UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Facultad de Ciencias

"El papel de las secuencias simples en los proteomas virales"

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

B I Ó L O G A

P R E S E N T A :

Irma Lozada Chávez

Director de Tesis:

Dr. Arturo Carlos II Becerra Bracho

2004

UNIVERSIDAD NACIONAL
AVENIDA II
MEXICO

Autorizo a la Dirección General de Bibliotecas de la UNAM a difundir en formato electrónico e impreso el contenido de mi trabajo recepcional.
NOMBRE: Irma Lozada Chávez
FECHA: 30/Julio/2004
FIRMA:

ACT. MAURICIO AGUILAR GONZÁLEZ
Jefe de la División de Estudios Profesionales de la Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

"El papel de las secuencias simples en los proteomas virales".

realizado por Irma Lozada Chávez

con número de cuenta 9957489-1 , quien cubrió los créditos de la carrera de: Biología

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis
Propietario Dr. Arturo Carlos II Becerra Bracho

Propietario Dr. Antonio Eusebio Lazcano-Araujo Reyes

Propietario Dr. Ernesto Pérez Rueda

Suplente Dr. Víctor Manuel Valdés López

Suplente Biol. Luis José Delaye Arredondo

Consejo Departamental de Biología

M. en C. Juan Manuel Rodríguez Chávez

FACULTAD DE CIENCIAS
UNIDAD DE ENSEÑANZA

Agradecimientos:

Quiero expresar primeramente mi total agradecimiento a mi hermano **Alejandro** porque ha sido mi familia y mi mejor amigo desde que han sido necesarios, porque me ha impulsado siempre a esperar más de mí y porque la fortaleza y lo bueno que pueda tener yo como persona se lo debo principalmente a él. ¡Para ti Chaparro!

Quiero agradecer especialmente a **Antonio Lazcano** por ser una de las personas que me inspiró para incursionar mi vida académica en la evolución temprana de la vida, constituyendo un sesgo impresionante por sobre muchas otras disciplinas hasta mis perspectivas futuras, por haberme permitido trabajar dentro de su excelente grupo de trabajo, por haberme ayudado incondicionalmente en el desarrollo y culminación de esta tesis, pero sobretodo por su amistad y porque siempre me hace ser mejor al mostrarme mis errores.

Así mismo quiero agradecer a **Arturo Becerra** por haber dirigido este trabajo de tesis y por haberme permitido aprender junto con él, de forma invaluable, sobre el quehacer de la investigación, por la amistad que tenemos y que es respetada por nuestra relación académica y porque siempre tuvo más de una manera para creer en mí.

De igual forma quiero agradecer a **Javier Díaz** por acompañarme en esta inquietud humana que prontamente se convierte en un estilo de vida, la ciencia. Le agradezco que todo este tiempo me haya llevado de la mano primero para aprender juntos, y después para aprender de él toda la dureza, formalidad, pasión, intuición y regocijo que trae la ciencia todos los días a las personas que como yo tratamos de ejercerla. Gracias Javier, porque has visto el brillo que llevo dentro y con toda paciencia has terminado de pulirme con tu joven experiencia y, sin ser intención mía, te ha llevado a ser mejor todo el tiempo. Finalmente, quiero agradecer tu amistad y tu amor, los cuales son de los principales motores que académica y personalmente me permitieron cerrar esta etapa de mi vida para pensar en las que vienen académicamente.

Quiero agradecer a **Sandra Ramírez** por ser una excelente investigadora que no necesita ver cargos académicos en mi persona para desarrollar todos los proyectos que tenemos, porque cree en mí y por su valiosa amistad. Porque ella y su apoyo se me han permitido ser feliz al pensar que las preguntas por muy difíciles y ambiciosas que parezcan pueden tener cavidad en su realización.

Quiero agradecer a **Luis Delaye** por su amistad, su inteligencia y sumo genio oculto discretamente por su sencillez, porque su humildad me ha dado el máximo ejemplo de la felicidad obtenida por la sabiduría y la naturaleza humana. Su vida académica tal cual, incluyendo sus enseñanzas, me han permitido ser mejor tanto académica como humanamente.

Finalmente quiero agradecer a **Ernesto Pérez Rueda, Sara Islas, Ervin Silva, Ana María Velasco** y **Víctor Valdés**, mis amigos y también formadores académicos, que el espacio y las palabras no me alcanzan para agradecerles todo lo que les admiro y respeto. Definitivamente, todos ustedes han influido en mi desarrollo académico y personal.

A mis compañeros y amigos del laboratorio: **Daniela, Claudia, Ana, Mariana, Ricardo, Germán, Diego** y **David** por aprender y compartir juntos días de trabajo en el laboratorio, haciendo muchos de ellos buenos momentos.

A mi parte artística y humanista: **Lucelina Nunes** y **Alberto Cabañas** por ser mis amigos bajo historias muy parecidas, porque me ayudaron a encontrar el artista que llevo dentro, pero sobretodo por compartirme sus experiencias, secretos y percepción de la vida.

A mis amigos: **Cynthia, Mónica Araujo, Leticia Ortega, Alejandro Carbajal, Mitzi Villajuana, Paola Colín, Lidia Leal, Lilia Montoya, Ricardo Menchaca, Maribel Hernández** y **Cristhian Ávila** con quienes comparto ciencia, experiencias y muy gratos momentos que llenan mi vida de forma muy peculiar.

A **Concepción, Daniel, Abraham, Moisés** y **Esaú** a quienes amo desde que entraron en mi vida.

Muy especialmente a quienes considero como mis familias y mis amigos, la **familia Araujo** y la **familia Díaz** que siempre han tenido un espacio especial entre su gente para mí.

A mi padre, **Nabor Lozada**, y a mi madre, **Ana María Chávez**, por mostrarme lo difícil que es la vida con sueños truncados, por ayudarme cuando estuvo en sus manos, por mostrarme lo fuerte que puedo ser estando sola y porque gracias a ellos, sin importar circunstancias o motivos me concibieron y cuidaron de mí bajo una familia los años que recuerdo como los más completos y felices de mi vida, mi niñez.

A mi **Universidad Nacional Autónoma de México** por haber permitido mi desarrollo académico y personal de forma inigualable a cualquier etapa previa de mi vida.

ÍNDICE

Resumen

En este trabajo, se toman en cuenta a los virus como buenos modelos para el análisis de las secuencias simples (secuencias que presentan un sesgo composicional), ya que hasta donde sabemos, poco o nada se ha propuesto sobre su posible papel en los proteomas virales.

Se analizó una base de 2 304 proteomas virales que constituyen los productos de 34 253 secuencias de aminoácidos distribuidas en 76 grupos taxonómicos, que incluye a 63 familias y 13 géneros, colectados de las bases de datos: *National Center for Biotechnology Information* (NCBI), *Kyoto Encyclopedia of Genes and Genomes* (KEGG), *Virus Database at University College London* (VIDA 2.0-UCL), *Human Immunodeficiency Virus Databases* (HIVD), *Universal Virus Database of the International Committee on Taxonomy of Viruses* (ICTVDdB)., en los cuales, se detectó la presencia de LCS en 7 617 CDS (22%) distribuidos en 62 familias y 12 géneros virales. La distribución de estas LCS pareciera que no depende del tamaño del proteoma, ya sea por residuos de aminoácidos o por número de CDS que los conforman y, en muchos casos, tampoco de un sesgo en la composición de sus aminoácidos. La diversa distribución de las LCS dentro de los CDS, no sólo en los extremos NH_2- y –COOH terminales, nos podría sugerir que su posición no perturba la función de las proteínas en las que se encuentran y que probablemente no sea importante que deba de conservarse la secuencia primaria de las LCS para su función, en caso de tener alguna. Una alta cantidad de dominios de función desconocida, *a priori*, fueron detectados en los proteomas virales aquí analizados; sin embargo, la mayoría de las LCS detectadas se encuentran dentro de dominios que podrían ser parte de proteínas involucradas en funciones estructurales y auxiliares.

Aunque esta base de proteomas virales presenta un sesgo por aquellos virus que infectan a especies de importancia médica y que no representa a toda la diversidad viral, los análisis hechos en este trabajo muestran que dada la distribución y frecuencia de las secuencias simples en los proteomas virales, así como su conservación y presencia en dominios funcionales, hacen posible sugerir que este tipo de secuencias tienen un papel relevante en la evolución de los virus.

1. INTRODUCCIÓN

Entender cómo evoluciona el genoma sigue siendo una pregunta abierta. Aunque la definición de **genoma** puede estar sujeta a varios contextos, de forma general se le puede definir como la molécula de DNA o RNA que contiene codificada la información genética total utilizada por un organismo, un virus, un plásmido, una mitocondria, los plásticos, etc. para dar lugar a nueva progenie. En los sistemas celulares el genoma es constituido por una cadena doble de DNA de uno a varios segmentos. En cuanto que en los sistemas virales el genoma consiste de uno a más segmentos de una cadena simple o doble de DNA o RNA organizado de forma lineal o circular. Independientemente de la composición o estructura del material genético, algunas de las unidades informativas codificadas en él, llamadas genes, deben ser expresados para dar lugar a todo el conjunto de proteínas que constituyen el **proteoma**, el cual es necesario para generar y mantener en los sistemas celulares los procesos bioquímicos para su sobrevivencia, reproducción y muerte y, en los sistemas virales garantizar su replicación, dispersión o latencia al entrar en las células.

