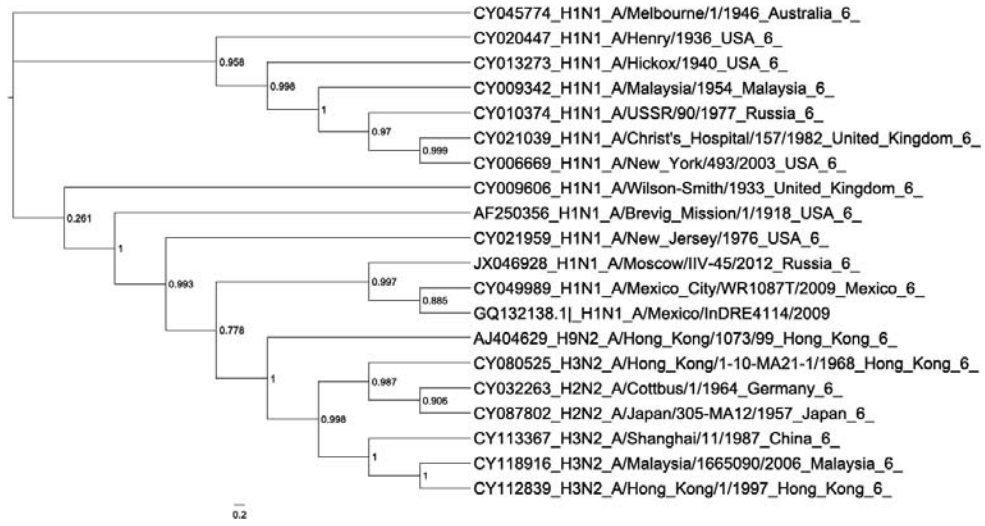


(**B2**)

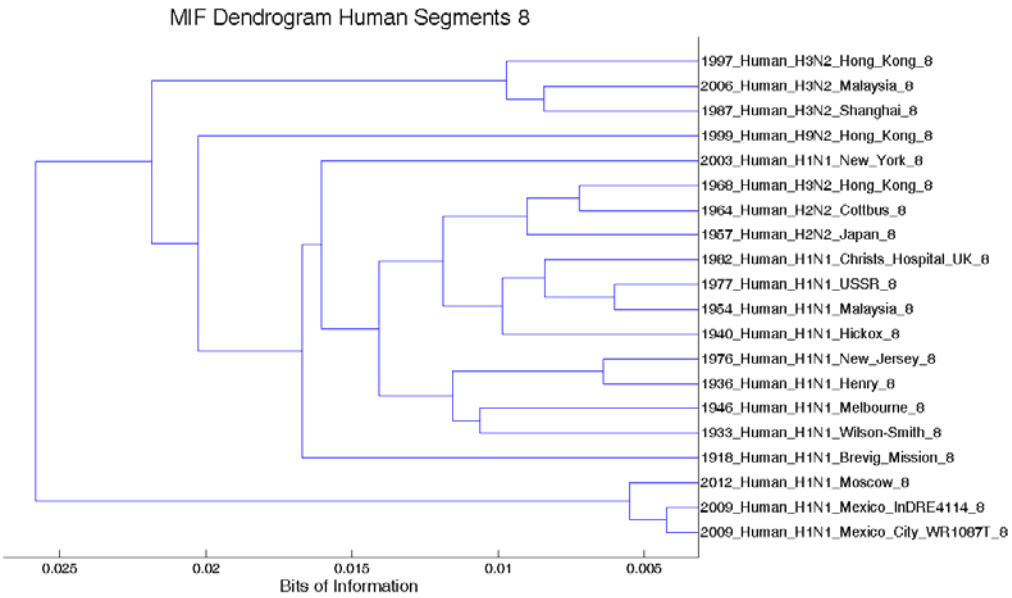


(**C1**)

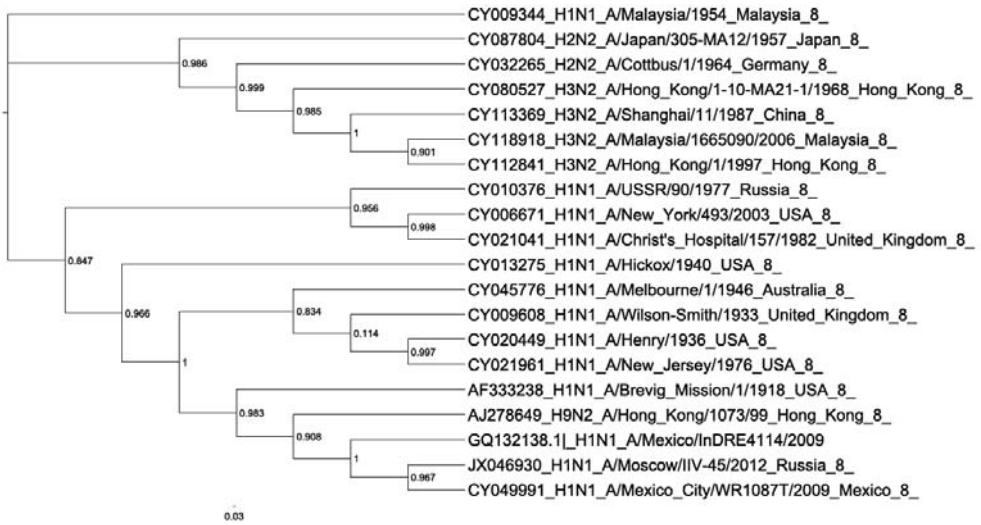
Figure 2. *Cont.*



(**C2**)



(**D1**)



(**D2**)

4.2.1. Dendrogram of S1

There is a clear separation between two groups: one group comprises H1 subtypes recently isolated. The 2nd group can still be subdivided into two subgroups: the subgroup of H1 subtypes isolated during the period 1918–1982, and the subgroup where all the non-H1 subtypes are encountered. The subtype 1999_H9N2 is located as the most dissimilar in this MI dendrogram, suggesting that PB2 of H9 is entirely different from the other sampled strains.

