INSTITUTO POLITÉCNICO NACIONAL

ESCUELA NACIONAL DE CIENCIAS BIOLÓGICAS

SECCIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

LifePrint: un nuevo método de distancia para construir árboles filogenéticos sin alineamientos múltiples

T E S I S

QUE COMO UNO DE LOS REQUISITOS

PARA OBTENER EL GRADO DE:

D O C T O R E N BIOMEDICINA Y

BIOTECNOLOGÍA MOLECULAR

P R E S E N T A:

M. EN C. FABIÁN REYES PRIETO

DIRECTORES DE TESIS:

DR. ROGELIO MALDONADO RODRÍGUEZ

DR. ALFONSO MÉNDEZ TENORIO

MÉXICO, D.F. ENERO 2011

Advances and Applications in Bioinformatics and Chemistry

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

LifePrint: a novel *k*-tuple distance method for construction of phylogenetic trees

This article was published in the following Dove Press journal:
Advances and Applications in Bioinformatics and Chemistry
31 December 2010
Number of times this article has been viewed

Fabián Reyes-Prieto[1]
Adda J García-Chéquer[1]
Hueman Jaimes-Díaz[1]
Janet Casique-Almazán[1]
Juana M Espinosa-Lara[1]
Rosaura Palma-Orozco[2]
Alfonso Méndez-Tenorio[1]
Rogelio Maldonado-Rodríguez[1]
Kenneth L Beattie[3]

[1]Laboratory of Biotechnology and Genomic Bioinformatics, Department of Biochemistry, National School of Biological Sciences, [2]Superior School of Computer Sciences, National Polytechnic Institute, Mexico City, Mexico; [3]Amerigenics Inc, Crossville, Tennessee, USA

Correspondence: Fabián Reyes-Prieto
Laboratory of Biotechnology and Genomic Bioinformatics, Department of Biochemistry, National School of Biological Sciences, National Polytechnic Institute, CP 11340, Mexico City, Mexico

Purpose: Here we describe LifePrint, a sequence alignment-independent *k*-tuple distance method to estimate relatedness between complete genomes.

Methods: We designed a representative sample of all possible DNA tuples of length 9 (9-tuples). The final sample comprises 1878 tuples (called the LifePrint set of 9-tuples; LPS9) that are distinct from each other by at least two internal and noncontiguous nucleotide differences. For validation of our *k*-tuple distance method, we analyzed several real and simulated viroid genomes. Using different distance metrics, we scrutinized diverse viroid genomes to estimate the *k*-tuple distances between these genomic sequences. Then we used the estimated genomic *k*-tuple distances to construct phylogenetic trees using the neighbor-joining algorithm. A comparison of the accuracy of LPS9 and the previously reported 5-tuple method was made using symmetric differences between the trees estimated from each method and a simulated "true" phylogenetic tree.

Results: The identified optimal search scheme for LPS9 allows only up to two nucleotide differences between each 9-tuple and the scrutinized genome. Similarity search results of simulated viroid genomes indicate that, in most cases, LPS9 is able to detect single-base substitutions between genomes efficiently. Analysis of simulated genomic variants with a high proportion of base substitutions indicates that LPS9 is able to discern relationships between genomic variants with up to 40% of nucleotide substitution.

Conclusion: Our LPS9 method generates more accurate phylogenetic reconstructions than the previously proposed 5-tuples strategy. LPS9-reconstructed trees show higher bootstrap proportion values than distance trees derived from the 5-tuple method.

Keywords: phylogeny, sequence alignment, similarity search, tuple, viroid

Introduction

The most used and widespread representations of the evolutionary history of biologic entities are phylogenetic trees. Typically, molecular phylogenetic tree construction starts from a set of sequences (DNA or proteins), computation of a multiple sequence alignment, and then, based on the multiple sequence alignment, construction of a tree using one or several optimization criteria, such as distance, maximum parsimony, minimum evolution, maximum likelihood, and Bayesian inference. Among these criteria, a distance-based method using neighbor-joining (NJ)[1] is frequently used because it is considerably faster than character-based methods such as maximum parsimony, maximum likelihood, and Bayesian inference. However, the requirement of using multiple sequence alignment carries some disadvantages for typical tree construction methods. One of the major limitations of multiple alignments arises from the heuristic

Esta investigación se desarrolló en el Laboratorio de Biotecnología y Bioinformática Genómica del Departamento de Bioquímica de la Escuela Nacional de Ciencias Biológicas del Instituto Politécnico Nacional, bajo la dirección del Dr. Rogelio Maldonado Rodríguez y del Dr. Alfonso Méndez Tenorio.

El alumno fue becario del CONACYT de agosto de 2007 a diciembre de 2010 (CVU 168916).

para **YULI** *y* **LEO**,

por su amor,

por la suma de talentos,

por ser esa alegría y motivación infinitas,

¡LOS AMO!

A mis padres, por su cariño y apoyo permanentes

A mi hermano, por ser mi mejor amigo, mi mentor, ¡MUCHAS GRACIAS MI SANGRE!

A mi cuñada y mis sobrinas, por la gran felicidad y orgullo que me representan

A mis abuelas, por seguir aquí, alimentando emocionalmente a sus familias

A mis abuelos in memoriam

A mi familia, por todos los buenos momentos, por su solidaridad

AGRADECIMIENTOS

A los doctores Rogelio y Alfonso, por su asesoría, apoyo y comprensión

A los doctores Guadalupe, Graciela, José Luis y Arturo, por sus preciadas sugerencias

A todo el equipo del Laboratorio de Biotecnología y Bioinformática Genómica, por el ambiente de amistad y ayuda, benéfico en todo momento

Al IPN, por fungir como una institución motivadora y vivificante, ha sido enorgullecedor crecer en ella

A México (oficialmente Estados Unidos Mexicanos), porque seguiré formando parte, junto con muchos otros, del esfuerzo por alcanzar un bienestar equitativo y colectivo

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS

MSA: *Multiple sequence alignment*; Alineamiento múltiple de secuencia

LPS9: *LifePrint set of 9-tuples*; Conjunto de 9-tuplos de *LifePrint*

NJ: Algoritmo *neighbor-joining*

SD: *Symmetric difference*; Diferencia simétrica

MP: *Maximum parsimony*; Máxima parsimonia

ML: *Maximum likelihood*; Máxima probabilidad

BI: *Bayesian inference*; Inferencia Bayesiana

HIV-1: *Human immunodeficiency virus type 1*; Virus de inmunodeficiencia tipo 1

VH: *Virtual hybridization*; Hibridación virtual

UFC: *Universal fingerprinting chip*; Sensor universal de huella genómica

NCBI: *National Center for Biotechnology Information*; Centro Nacional para la Información Biotecnológica

K2P: *Kimura 2 parameter*; 2-parámetros de Kimura

ICTV: *International Committee on Taxonomy of Viruses*; Comité Internacional en Taxonomía de Virus

RESUMEN

Este trabajo propone y caracteriza *LifePrint*, un método de distancia de *k*-tuplos sin alineamientos múltiples (MSA, por sus siglas en inglés) para estimar relaciones filogenéticas entre genomas completos.

Se diseñó una muestra representativa de todos los tuplos posibles de DNA con una longitud de 9 nucleótidos (9-tuplos). Dicho conjunto comprende 1878 tuplos (el Conjunto de 9-tuplos de *LifePrint*; LPS9, por sus siglas en inglés), diferentes cada uno en por lo menos dos diferencias nucleotídicas internas y no contiguas. Para validar *LifePrint* se analizaron varios genomas reales y simulados de viroides. Usando diferentes métricas de distancia, se escrutaron diversos genomas para estimar las distancias de *k*-tuplos entre sus secuencias. Posteriormente, se usaron las distancias de *k*-tuplos estimadas, para construir árboles filogenéticos, usando el algoritmo *neighbor-joining* (NJ). La precisión comparada entre *LifePrint* y un método de 5-tuplos (que utiliza el conjunto completo de tuplos de dicha longitud), reportado previamente, fue evaluada usando la diferencia simétrica (SD, por sus siglas en inglés) entre los árboles estimados por cada método y un árbol filogenético ("árbol verdadero") construido con genomas simulados.

El esquema óptimo de búsqueda de similitud identificado para el LPS9, permite hasta dos diferencias nucleotídicas entre cada tuplo y el genoma escrutado. Los resultados de las búsquedas de similitud en genomas simulados, indicaron que, en la mayoría de los casos, el LPS9 es capaz de detectar eficientemente sustituciones de una sola base entre los genomas. El análisis de variantes genómicas simuladas, con una proporción alta de sustituciones, indica que el LPS9 es capaz de discernir relaciones entre variantes hasta con un 40% de sustituciones nucleotídicas.

LifePrint estimó filogenias más precisas que el método de 5-tuplos reportado previamente. Los árboles construidos con *LifePrint* presentan valores de proporción de *bootstrap* mayores que los árboles construidos con el método de 5-tuplos.

ABSTRACT

Here we describe *LifePrint*, a sequence alignment-independent *k*-tuple distance method to estimate relatedness between complete genomes.

We designed a representative sample of all possible DNA tuples of length 9 nucleotides (9-tuples). The final sample comprises 1878 tuples (called the *LifePrint* set of 9-tuples; LPS9) that are distinct from each other by at least two internal and noncontiguous nucleotide differences. For validation of our *k*-tuple distance method, we analyzed several real and simulated viroid genomes. Using different distance metrics, we scrutinized diverse viroid genomes to estimate the *k*-tuple distances between these genomic sequences. Then we used the estimated genomic *k*-tuple distances to construct phylogenetic trees using the neighbor-joining algorithm (NJ). A comparison of the accuracy of LPS9 and the previously reported 5-tuple method was made using symmetric difference (SD) between the trees estimated from each method and a true phylogenetic tree constructed with simulated genomes.

The identified optimal search scheme for LPS9 allows only up to two nucleotide differences between each 9-tuple and the scrutinized genome. Similarity search results of simulated genomes indicate that, in most cases, LPS9 is able to detect single-base substitutions between genomes efficiently. Analysis of simulated genomic variants with a high proportion of base substitutions indicates that LPS9 is able to discern relationships between genomic variants with up to 40% of nucleotide substitution.

LifePrint generates more accurate phylogenetic reconstructions than the previously proposed 5-tuples strategy. LPS9-reconstructed trees show higher bootstrap proportion values than distance trees derived from the 5-tuple method.

ÍNDICE

I. INTRODUCCIÓN

ANTECEDENTES

En un sentido amplio, la filogenética consiste en la reconstrucción de la historia evolutiva de los seres vivos, lo que implica el uso de métodos para inferir el pasado a partir de características presentes en especies actuales. Un árbol filogenético es la representación gráfica más popular de dicha historia. El objetivo principal de la mayoría de los estudios filogenéticos, y un importante subproducto de otros, es construir el árbol filogenético que describa mejor el devenir evolutivo.[1]

A lo largo de este texto, cualquier mención de los términos árbol filogenético y árboles filogenéticos, se hace escribiendo sólo árbol o árboles, respectivamente.

Hasta antes de los 1970s, la filogenética estuvo basada en el análisis de características morfológicas y/o citológicas. El uso de información molecular en filogenética, originó una revolución y, hacia finales de los 1980s, el acceso a secuencias de DNA incrementó el número de caracteres que podían ser comparados de menos de 100 a más de 1000, lo que mejoró considerablemente el poder resolutivo de la inferencia filogenética. Algunos genes se convirtieron en marcadores de referencia. En particular, y debido a

su considerable grado de conservación en todos los seres vivos, el gen que codifica para la subunidad pequeña del RNA ribosomal fue usado extensivamente, y permitió el reconocimiento de *Archaea* como un tercer dominio en el árbol de la vida. No obstante, al considerar un mismo grupo de especies, comenzaron a observarse diferencias topológicas relevantes entre árboles construidos con diferentes genes. También, se evidenció que la información de un solo gen, con frecuencia resultaba insuficiente para obtener un soporte estadístico sólido en ciertos nodos dentro de los árboles.[1]

Típicamente, un árbol se construye a partir de un conjunto de secuencias relacionadas evolutivamente, el cálculo de un alineamiento múltiple (MSA, por sus siglas en inglés) y luego, con base en éste, la construcción del árbol empleando criterios como distancia, máxima parsimonia (MP), o probabilidad (por ejemplo, máxima probabilidad o inferencia Bayesiana; ML o BI, por sus siglas en inglés, respectivamente). Los métodos basados en distancia que usan el algoritmo *neighbor-joining* (NJ),[2] se emplean frecuentemente porque son considerablemente más rápidos que los métodos de caracteres (basados en MP o probabilidad).