Antes del reporte del primer genoma de un organismo secuenciado en 1995 (Fleischmann *et al*, 1995; Casari *et al*, 1995) ya se encontraban en los bancos de datos las secuencias de un gran número de genomas virales disponibles. Para finales del 2002 estaban depositadas las secuencias de más de 100 genomas celulares (Janssen *et al*, 2003) y más de 3000 virales (Mills *et al*, 2003), los cuales han proporcionado a la genómica la materia prima para su consolidación como una disciplina comparativa que permite el análisis de la información codificada en estas secuencias.

Los análisis de los genomas completos han puesto en evidencia una asombrosa variabilidad en ellos en términos del contenido y orden de sus genes. Una de las observaciones derivadas de estas comparaciones es que a pesar de la amplia diversidad en los tamaños de los genomas celulares, incluso entre genomas de especies del mismo dominio celular (Arquea, Bacteria y Eucaria), el promedio del contenido de genes en éstos resulta relativamente constante, mas no así el tipo de genes, puesto que éstos pueden variar en correlación con los requerimientos energéticos y metabólicos de cada especie (Waters *et al*, 2003; Mira *et al*, 2002). Por otro lado, los genomas virales tienden a mantener tamaños pequeños con un contenido de genes proporcional a ello y, aunque presentan gran diversidad en la composición y estructura de sus genomas, éstos codifican para pocas clases funcionales (Mills *et al*, 2003). De esta forma, entender cómo se incrementa el contenido genético y qué mecanismos son los responsables para que a partir de él se generen nuevas funciones que serán expresadas en el proteoma, no ha constituido una tarea fácil.

De manera general, podemos decir que los principales mecanismos involucrados en la evolución del genoma son, para la *ganancia de genes*, la duplicación génica, la fusión/fisión, y la transferencia horizontal de genes (THG);

mientras que para la *pérdida de genes*, podemos mencionar al origen de pseudogenes, también la fusión/fisión de genes y la THG. Estudios previos (Wolf *et al*, 2001; Snel *et al*, 2002), señalan que los mecanismos que influyen mayoritariamente en la modificación del contenido génico en los procariontes son la duplicación, la pérdida y la transferencia horizontal de genes. Se estima que la pérdida es la fuerza más influyente, al ser tres veces más frecuente que la THG, seguida por la duplicación, la cual contribuye dos veces más que la THG (Kunin y Ouzounis, 2003). La figura 1.1 señala algunos de los diferentes procesos y fenómenos mutacionales que se han observado en la evolución del genoma.

Existen otros procesos mutacionales al nivel de la replicación y reparación del DNA, de forma puntual o en *tandem,* que en primera instancia generan la materia prima sobre la cual la evolución puede actuar para generar nuevas funciones. A continuación se hablará del fenómeno propuesto aquí como otro mecanismo más que puede influir en la generación de nuevas funciones en las secuencias de las proteínas.

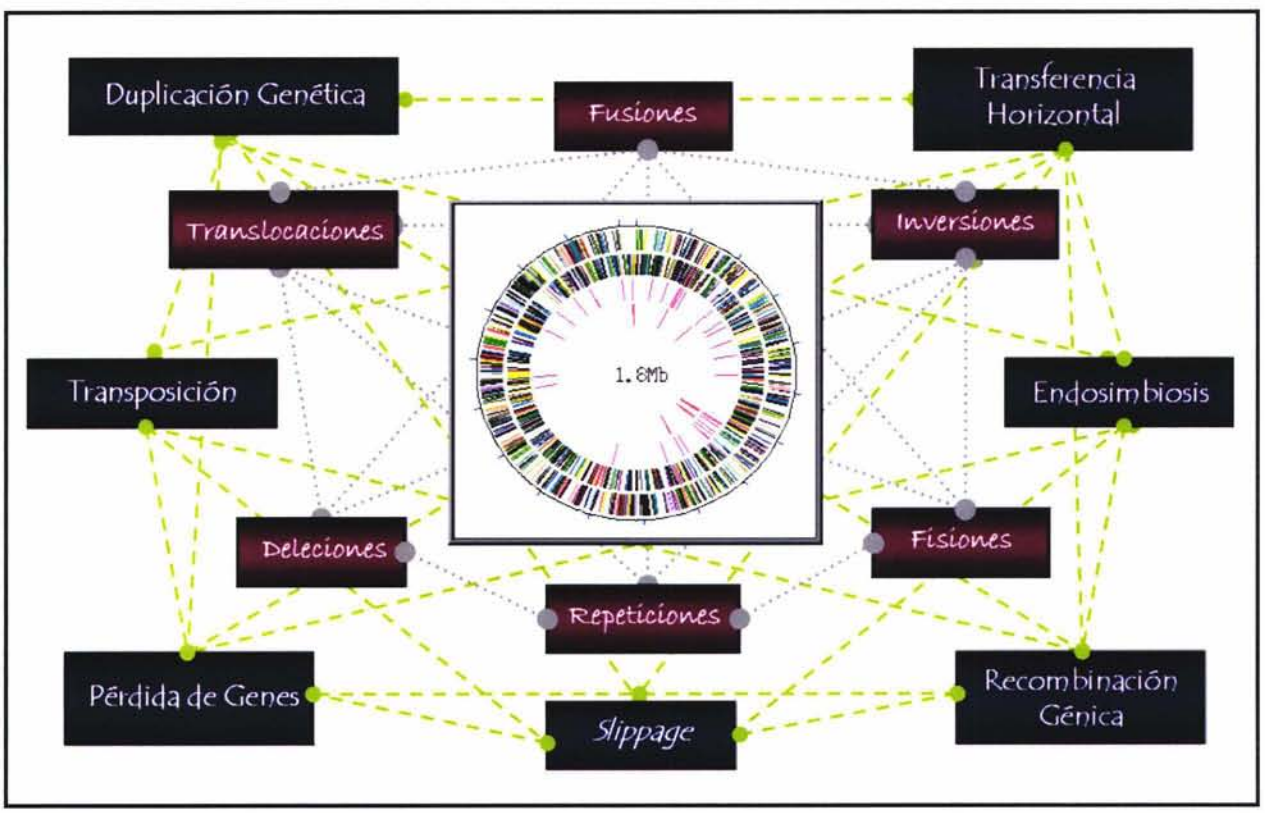


Fig. 1.1. Diagrama que señala los principales procesos involucrados en la evolución del genoma y algunos de los fenómenos que se derivan de ellos. Al centro, el esquema del mapa físico de un genoma hipotético; las leyendas en los cuadros azules representan los principales mecanismos estudiados que influyen en el tamaño del genoma; las leyendas en los cuadros en rojo muestran algunos de los fenómenos analizados en las secuencias que son causados por uno o más de estos mecanismos; las líneas con diferente formato sólo marcan los diversos niveles de interacción entre los procesos y los fenómenos a los que se da lugar, en el genoma, en diversas circunstancias y tiempo evolutivo.

1.1 Las Secuencias Simples.

Existen diversos procesos que generan mutaciones en las secuencias genómicas, algunas de las cuales se pueden reflejar en las secuencias de las proteínas e influir en la función de las mismas. Si bien, la duplicación, recombinación y transferencia horizontal son de los procesos más estudiados en este sentido, es necesario tomar en cuenta que en los procesos de replicación del material genético, al generar cambios puntuales y en *tandem* en regiones de las secuencias de los genomas, también pueden ser fuentes de variabilidad funcional.

Cuando estas regiones presentan un sesgo composicional en los residuos de nucleótidos o de aminoácidos que los conforman, ya sea como unidades repetidas de un solo elemento o como arreglos en *tandem* se les llama **secuencias simples** o **secuencias de baja complejidad (*LCS* - *Low Complexity Sequences*)** (Wootton y Federhen, 1993; Tautz y Schlötterer, 1994). Este tipo de secuencias ha sido un fenómeno ampliamente estudiado desde una perspectiva biológica y matemática por diversas disciplinas e intereses.

Desde el punto de vista de la biología molecular, las LCS de DNA fueron los primeros DNA's sintéticos que se produjeron *in vitro* (Kornberg *et al*, 1964; Byrd *et al*, 1965). Su *status* como secuencias repetitivas universales en los genomas eucariontes fue descubierta después. La razón de ello, es que los primeros métodos de análisis de genomas, tales como la ultracentrifugación y la reasociación cinética no permitieron su inmediata detección. Estas fueron encontradas después por la secuenciación de clones y por experimentos de hibridación por *southern blot*. Posteriormente, fueron usadas para mostrar que hay elementos repetitivos presentes en todos los genomas eucariontes (Hamada *et al*, 1982; Tautz y Renz, 1984; Levinson y Gutman, 1987). A partir de ello, las LCS han recibido un gran número de sinónimos, tales como: microsatélites, VNTR's, DNA satélite, secuencias teloméricas, islas de CpG's, *hotspots* de recombinación, cuasi-periodicidad de trinucleótidos en secuencias codificantes, secuencias de baja complejidad irregulares en regiones intergénicas, entre otros (Wootton, 1999).