4.2.2. Phylogenetic Analysis of S1

The corresponding phylogeny also discerns between two groups: one group comprises N2 subtypes, meaning H3N2 and H2N2 subtypes with the exception of the H1N1_1954, which will repeat this grouping at the S8 phylogeny. The 2nd group comprises the H1N1 subtypes with the exception of the H9N2_1999 subtype.

It is remarkable how the MI dendrogram resembles the accurate discrimination between subtypes, in this case based at N1 and N2 viral classification. Also interesting is the similar clusterization of the most recent viral isolates, meaning that MI dendrogram may also capture vestiges of ancestry even when no evolutionary model is taken into account.

4.2.3. Dendrogram of S4

The similarities among the genomic segments of human influenza A that codes for HA are illustrated in a MI dendrogram (Figure 2B1). We remark that in this work we use the whole segments and not only the coding sequence. HA is an important viral protein, localized at the surface of the virion and it participates on the recognition of host immune cells and in cell invagination. It constitutes the 25% of the viral mass. Note that in the MI dendrogram (Figure 2B1) there is a clear distinction between the subtypes classified as H1 and the remaining ones. The MI dendrogram can discern between the types of HA by grouping exactly all H2 subtypes in one group, all H3 subtypes in another group, and all H1 subtypes still in another group. It is of interest that H9 is linked with the groupings of H1. On the other hand, the subtype H1_1918 is the most dissimilar sequence and it is located in the most outer part of the dendrogram. This could be attributable to the large number of carbohydrates that are expressed in the subtype H1_1918 in comparison with the remaining ones [46]. It is of notice the grouping of the subtypes 2009-2012 which is consistent with the dendrogram of S1.

4.2.4. Phylogenetic Analysis of S4

Overall, the topology of the phylogenetic trees for the S1, S2, S6, and S8 is ladder-type except for the vicinity of the most recent subtypes with the isolates during 1933–1946 (Figure 2A2, B2, C2, D2). Note also that the topology of the phylogenetic trees for S1, S4, S6, and S8, are not identical among them. In the MI dendrogram of S4 coding for HA, (Figure 2B1), we can observe that the 1918 H1N1 influenza A strain is arranged as the most dissimilar sequence, in this case the most ancestral one.

4.2.5. Dendrogram of S6

In Figure 2C1 the dendrogram corresponding to S6 containing the gene for the NA molecule is shown. This protein has also antigenic binding sites. The MI dendrogram discerns between the 2 main groups associated to N1 and N2. One group contains solely the N1 subtypes that appeared during 1918 to 1982. The 2nd group can be subdivided into two subgroups: to wit, those of subtype N2 that are linked with H1N1_2003. There is an interesting progression in this group which goes from 1957_H2N2 until 2006_H3N2 and then it culminates with the incorporation of the subtype 1999_H9N2. Interestingly enough, the recent H1 subtypes (2009–2012) are clustered in the 2nd subgroup and this is in harmony with the MI dendrograms of S2 and S4. The subtype 1976_H1N1 appears as the most dissimilar one.

4.2.6. Phylogenetic Analysis of S6

The comparison between the MI dendrogram of S6 (Figure 2C1) with its cladogram (Figure 2C2), as derived from bioinformatics methodologies, shows that the former resembles crisply and reasonable, similar subtype grouping. This is also observed in the dendrograms of S1, S4, S6, and S8. Bioinformatics methodologies organize subtypes adequately by homology (Figure 2A2, B2, C2, D2).

4.2.7. Dendrogram of S8

The segment S8 is interesting per se since it is bi-cistronic, *i.e.*, it encodes for 2 different proteins: to wit, NS1 which is the only Non-Structural protein of the virion which inhibits the mRNA splicing and the nuclear export of cellular and viral mRNA; it also plays a fundamental role in evading the immune response since it is an antagonist of interferon γ; and NEP which is a protein with a signal of nuclear export and it is expressed from edited mRNA by splicing mechanisms. The topology of this dendrogram (Figure 2D1) is more complex than the previous ones: notwithstanding the MI dendrogram separates H1 and N2 subtypes, there are four groups and three sequences that are left without a pairwise similar subtype. Once more, a group is formed that comprises the most recent subtype 2009–2012 in agreement with the MI dendrograms of S1, S4, and S6. This group of recent ones is the most dissimilar of the whole dendrogram. The subtype that corresponds to influenza A from 1918 is set alone, sharing similarity with the 20^{th} century H1N1 subtypes. The central group is the most complex as it generates two subgroups of H1 (1982, 1977, 1954, 1940, and 1933, 1946, 1936, 1976) flanking the subgroup formed by the subtypes N2 (1968, 1964, 1957). There is another subtype N2 (1999_H3N2) situated as an out-group-like of the central group. On the other hand, there is a group of the subtype N2 that turns out to be the 2nd less similar of the whole dendrogram and that involves the subtypes H3N2 of 1997, 2006, and 1987.