Sin embargo, la necesidad de MSA conlleva desventajas para la construcción típica de árboles. Una de las mayores limitaciones está dada por los métodos heurísticos usados para calcular los MSA. Estos métodos heurísticos pueden enfrentar problemas para manejar

secuencias grandes, dado que los algoritmos subyacentes tienen una complejidad computacional de orden cuadrático (es decir, pequeños aumentos en la longitud de las secuencias, implican grandes incrementos en el tiempo requerido para procesar los MSA), lo que resulta impráctico en algunos casos, por ejemplo, analizando relaciones entre genomas completos. Adicionalmente, ya que los MSA con frecuencia contienen ambigüedades de homología, la inferencia filogenética basada en análisis de MSA puede producir árboles equivocados.[1,3] Ciertos tipos de eventos evolutivos, como translocaciones e inversiones, difícilmente son considerados e incluidos en un análisis de MSA. Una desventaja más en el caso de los métodos basados en distancia, es que consideran sólo el número de diferencias entre las secuencias, sin considerar su posición. Algunos programas comunes para calcular MSA son *MUSCLE*,[4] *DIALIGN 2*,[5] *T-Coffee*,[6] *CLUSTAL W*,[7] y *Kalign*.[8] En la Tabla 1 se muestra una breve descripción de los programas antes mencionados.

Los estudios filogenéticos típicos a nivel genómico, son computacionalmente sobredemandantes y, en algunos casos, pueden resultar imprácticos.

Tabla 1. Características generales de algunos programas para calcular MSA

Nombre	Descripción	Tipo de secuencia	Tipo de alineamiento
MUSCLE	Usa el algoritmo de alineamiento progresivo. Realiza un refinamiento del alineamiento. Considerado por lograr tanto mejor precisión promedio como mayor velocidad que los otros programas	Ambos	Local o Global
DIALIGN 2	Método basado en la comparación de segmentos. Realiza un refinamiento del alineamiento	Ambos	Incorpora alineamiento local además del alineamiento global
T-Coffee	Usa el algoritmo de alineamiento progresivo. Permite combinar resultados obtenidos con varios métodos de alineamiento	Ambos	Incorpora alineamiento local además del alineamiento global
CLUSTAL W	Usa el algoritmo de alineamiento progresivo. Programa de propósito general	Ambos	Local o Global
Kalign	Usa el algoritmo de alineamiento progresivo	Ambos	Global

Notas: Breve descripción de algunos programas comunes para calcular MSA. Tipo de secuencia hace referencia a nucleótidos y/o aminoácidos.

Para sobreponerse a estas limitaciones prácticas, se han propuesto métodos alternativos independientes de MSA, por ejemplo, los métodos basados en el contenido de genes, definen la distancia entre dos genomas según el porcentaje de genes homólogos compartidos.[9-11] Recientemente, se ha probado con éxito el uso de genes firma correspondientes a distintas jerarquías taxonómicas.[12] Los métodos de compresión realizan la búsqueda de repetidos exactos, aproximados, directos o invertidos, y miden la similitud de genomas completos con base en su tasa de compresión relativa.[13-15] El vector de composición es un método que emplea cadenas (es decir, secuencias nucleotídicas cortas) informativas para inferir filogenias. Funciona a partir de una estrategia de selección optimizada, la cual extrae las cadenas con la mejor entropía relativa de un grupo de secuencias cuidadosamente seleccionadas, utilizándolas después en el cálculo de las distancias evolutivas. Este procedimiento fue

aplicado exitosamente para subtipificar el virus de inmunodeficiencia humana tipo 1 (HIV-1, por sus siglas en inglés), usando cadenas de 5 a 9 nucleótidos.[16] El método de selección ha sido mejorado estadísticamente[17] y usado para analizar virus de DNA de doble cadena.[18]

Otro método de distancia independiente de MSA, implica la estimación de las distancias de *k*-tuplos entre secuencias. La distancia de *k*-tuplos entre dos secuencias, se refiere a la suma total de diferencias en frecuencia, sobre todos los tuplos posibles de longitud *k*, entre las secuencias. La frecuencia de 2-tuplos permitió construir un árbol biológicamente plausible con genomas mitocondriales.[19] Esta estrategia también ha sido aplicada empleando cadenas de aminoácidos.[20-21] Debido a la cantidad de información a procesar, en este enfoque se requieren relativamente cantidades grandes de memoria y capacidad de procesamiento. Consecuentemente, en la práctica los valores de *k* utilizados han sido establecidos en longitudes relativamente pequeñas, como 5 y 6. Varios programas que calculan MSA (por ejemplo, *MUSCLE*, *CLUSTAL W* y *Kalign*) primero computan una matriz de distancias de *k*-tuplos con las secuencias de interés y, posteriormente, usan algoritmos como NJ para construir rápidamente un "árbol guía", el cual determina el orden en que las secuencias serán alineadas. No obstante, los árboles guía raramente se usan como árboles finales, y

para tal propósito regularmente se emplean otros programas como *PHYLIP*[22] o *PAUP*.[23]

Recientemente, se comprobó que un método de distancia de 5-tuplos (que utiliza el conjunto completo de tuplos de dicha longitud) mejora, en la mayoría de los casos, el desempeño de otros estimadores de distancia, duplicando o más la precisión de éstos. En dicho estudio se emplearon 1470 conjuntos simulados de secuencias (generadas bajo diferentes escenarios evolutivos), así como los árboles NJ construidos con tales conjuntos, y se comparó el desempeño de la distancia de *k*-tuplos con el de cuatro estimadores de distancia comúnmente empleados (Jukes-Cantor, Kimura, F84 y Tamura-Nei). Los resultados mostraron que los árboles construidos a partir de la distancia de *k*-tuplos son más precisos, la mayoría del tiempo, que aquellos construidos con las otras distancias. [3]

Por otro lado, la hibridación virtual (VH, por sus siglas en inglés) se fundamenta en algunas consideraciones relacionadas con el DNA como complementariedad, desnaturalización y renaturalización, síntesis secuencia-específica de oligonucleótidos, entre otras. El concepto general de la VH implica la simulación por computadora de eventos de hibridación entre moléculas de DNA.[24] Previamente, se verificó una correlación razonable entre valores de ΔG^{o}, calculados por VH para los rDNA 16S de 7 diferentes especies microbianas (1 *Bacillus* y 6 *Pseudomonas*), y la intensidad de las señales de

hibridación, obtenidas para estos mismos rDNA 16S, a través de un sistema de hibridación por chips de DNA.[25] Se concluyó que las hibridaciones, tanto perfectas como ambiguas, contribuyen a la identificación microbiana a través de la obtención de la huella genómica por hibridación. En otro trabajo, se identificaron exitosamente tipos y subtipos de papillomavirus humano empleando VH.[26]

En este estudio, se propone y caracteriza *LifePrint*, un método de distancia de *k*-tuplos independiente de MSA, que sólo usa una muestra representativa de todos los tuplos posibles de una determinada longitud *k*.

JUSTIFICACIÓN

La construcción típica de árboles a través de métodos de distancia requiere de MSA, lo que conlleva algunas desventajas. Proponer un método de distancia sin MSA, supone una alternativa para conservar las ventajas de implementación y rapidez inherentes al criterio de distancia, acotando algunas de las desventajas que suponen los MSA.

HIPÓTESIS

Usando un método de distancia de *k*-tuplos, que sólo use una muestra representativa de todos los tuplos posibles de una determinada longitud *k*, es posible inferir filogenias precisas de genomas completos sin calcular MSA.

OBJETIVO GENERAL

Proponer y caracterizar un método de distancia de *k*-tuplos que estime con precisión filogenias de genomas completos usando solamente una muestra representativa de todos los tuplos posibles de una determinada longitud *k*.

OBJETIVOS ESPECÍFICOS

- Diseñar el Conjunto de 9-tuplos de *LifePrint* (LPS9, por sus siglas en inglés)
- Identificar el esquema de búsqueda óptimo para usar el LPS9
- Evaluar la cobertura genómica que realiza el LPS9
- Evaluar la capacidad del LPS9 para detectar repetidos nucleotídicos

- Calcular las distancias de *k*-tuplos entre genomas de viroides, usando tres diferentes métricas de distancia
- Determinar el intervalo dinámico del LPS9
- Comparar la precisión de los árboles construidos con *LifePrint* y el método de 5-tuplos reportado previamente

II. MATERIALES Y MÉTODOS

LPS9

Con base en análisis previos se observó que los 9-tuplos muestran un funcionamiento óptimo para estudiar genomas de viroides.[27] Para calcular las distancias de *k*-tuplos se escrutaron genomas reales y simulados de viroides a través de búsquedas de similitud usando un conjunto de 1878 9-tuplos (el LPS9). Cada tuplo de dicho conjunto es distinto por lo menos en dos diferencias nucleotídicas internas y no contiguas. El LPS9 es una muestra representativa de todos los tuplos posibles de longitud 9, es decir, 262,144 tuplos (4^9 tuplos). La Figura 1 ilustra la distribución del LPS9 a lo largo del conjunto completo de 262,144 9-tuplos. El LPS9 está disponible para su uso en la página electrónica del *UFC Applications Server*.[28]

Se diseñó el LPS9 con el programa *UFC designer* (Méndez-Tenorio y colaboradores, sin publicar). El programa selecciona tuplos con secuencias altamente discriminatorias a través de una estrategia de agrupamiento propuesta previamente.[24]

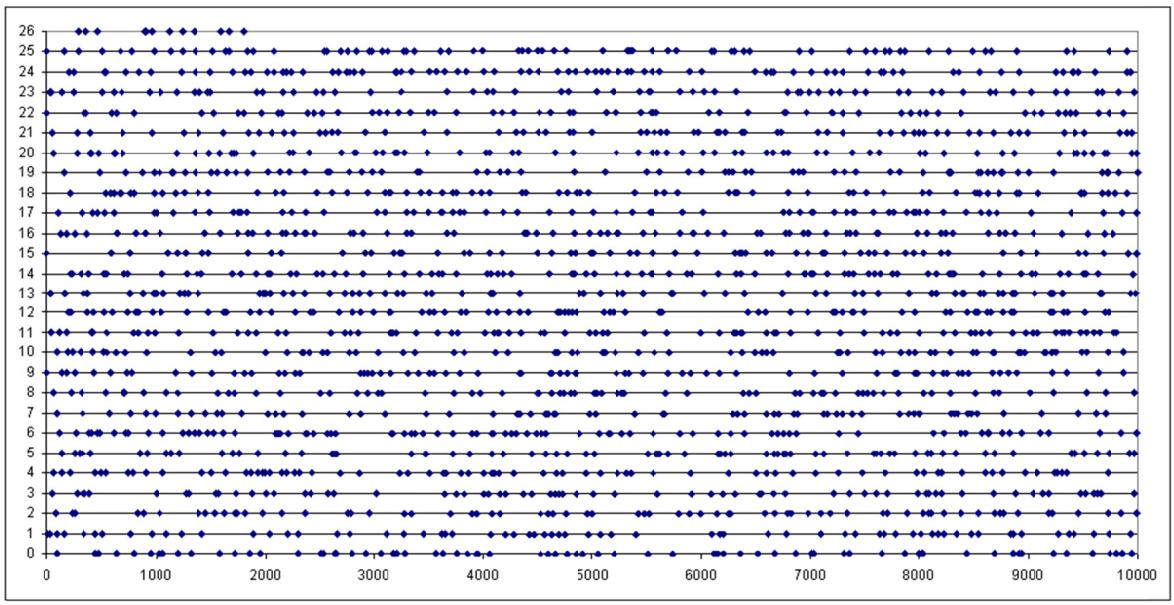

Figura 1. Distribución del LPS9 dentro del conjunto completo de 9-tuplos. Los 1878 tuplos del LPS9 están representados gráficamente (puntos negros) de acuerdo con sus posiciones dentro de la lista original de todos los 9-tuplos posibles (262,144). Cada línea representa 10,000 tuplos. Elaborada por el Dr. Hueman Jaimes-Díaz.