Se cree que las LCS son generadas principalmente por un mecanismo conocido como ***slippage*** (Levinson y Gutman, 1987; Wootton, 1999), aunque pueden verse originadas, e incluso incrementadas, por otros procesos mutacionales, tales como la duplicación genética, la recombinación y la transferencia horizontal mencionadas anteriormente (ver figura 1.1). El proceso que explica cómo se lleva a cabo el *slippage* fue establecido hace 40 años aproximadamente. En su forma más sencilla, este proceso involucra la apertura y desplazamiento local de las hebras del dúplex del DNA seguido de un mal apareamiento de las bases complementarias, cuando son seguidas de replicación o reparación, permitiendo inserciones o deleciones de una o varias de las unidades repetidas cortas según la dirección de la hebra sobre la que se lleve a cabo (Levinson y Gutman, 1987; Bzymek y Lovett, 2001). Un esquema que muestra cómo se llevaría a cabo este proceso mutacional se observa en la figura 1.2.

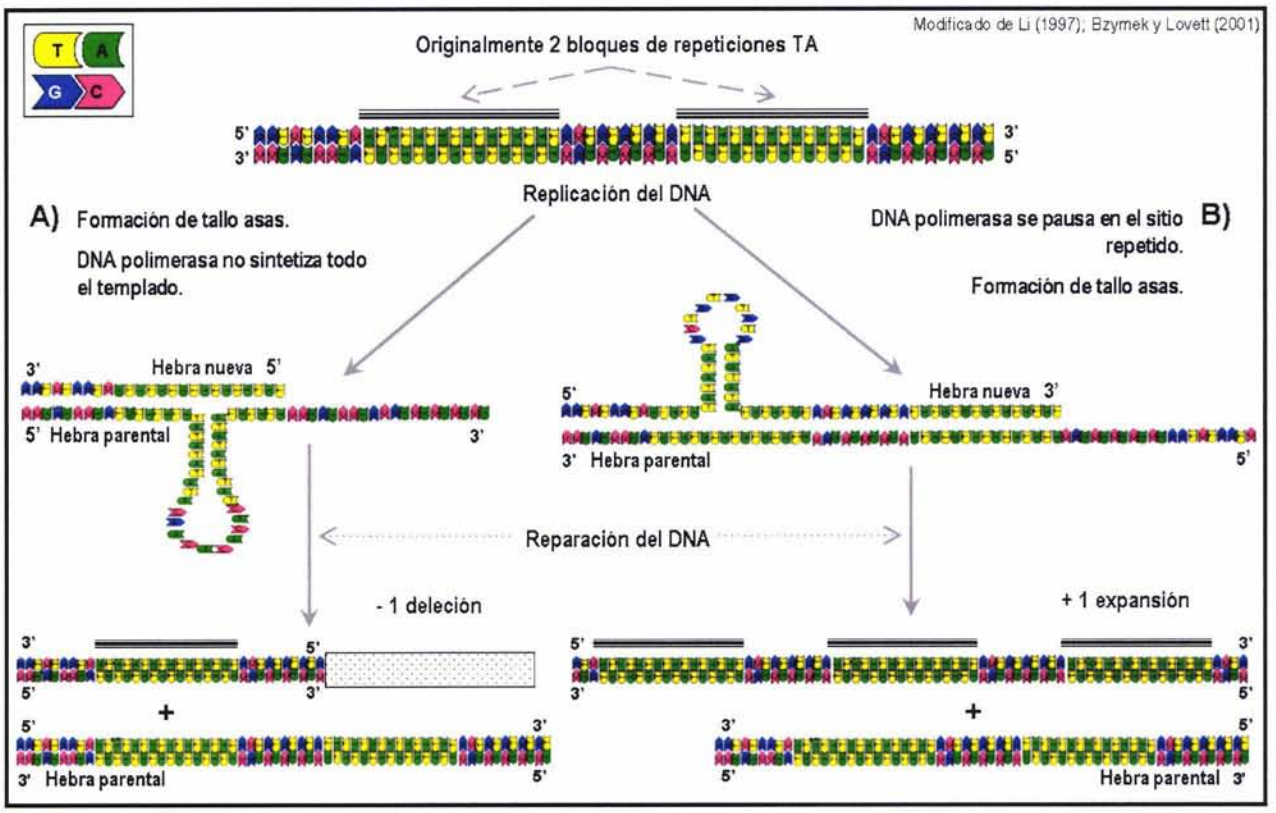


Fig. 1.2. Esquema que ilustra el proceso de mutación llamado *slippage*. Originalmente, en algunas regiones del genoma existen segmentos o bloques de repeticiones, tal como se muestra en este esquema con las repeticiones del tipo TA en la cadena dúplex del DNA. El *slippage* en replicación implica que la hebra nueva que se está sintetizando (la naciente) se disocia de la hebra del templado durante la replicación de las regiones repetidas y la hebra naciente puede re-alinearse fuera de fase con la hebra templado, generalmente acompañado con la formación de tallos asa. Cuando continúa la replicación, la hebra naciente será más larga **(B)** o más corta **(A)** que la hebra templado, dependiendo sobre cual hebra se formaron las estructuras tallo asa, si sobre la naciente o la templado. Las hebras de DNA son representadas con cadenas de complementariedad Watson-Crick entre los cuatro nucleótidos, donde: T es timina, A es adenina, G es guanina y, C es citosina; las unidades repetidas están representadas por bloques repetidos de nucleótidos TA, sobre los cuales se dibujaron líneas negras; la dirección de la replicación es 5' → 3' en ambas hebras del dúplex de DNA.

La recurrente detección de las LCS en los bancos de secuencias sugirió, originalmente, que pueden tener una función general, por ejemplo, en la recombinación y conversión de genes (Levinson y Gutman, 1987), en la regulación de genes en la morfogénesis y en el desarrollo embrionario, y en la transducción de señales (Wootton, 1999); así como también, formando parte de proteínas estructurales, tales como keratinas, colágenas, miosinas, fibrinas, mucinas y elastinas (Wootton, 1993), o como parte de enzimas muy antiguas como ATP sintetasa, proteínas ribosomales L7/L12, la RNA helicasa ATP-dependiente, entre otras (Tautz, 1989; Hancock, 1995; Wahl *et al*, 2000; Becerra *et al*, enviado). Además, muchas de estas LCS presentan interacciones moleculares con importantes consecuencias biológicas, tales como las variaciones de los desórdenes neurológicos en humanos y la localización de epítopes auto inmunes (Wootton y Federen, 1993; Wootton, 1999; Bremner *et al*, 2001). Aún cuando estos hallazgos han provisto evidencias para asignar funciones a algunos tipos de LCS, la importancia biológica de otras no es clara.

Debido a que las LCS se han detectado tanto en las secuencias codificantes como en las no codificantes, el papel evolutivo de éstas en los sistemas celulares ha sido ampliamente estudiado por varios grupos de trabajo (Hamada *et al*, 1982; Tautz y Renz, 1984; Levinson y Gutman, 1987; Tautz y Schlötterer, 1994; Wootton y Federhen, 1993; Wootton, 1994; Hancock, 1996; Albá *et al*, 1999; Huntley y Golding, 2000; Katti *et al*, 2001; Huntley y Golding, 2002; Tompa, 2003; Becerra *et al*, enviado). Los cuales coinciden en señalar que existe una importante cantidad de LCS en las bases de secuencias, tales como el GenBank (Benson *et al*, 2003) y el SwissProt (O'Donovan *et al*, 2002) (Wootton y Federhen, 1994; Wootton, 1994; Marcotte *et al*, 1998), mas no así en las bases de estructuras, como el Protein Data Bank (Berman *et al*, 2000) (Wootton, 1994; Huntley y Golding, 2002), donde se detectó que las LCS que se encuentran dentro de un pequeño número de proteínas con una estructura terciaria cristalizada, forman parte de estructuras helicoidales y de conectores irregulares dentro de regiones físico-químicamente definidas como no globulares, y que tienden a formar estructuras desordenadas en las proteínas, lo cual dificulta la determinación de sus coordenadas cristalográficas. En este sentido, Tompa (2003) al analizar un conjunto de 126 proteínas no estructuradas, es decir, proteínas que no necesitan un plegamiento bien definido para llevar a cabo su función, por lo tanto, su estado nativo y funcional es intrínsicamente no estructurado (Tompa, 2002), detecta que éstas poseen una mayor frecuencia de LCS que lo propio para las proteínas estructuradas contenidas en el SwissProt, en *Sacharomyces cereviseae* y en *Homo sapiens*, previamente reportado.