Overall, for the four dendrograms there are $\sim 4.5-, \sim 6.6-, \sim 16-,$ and $\sim 10-$ fold dissimilarities with S1, S4, S6, and S8, respectively, with respect to the amount of information. In the MI dendrogram of S1 and S8 (Figure 2A1, D1), we are neither dealing with HA nor NA, but with PB2 and NS1-NEP. Yet, it can be observed in the latter figures that there are three informational clades with the nomenclature for HA and NA. One formed by the N2 subtypes from 1957 to 2006, the other formed by H1 and H9 subtypes from 1918 to 2009, and the third one comprises H9 as well as N2. Albeit there is

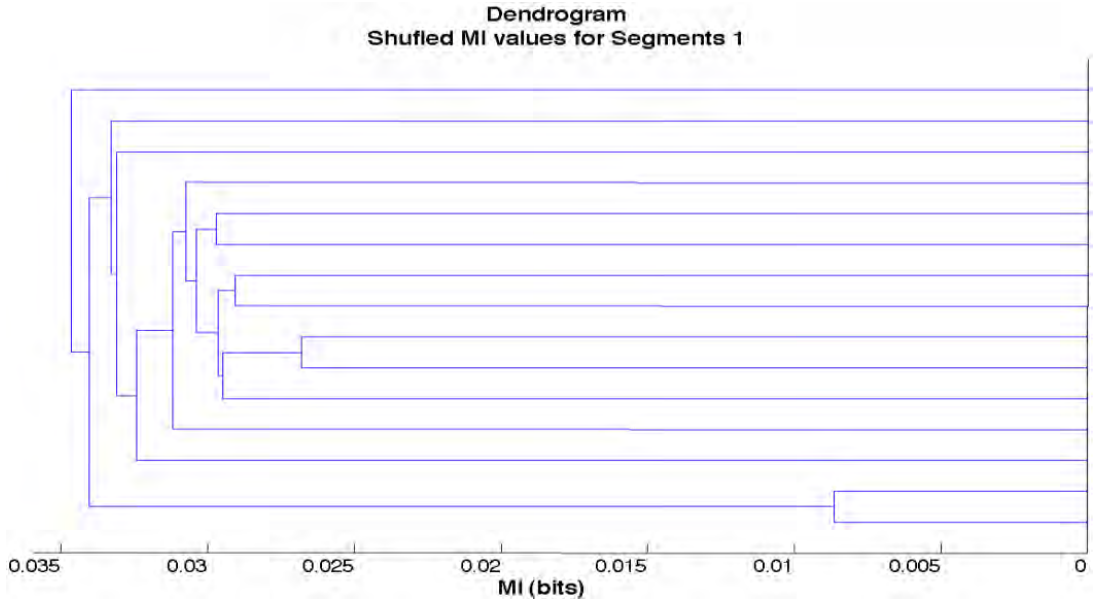
no nomenclature for polymerase subtypes of influenza A nor for the bi-cistronic gene, there seems to be a correlation with the canonical classification according to the types of HA's and NA's. By observing the corresponding phylogenetic reconstructions, there is a striking similarity to the informational cluster of the MI dendrogram. *In the human case the informational groups of the MI dendrogram correspond to the phylogenetic clades* (Figure 2A2, B2, C2, D2). Therefore, MI dendrograms capture evolutionary relevant information just as phylogenetic reconstructions do. Further statistical analysis show how the presence of the canonical subtypes HA and NA are related to subtypes of the remaining proteins (not shown).

In summary, all dendrograms of the MIs and the phylogenetic trees distinguish groups between subtypes and isolation times. Both clusterizations are surprisingly similar showing the capabilities of the MI to extract evolutionary important aspects of the actual RNA viral sequences. MI dendrograms for each segment reflected biologically coherent informational groups and were consistent with their corresponding clusters in the phylogenetic reconstructions. The MI dendrograms for the avian segments S1, S4, S6, and S8, as well as their corresponding phylogenetic reconstructions reinforce the same conclusions (not shown).