Dicha estrategia considera un grupo como un conjunto formado por todos los tuplos que comparten una característica en particular (por ejemplo, un determinado número de diferencias) con relación a un tuplo de referencia (la "marca" del grupo) seleccionado al azar. A continuación, se describe brevemente la secuencia de tres criterios implicada en la estrategia de agrupamiento con la que se diseñó el LPS9. 1) Criterio de sustitución. Se generaron grupos con tuplos que sólo compartían una diferencia con las marcas. Posteriormente se seleccionó el conjunto que incluía las marcas del tipo mencionado, es decir, todos los tuplos con al menos dos diferencias entre ellos. 2) Criterio de bloque. Se generaron grupos y, se excluyeron los tuplos

que compartían dos diferencias, ubicadas en los extremos con relación a las marcas. Para ejecutar este procedimiento, primero se agruparon los tuplos muy similares a las marcas y, finalmente, se seleccionaron aquellos que compartían con ellas un mismo bloque interno de 7 nucleótidos. El resultado fue el conjunto con las marcas que presentaban diferencias internas. 3) Criterio de refinación. Bajo este criterio, se agruparon en primera instancia, los tuplos muy similares a las marcas y posteriormente, se excluyeron los que compartían con ellas diferencias contiguas, seleccionándose de esta manera las marcas con diferencias no contiguas.

Antes de aplicar cada uno de los tres criterios, el programa *UFC designer* aleatoriza los conjuntos de tuplos para disminuir los potenciales errores (*bias*) de muestreo durante la ejecución de la estrategia de agrupamiento.

Las secuencias y el número de tuplos seleccionados dependen del proceso de aleatorización; en consecuencia el LPS9 diseñado no es único. Sin embargo, cualquier otro LPS9 diseñado usando la estrategia de agrupamiento antes mencionada, conduciría a resultados similares.

SECUENCIAS GENÓMICAS

En este trabajo se usaron genomas reales y simulados de viroides. La pequeña longitud (≈300 nucleótidos) de dichos genomas facilitó llevar a cabo los análisis descritos más adelante. Ver el Apéndice 1 para los números de acceso del NCBI (*National Center for Biotechnology Information*) de los 36 genomas reales.

Para comparar la precisión entre *LifePrint* y el método de 5-tuplos se utilizó un conjunto de genomas simulados obtenido como se describe a continuación. Usando el programa *EvolSeq* (disponible para su uso en la página electrónica del *UFC Applications Server*) fue simulada la evolución de 32 genomas (nombrados desde CVII31 hasta CVII62) derivados de un ancestro común (*Citrus viroid II*). La simulación consideró un esquema evolutivo de 5 generaciones basado en un modelo de sustitución con una proporción transiciones/transversiones de dos, como la definida en el modelo de 2-parámetros de Kimura (K2P, por sus siglas en inglés).

BÚSQUEDA DE SIMILITUD

Se usó el programa de VH para escrutar los genomas reales y simulados con el LPS9 en busca de identidad y/o similitud (detección), es decir, para llevar a cabo las búsquedas de similitud. El

programa de VH localiza sitios termodinámicamente estables para la hibridación de los *k*-tuplos dentro de las secuencias genómicas.[25] Dado que la información termodinámica no es relevante para este estudio, se estableció un valor de corte de 0 para la energía libre. La VH produce dos diferentes salidas: una lista detallada de las posiciones en las cuales cada tuplo es localizado en la secuencia genómica y una tabla global con la frecuencia con la que cada tuplo es localizado en la secuencia genómica (tabla global de frecuencia). Alternativamente, la tabla global puede mostrar sólo la presencia (1) o ausencia (0) de los tuplos en las secuencias (tabla global binaria). El programa de VH está disponible para su uso en la página electrónica del *UFC Applications Server*.

Para identificar el esquema óptimo de búsqueda de similitud, se compararon cuatro condiciones permitiendo diferentes números de diferencias entre las secuencias de los tuplos del LPS9 y los 36 genomas reales: a) sin diferencias, b) sin diferencias y permitiendo una diferencia, c) sin diferencias y permitiendo hasta dos diferencias, y d) sin diferencias y permitiendo hasta tres diferencias.

El esquema óptimo identificado fue utilizado para el resto de los análisis llevados a cabo con el LPS9. En el caso del método de 5-tuplos sólo se permitieron identidades.

COBERTURA GENÓMICA

Se usó el LPS9 como *query* para llevar a cabo búsquedas de similitud y evaluar la capacidad del LPS9 para cubrir completamente las secuencias genómicas. La cobertura genómica del LPS9 que se muestra en la Figura 2 evidencia la secuencia consenso producida al apilar los 9-tuplos de acuerdo a su posición de detección a lo largo de los primeros 80 nucleótidos localizados en el extremo 5' del genoma del *Hop stunt viroid* (longitud de 302 nucleótidos). Las identidades y diferencias se muestran en letras mayúsculas y minúsculas, respectivamente. Para este análisis representativo se seleccionó arbitrariamente el genoma del *Hop stunt viroid*, pero fueron obtenidos resultados similares con el análisis de otros genomas.

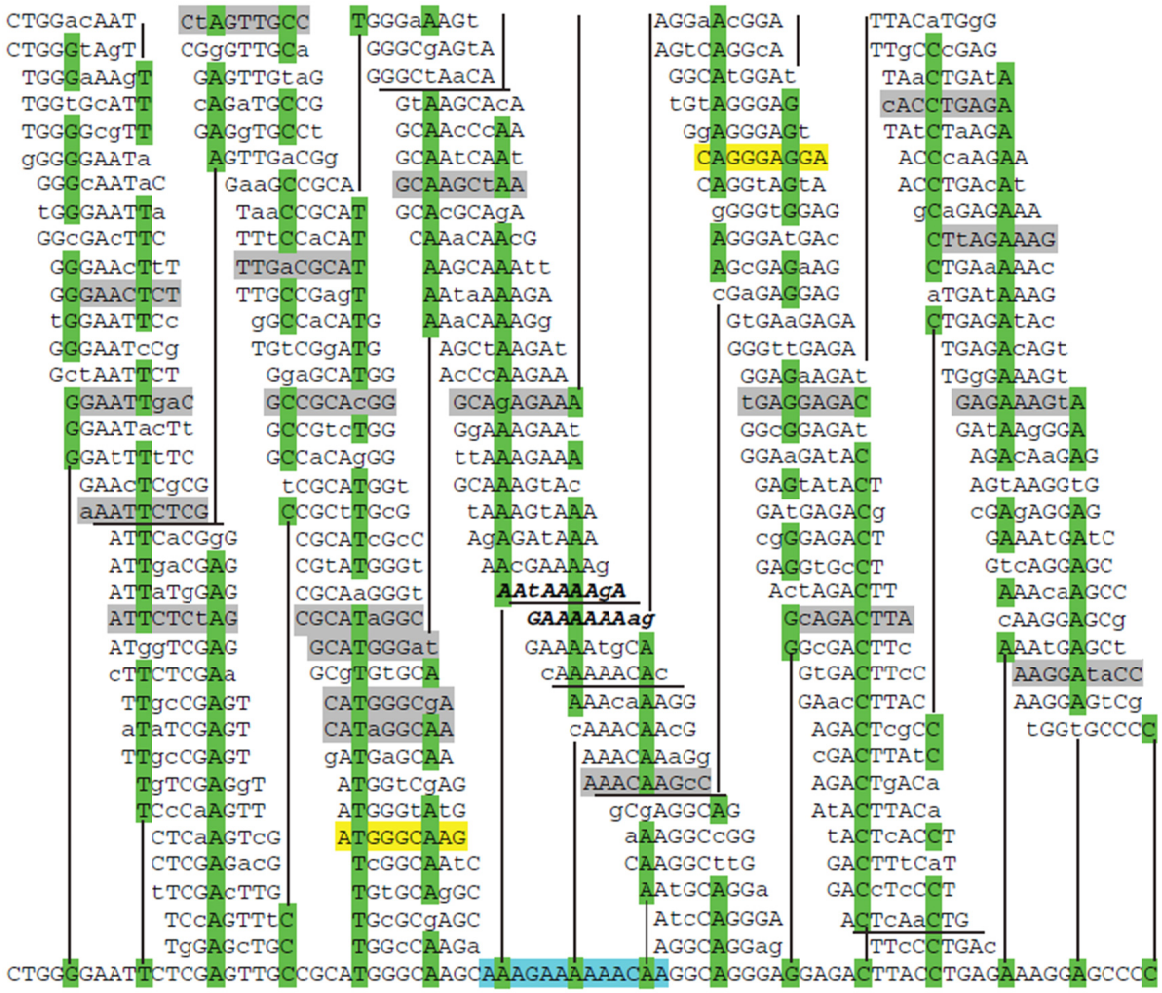

Figura 2. Cobertura genómica. De acuerdo a su posición de detección se apilaron los tuplos del LPS9 en los primeros 80 nucleótidos (extremo 5') del genoma del *Hop stunt viroid*. Las coincidencias entre los nucleótidos más frecuentes y su respectiva posición genómica son indicadas en cada columna. Cada cinco nucleótidos se hizo una marca verde como referencia al número del nucleótido. Las identidades y diferencias se muestran en letras mayúsculas y minúsculas, respectivamente. Los tuplos que hallaron identidades o sitios con una diferencia, están marcados en amarillo y gris, respectivamente. Las seis cadenas que no fueron detectadas directamente por algún tuplo (comenzando en los nucleótidos 7, 27, 36, 39, 42 y 60) están subrayadas. Tres de ellas (nucleótidos 36, 39 y 42) están localizadas en la región con elevada composición de adeninas, la cual está marcada en azul. Elaborada por el Dr. Rogelio Maldonado-Rodríguez.

DETECCIÓN DE REPETIDOS NUCLEOTÍDICOS

Con el propósito de evaluar la capacidad del LPS9 para detectar repetidos nucleotídicos (homopolímeros) se usó como modelo la

secuencia

AAAAAAAAATTTTTTTTTCCCCCCCCCGGGGGGGGG. La Figura 3 muestra los 9-tuplos del LPS9 de acuerdo a su posición de detección a lo largo del modelo antes mencionado.

BOOTSTRAP DE LAS DISTANCIAS DE *K*-TUPLOS

Usando como entrada las tablas (globales de frecuencia y binaria) generadas por el programa de VH, el programa *Characters* calcula independientemente diferentes tipos de distancias de *k*-tuplos. *Characters* produce una matriz con las distancias de *k*-tuplos originales (archivo *original*) y una segunda salida que comprende las réplicas de *bootstrap* generadas a partir de la matriz original (archivo *bootstrap*). El programa *Characters* está disponible para su uso en la página electrónica del *UFC Applications Server*. Las estrategias generales para el *bootstrap* y la construcción de árboles están ilustradas en la Figura 4.

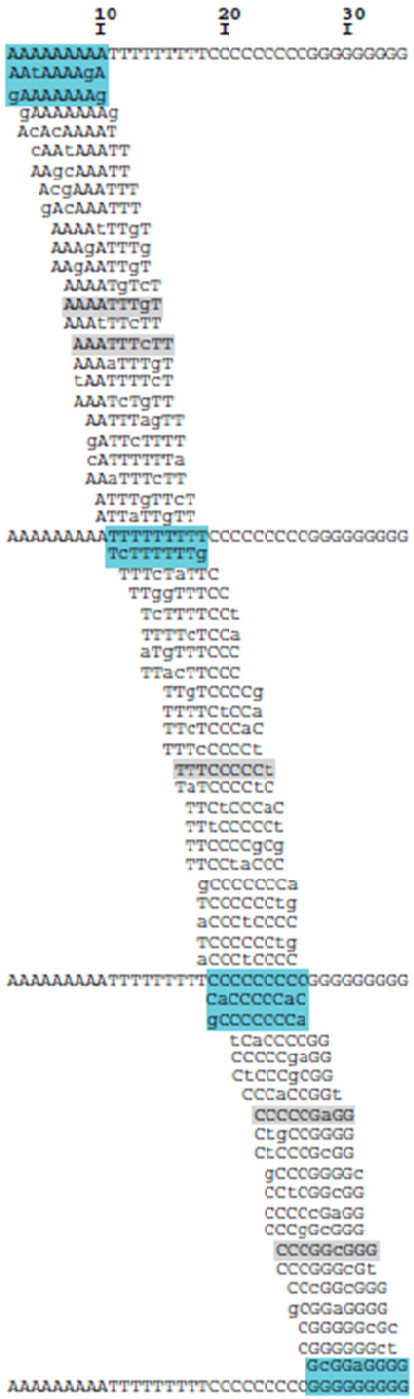

Figura 3. Detección de repetidos nucleotídicos por el LPS9. Se llevó a cabo una búsqueda de similitud usando un modelo de 36 nucleótidos que contiene cuatro diferentes repetidos de un sólo nucleótido. De acuerdo a sus posiciones de detección se apilaron los tuplos del LPS9 a lo largo del modelo. Las identidades y diferencias se muestran en letras mayúsculas y minúsculas, respectivamente. Los tuplos capaces de detectar directamente el repetido en el modelo y la cadena detectada están marcados en gris. Elaborada por el Dr. Rogelio Maldonado-Rodríguez.