Se ha sugerido que las LCS juegan un papel importante como fuente de variabilidad genética y en la evolución del tamaño del genoma (Tauzt *et al*, 1986; Hancock, 1995). Debido a su hipermutabilidad, son reconocidas como una fuente importante de variación fenotípica, de forma especial dentro de los procariontes patógenos (Moxon *et al*, 1994; Moxon, 1999; Bayliss *et al*, 2004). Además, se ha visto que hay una buena relación entre el DNA repetitivo, la presencia de LCS, y la longitud del genoma (Hancock, 1995). Correlacionado a ésto, la frecuencia de las LCS

presentes en las secuencias de DNA y de proteínas de los genomas procariontes es mucho menor que en la de los eucariontes (Huntley y Golding, 2000; Becerra *et al*, enviado).

Sobre la composición de las LCS detectadas en los genomas de eucariontes, tales como *Arabidopsis thaliana, Dictyostelium discoideum, Drosophila melanogaster, Caenocarbitis elegans, Saccharomyces cereviseae, Plasmodium falciparum, Mus musculus, Homo sapiens*, entre otros (Karlin y Burge, 1996; Marcotte *et al*, 1998; Liu y Lindquist, 1999; Albá *et al*, 1999; Pupko y Graur, 1999; Michelitsch y Weissman, 2000; Meeds *et al*, 2001; Katti *et al*, 2001; Albá *et al*, 2001; Huntley y Golding, 2000; Huntley y Golding, 2002; Wickstead *et al*, 2003) y un gran número de especies de los dominios Bacteria y Archaea (Karlin y Burge, 1996; Marcotte *et al*, 1998; Kyrpides *et al*, 1998; Parkhill *et al*, 2000; Michelitsch y Weissman, 2000; Huntley y Golding, 2000; Albá *et al*, 2001; Bayliss *et al*, 2004; Becerra *et al*, enviado), se ha encontrado que las LCS están principalmente constituidas por los aminoácidos: S, E, L, Q, K, H, D, G, V y N. Se han reportado diferencias sutiles en el contenido y tipo de aminoácidos que las componen, pero entre especies no hay cambios significativos aparentes (Huntley y Golding, 2000; Becerra *et al*, enviado). Además, se han detectado con mucha menor frecuencia secuencias de proteínas con LCS en *S. cereviseae, C. elegans* y *A. thaliana* que presentan una composición de aminoácidos muy peculiar, tales como T, P y R (Albá *et al*, 1999; Huntley y Golding, 2000).

Se han detectado LCS de longitud variable, donde la más pequeña es de 5 aminoácidos (Wootton y Federhen, 1994) y la más grande de poco más de 300 (Huntley y Golding, 2000). Por otro lado, se ha podido detectar que existe, en los sistemas celulares, cierta tendencia de las LCS a ubicarse hacia alguna de la partes amino o carboxilo terminal de la secuencia codificante (Becerra *et al*, enviado), donde las razones biológicas de este comportamiento permanecen poco claras, aunque esta ubicación particular de las LCS podría evitar una interferencia desfavorable, a nivel estructural, con la función de la proteína que la contiene. En cuanto a su patrón de conservación en proteínas homólogas de eucariontes, se encontró que la proporción de proteínas que contienen LCS cambia dramáticamente de especie a especie; además, la composición de aminoácidos de las LCS detectadas también cambia entre especies, lo cual sugiere que la secuencia primaria de las LCS podría no ser importante para su función (Huntley y Golding, 2000).

Por otro lado, podemos observar que la mayoría de las secuencias de los ácidos nucleicos y proteínas son diferentes de una cadena de letras generada aleatoriamente, y con ello, los avances teóricos han generado un mejor entendimiento de los atributos de una secuencia, tales como complejidad, patrón y periodicidad (Wootton, 1999). De esta forma, Wootton y sus colegas desarrollaron un método, el algoritmo *Segment Sequence(s) by Local Complexity* (SEG), basado en el contenido informacional para detectar esta baja complejidad, regiones de secuencias simples (ver apéndice I). La complejidad en una secuencia es una medida basada solamente en la composición de residuos

(Wootton, 1993). Así, tres términos que pueden definir matemáticamente a una secuencia de baja complejidad como tal, según Wootton, son:

1) La **complejidad,** la cual provee una representación general del sesgo composicional que es independiente de los otros términos, el patrón y la periodicidad.

2) El **patrón**, que incluye repeticiones irregulares, usualmente analizado por su contenido y espaciamiento de residuos, y por k-grams (donde ***k-grams*** son k-letras-palabras, por ejemplo: ATTG es un 4-gram).

3) La **periodicidad** que es la repetición del tipo de residuos k-grams a un intervalo constante (período, módulo o distancia). Para secuencias de DNA, es habitual distinguir la verdadera periodicidad (repeticiones en tandem, exactas o con variaciones, de un patrón de secuencia de longitud constante) de la quasi-periodicidad, donde las repeticiones se generan como una consecuencia secundaria de diferentes sesgos composicionales en diferentes fases, por ejemplo, periodicidad de módulo 3 para las secuencias codificantes en proteínas.

Un ejemplo de la presencia de estos atributos en una secuencia se presenta a continuación, las tres secuencias mostradas en la tabla presentan baja complejidad, la cual es idéntica en ellas porque tienen la misma composición (G_8, A_8), mas no así, los otros dos atributos antes descritos:

COMPLEJIDAD DE LA SECUENCIA	ATRIBUTOS
1) G A A G G A A A G G G A G A G A	No tiene ni patrón significativo ni periodicidad.
2) **G G A** G G A A A **A G G A** A G G A	Presenta patrones k-gram (GGA y AGGA), distribuidos de forma irregular y no muestran periodicidad.
3) **G A** G A G A G A G A G A G A G A	Tiene periodicidad de módulo 2 (GA), y como una consecuencia, patrones k-gram.

El término "simple o baja complejidad" entonces, hace referencia solamente a una parte de la gran diversidad y riqueza de la variación en las secuencias naturales. El que se pueda contar con atributos que midan esta condición tiene profundas implicaciones para el entendimiento molecular, estructural, funcional y evolutivo de las LCS.

1.2 Los genomas virales.

Desde el punto de vista molecular, los **virus** consisten básicamente de una o varias moléculas de ácido nucleico (su genoma) protegido(s) por una cubierta proteica o cápside (Davis *et al*, 1985) (ver figura 1.3). La gran diversidad de tipos de virus que existe en la naturaleza es, probablemente, sólo un reflejo de la diversidad de tipos de genomas virales que han surgido durante la evolución de la vida (Pedulla *et al*, 2003). Por ejemplo, los virus pueden tener su genoma codificado en moléculas de DNA o RNA y pueden ser de cadena sencilla o de cadena doble; el genoma puede ser lineal o circular y, como se mencionó antes, estar formado en una sola molécula (genoma monopartita) o en varias (genoma multipartita) (Murphy *et al*, 1995). Estos genomas son secuencias de nucleótidos altamente organizados con distintos genes que codifican la información para la producción de diversas proteínas (proteoma) involucradas en la replicación del genoma viral y en su dispersión.

Los genomas virales fueron los primeros sistemas secuenciados (Vega y Rivera, 2001) hace más de 25 años a partir del avance tecnológico de la biología molecular (bacteriófago ϕ174 en 1977), a la fecha, forman una parte importante de las bases de datos de secuencias y genomas completos disponibles (Mills, 2003). Por sus características infectivas se les ha tomado como los vehículos modelo para entender muchos de los procesos celulares, tales como los de replicación, transcripción y traducción de la información. Se distinguieron originalmente por ser pequeños (de ahí, el término original de "virus filtrables") y por ser parásitos intracelulares obligados (Flint *et al*, 2000). Sin embargo, estas propiedades las comparten algunas bacterias. Aún cuando los genomas de los virus son de tamaños diversos, de forma general son pequeños, entre 2.5 y 50 kbs (1 kb = 1 x 1000 nucleótidos), y sólo algunas familias de virus con genomas de DNA de cadena doble son los que poseen los genomas más grandes, entre 200 y 600 kbp (Regenmortel *et al*, 2000). Incluso, se ha encontrado que los genomas virales de Pyramimonas (560 kbp) (**kbp**–*kilobase pairs*) (Sandaa, *et al*, 2001), del bacteriófago D (670 kbp) de *Bacillus megaterium* (Hutson *et al*, 1995), y del Mimivirus (800 kbp) (La Scola *et al*, 2003) son cercanos o más grandes en comparación al tamaño de los genomas de algunas bacterias pequeñas, algunas de ellas parásitas también, tales como *Mycoplasma genitalium* (580 kbp), *Ureaplasma urealyticum* (752 kbp), *Buchnera* sp. (641 kbp), y *Wigglesworthia brevipalpis* (698 kbp) (www.ncbi.nlm.nih.gov/PMGifs/Genomes/eub_g.html).

Actualmente, en los rasgos distintivos de los virus se retoman su organización simple y su mecanismo de replicación. De hecho, una partícula vírica completa o ***virión*** puede ser considerada, principalmente, un bloque de material genético (ya sea de DNA o RNA) capaz de replicación autónoma, rodeado de una capa proteica y a veces, de una cubierta membranosa adicional que le protege del medio y sirve de vehículo para su transmisión desde una célula hospedera a otra (Davis *et al*, 1985).