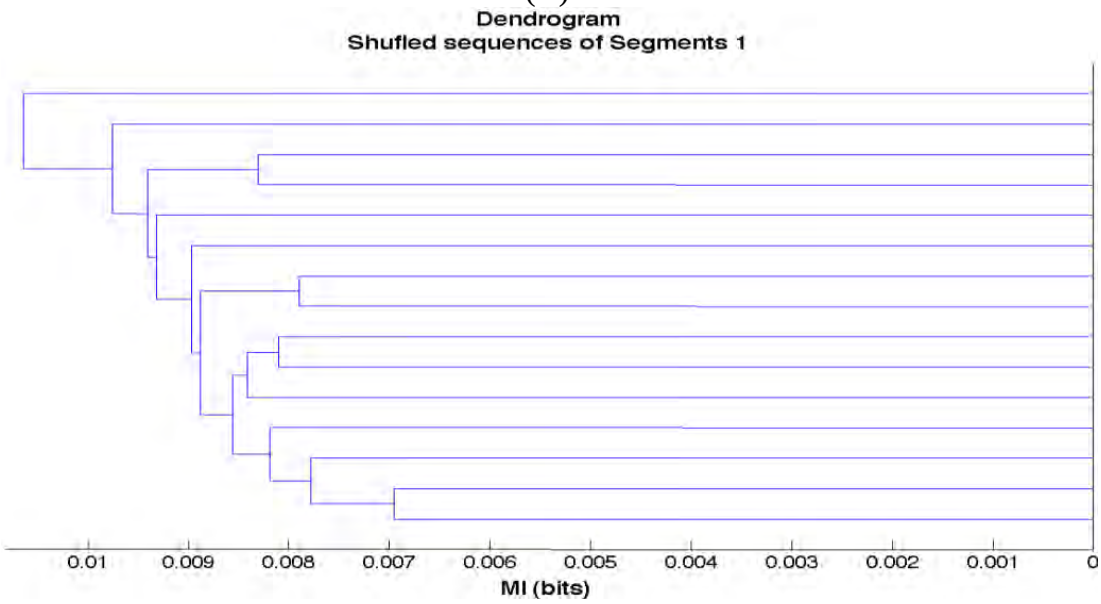
4.3. Random Controls

To test if the dendrograms of the MI could be the result of randomness we used two different controls. In the first one, we shuffled the actual biological sequences and then we calculated the MI dendrogram (Figure 3A). In the second one, the MI values obtained from the original viral sequences were randomized (Figure 3B). In both cases, we have disrupted the correlations that are detected in the actual biological sequences by the MI and therefore the autocorrelations selected during evolutionary pressures in these random controls disappear. Note that in both cases the hierarchies of the splits are soon generated in the most dissimilar possible zones, which lie in the truly random region. Henceforth all clusters are not alike to each other and these results validate the fact that our previous dendrograms cannot be the result of either randomness or of artifacts. We included these random controls to test whether the different subtypes of influenza A show evolutionary selection either for or against sequences that covary among them and favor particular informational groups. As the randomized controls do not possess biological meaningful information, we presumed that these controls are evolutionarily neutral with respect to viral evolution or genotype, such that preferred groupings would occur by chance. These results strongly argue that the viral genomes have evolved to favor particular informational groups according to the viral subtype.

Figure 3. Control experiments: (**A**) MI dendrogram from randomization of the actual viral sequences; (**B**) MI dendrogram from randomization of the MI values from the original viral sequences.



(**A**)



(**B**)

4.4. Flow of Information During Viral Evolution

Since evolutionary processes entail loss and gain of relevant information for survival, we estimated the flow of information during a century of evolution of influenza by determining the average information content for each segment in each influenza A subtype as a function of time irrespective of the subtype (Figure 4). Note that the content of information of PB2, for both human and avian strains, has oscillated throughout the sample years of registry (Figure 4A1). For human strains the range of oscillation is of the order of ~1.26 whereas for the avian strains this value is of the order of ~1.16. Notice that the average value of information at the beginning of the XX century lies in the range of the

most recent values of information. Each trough and peak, in principle, is associated to antigenic shifts originating a pandemic at its time of appearance. In regard to the average extent of information of S4, its value in 1918 and 1988 was the highest among all human and avian subtypes, respectively, all along a century (Figure 4B1). There is an oscillatory behavior in both avian and human strains with most frequent oscillations in the latter than in the former (Figure 4B1 and inset). Note that over almost a century, the average information content of S6, containing the NA gene, appears to have cycles of small amplitude for the human strains whereas for the avian subtypes the pattern is also oscillating but with a large peak in 1963 (Figure 4C1). This abrupt increase is of the order of ~1.57 in regard to the remaining oscillating values. In regard to S8 of human subtypes its information content was the highest in 1918 and then oscillated in a lower range whereas in the avian subtypes the oscillations were in a still lower range except at peaks in 1988 and 2008. The prototypic pandemic H1N1 influenza virus emerged in 1918 and then gave rise to periodic seasonal strains that began to diminish in frequency during the late 1950. In all segments it can be observed a peak in 1977 indicating a resurgence of H1N1 viruses, thus reestablishing the H1N1 seasonal strains that are currently in circulation. Pandemic viruses typically evolve into seasonal forms (not shown in Figure 4) that develop resistance to antibody neutralization. These seasonal strains would form a slow oscillating trunk from which the pandemic strains depart as branches. This episodic pattern is consistent with a "cactus-like" structure of evolution [15] in which there are periods of neutral mutation during inter-pandemic periods interrupted by sudden bursts of change during antigenic shifts.

There seems not to be a singular characteristic which could herald the large change that was about to happen in the Mexican pandemic of 2009 in spite of its corresponding antigenic shift. The same holds true for any other pandemic.

Figure 4. Average amount of information for: S1 (**A1**); S4 (**B1** and inset at other scale); S6 (**C1**); and S8 (**D1**) as a function of time from 1918 to 2009 in human (red curves) and from 1902 to 2011 in avian (blue curves) subtypes of the RNA genome of influenza A.