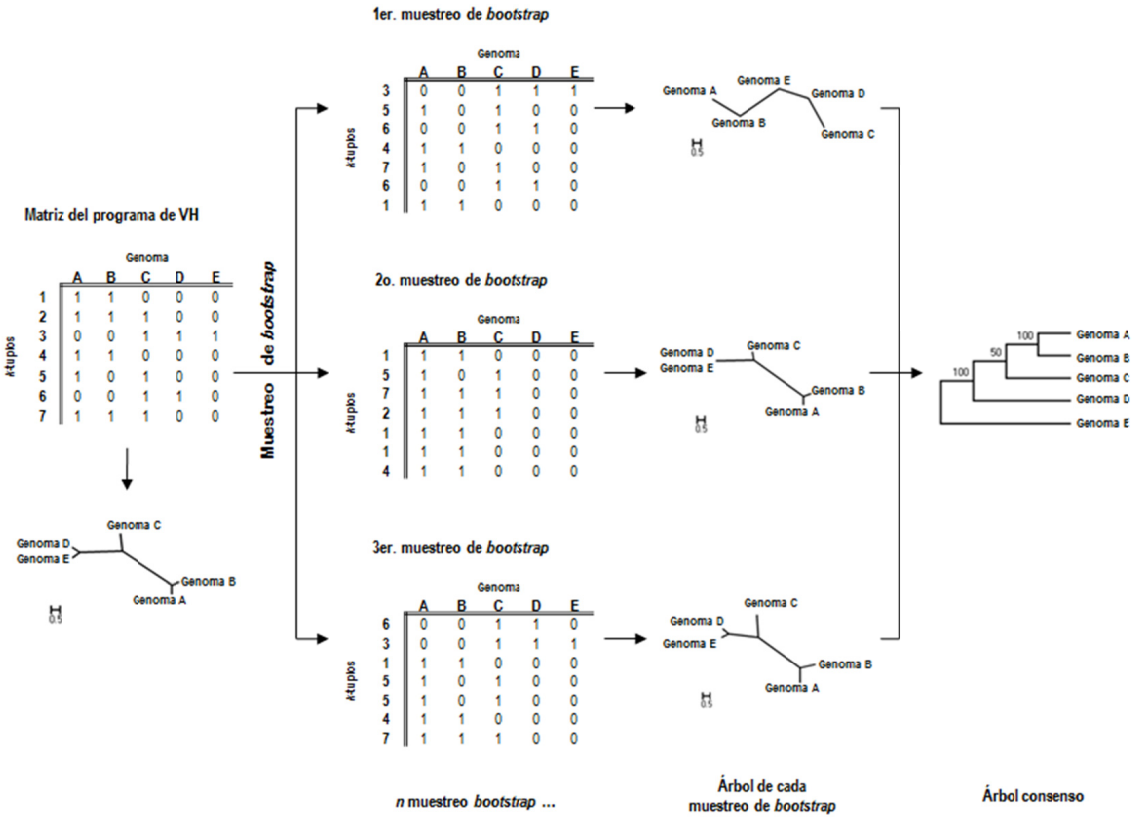

Figura 4. Esquema del *bootstrap* y la construcción de árboles en *LifePrint*. El programa de VH genera una matriz con la detección que cada tuplo (renglones 1 a 7) hace dentro de cada genoma (columnas A a E). Posteriormente los renglones completos de la matriz original son muestreados aleatoriamente con sustitución para producir una nueva matriz de *bootstrap* con el mismo número de renglones que la matriz original. Para cada matriz de *bootstrap* se calcula una tabla de distancias de *k*-tuplos que se usa para construir un árbol. Finalmente se calcula un árbol consenso a partir de los árboles de *bootstrap*. Los números en el árbol consenso muestran el porcentaje de abundancia de los grupos en las muestras de *bootstrap*, es decir, los valores de soporte del *bootstrap*. Modificada a partir de una figura elaborada por la M. en C. Janet Casique-Almazán.

Se asume que diferentes métricas de distancia tendrán diferentes precisiones inherentes para la estimación filogenética. Se usaron tres diferentes métricas para calcular las distancias de *k*-tuplos entre los genomas de viroides. 1) La distancia logarítmica de *k*-tuplos basada en el índice de Jaccard (dLog). En este caso las distancias basadas en el índice de Jaccard sólo consideran los tuplos compartidos entre las

secuencias genómicas, y las distancias son independientes de la frecuencia de tuplos en dichas secuencias. 2) En contraste la distancia de *k*-tuplos basada en el coeficiente de correlación de Pearson (dPear) toma en cuenta la frecuencia de los tuplos. 3) Finalmente, la distancia de *k*-tuplos típica (dk) está basada en la frecuencia de aparición de los tuplos en las secuencias genómicas y considera la longitud de éstas.[2]

La dLog dados dos genomas, A y B, fue calculada en dos pasos. Primero, se calculó el índice de Jaccard (también conocido como coeficiente de similitud de Jaccard) usando la fórmula:

$$J = \frac{M_{11}}{M_{01}+M_{10}+M_{11}},$$

donde J es el índice de Jaccard; M_{11} es el número total de tuplos que aparecen en ambos genomas A y B; M_{01} es el número total de tuplos distintivos para B; M_{10} es el número total de tuplos distintivos para A. Segundo, el valor de la dLog fue computado usando la fórmula:

$$\text{dLog}\,(A,B) = -\ln\frac{J}{k},$$

donde J es el índice de Jaccard; k es la longitud del tuplo.

La dPear dados dos genomas, A y B, fue calculada usando la fórmula:

$$\text{dPear} = 1 - r = 1 - \sum_i (A_i - \overline{A})\,(B_i - \overline{B}) / (\sqrt{\sum_i (A_i - \overline{A})^2}\,\sqrt{\sum_i (B_i - \overline{B})^2}),$$

donde r es el coeficiente de correlación de Pearson; A_i y B_i corresponden a las frecuencias del tuplo i en los genomas A y B, respectivamente. $\bar{A}$ y $\bar{B}$ corresponden al promedio de las frecuencias en los genomas A y B, respectivamente.

La dk dados dos genomas, A y B, fue calculada usando la fórmula:

$$\mathrm{dk}(A,B) = \sum_{i=1}^{4^k} |A_i - B_i|^2,$$

donde A_i y B_i corresponden a las frecuencias del tuplo i $(= counts/n - k + 1)$ en los genomas A y B, respectivamente; n es la longitud de secuencia de cualesquiera de los genomas A o B; k es la longitud del tuplo.

En los casos del LPS9 y los 5-tuplos se usó el programa *Characters* con la finalidad de calcular los tres diferentes tipos de distancias de *k*-tuplos entre los 32 genomas simulados, lo que se llevó a cabo utilizando las tablas globales (frecuencia y binaria) provenientes del programa de VH.

Utilizando la tabla global de frecuencia de los 36 genomas reales y calculando las dPear, se generaron los archivos de *bootstrap* (1000 réplicas) con el LPS9 y los 5-tuplos.

INTERVALO DINÁMICO

En este análisis se seleccionó la dLog para identificar los tuplos compartidos entre los genomas independientemente de la frecuencia en la que están presentes en los genomas y de la longitud de éstos.

Se evaluó el intervalo dinámico del LPS9, es decir, la capacidad de dicho conjunto para calcular valores de dLog que permitieran establecer una distinción clara entre un grupo de genomas con un amplio intervalo de similitud. Con esta finalidad se usó el genoma del *Citrus viroid II* como referencia para simular conjuntos con 100 variantes genómicas cada uno.

La capacidad de distinguir variantes se evaluó utilizando dos enfoques de sustitución simulada: el "independiente" y el "sucesivo". Las sustituciones fueron simuladas usando *scripts* implementados con el lenguaje de programación *Perl* (*Active Perl 5.8*). En el enfoque independiente después de introducirse las sustituciones en el genoma de referencia se calcularon las dLog entre cada una de las variantes generadas y el genoma de referencia. En el segundo enfoque se introdujeron de manera sucesiva y acumulada las sustituciones simples, y posteriormente se calcularon las dLog entre cada par formado por las variantes nueva y previa. Para ambos enfoques fueron registrados los valores mínimo, máximo y promedio de las dLog.

Para entender mejor los casos en los cuales las sustituciones están localizadas en los extremos 5' y/o 3', se calcularon las dLog para dos conjuntos de variantes. El primer conjunto estaba formado por variantes con sustituciones simples en cada una de las nueve posiciones a partir de los extremos 5' o 3'. El segundo conjunto estaba formado por variantes con eliminaciones acumuladas en el extremo 3'.

El intervalo dinámico también hace referencia a los valores límite de similitud entre dos secuencias que pueden ser interpretados para distinguirlas. Para evaluar esta propiedad se simularon 15 grupos de 100 variantes genómicas del *Citrus viroid II*, con una proporción creciente de sustituciones. Cada variante fue simulada con el efecto acumulado de sustituciones simples sucesivas y al azar. Los grupos fueron simulados con la introducción de 1(0.5%), 3 (1%), 6 (2%), 9 (3%), 12 (4%), 24 (8%), 36 (12%), 48 (16%), 60 (20%), 72 (24%), 84 (28%), 96 (32%), 120 (40%), 150 (50%) y 200 (66%) sustituciones, respectivamente. Los números entre paréntesis son porcentajes de sustitución en relación con la longitud de las variantes (300 nucleótidos). También se calculó el número real de sustituciones acumuladas en cada variante empleando otro *script* de *Perl* que es capaz de llamar a los módulos *water* y *needle* del programa *EMBOSS*.[29] Después calculamos el número promedio de sustituciones acumuladas para cada grupo. Finalmente, de los 15

grupos se registraron los valores mínimo, máximo y promedio de las dLog.

CONSTRUCCIÓN DE ÁRBOLES

Como se describió anteriormente en la sección de *SECUENCIAS GENÓMICAS*, se simuló la evolución de 32 genomas utilizando como ancestro común al *Citrus viroid II*. Después se usaron los resultados derivados de esta simulación para preparar manualmente la representación *Newick* del "árbol verdadero", la cual fue requerida para otros análisis. Se visualizó y editó dicho árbol (Figura 5) usando el programa *MEGA 4.0*.[30]

Los 32 genomas simulados se usaron para construir árboles con *LifePrint* y el método de 5-tuplos. Todos los árboles en este trabajo se construyeron sin raíz. Los árboles fueron estimados para cada tabla global (de frecuencia y binaria) usando el algoritmo NJ implementado en el módulo *neighbor* del programa *PHYLIP 3.69*.

Utilizando los dos archivos de *bootstrap* de los 36 genomas reales se construyeron los árboles NJ usando el módulo *neighbor* del programa *PHYLIP 3.69*. Posteriormente usando el programa *MEGA 4.0* se estimaron los dos árboles consenso (Figuras 6-7) y se llevó a cabo su visualización y edición.