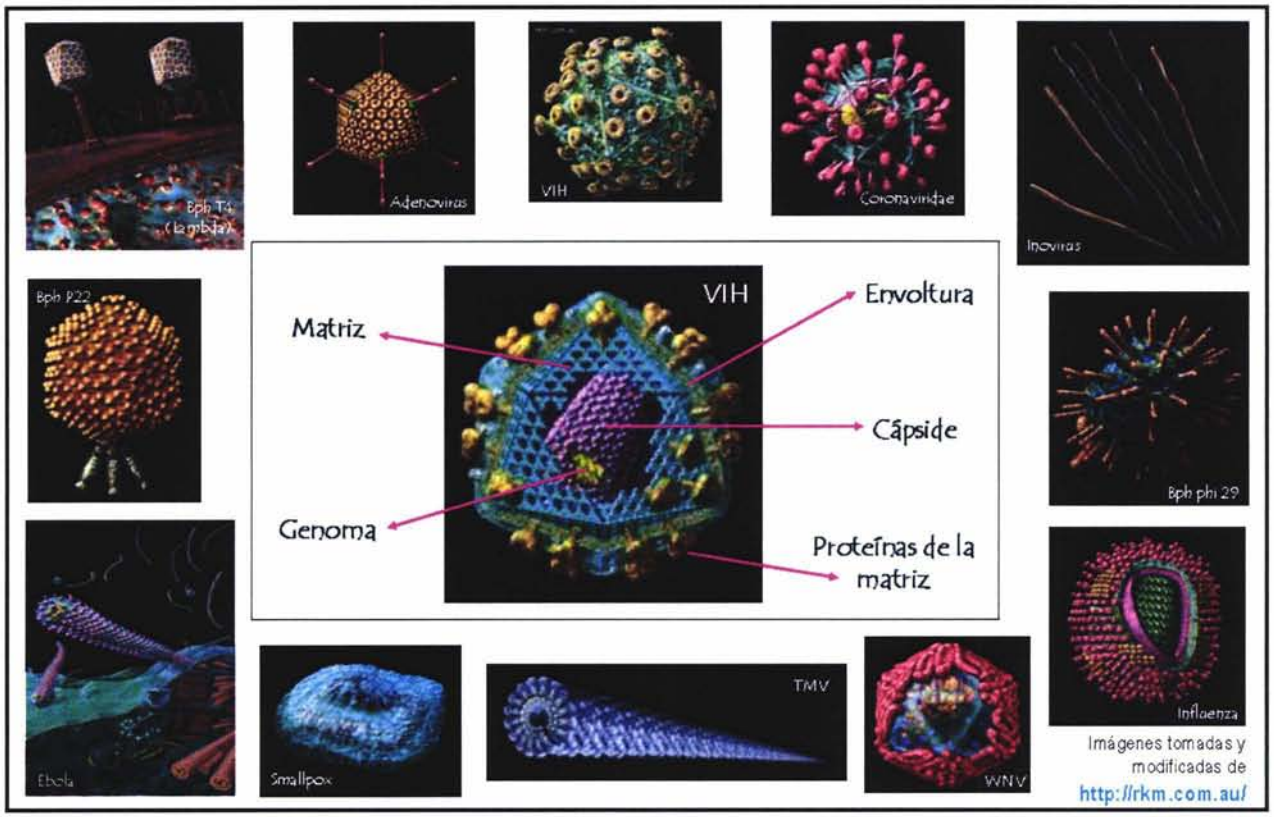


Fig 1.3. Estructura general de un virus. Dada la diversidad de familias virales que existen, el consenso para las características en la estructura de los mismos no es generalizable. Los virus tienen ácidos nucléicos, **RNA** o **DNA**, los cuáles constituyen el genoma viral. Es importante enfatizar que: a) el ácido nucleico puede tener una sola cadena (**ss**, por *single stranded*), doble cadena (**ds,** por *double stranded*), ser lineal o circular, contínuo o segmentado, b) los virus poseen un solo tipo de ácido nucléico. Los componentes básicos de un virus son: a) **proteínas estructurales**, que forman a la partícula viral, y b) **proteínas no estructurales**, tales como las enzimas, c) **cápside**, la cubierta externa, constituída por capsómeros, que son hilos de polipéptidos entretejidos. Esta protección también es útil al virus en la penetración de las células, d) **cápside + ácido nucléico = nucleocápside**. Algunos virus tienen una envoltura lipídica cuyo origen es la misma membrana plasmática de la célula huésped, y que es adquirida al salir las nuevas partículas virales de la célula en un proceso de gemación. Los capsómeros atraviesan esta envoltura como proyecciones tridimensionales de diversas formas y con diferentes funciones y, e) **nucleocápside + envoltura lipídica = virión**. La forma de la nucleocápside determina las diferentes clases de simetría de los virus. Existen virus con **simetría helicoidal**, en la que el virus se aprecia como una espiral con el ácido nucléico en el eje central, como el virus del Mosaico del Tabaco (TMV) con genoma de ssRNA+ o el virus de Ébola y de la Influenza, ambos con genoma de ssRNA-. Otro tipo de **simetría** es la **icosahédrica**. En esta forma geométrica la partícula viral presenta 20 caras con 12 ángulos, por ejemplo el virus de Inmunodeficiencia Humana (VIH) cuyo genoma es de RNA y utiliza una reverso transcriptasa para su replicación, el virus del West Nile y probablemente el Coronavirus, estos últimos con genomas de ssRNA+. Otros virus tienen una combinación de estos últimos dos tipos de simetría, generalmente bacteriófagos de DNA de cadena doble o sencilla, tales como el bacteriófago phi-29, bacteriófago p22, bacteriófago T4 (lambda). Algunos virus con un gran genoma (generalmente Poxvirus, tales como el virus de la Varicela, Smallpox con genoma de dsDNA), tienen lo que se denomina **simetría compleja** (no helicoidal ni icosahédrica), con lípidos tanto en la envoltura como en las membranas externas. Al centro de la figura, se muestra la estructura general del Virus de Inmunodeficiencia Humana que pertenece a la familia Retroviridae, el cual posee un genoma de RNA (en amarillo), una cápside (en morado), una matriz (en azul), y ésta última está compuesta por ciertas proteínas que le proporcionan especificidad infectiva (en café), además presenta una envoltura compuesta por lípidos y fosofolípidos (franja color dorado que cubre la matriz). Información tomada de Regenmortel *et al* (2000).

La división de los virus no es de tipo celular, ya que poseen en su cubierta pocas o ninguna de las enzimas biosintéticas necesarias para su replicación (Flint *et al*, 2000). Por esta razón, los virus se multiplican por síntesis y luego por reunión de sus componentes (Murphy *et al*, 1995). Así, el ácido nucleico vírico, tras desprenderse de sus cubiertas, entra en contacto con la maquinaria celular apropiada, donde especifica la síntesis de las proteínas requeridas para la reproducción virica. El ácido nucleico vírico se replica entonces a sí mismo a través del uso de enzimas víricas y celulares, con lo cual, se forman los componentes de la capa vírica, y estos componentes finalmente se unen al final del ciclo (Murphy *et al*, 1995). En algunos virus, la replicación es iniciada por las enzimas presentes en los viriones.

Hay familias virales de DNA y familias que contienen RNA y que pueden estar formados por una sola cadena (**ss**, por *single stranded*), doble cadena (**ds,** por *double stranded*), en el caso de los virus de DNA, éstos no se encargan de forma directa de la síntesis de proteínas. Las copias de algunos segmentos del ácido nucleico (RNA) dirigen dicha síntesis, algunos virus tienen enzimas específicas, principalmente polimerasas y transcriptasas. Cuando el RNA de un virus puede emplearse directamente como RNA mensajero (mRNA), decimos que tiene "polaridad positiva" (ssRNA+); en cambio, cuando requiere de una transcriptasa para hacer copias (complementarias) en sentido positivo, se habla de "polaridad negativa" (ssRNA-) (ver figura 1.3).

Los virus están distribuidos en los tres dominios celulares. En cada clase taxonómica cada virus es capaz de infectar solamente a ciertas especies de células (Davis *et al*, 1985). Los tipos de hospederos están determinados por la especificidad de adhesión a las células, lo que depende tanto de las propiedades de la capa del virión como de los receptores específicos situados en la superficie celular (Murphy *et al*, 1995). Estas limitaciones desaparecen cuando se produce la transfección, es decir, cuando la infección es llevada a cabo por el ácido nucleico vírico desnudo, cuya entrada no depende de receptores específicos de virus. Por otro lado, el tipo de hospedero para los virus depende de la existencia de factores celulares necesarios para replicación vírica (Murphy *et al*, 1995).