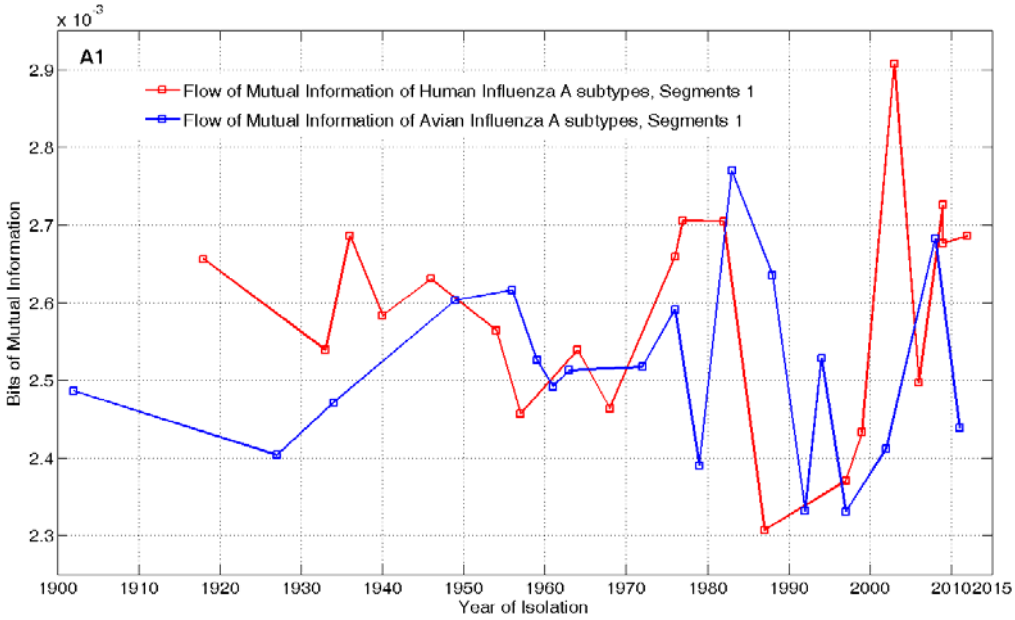
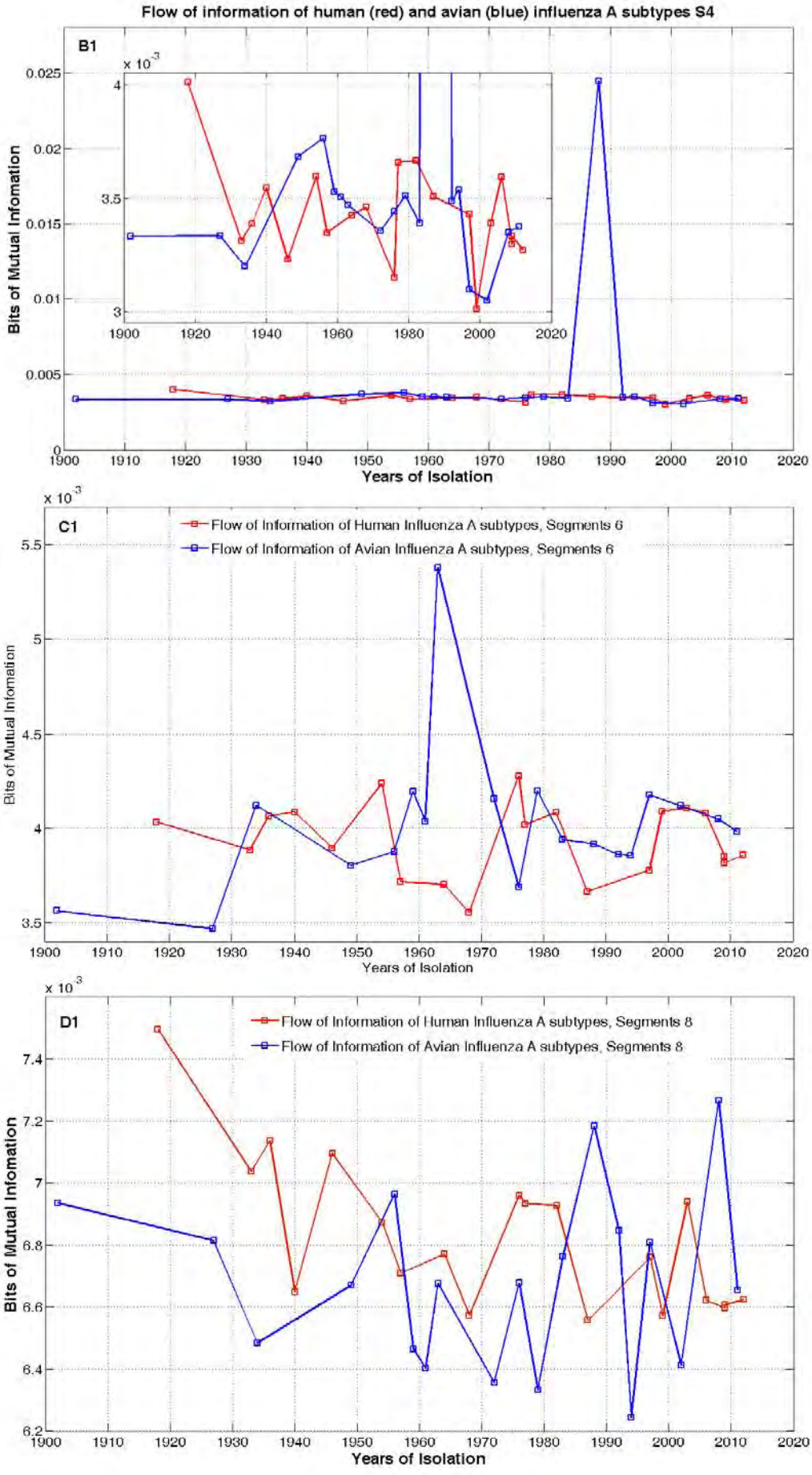


Figure 4. *Cont.*



5. Discussion and Conclusions

In this work, we have ushered in the use of the MI in symbolic series of genomic RNA sequences for capturing evolutionary relevant changes in information. The use of information theory in biology is not new. For example, the use of entropy and mutual information has been used for detecting gene-gene associations [47], multivariate entropy distance method has been developed for prokaryotic gene identification [48], the construction of phylogenetic trees using relative information between DNA sequences based on Lempel-Ziv complexity (particularly useful when multiple alignment based strategies fail) [49]; for detecting heart arrhythmias [50] and for determining nucleosome positioning motifs [51]. We use as an example 39 viral genomes comprising 312 sequences of influenza A that span over almost a century. We have applied standard information theory techniques to actual data on the history of influenza pandemics and their respective RNA symbolic sequences, and we have compared the results with Maximum Likelihood Phylogenetic methodology for determining evolutionary reconstruction histories of influenza A. We here report the use of the MI for detecting relevant informational changes which were known to be directly associated to the emergence of influenza pandemics associated to antigenic shifts in the eight segments of influenza A. In contrast, the analysis of more than 4,000 influenza A/H3N2 HA sequences from 1968 to 2010 using mutual information (MI)-based machine-learning model to design a site transition network for each amino acid site of HA to predict antigenic drifts has been reported [18]. The rate of prediction accuracy of this method is of the order of 70% [18]. In our present work we note that during antigenic shifts episodic changes in MI occur in all segments of influenza A. It has been found that positive selections are ongoing most of the time (*i.e*. not sporadic), and multiple mutations at antigenic sites cumulatively enhance antigenic drift [18,52]. It is worthwhile to mention that pandemic flu varieties evolve into seasonal flu varieties that are the ones that undergo antigenic drift.