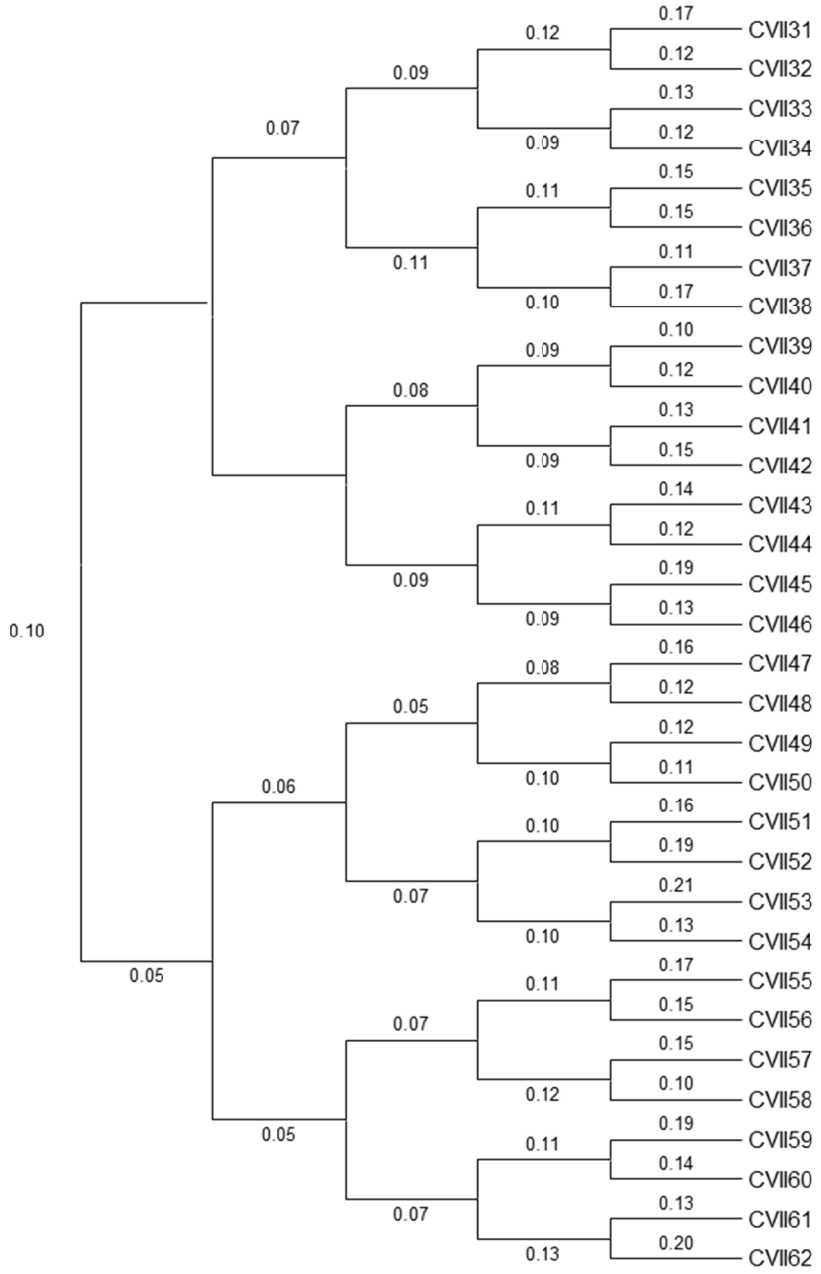

Figura 5. Árbol verdadero. El árbol verdadero fue manualmente construido usando como referencia la evolución simulada de 32 genomas derivados del *Citrus II viroid*. La simulación se hizo siguiendo un esquema de 5 generaciones y considerando un modelo de sustitución con una proporción transiciones/transversiones de dos, como la definida en el modelo K2P. El valor sobre las ramas corresponde a la distancia real entre los clados. Elaborada por el Dr. Alfonso Méndez-Tenorio

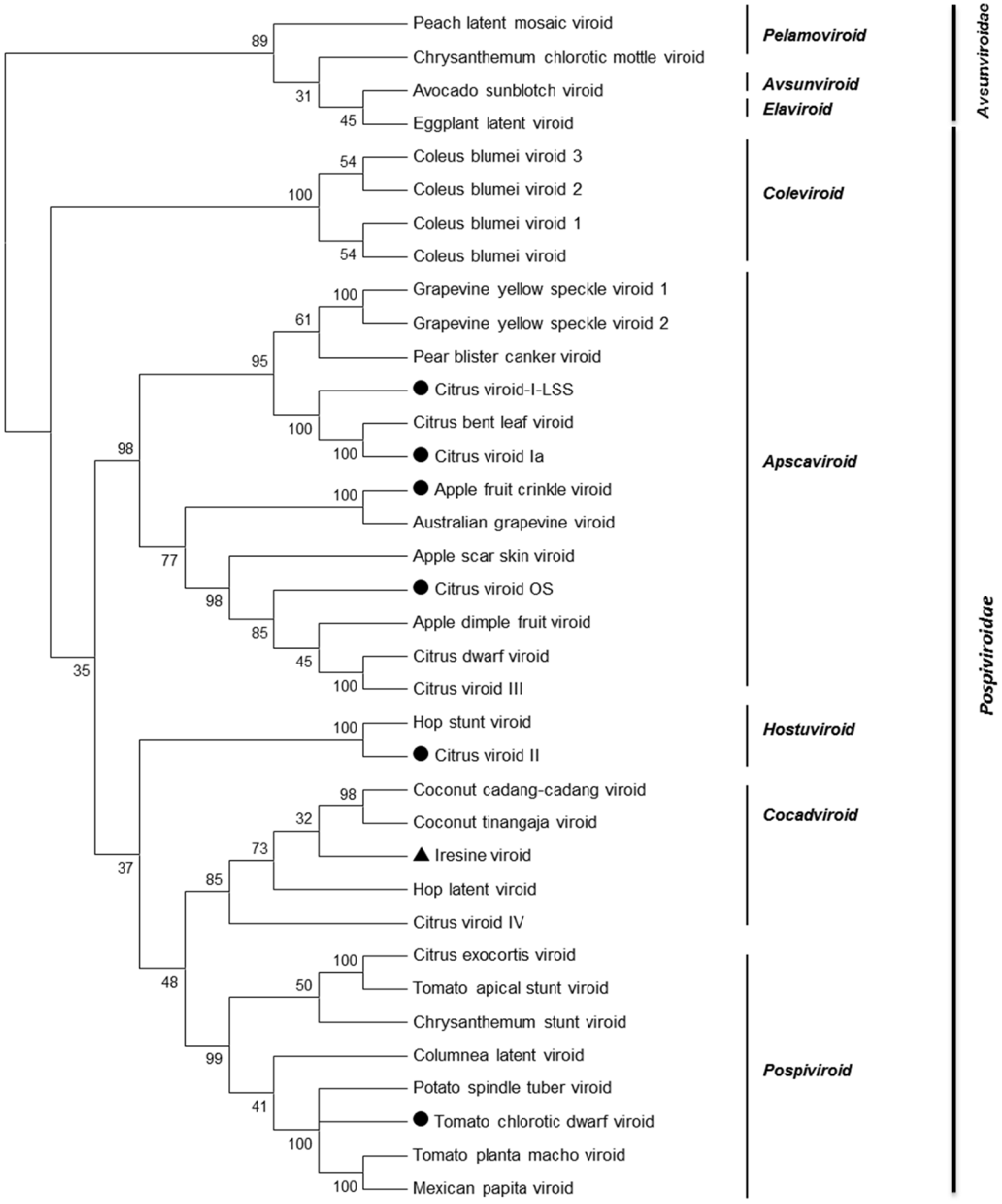

Figura 6. Árbol consenso construido con *LifePrint* para 36 genomas reales (dPear, 1000 réplicas). Las familias fueron asignadas de acuerdo a la filogenia propuesta por el Comité Internacional de Taxonomía de Virus (ICTV, por sus siglas en inglés). Los números representan valores de confianza de bootstrap para los clados inferidos. Los círculos negros corresponden a genomas de viroides no clasificados. El triángulo negro corresponde a un genoma que debía agruparse correctamente en la subfamilia *Pospoviroid*. Elaborada por la M. en C. Janet Casique-Almazán.

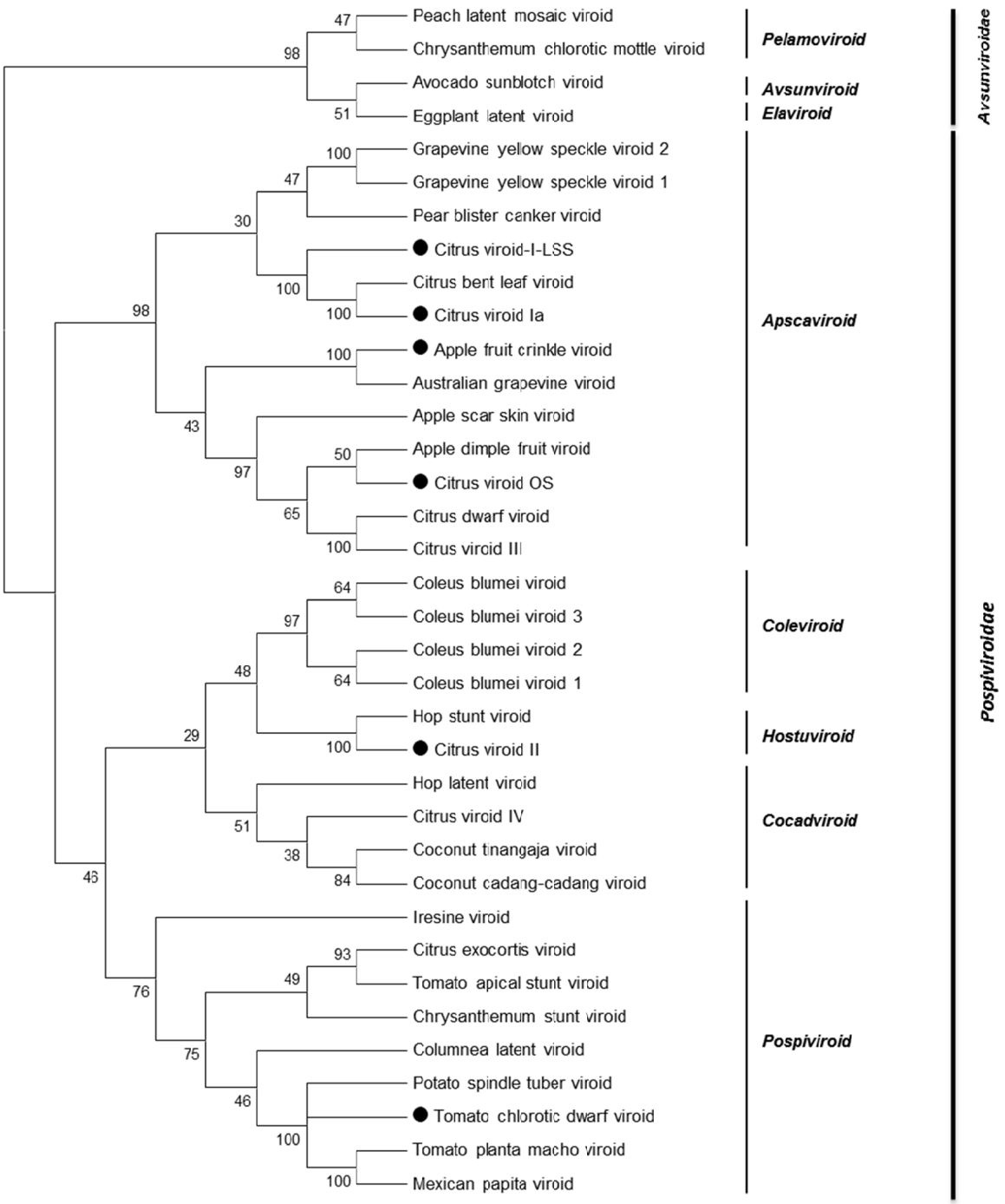

Figura 7. Árbol consenso del método de los 5-tuplos para 36 genomas reales (dPear, 1000 réplicas). Las familias fueron asignadas de acuerdo a la filogenia propuesta por el Comité Internacional de Taxonomía de Virus (ICTV, por sus siglas en inglés). Los números representan valores de confianza de bootstrap para los clados inferidos. Los círculos negros corresponden a genomas de viroides no clasificados. Elaborada por la M. en C. Janet Casique-Almazán.

EVALUACIÓN DE LA PRECISIÓN

Se usó la diferencia simétrica (SD, por sus siglas en inglés) entre árboles para evaluar la precisión de diferentes métodos de construcción de árboles. La SD o distancia topológica es una métrica ampliamente utilizada para evaluar diferencias entre árboles. Debe mencionarse que la estimación de la SD es independiente de la longitud de las ramas de los árboles. Dados dos árboles bifurcados sin raíz la SD es el doble del número de ramificaciones interiores en el cual las particiones son diferentes entre los árboles comparados.[31] Otra definición equivalente de la SD es la de la distancia entre pares de árboles basada en el número de ramificaciones en las que difieren ambos árboles. Así, la SD es simplemente una cuantificación de las particiones entre los dos árboles comparados.[32] La SD entre los árboles se calculó usando el módulo *treedist* incluido en el programa *PHYLIP 3.69*.

Brevemente, se condujo una comparación pareada de la topología del árbol verdadero, contra cada una de las topologías de los 12 árboles NJ construidos con los 32 genomas simulados.