El estudio del origen y evolución de los virus ha dado lugar a no pocos debates. El hecho de que utilicen el mismo código genético que los organismos celulares es un reflejo de que dicho código surgió antes de que aparecieran los virus, aunado a la dependencia del virus a la maquinaria celular para la síntesis de proteínas y la replicación de los ácidos nucleicos; y por otro lado, la observación de que los aminoácidos de proteínas virales ocurren en proporciones similares a las proteínas de otros organismos nos lleva a pensar que a pesar de la simplicidad estructural de los virus es muy poco probable que éstos hayan precedido a los primeros organismos celulares (McGeoch y Davison, 1995). Existen varias hipótesis, algunas más sustentadas que otras, sobre el origen de los virus, en una revisión de Campbell (2001) se propone que los virus: i) son parásitos intracelulares

degenerados, ii) son relictos de vida pre-celular, o iii) son genes celulares que escaparon. Un mejor desarrollo sobre estas hipótesis puede observarse en el apéndice II.

Existen diversos mecanismos por los cuales los virus adquieren variabilidad genética que los convierte en grandes exploradores del proceso evolutivo (Domingo y Novella, 2000). La comparación de las secuencias virales ha llevado a establecer, en los últimos años, las relaciones evolutivas entre genes virales y no virales, construyendo así algunos posibles eventos de la evolución viral (Flint *et al*, 2000; Mills *et al*, 2003). Adicionalmente, se ha observado que los análisis comparativos de las secuencias nucleotídicas de los genomas virales y/o de las secuencias de aminoácidos confirman la clasificación tradicional de los virus congruentemente (Mills *et al*, 2003; Mayo y Pringle, 1998). Dichas comparaciones han mostrado además homologías inesperadas entre genes de virus de diferentes grupos que habían sido considerados poco relacionados entre sí, así como homologías entre genes virales y no virales (Flint *et al*, 2000). De esta manera, por medio del análisis comparativo de secuencias de los virus que existen actualmente, podemos inferir algunas propiedades y características de sus ancestros.

En este sentido, la diversidad viral se ha podido ordenar, estimar, calcular y medir de diversas formas; en este trabajo en particular, se trabajó con la clasificación viral de Baltimore (Mayo, 1997), en la cual los virus están clasificados según el tipo de genoma que poseen y, dentro de cada grupo, a las familias que los representan (ver figura 1.4 de a-f), también se utiliza una segunda categorización de los virus reportada en *The International Committee on Taxonomy of Viruses* (**ICTV**), en la cual se les agrupa según el tipo de hospedero celular que presentan, puede ser eucarionte o procarionte (Mayo y Pringle, 1995). Las características principales de los virus según estas clasificaciones se pueden observar en las tablas II a VII en el anexo 1, las cuales son el resultado del análisis de la información bibliográfica y de las secuencias de los genomas completos, por lo cual habrá familias de la clasificación original de la figura 1.4 que no se encuentren en estas tablas, o grupos representados por un solo integrante, pero que no consideramos necesarios incluir en este trabajo dado los objetivos que se analizan aquí.

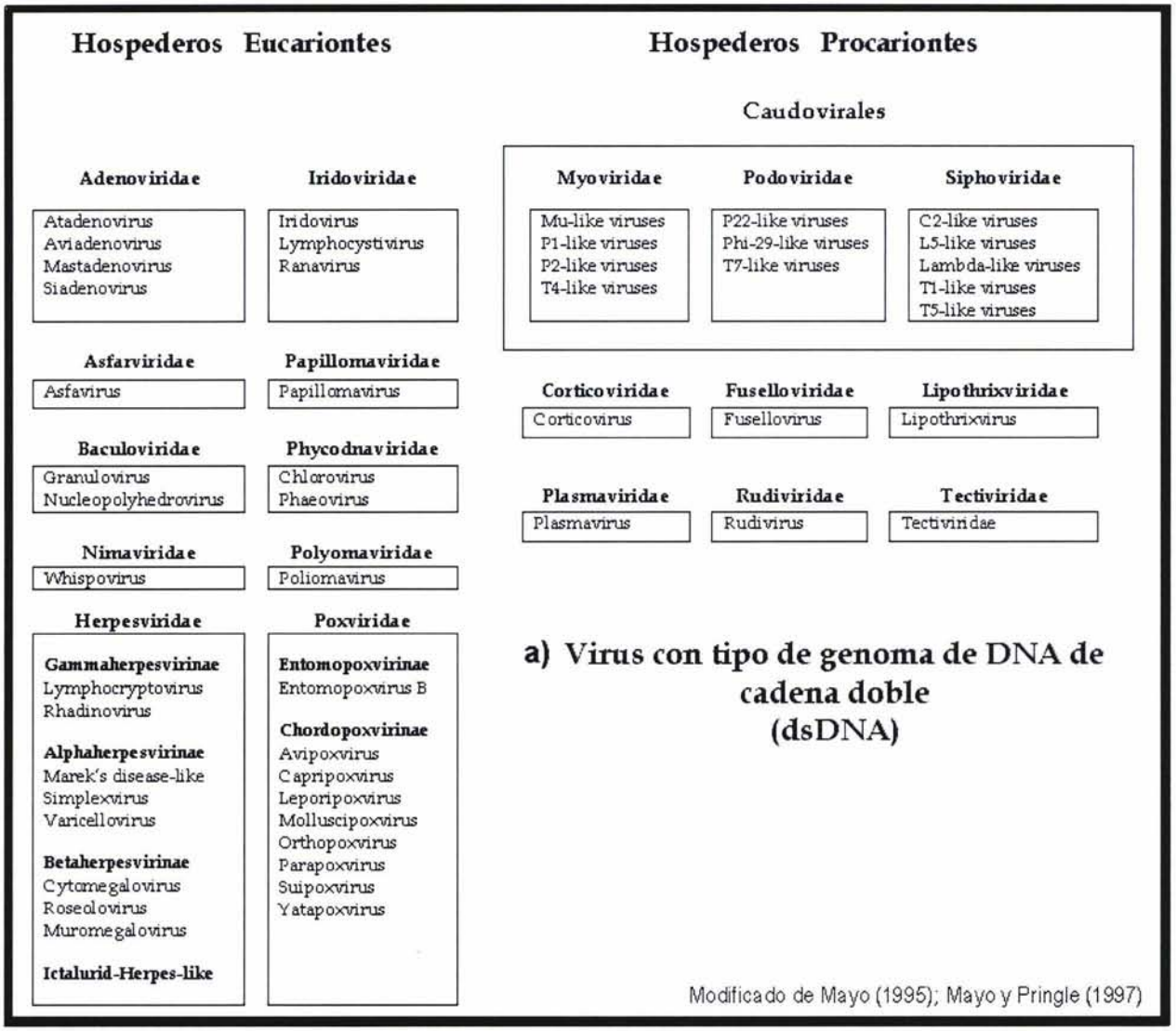


Fig. 1.4. Clasificación taxonómica de los virus según el tipo de genoma que posean y el hospedero celular que parasiten según el Internacional Committee on Nomenclature of Virases (Mayo y Pringle, 1998). Ver desarrollo en la página 19.