The MIF is sensible enough to detect punctuated changes in a given viral sequence. The use of MI dendrograms certainly opens up new research avenues, such as elaborating a possible taxonomic classification of DNA and RNA viruses from an informational approach since it can clearly distinguish among them (ongoing work). MI dendrograms of S1, S2, S3, and S4, turned out to be practically identical or very similar in the clusterization, but not to the topology, of their corresponding Maximum Likelihood Phylogeny. A dendrogram is equivalent to a similarity analyses in evolutionary biology and is useful when the evolutionary dynamics of all the sequences are not fully understood. The combined use of the MI profiles with a simple hierarchical algorithm casted light upon the importance of relevant mutations associated to antigenic shifts. This was clearly illustrated for the case of influenza A (H1N1) of 1918 and 2009 (Figures 1 and 2). There are several evolutionary processes that might give rise to influenza strains with pandemic potential, and evaluating the significance of each of these has attracted much research effort. Molecular analyses suggest that the two previous pandemics involved reassortments between human-adapted and avian-adapted influenza viruses [43,44,53]. The 1957 pandemic probably involved reassortment between an avian H2N2 influenza virus and a human H1N1 influenza virus [54–56] and the 1968 pandemic probably involved reassortment between an avian H3Nx influenza virus and a human H2N2 influenza virus [54,57]. Even in these relatively well-documented cases, however, the species in which reassortment took place is not known. In this work we show that the MI profiles shed light upon the evolutionary process of influenza A. In particular we used segments

S1, S4, S6, and S8 which are clinically, pharmacologically, and immunologically of great interest. We highlight that the dendrograms of the MI profiles, constructed upon different strains of influenza A over time, indicate how the information flows among different viral subtypes during the evolution of these segments. It is clear that the evolutionary pattern that is observed in one segment is not equivalent to the rest of the viral genome. This means that focusing in changes in only one segment or a gene does not represent the whole evolutionary processes of the whole organism. For example, S1 and S8 show significant changes in information during a century of evolution (Figures 2 and 4). In other words, it is necessary to include all segments of influenza A genome to discern an accurate and more comprehensive picture of all evolutionary aspects that can act in part or in concert. The phylogenies of the HA genes of *a given specific subtype* like the human influenza A (H3N2) and human measles virus appear immediately distinct, with influenza exhibiting a ladder-like tree with a long trunk and very short side branches and measles presenting a bushier tree with deeper branchings [58]. For a specific subtype of influenza, the trunk branch corresponds to the progenitor lineage; mutations that occur along the trunk are eventually fixed, persisting until "over-written" by subsequent mutations. In contrast, mutations that appear on side branches are eventually lost from the population. Lack of powerful technologies and cutting-edge software programs are no longer an excuse for avoiding evolutionary analysis for whole genomes.

The flows of information over a century for the four segments were in agreement with their dendrograms in particular the dendrogram of S6. There is a significant abrupt increase in the information level of S6 of A (H1N1) subtype 2009 when compared with all the others (Figure 1C). This may have far-reaching implications in regard to the immune response, its contribution to the appearance of the pandemic, and for antiviral drug efficacy. Neuraminidase (NA; 1,407 nucleotides, S6), is involved in budding new virions from infected cells. Antiviral drugs are available for prevention and treatment of influenza. The influenza virus neuraminidase inhibitors zanamivir and oseltamivir were introduced into clinical practice in various parts of the world from 1999 through 2002 [58]. Oseltamivir limits replication of both influenza A and B viruses [59]. In most European countries, neuraminidase inhibitors are not widely used to treat seasonal influenza, but they are being stockpiled in many countries as part of their pandemic influenza preparedness. The drugs include amantadine, which inhibits the uncoating of virions by interfering with M2, and oseltamivir, which inhibits the release of virions from infected cells by interfering with NA [60]. According to our results, we anticipate that the efficacy of oseltamivir for A (H1N1) 2009 treatment might have changed.

The flows of information over a century for the 4 segments were in agreement with their dendrograms. *Mutual information may reflect but not predict relevant mutations of antigenic shifts driving influenza virus pandemics.* Every phylogeny depicts these processes and our analysis shows that MI dendrograms captures remarkably well the same evolutionary dynamics. Therefore, we are successfully connecting an evolutionary process with gains and losses of information. For each segment of the virus, we can say that during its evolution, new correlations arise or old correlations vanish, and they are all relevant for the evolutionary strategies of the virus. Correlations appear and disappear over time, and this is information conveyed through evolutionary processes. To our knowledge no other evolutionary tests are capable of determining the gains and losses in terms of bits of mutual information. The approach of the MI dendrograms was validated by phylogenetic reconstructions and both can certainly enhance our understanding of them. Whole-genome phylogenies

show the coexistence of multiple viral lineages, particularly on a limited spatial and temporal scale. This indicates that the transitions among antigenic types do not always proceed in a simple linear manner, that reassortments among coexisting lineages is relatively frequent, and that, for these reasons, predicting the path of influenza virus evolution from sequence data alone is inherently difficult. Interestingly, the resulting clusters of MI dendrograms turned out to be consistent with methodologies such as neighbor joining using MEGA5 (not shown) [61–63] as well as Maximum Likelihood methodologies [40,41]. Therefore we propose the use of the MI as a complement of subsequent bioinformatics methodologies.