Se utilizó el programa *phylo-comparison*[33] para comparar las topologías del árbol verdadero y las de los dos árboles NJ (*LifePrint* y método de los 5-tuplos) construidos con las dPear calculadas entre los 32 genomas simulados.

Finalmente, se discute acerca de los valores de soporte del *bootstrap* entre los árboles NJ consenso obtenidos para los 36 genomas reales y la filogenia de los viroides propuesta por el Comité Internacional de Taxonomía de Virus (ICTV, por sus siglas en inglés).[34]

III. RESULTADOS

LPS9

Los criterios de agrupamiento generaron los siguientes conjuntos. Después del criterio de sustitución el número de tuplos disminuyó de 262,144 a 29,868. El criterio de bloque redujo el número hasta 4206 y finalmente el criterio de refinamiento identificó el conjunto final de 1878 tuplos, es decir, el LPS9.

En la Figura 1 cada línea representa 10,000 tuplos. El número de tuplos del LPS9 en cada línea varía desde 64 hasta 84. En promedio un tuplo del LPS9 es seleccionado por cada 144 del conjunto completo de 262,144.

BÚSQUEDA DE SIMILITUD

La Tabla 2 compila el número promedio de tuplos que hicieron detección con los 36 genomas reales bajo cuatro diferentes condiciones. El esquema óptimo para el LPS9 fue la condición c (permitiendo hasta dos diferencias).

Tabla 2 Número de tuplos del LPS9 que reconocieron zonas en los 36 genomas reales, bajo cuatro diferentes esquemas de búsqueda de similitud

Condiciones	Diferencias permitidas	Número promedio de tuplos del LPS9 que hicieron detección	Proporción en relación al número total de tuplos del LPS9
a	0	3	0.2
b	0 y 1	64	3.4
c	0, 1 y 2	605	32.2
d	0, 1, 2 y 3	1705	90.8

Notas: Se llevaron a cabo búsquedas de similitud entre el LPS9 y los 36 genomas reales permitiendo diferente número de diferencias entre sus secuencias. Se calculó el promedio de tuplos que hicieron detección bajo cuatro diferentes condiciones.

Cada uno de los tuplos del LPS9 escruta 352 diferentes cadenas 9-mer (1 idéntica, 27 permitiendo 1 diferencia y 324 permitiendo 2 diferencias) dentro de los genomas, lo que significa un número total de 661,056 (1878*352) cadenas. Ya que todas las posibles 9-mer son 262,144, es de esperarse que cada cadena sea escrutada en promedio 2.5 veces. De hecho se observó que un genoma con longitud de 300 nucleótidos, es decir, que contiene 292 cadenas 9-mer, fue cubierto en promedio por 605 tuplos del LPS9, lo que corresponde a 2.05 tuplos por cadena.

COBERTURA GENÓMICA

Dados el esquema óptimo de búsqueda de similitud y el LPS9, con *LifePrint* los genomas analizados son cubiertos completamente.

Observando la Figura 2 puede apreciarse que en los primeros 80 nucleótidos del *Hop stunt viroid* seis cadenas 9-mer no fueron

detectadas directamente por el LPS9. La región con menor cobertura (del nucleótido 34 al 46, marcada en azul) tiene una elevada composición de A. Además, esta región tiene tres de las seis cadenas que no fueron detectadas directamente por el LPS9 (comenzando en los nucleótidos 32, 39 y 42). Sólo dos tuplos (AATAAAAGA y GAAAAAAAG) compartieron similitud con esta región.

DETECCIÓN DE REPETIDOS NUCLEOTÍDICOS

Los resultados obtenidos con el modelo de repetidos nucleotídicos muestran (Figura 3) que dos tuplos con siete A, 1 con siete T, 2 con siete C y 1 con siete G (todos en negritas en la Figura 3) fueron capaces de detectar directamente cadenas de 9 nucleótidos consecutivos e idénticos.

INTERVALO DINÁMICO

La Tabla 3 muestra los resultados obtenidos para los enfoques independiente y sucesivo. Bajo el enfoque independiente obtuvimos un valor de 0 en una variante con una sustitución en el extremo 3'.

El análisis de variantes con sustituciones o eliminaciones ubicadas en los extremos 5' o 3', nos reveló que sólo en pocas ocasiones se presentan cambios puntuales no detectados directamente por el

LPS9. De esta manera, la distinción de variantes no fue afectada. Los resultados de las variantes recién mencionadas se muestran en la Tabla 4, en donde la dLog se calculó entre las variantes y el genoma del *Citrus II viroid*.

Tabla 3 Valores de dLog para variantes de sustituciones simples

Valores de dLog	Enfoque independiente	Enfoque sucesivo
Mínimo	0.000000	0.00040
Máximo	0.005894	0.00580
Promedio	0.003780	0.00390

Notas: Se calcularon los valores mínimo, máximo y promedio de la dLog para variantes del *Citrus II viroid* obtenidas empleando los enfoques independiente y sucesivo.

Tabla 4 Valores de dLog para variantes con sustituciones simples o eliminaciones ubicadas en los extremos de sus secuencias

Posición a partir del extremo	Valores de dLog			Número de nt eliminados en el extremo 3'
1	0.00000	0.00069	0.000000	0
2	0.00069	0.00117	0.000380	1
3	0.00082	0.00179	0.000580	2
4	0.00137	0.00248	0.000580	3
5	0.00145	0.00331	0.000960	4
6	0.00255	0.00324	0.000962	5
7	0.00234	0.00441	0.001349	6
8	0.00381	0.00531	0.001543	7
9	0.00426	0.00552	0.002128	8
	Extremo 5'	**Extremo 3'**		

Notas: Se calcularon las dLog entre cada variante y el genoma del *Citrus II viroid*. En la segunda y tercer columna mostramos los resultados para las tres posibles sustituciones simples en los extremos 5' o 3' donde se realizó la búsqueda de similitud con los tuplos del LPS9. En la cuarta columna están los resultados para el efecto combinado de eliminaciones sucesivas en el extremo 3' y las sustituciones resultantes en los nuevos extremos.

En la Tabla 3 la dLog promedio para una sustitución presenta valores de 0.00378 a 0.00390. Es revelador cuando examinamos la lista de tuplos involucrados en detectar una sustitución simple que implique

una dLog dentro del intervalo mencionado. Con dicha finalidad usamos la variante 144 A→G la cual mostró una dLog de 0.00390 en relación al genoma del *Citrus II viroid*. En la Figura 8 enlistamos los tuplos que detectaron la sustitución de A por G en la posición 144. Obsérvese que 20 tuplos son distintivos en esa posición, 15 para A y 5 para G. Dicha figura ilustra como el LPS9 detecta eficientemente sustituciones simples.

Con el objetivo de estimar los límites en el grado de proximidad que hay entre dos secuencias de tal manera que puedan ser claramente distinguidas, los resultados en la Tabla 5 muestran que la dLog alcanza la saturación aproximadamente cuando un 40% de las secuencia ha experimentado sustituciones.

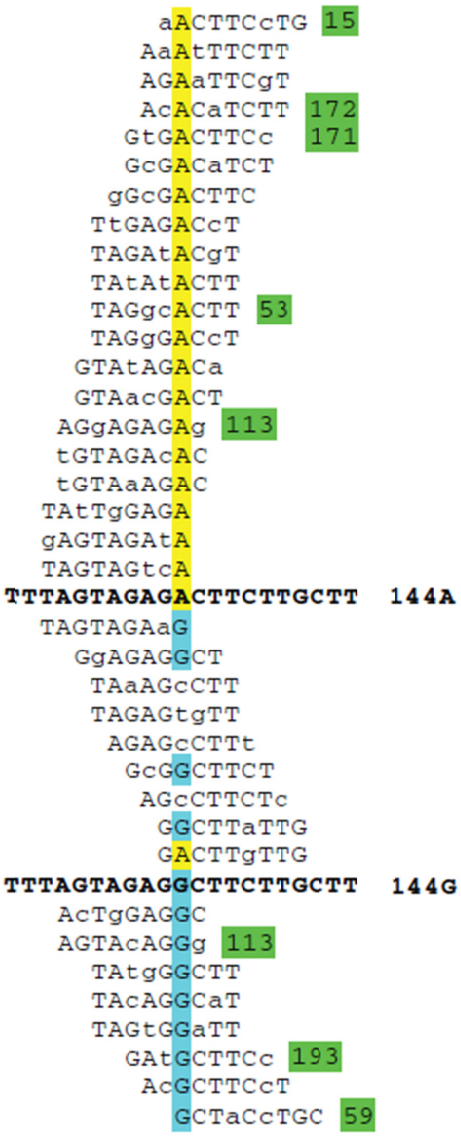

Figura 8. Detección diferencial de una variante con una sustitución simple con un valor de dLog promedio. Se calculó la dLog entre la variante 144 A→G y el genoma del *Citrus II viroid*. Los tuplos del LPS9 que hicieron detección en una cierta región de ambas secuencias están posicionados entre ellas (en negritas), mientras que los tuplos distintivos están colocados arriba o abajo de su respectiva secuencia. La posición de la sustitución está marcada en amarillo cuando hay identidad con A y en azul cuando la identidad es con G. Con verde se marcaron los números de otras posiciones del genoma donde algunos tuplos hicieron detección, lo que impediría considerarlos sin ambigüedad como distintivos. Elaborada por el Dr. Rogelio Maldonado-Rodríguez.

Tabla 5 Capacidad del LPS9 para distinguir entre secuencias con diferente grado de proximidad

Sustituciones simples			Valores de dLog			
Hechas	**Observadas**		**Mínimo**	**Máximo**	**Promedio**	**σ**
1	1.00	0.334	0.00134	0.00589	0.00382	0.00098828
3	2.97	0.993	0.00423	0.01575	0.01150	0.00187987
6	5.96	1.993	0.01437	0.02847	0.02208	0.00287570
9	8.89	2.973	0.02312	0.04112	0.03246	0.00361394
12	11.82	3.953	0.03117	0.04995	0.04205	0.00401709
24	22.82	7.632	0.06276	0.09471	0.07808	0.00704353
36	32.91	11.006	0.09334	0.13755	0.11449	0.01101853
48	42.99	14.378	0.11396	0.21124	0.16224	0.01946713
60	51.61	17.261	0.14709	0.29580	0.21528	0.03098441
72	60.87	20.358	0.17289	0.42443	0.28682	0.04982284
84	68.55	22.927	0.22703	0.76171	0.36894	0.08921467
96	75.29	25.181	0.26546	0.76253	0.45189	0.12901015
120	87.86	29.385	0.30701	0.76664	0.61653	0.14854929
150	99.86	33.398	0.35974	0.76852	0.72943	0.07186317
200	116.55	38.980	0.39897	0.77006	0.76087	0.01373281
	Número	**Porcentaje**				

Notas: Se simularon 15 grupos de 100 variantes del *Citrus II viroid* con un número promedio de sustituciones simples de 1 hasta 116. Calculamos las dLog mínima, máxima y promedio de cada variante con relación al genoma del *Citrus II viroid*. La columna de Porcentaje se calculó dividiendo la columna de Número entre 299, que es la longitud de los genomas.

EVALUACIÓN DE LA PRECISIÓN

En la Tabla 6 se resumen los resultados de las comparaciones de SD entre el árbol verdadero y 12 diferentes árboles NJ.

Tabla 6 Valores de SD entre el árbol verdadero y los árboles NJ construidos a partir de la distancia de *k*-tuplos basada en tres diferentes métricas de distancia

Tipos de distancias de *k*-tuplos	**SD Árbol verdadero *VS LifePrint* (a partir de la tabla global binaria)**	**SD Árbol verdadero *VS LifePrint* (a partir de la tabla global de frecuencia)**	**SD Árbol verdadero *VS* Método 5-tuplos (a partir de la tabla global binaria)**	**SD Árbol verdadero *VS* Método 5-tuplos (a partir de la tabla global de frecuencia)**
dLog	10	10	18	18
dPear	14	6	18	26
dk	14	8	18	26

Notas: Se midió la precisión de los métodos de *LifePrint* (empleando LPS9) y el conjunto completo de 5-tuplos a través de comparar cada árbol NJ con el árbol verdadero usando la SD. dLog, distancia de *k*-tuplos basada en el índice de Jaccard; dPear, distancia de *k*-tuplos basada en el coeficiente de correlación de Pearson; dk, distancia de *k*-tuplos típica.

A continuación se presentan las imágenes obtenidas con el programa *phylo-comparison* de la comparativa topológica entre el árbol verdadero y los árboles NJ construidos con dPear a través de *LifePrint* y el método de 5-tuplos.