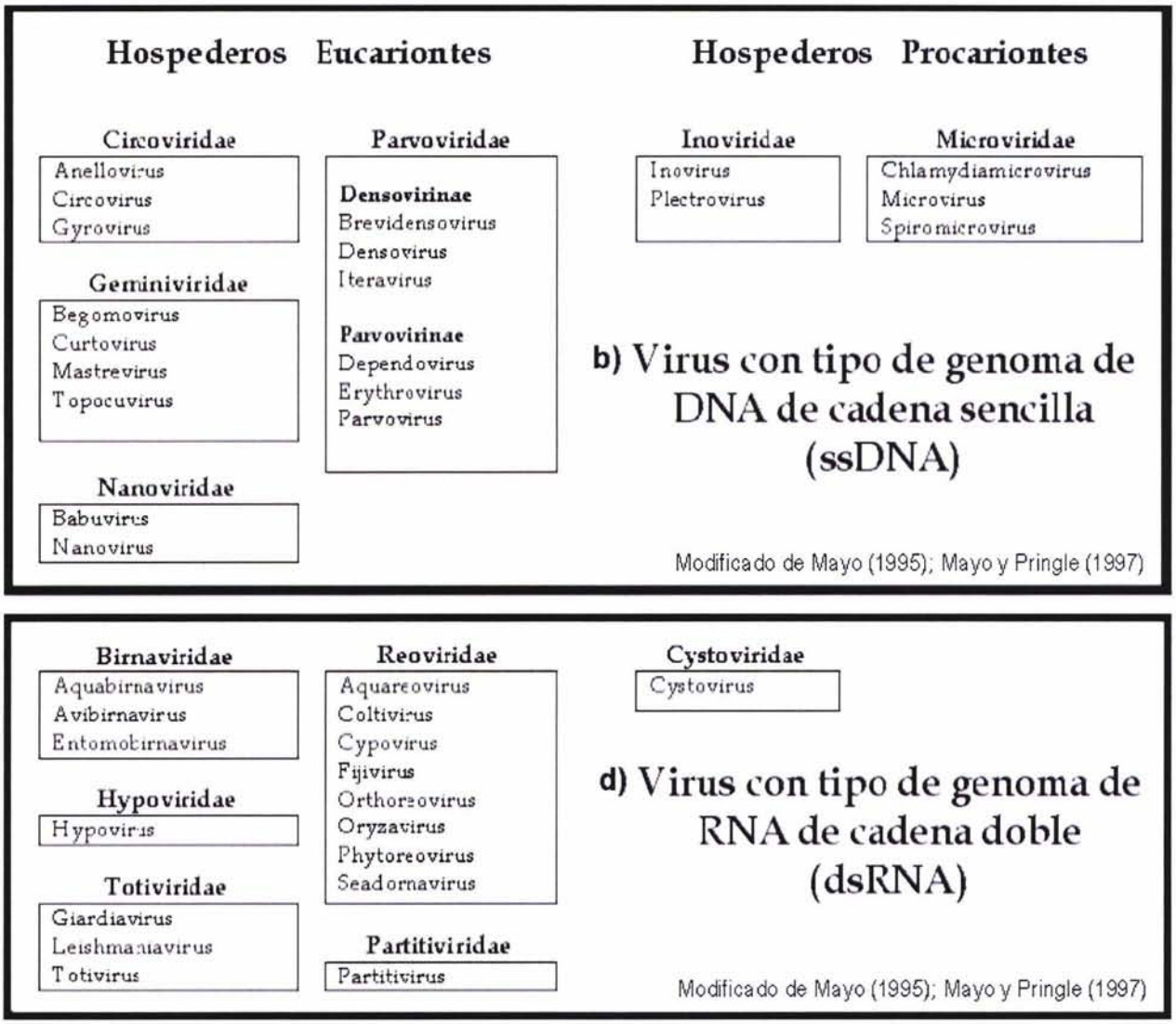


Fig. 1.4. Clasificación taxonómica de los virus según el tipo de genoma que posean y el hospedero celular que parasiten según el Internacional Committee on Nomenclature of Virases (Mayo y Pringle, 1998). Ver desarrollo en la página 19.

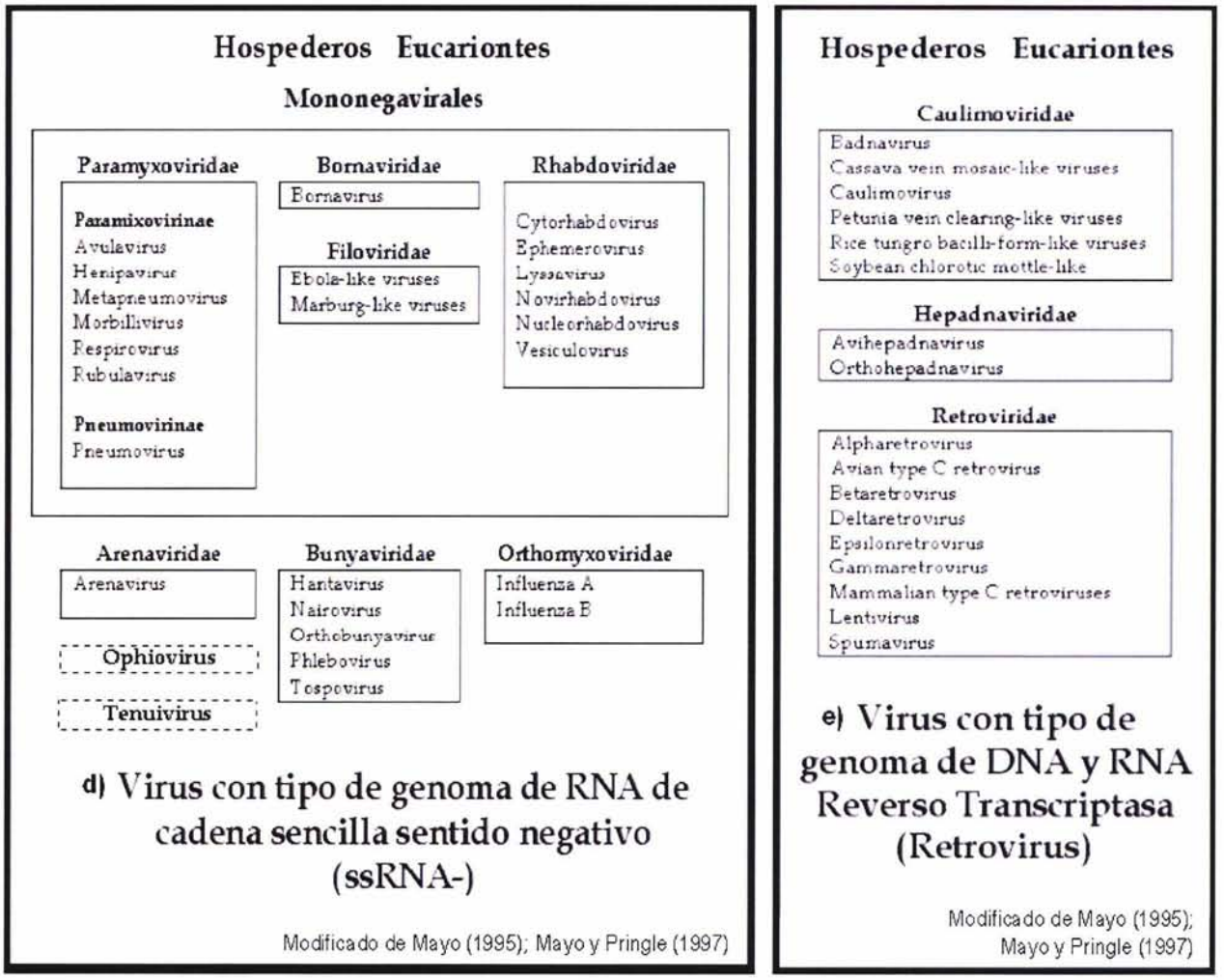


Fig. 1.4. Clasificación taxonómica de los virus según el tipo de genoma que posean y el hospedero celular que parasiten según el Internacional Committee on Nomenclature of Virases (Mayo y Pringle, 1998). Ver desarrollo en la página 19.

Fig. 1.4. Clasificación taxonómica de los virus según el tipo de genoma que posean y el hospedero celular que parasiten según el Internacional Committee on Nomenclature of Virases (Mayo y Pringle, 1998). Las figuras son una lista de los taxa de virus por tipo de genoma (del inciso a al f). Los géneros están listados en cajas que agrupan a las familias correspondientes. Las familias están indicadas por bloques con recuadros de líneas negras, al interior de ellas se encuentran los órdenes taxonómicos que las componen. Los órdenes que se encuentran en bloques con líneas interrumpidas son aquellos que no se encuentran asignados a alguna familia actualmente, sin embargo, en este trabajo también se tomaron estos grupos taxonómicos dada su relevancia en el número de genomas completamente secuenciados disponibles en los bancos de secuencias del National Center for Biotechnology Information (NCBI).

1.3. Las LCS en los proteomas virales.

Aún cuando el fenómeno de las LCS ha sido ampliamente estudiado en los sistemas celulares, no ha tenido el mismo desarrollo en los sistemas virales. Dado que a la fecha los proyectos de secuenciación presentan un interés industrial y de salud prioritariamente, la colección actual de genomas virales completamente secuenciados presenta un sesgo por aquellas especies o cepas de virus que están involucradas a algún proceso infectivo en un pequeño grupo de hospederos con respecto a la gran diversidad de especies en los tres dominios de la vida. A pesar de ello, actualmente, los sistemas virales presentan un gran número de genomas completamente secuenciados con respecto a los celulares, por arriba de 3000 (Mills *et al*, 2003). De esta forma, el análisis de algunas de las secuencias virales ha reflejado una característica compartida con las celulares, que es la presencia de LCS. Aunque son pocos trabajos enfocados en este sentido. A continuación mencionaré el aporte que han hecho algunos de ellos al fenómeno de LCS en virus:

Los primeros trabajos donde se reporta la presencia de las LCS en las secuencias de virus (Wootton y Federhen, 1993; Wootton, 1994) se realizó bajo el análisis de bancos de secuencias completos, tanto de secuencias virales como celulares, tales como el SWISSPROT y el PDB, donde se observó que las proteínas virales Tat y Rev del Virus de Inmunodeficiencia Humana (VIH) presentan LCS compuestas principalmente del aminoácido arginina y cuya función es unir la proteína TAR a la estructura tallo asa del RNA durante cierto estadio del ciclo replicativo del virus. También, se detectaron LCS en la proteína de la cápside VP2 del Parvovirus Canino, la cual está compuesta del aminoácido glicina principalmente. Esta LCS causa desórdenes en el mapa de densidad electrónica cuando trata de resolverse su estructura cristalográfica y su posición dentro de la proteína se encuentra en la región N-terminal, que corresponde al sitio de unión al DNA (ver figura 1.5).