Acknowledgments

MVJ was financially supported by PAPIIT-IN107112, UNAM, México. VSS was granted a doctoral fellowshisp from CONACYT and he also was partially supported by Coordinación de Estudios de Posgrado, UNAM.

Conflict of Interest

The author has no conflict of interest to declare.

References

1. Breitbart, M.; Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **2005**, *13*, 278–284.
2. Forterre, P. The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res.* **2006**, *117*, 5–16.
3. Cornelis, G.; Heidmann, O.; Bernard-Stoecklin, S.; Reynaud, K.; Vérond, G.; Mulote, B.; Dupressoir, A.; Heidmann, T. Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E432–E441.
4. Norrby, E. Nobel Prizes and the emerging virus concept. *Arch. Virol* .**2008**, *153*, 1109–1123.
5. Miller, E.S.; Kutter, E.; Mosig, G.; Arisaka, F.; Kunisawa, T.; Rüger, W. Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.* **2003**, *67*, 86–156.
6. Forterre, P.; Gribaldo, S.; Gadelle, D.; Serre, M.C. Origin and evolution of DNA topoisomerases. *Biochimie* **2007**, *9*, 427–446.
7. Duffy, S.; Shackelton, L.A.; Holmes, E.C. Rates of evolutionary change in viruses: Patterns and determinants. *Nat. Rev. Genet.* **2008**, *9*, 267–276.
8. Eigen, M. The origin of genetic information: Viruses as models. *Gene* **1993**, *135*, 37–47.
9. Smith, R.A.; Loeb, L.A.; Preston, B.D. Lethal mutagenesis of HIV. *Virus Res.* **2005**, *107*, 215–228.
10. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **1999**, *401*, 877–884.
11. Sharp, P.M. Origins of human virus diversity. *Cell* **2002**, *108*, 305–312.
12. Rambaut, A.; Posada, P.; Crandall, K.A.; Holmes, E.C. The causes and consequences of HIV evolution. *Nat. Rev. Gen.* **2004**, *5*, 52–61.

13. Pybus, O.G.; Rambaut, A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Gen.* **2009**, *10*, 540–550.
14. Fitch, W.M.; Leiter, J.M.E.; Li, X.; Palese, P. Positive Darwinian evolution in human influenza A viruses. *Proc.Natl. Acad. Sci. USA* **1991**, *88*, 4270–4274.
15. Wolf, Y.I.; Viboud, C.; Holmes, E.C.; Koonin, E.V.; Lipman, D.J. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biol. Direct* **2006**, *1*, 34.
16. Nelson, M.I.; Simonsen, L.; Viboud, C.; Miller, M.A.; Taylor, J.; St George, K.; Griesemer, S.B.; Ghedin, E.; Sengamalay, N.A.; Spiro, D.J.; *et al.* Stochastic processes are key determinants of short-term evolution in influenza A virus. *PLoS Pathog.* **2006**, *2*, e125.
17. Eigen, M. Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **1971**, *58*, 465–523.
18. Eigen, M.; Schuster, P. The hypercycle A principle of natural self-organization; Springer-Verlag: Heidelberg, Gemany, 1979.
19. Stich, M.; Briones, C.; Manrubia, S.C. Collective properties of evolving molecular quasispecies. *BMC Evol. Biol.* **2007**, *7*, 110–123.
20. Wilke, C.O. Quasispecies theory in the context of population genetics. *BMC Evol. Biol.* **2005**, *5*, 44–52.
21. Kimura, M.; Maruyama, T. The mutational load with epistatic gene interactions in fitness. *Genetics* **1966**, *54*, 1337–1351.
22. Nelson, M.I.; Holmes, E.C. The evolution of epidemic influenza. *Nat. Rev. Gen.* **2007**, *8*, 196–205.
23. Hilleman, M.R. Realities and enigmas of human viral influenza: Pathogenesis, epidemiology and control. *Vaccine* **2002**, *20*, 3068–3087.
24. Murphy, B.R.; Webster, R.G. Orthomyxoviruses. In *Fields Virology*, 3rd ed.; Fields, B.N., Knipe, D.M., Howley, P.M., Eds.; Lippincott-Raven Publishers: Philadelphia, PA, USA, 1996; pp. 1397–1445.
25. De Jong, J.C.; Rimmelzwaan, G.F.; Fouchier, R.A.; Osterhaus, A.D. Influenza virus: A master of metamorphosis. *J. Infect.* **2000**, *40*, 218–228.
26. Ferguson, N.M.; Galvani, A.P.; Bush, R.M. Ecological and immunological determinants of influenza evolution. *Nature* **2003**, *422*, 428–433.
27. Shannon, C.E.A. Mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656.
28. Xia, Z.; Jin, G.; Zhu, J.; Zhou, R. Using a mutual information-based site transition network to map the genetic evolution of influenza A/H3N2 virus. *Bioinformatics* **2009**, *25*, 2309–2317.
29. Influenza Virus Resource. Available online: http://www.ncbi.nlm.nih.gov/genomes/FLU/ (accessed on 15 November 2012).
30. Avian influenza A (H5N1)-update 31: Situation (poultry) in Asia: Need for a long-term response, comparison with previous outbreaks. Available online: http://www.who.int/csr/don/2004_03_02/en/ (accessed on 15 November 2012).
31. Influenza A virus (A/Brevig Mission/1/1918(H1N1)). Available online: http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=88776/ (accessed on 15 November 2012).
32. Kullback, S. *Information Theory and Statistics*; John Wiley and Sons: New York, NY, USA, 1959.
33. Blahut, R.E. *Information Theory and Statistics*; Addison-Wesley: Reading, MA, USA, 1987.