La Figura 9 es la comparativa entre el árbol verdadero (Tree A) y el árbol construido con *LifePrint* (Tree B). La Figura 10 es la comparativa entre el árbol verdadero (Tree A) y el árbol construido con el método de 5-tuplos (Tree B).

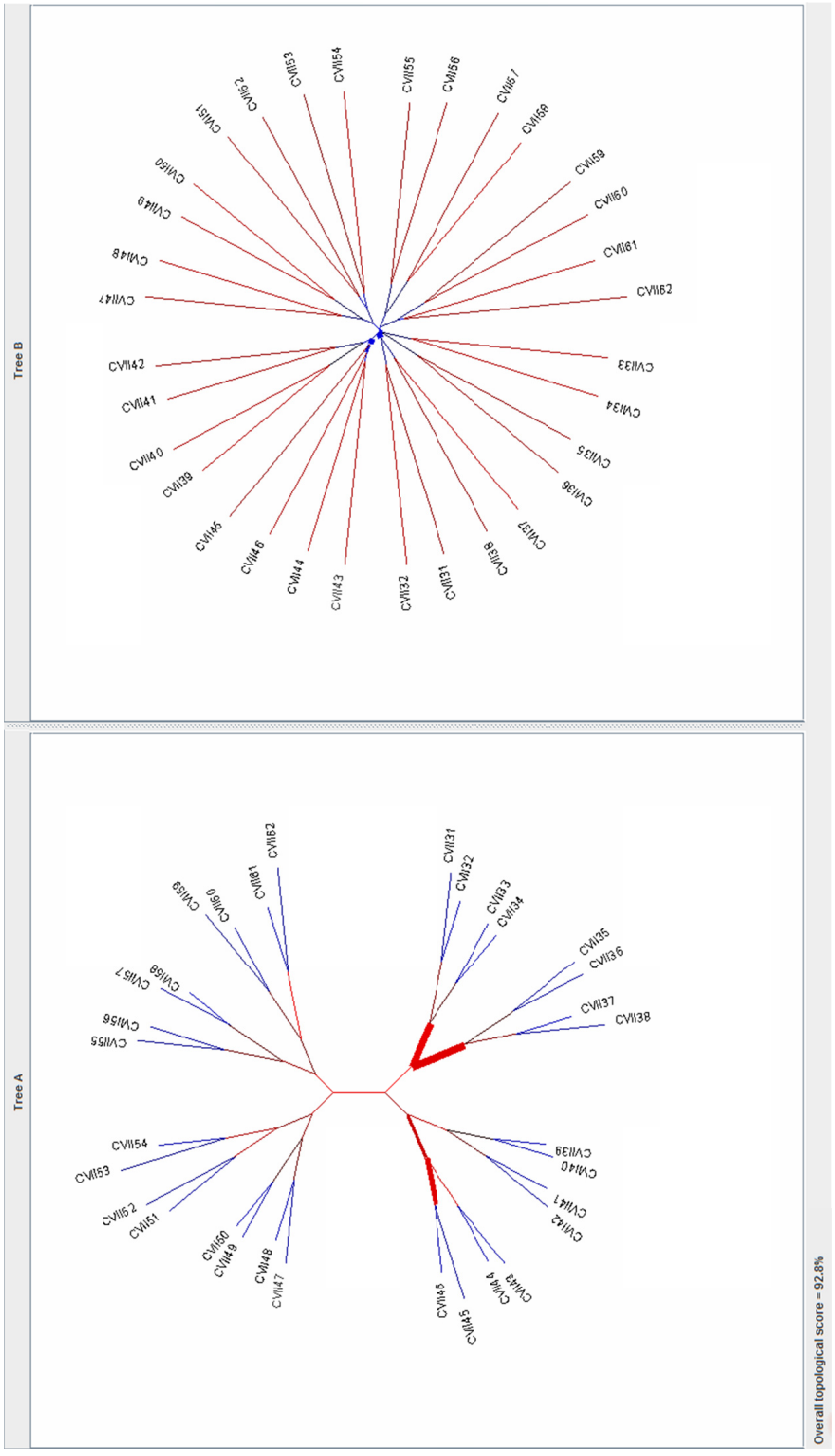

Figura 9. Comparativa entre el árbol verdadero y el árbol construido con *LifePrint*. Tree A = Árbol verdadero. Tree B = Árbol construido con *LifePrint*. Las líneas gruesas muestran las peores coincidencias. La puntuación topológica es proporcional al grosor de las líneas, es decir, mientras a mayor grosor la diferencia en el clado es más grande. En la parte inferior aparece el valor numérico de la puntuación topológica global.

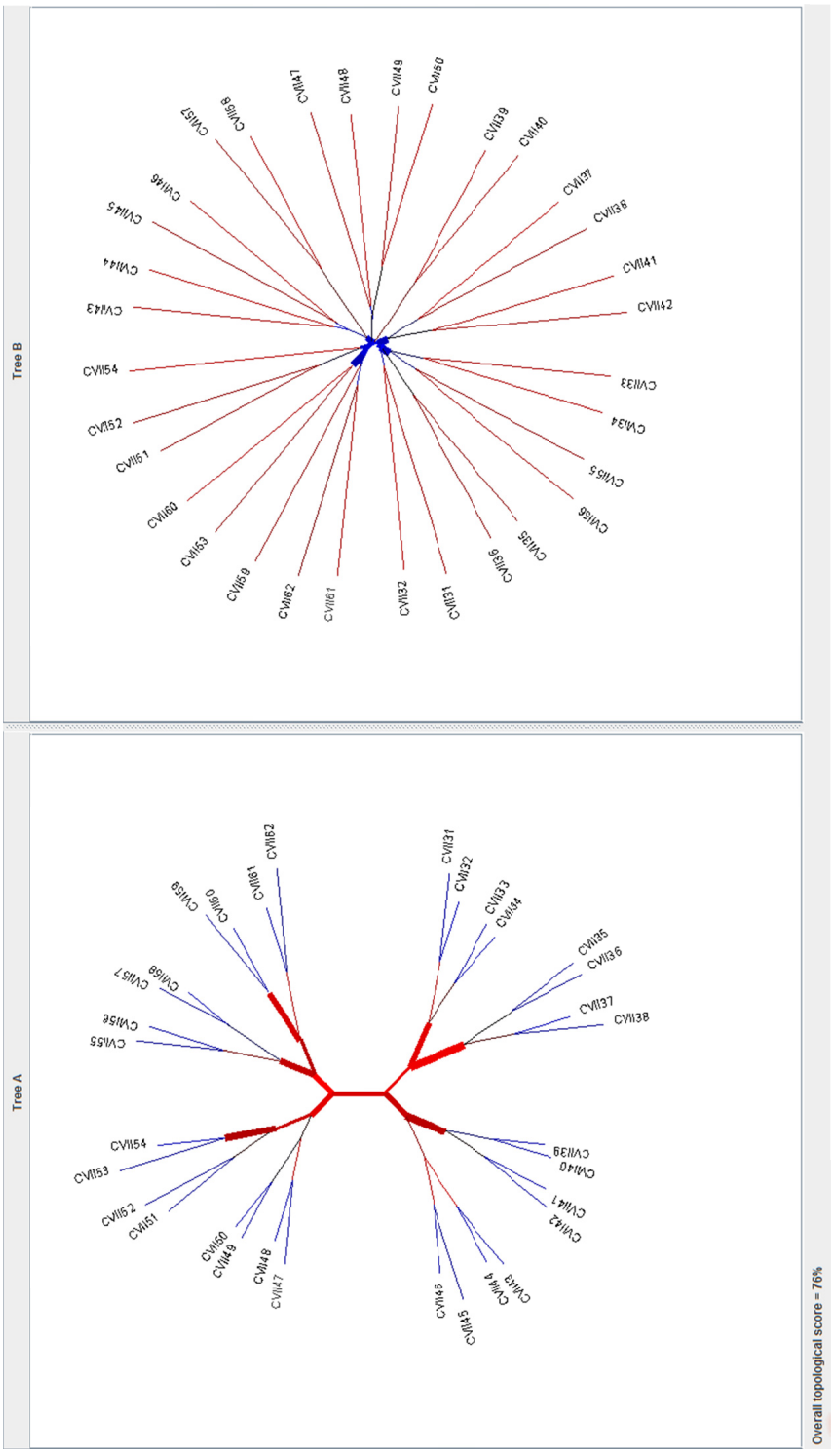

Figura 10. Comparativa entre el árbol verdadero y el árbol construido con el método de 5-tuplos. Tree A = Árbol verdadero. Tree B = Árbol construido con el método de 5-tuplos. Las líneas gruesas muestran las peores coincidencias. La puntuación topológica es proporcional al grosor de las líneas, es decir, mientras a mayor grosor la diferencia en el clado es más grande. En la parte inferior aparece el valor numérico de la puntuación topológica global.

IV. DISCUSIÓN

En el estudio de los 5-tuplos,[2] así como en los programas (*MUSCLE*, *CLUSTAL W* y *Kalign*) que primero computan una matriz de distancias de *k*-tuplos antes de calcular MSA, se emplean conjuntos completos de tuplos. En el caso de *LifePrint* la distribución homogénea del LPS9 dentro del conjunto completo de 9-tuplos permite considerarlo como una muestra representativa, lo que posibilitó no sólo coberturas completas de los genomas, sino construcción de árboles con valores de proporción de *bootstrap* mayores.

La condición c se consideró como la condición de búsqueda óptima para emplear el LPS9, dado que la proporción de tuplos que aparecen en los genomas analizados es de entre 20 y 80% del LPS9. En las condiciones a y b el LPS9 se subutiliza, mientras que en la condición d se produce una saturación de señales.

El LPS9 no detectó directamente todas las posibles cadenas 9-mer, pero sus tuplos sí cubrieron completamente los genomas escrutados. Cada posición dentro de los genomas es reconocida por varios tuplos, lo que incrementa la sensibilidad para detectar cambios simples en sus secuencias. Este es el caso de las sustituciones simples.

Dado que el método de *LifePrint* implica la selección de tuplos con un mínimo de entropía, el LPS9 no es capaz de detectar directamente regiones de baja complejidad (es decir, repetidos nucleotídicos). ¿Qué tan importantes son las regiones de baja complejidad para una inferencia filogenética precisa? ¿Qué ocurre si estas regiones no son incluidas en el análisis? Para responder estas preguntas es relevante distinguir si estas regiones son o no codificantes de proteínas. Las regiones que son codificantes y las que no lo son tienden a evolucionar por mecanismos diferentes. Cuando las regiones son codificantes, aún regiones con repetidos, los cambios se limitan a sustituciones o mutaciones sinónimas que producen aminoácidos conservados. En estos casos es importante considerar tales regiones en el cálculo de distancias evolutivas, aunque su impacto también estará relacionado con la proporción del genoma que representan y su presencia en otros genomas. Sin embargo, las secuencias no codificantes no están expuestas a las mismas presiones evolutivas que las regiones codificantes. La mayoría de las mutaciones son neutrales en las regiones no codificantes, pero algunas sustituciones pueden seguir mecanismos evolutivos complejos (por ejemplo, covariación), tal es el caso de secuencias no codificantes que son importantes para otras funciones (por ejemplo, regulación de la expresión génica).

Se espera que cuando un número crítico de variantes sean incluidas en el estudio filogenético, una determinada variante que sea considerablemente distante a otra(s), será más cercana (por ejemplo, más similar) a otra(s). Por lo tanto la saturación de la distancia de *k*-tuplos no debería ser una limitante para la construcción de árboles cuando se incluyen en el análisis un número crítico de secuencias.

Los resultados de las comparaciones de SD indican que *LifePrint* y el método de 5-tuplos no pueden recuperar el árbol verdadero. Estos resultados implican que la métrica usada para la distancia de *k*-tuplos conduce a diferentes precisiones en la reconstrucción de los árboles.

La inspección visual de los árboles NJ de los 32 genomas simulados indica que los valores de soporte del *bootstrap* son mayores para los árboles construidos por *LifePrint* que para los construidos con el método de los 5-tuplos. Para los árboles construidos por ambos métodos se observaron valores de proporción de *bootstrap* bajos en aquellos clados inconsistentes con el árbol verdadero. De esta manera parece que la prueba de *bootstrap* puede ser usada como una aproximación confiable para evaluar la precisión de la reconstrucción.