34. Li, W. Mutual information function *vs.* correlation functions. *J. Stat. Phys.* **1990**, *60*, 823–837.
35. Hartl, D.L. *A Primer of Population Genetics*, 3rd ed.; Sinauer Associates: Sunderland, MA, USA, 2000.
36. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797.
37. Gouy, M.; Guindon, S.; Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol*. **2010**, *27*, 221–224.
38. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **1986**, *17*, 57–86.
39. Posada, D. jModelTest: Phylogenetic model averaging. *Mol. Biol. Evol*. **2008**, *25*, 1253–1256.
40. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **1981**, *17*, 368–376.
41. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **2010**, *5*, e9490.
42. FigTree. Available online: http://tree.bio.ed.ac.uk/software/figtree/ (accessed on 10 June 2013).
43. Reid, A.H.; Fanning, T.G.; Janczewski, T.A.; Taubenberger, J.K. Characterization of the 1918 "Spanish" influenza virus neuraminidase gene. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 6785–6790.
44. Reid, A.H.; Taubenberger, J.K. The origin of the 1918 pandemic influenza virus: A continuing enigma. *J. Gen. Virol.* **2003**, *84*, 2285–2292.
45. Garten, R.J.; Davis, C.T.; Russell, C.A.; Shu, B.; Lindstrom, S.; Balish, A.; Sessions, W.M.; Xu, X.; Skepner, E.; Deyde, V.; *et al.* Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* **2009**, *325*, 197–201.
46. Wei, C.-J.; Boyington, J.C.; Dai, K.; Houser, K.V.; Pearce, M.B.; Kong, W.-P.; Yang, Z.-Y.; Tumpey, T.M.; Nabel, J.G. Cross-neutralization of 1918 and 2009 influenza viruses: Role of glycans in viral evolution and vaccine design. *Sci. Transl. Med.* **2010**, *2*, 24ra21.
47. Butte, A.J.; Kohane, I.S. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomp.* **2000**, *5*, 415–426.
48. Ouyang, Z.Q.; Zhu H.Q.; Wang J.; She Z.S. Multivariate entropy distance method for prokaryotic gene identification. *J. Bioinf. Comp. Biol.* **2004**, *2*, 353–373.
49. Out, H.H.; Sayood, K. A new sequence distance measure for phylogenetic tree reconstruction. *Bioinformatics* **2003**, *16*, 2122–2130.
50. Núñez-Acosta, E.; Lerma, C.; Márquez, M.F.; José, M.V. Mutual information analysis reveals bigeminy patterns in Andersen-Tawil syndrome and in subjects with history of sudden cardiac death. *Physica A: Statist. Mech. Appl.* **2012**, *391*, 693–707.
51. Sosa, D.; Miramontes, P.; Li, W.; Mireles, V.; Bobadilla, J.R.; José, M.V. Periodic distribution of a putative nucleosome positioning motif in human, non-human primates, and Archaea: Mutual information analysis. *Int. J. Genomics* **2013**, 963956.
52. Shih, A.C.; Hsiao, T.C.; Ho, M.S.; Li, W.H. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proc. Nat. Acad. Sci. USA* **2007**, *104*, 6283–6288.
53. Webster, R.G.; Bean, W.J.; Gorman, O.T.; Chambers, T.M.; Kawaoka, Y. Evolution and ecology of influenza A viruses. *Microbiol. Rev.* **1992**, *56*, 152–179.
54. Scholtissek, C.; Burger, H.; Bachmann, P.A.; Hannoun, C. Genetic relatedness of hemagglutinins of the H1 subtype of influenza A viruses isolated from swine and birds. *Virology* **1983**, *129*, 521–523.

55. Kawaoka, Y.; Krauss S.; Webster, R.G. Avian-to-human transmission of the PB1gene of influenza A viruses in the 1957 and 1968 pandemics. *J. Virol.* **1989**, *63*, 4603–4608.
56. Schafer, J.R.; Kawaoka, Y.; Bean, W.J.; Suss, J.; Senne, D.; Webster, R.G.; et al. Origin of the pandemic 1957 H2 influenza A virus and the persistence of its possible progenitors in the avian reservoir. *Virology* **1993**, *194*, 781–788.
57. Bean, W.J.; Schell, M.; Katz, J.; Kawaoka, Y.; Naeve, C.; Gorman, O.; Webster, R.G.; *et al.* Evolution of the H3 influenza virus hemagglutinin from human and non-human hosts. *J. Virol.* **1992**, *66*, 1129–1138.
58. Bedford, T.; Cobey, S.; Pascual, M. Strength and tempo of selection revealed in viral gene genealogies. *BMC Evol. Biol.* **2011**, *11*, 220.
59. Hayden, F.G. Pandemic influenza: Is an antiviral response realistic? *Pediatr. Infect. Dis. J.* **2004**, *23*, S262–S269.
60. Moscona, A. Neuraminidase Inhibitors for Influenza. *N. Engl. J. Med.* **2005**, *353*, 1363–1373.
61. Tamura, K.; Nei, M.; Kumar, S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Nat. Acad. Sci. USA* **2004**, *101*, 11030–11035.
62. Saitou, N.; Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **1987**, *4*, 406–425.
63. Tamura, K.; Peterson, D.; Peterson, N.; Stecher, G.; Nei, M.; Kumar, S. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **2011**, *28*, 2731–2739.