Las Figuras 6 y 7 indican que aunque las dos reconstrucciones de los 36 genomas reales son consistentes entre sí, los valores de proporción de *bootstrap* son mayores para el árbol construido con *LifePrint*. Al comparar estos mismos árboles con la filogenia

propuesta por el ICTV, se observó que las principales familias de viroides fueron identificadas por *LifePrint*, aunque algunos grupos están organizados de manera diferente, como es el caso de los miembros de *Avsunviroidae*. Sin embargo, tales conflictos están asociados con los valores de confianza de *bootstrap* relativamente bajos.

Se opina que una comparativa entre los árboles considerando la longitud de las ramas es crítica cuando topologías se han estimado utilizando los mismos criterios de optimización. Sin embargo, en este caso particular, estamos evaluando sólo diferentes métodos de distancia de *k*-tuplos (*LifePrint* frente al método de 5-tuplos), por lo que incluir árboles de métodos basados en caracteres, como MP o ML, produciría comparaciones inciertas, dado que cada criterio de optimización implica diferentes mediciones en las longitudes de las ramas. Para el objetivo específico de este trabajo se consideró adecuado limitar la comparativa topológica sólo a los árboles obtenidos con métodos de distancias de *k*-tuplos.

Al observar las Figuras 9 y 10, así como los valores numéricos de la puntuación topológica global puede establecerse que la diferencia topológica es evidentemente menor entre el árbol verdadero y el construido con *LifePrint*.

Estudios previos[3] sugieren que las reconstrucciones filogenéticas basadas en 5-tuplos funcionan mejor que aquellas basadas en

9-tuplos. Sin embargo, el enfoque de *LifePrint* es diferente del método de 5-tuplos descrito previamente ya que permite definir el número de diferencias entre las secuencias analizadas, mientras que los métodos de dk sólo buscan coincidencias perfectas (es decir, identidades). Por lo tanto en principio el rendimiento es diferente.

También debe señalarse que los resultados no proveen una evidencia definitiva para distinguir el método de *k*-tuplos más preciso que se haya analizado en este estudio. Como se mencionó antes, el hecho de que *LifePrint* permita un determinado número de diferencias entre las secuencias usadas, en contraste con el uso exclusivo de identidades en el método de los 5-tuplos, puede explicar las discrepancias entre estos enfoques.

Los resultados son consistentes con los estudios previos que sugieren que las inferencias filogenéticas de los métodos basados en la distancia de *k*-tuplos son más precisas que aquellas basadas en MSA.[3]

Los resultados indican que la precisión de determinados métodos de reconstrucción basados en *k*-tuplos depende tanto de la longitud de las secuencias analizadas como de la similitud entre éstas.

V. CONCLUSIONES

- Se obtuvo el LPS9, una muestra representativa de todos los tuplos posibles de DNA con una longitud de 9 (9-tuplos). Dicho conjunto comprende 1878 tuplos diferentes cada uno por lo menos en dos diferencias nucleotídicas internas y no contiguas.
- El esquema óptimo de búsqueda de similitud empleando el LPS9 fue el que permitió hasta dos diferencias. Dicho esquema se consideró como óptimo dado que la proporción de tuplos que aparecen en los genomas analizados fue de entre 20 y 80% del LPS9, es decir, no es subutilizado ni se alcanza la saturación al emplearlo.
- Con el LPS9 y el esquema óptimo de búsqueda de similitud que fue identificado, a través de *LifePrint* los genomas analizados son cubiertos completamente.
- Algunos de los tuplos del LPS9 fueron capaces de detectar directamente cadenas de 9 nucleótidos consecutivos e idénticos.
- Se calcularon las dLog, dPear y dk entre 32 genomas simulados derivados del *Citrus viroid II*. Se identificó que los árboles más precisos se construyeron con la dPear en el caso de *LifePrint* y con la dLog en el caso del método de los 5-tuplos.
- El LPS9 detecta directa o indirectamente los cambios puntuales entre las secuencias analizadas. Las distancias calculadas no

permiten una distinción efectiva entre los genomas cuando aproximadamente un 40% de la secuencia ha experimentado sustituciones.

- Las distancias basadas en el LPS9 (*LifePrint*) conducen a reconstrucciones más precisas que las distancias estimadas por el método de 5-tuplos. Los valores de soporte de *bootstrap* también fueron mayores en lo árboles construidos por *LifePrint* que en aquellos construidos con el método de 5-tuplos. Además, en el caso de *LifePrint* las dLog y dPear funcionan mejor con las tablas globales binaria y de frecuencia, respectivamente.

VI. PERSPECTIVAS

Es importante destacar que *LifePrint* usa eficientemente una muestra representativa de todos los 9-tuplos posibles para estimar la filogenia entre genomas completos. Esta característica permitirá implementar enfoques que disminuyan considerablemente el tiempo necesario para estimar filogenias usando la información completa disponible a escala genómica.

En estudios posteriores será importante evaluar exhaustivamente la relación entre la longitud de las secuencias analizadas como de la similitud entre éstas para determinar la longitud y características de los tuplos de un método optimizado de *k*-tuplos, explorando en paralelo su precisión intrínseca.

Otras áreas que deberán ser exploradas usando el método *LifePrint* incluyen la incorporación de estrategias sensibles a la posición de detección, y el empleo de criterios de optimización de parsimonia para estimar filogenias a partir del esquema de búsqueda de similitud propuesto. Adicionalmente, el uso de métodos de *k*-tuplos empleando datos basados en aminoácidos serían importantes para analizar regiones codificantes y/o genomas relacionados muy divergentes.

VII. REFERENCIAS

1. Whelan S. Inferring Trees. En: Keith JM, editor. Bioinformatics, Volume 1: Data, Sequence Analysis and Evolution. New Jersey: Humana Press; 2008: 287-290.

2. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4: 406-425.

3. Yang K, Zhang L. Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. Nucleic Acids Res. 2008; 36 (5): e33.

4. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32 (5): 1792-1797.

5. Morgenstern B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics. 1999; 15 (3): 211-218.

6. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000; 302 (1): 205-217.

7. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through

sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22: 4673-4680.

8. Lassmann T, Sonnhammer EL. Kalign--an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics. 2005; 6: 298.

9. Snel B, Bork P, Huynen M. Genome evolution. Gene fusion versus gene fission. Trends Genet. 2000; 16 (1): 9-11.

10. Herniou EA, Luque T, Chen X, Vlak JM, Winstanley D, Cory JS, O'Reilly DR. Use of whole genome sequence data to infer baculovirus phylogeny. J Virol. 2001; 75 (17): 8117-8126.

11. House CH, Fitz-Gibbon ST. Using homolog groups to create a whole-genomic tree of free-living organisms: an update. J Mol Evol. 2002; 54 (4): 539-547.

12. Dutilh BE, Snel B, Ettema TJ, Huynen MA. Signature genes as a phylogenomic tool. Mol Biol Evol. 2008; 25 (8): 1659-1667.

13. Milosavljević A. Discovering sequence similarity by the algorithmic significance method. Proc Int Conf Intell Syst Mol Biol. 1993; 1: 284-291.

14. Chen X, Kwong S, Li M. A compression algorithm for DNA sequences and its applications in genome comparison. Genome Inform Ser Workshop Genome Inform. 1999; 10: 51-61.

15. Chen X, Li M, Ma B, Tromp J. DNA Compress: fast and effective DNA sequence compression. Bioinformatics. 2002; 18 (12): 1696-1698.

16. Wu X, Cai Z, Wan XF, Hoang T, Goebel R, Lin G. Nucleotide composition string selection in HIV-1 subtyping using whole genomes. Bioinformatics. 2007; 23 (14): 1744-1752.

17. Lu G, Zhang S, Fang X. An improved string composition method for sequence comparison. BMC Bioinformatics. 2008; 9 (Suppl 6): S15.

18. Gao L, Qi J. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evol Biol. 2007; 7: 41.

19. Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 1995; 11 (7): 283-290.

20. Stuart GW, Moffett K, Baker S. Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics. 2002; 18 (1): 100-108.

21. Qi J, Wang B, Hao BI. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J Mol Evol. 2004; 58 (1): 1-11.

22. Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2).Cladistics. 1989; 5: 164–166.

23. Swofford DL. PAUP*.Phylogenetic Analysis Using Parsimony (*and Other Methods).Version 4. Massachusetts: Sinauer Associates; 2003.

24. Méndez-Tenorio A. Elaboración de software relevante para el diseño y evaluación de microarreglos de DNA. Tesis de Doctorado; México, D.F.: Departamento de Bioquímica, ENCB, IPN; 2006.

25. Reyes-Lopez MA, Méndez-Tenorio A, Maldonado-Rodríguez R, Doktycz MJ, Fleming JT, Beattie KL. Fingerprinting of prokaryotic 16S rRNA genes using oligodeoxyribonucleotide microarrays and virtual hybridization. Nucleic Acids Res. 2003; 31 (2): 779-89.

26. Méndez-Tenorio A, Flores-Cortés P, Guerra-Trejo A, Jaimes-Díaz H, Reyes-Rosales E, Maldonado-Rodríguez A, Espinosa-Lara M, Maldonado-Rodríguez R, Beattie KL. In silico evaluation of a novel DNA chip based fingerprinting technology for viral identification. Revista Latinoamericana de Microbiología 2006; 48 (2): 56-65.

27. Casique-Almazán, J. Evaluación in silico de la capacidad de identificación viral del sensor universal de huella genómica (UFC). Tesis de Maestría. México, D.F.: Departamento de Bioquímica, ENCB, IPN; 2008.

28. UFC Applications Server [http://biomedbiotec.encb.ipn.mx/UFCVH/]

29. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000; 16 (6): 276-277.

30. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Mol Biol Evol. 2007; 24: 1596-1599.

31. Nei M, Kumar S. Molecular Evolution and Phylogenetics. New York: Oxford University Press; 2000: 81-83, 171-177.

32. Felsenstein J. Inferring Phylogenies. Massachusetts: Sinauer Associates; 2003: 335-349, 528-532.

33. Nye TM, Liò P, Gilks WR. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics. 2006; 22 (1): 117-9.

34. Flores R, Randles JW, Owens RA, Bar-Joseph M, Diener TO. Subviral Agents: Viroids. En: Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA, editores. Virus Taxonomy. VIIIth Report of the International Committee on Taxonomy of Viruses. New York: Elsevier Academic Press; 2005: 1151-1152.

VIII. APÉNDICE 1

Números de acceso del NCBI (*National Center for Biotechnology Information*) de los 36 genomas reales.

Apple dimple fruit viroid, complete genome	NC_003463
Apple fruit crinkle viroid, complete genome	NC_003777
Apple scar skin viroid, complete genome	NC_001340
Australian grapevine viroid, complete genome	NC_003553
Avocado sunblotch viroid, complete genome	NC_001410
Chrysanthemum chlorotic mottle viroid, complete genome	NC_003540
Chrysanthemum stunt viroid, complete genome	NC_002015
Citrus bent leaf viroid, complete genome	NC_001651
Citrus dwarf viroid, complete genome	NC_005821
Citrus exocortis viroid, complete genome	NC_001464
Citrus viroid Ia, complete genome	NC_001907
Citrus viroid II, complete genome	NC_003881
Citrus viroid III, complete genome	NC_003264
Citrus viroid IV, complete genome	NC_003539
Citrus viroid OS, complete genome	NC_004359
Citrus viroid-I-LSS, complete genome	NC_004358
Coconut cadang-cadang viroid, complete genome	NC_001462
Coconut tinangaja viroid, complete genome	NC_001471
Coleus blumei viroid 1, complete genome	NC_003681
Coleus blumei viroid 2, complete genome	NC_003682
Coleus blumei viroid 3, complete genome	NC_003683
Coleus blumei viroid, complete genome	NC_003882
Columnea latent viroid, complete genome	NC_003538
Eggplant latent viroid, complete genome	NC_004728
Grapevine yellow speckle viroid 1, complete genome	NC_001920
Grapevine yellow speckle viroid 2, complete genome	NC_003612
Hop latent viroid, complete genome	NC_003611
Hop stunt viroid, complete genome	NC_001351
Iresine viroid, complete genome	NC_003613
Mexican papita viroid, complete genome	NC_003637
Peach latent mosaic viroid, complete genome	NC_003636
Pear blister canker viroid PBCVd, complete genome	NC_001830
Potato spindle tuber viroid, complete genome	NC_002030
Tomato apical stunt viroid, complete genome	NC_001553
Tomato chlorotic dwarf viroid, complete genome	NC_000885
Tomato planta macho viroid, complete genome	NC_001558