CODE	RESIDUES	NAME	SEQUENCE	COMMENTS
DISORDERED REGIONS OR PARTS NOT CRYSTALLIZED:				
2DPV	21-44	Canine Parvovirus VP2 coat protein	tggngngggggggsggvgistgt	Disordered in electron density map. Part of N-terminal DNA-binding region
3GRS	6 17	Glutathione reductase	pqpqqppmaga	Disordered in electron density map. N-terminal region
1PHS	201-217	Phaseolin	kelskhaksarkslsk	Partly disordered in electron density map
1COL	185-187	Colicin A	lgvpsiavglagilliaavvqali	Part of N-terminal domain which was not crystallized

Modificado de Wootton (1994)

Fig. 1.5. Análisis de LCS en el banco de secuencias del PDB y SWISS-PROT. Donde se observó que las LCS detectadas en el PDB se encuentran en regiones desordenadas estructuralmente, en partes que no pueden ser cristalizadas o en proteínas no-globulares. Dentro de las secuencias analizadas en este trabajo se encontró una LCS ricas en glicinas en la proteína de la cápside VP2 del Canine Parvovirus, la cual se esquematiza dentro del recuadro rojo. De forma similar se detectaron también las LCS en las proteínas Tat y Rev del virus de Inmunodeficiencia Humana principalmente compuestas de arginina como se menciona arriba en el texto. Los parámetros de SEG utilizados para este análisis fueron L(1)= 21, $K_2(1)$= 2.5,$K_2(2)$= 2.7.

Por otro lado, en el grupo de los alfavirus (virus de ssRNA+ y miembros de la familia Togaviridae) hay reportes de LCS localizadas en la región N-terminal de las proteínas de la cápside (Perera *et al*, 2001). Estas LCS están compuestas principalmente del aminoácido prolina y, aún cuando no se le ha asignado una función definida, se ha visto que estas regiones permiten el reconocimiento no específico al RNA. Las regiones conservadas que se encuentran a los lados de estas LCS son muy importantes para el ensamblamiento de la nucleocápside a través de las interacciones de las estructuras terciarias del tipo espiral-vuelta, las cuales pueden estabilizar a los intermediarios formados entre las interacciones del dominio C-terminal de la proteína de la cápside al RNA genómico, contribuyendo así a la estabilidad del virión (ver figura 1.6); además, se demostró que deletando parcial o completamente estas regiones se causa un decremento significativo en la replicación del virus.

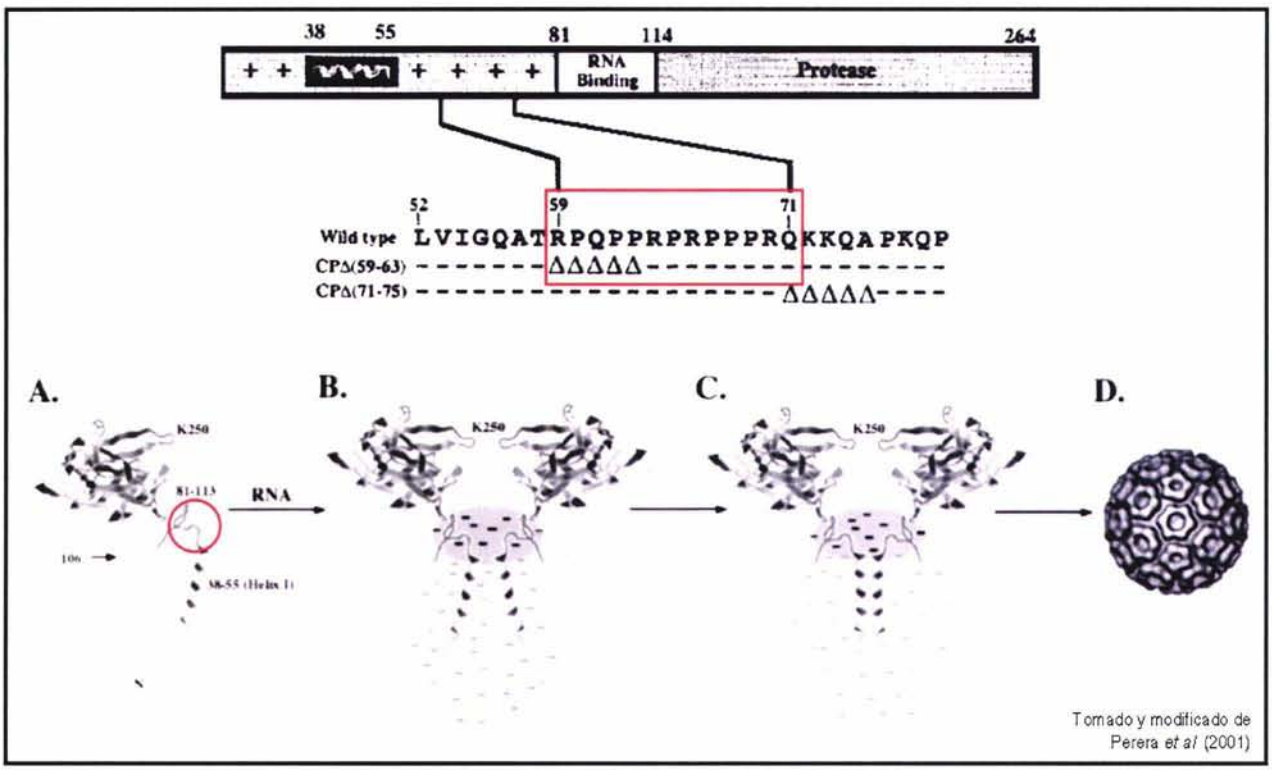


Fig. 1.6. Estructura del CDS que contiene a la LCS de prolina en Sindbis virus (SINV) (alfavirus). El diagrama de arriba esquematiza el CDS que contiene a la LCS de 10 residuos de aminoácidos rica en prolina (recuadro rojo) localizada en la región carboxilo terminal. Los residuos de 1 a 80 tienen un alto grado de carga positiva y están implicados en el reconocimiento no específico del RNA, con la excepción de los residuos del 38 al 55, los cuales no se encuentran cargados y forman la hélice I. Los residuos del 81 al 113 contienen importantes determinantes que están involucrados en el reconocimiento específico al RNA genómico. Residuos del 114 al 264 forman el dominio de la proteasa y también están involucrados en el contacto con los capsómeros y con el dominio citoplasmático de la proteína E2. El diagrama de abajo esquematiza un modelo para las interacciones del enrollamiento en espiral del ensamblamiento del alfavirus. A) Un monómero de la cápside (CP) de SINV, la información estructural está disponible solamente para los residuos del 106 al 264 y el resto se diseñó bajo modelación cristalográfica a partir de las secuencia, los residuos del 38 al 55 son mostrados como una α-hélice. B) Bajo la presencia del RNA (área gris con cargas negativas), el CP forma dímeros al unirse al ácido nucleico iniciado por las interacciones de los residuos del 81 al 264, C) cuya estabilización es mediada por la hélice I, D) hasta juntar 120 copias de estos dímeros y formar a la nucleocápside (NC) del SINV.

Otras LCS han sido identificadas en las secuencias de diversos grupos virales con diferente composición de aminoácidos, formando parte de una familia de secuencias, donde no todas se definen como LCS, a las cuales se les ha atribuido una función específica en el transporte de proteínas nucleares (llamadas *nuclear localization sequences* - **NLSs)** debido a que el correcto reconocimiento de las proteínas nucleares es mediada por este tipo de secuencias al permitir la unión específica al núcleo y subsecuente translocación en la envoltura nuclear vía una importina que reside en el citoplasma y cerca del complejo del poro nuclear (Stochaj *et al*, 1993). Además, la presencia de estas secuencias se han asociado a enfermedades, tales como el Parkinson, el Hutchintong, entre otras (Bremner *et al*, 2001).

Las NLSs fueron primero encontradas en el polipéptido del antígeno T del Simian Virus 40 como un pequeño motivo de 5 residuos continuos cargados positivamente en la secuencia PKKKRKV, esta secuencia pertenece a la familia de NLS simples, caracterizadas por ser LCS de 4 a 8 residuos constituidos principalmente de los aminoácidos de lisina y arginina (Kalderón *et al*, 1984). A partir de ello, se han detectado varias NLSs en las regiones codificantes de varias secuencias; sin embargo, han sido pocos los trabajos donde se han detectado NLSs en las secuencias de virus y que se encuentren definidas como LCS, dado su sesgo composicional. Una de estas NLSs definida como LCS se encuentra ubicada en la región carboxilo terminal del polipéptido caracterizado como el núcleo del antígeno (HBcAg) del virus de la Hepatitis B, el cual está constituido principalmente por 4 grupos del aminoácido arginina (ver figura 1.7). Un análisis de deleción puntual sobre las 4 regiones de este polipéptido revelan que los grupos número 4 (localizado en la parte amino terminal) y 1 (localizado en la región carboxilo terminal) constituyen distintas e independientes NLSs, mientras que los grupos 2 y 3 (localizados en la parte media) no parecen tener influencia en la localización celular del polipéptido HBcAg (Eckhardt *et al*, 1991). La sustitución del segundo residuo de arginina del grupo 4 por uno de treonina inactiva la señal de localización nuclear en esta región del polipéptido HBcAg demostrando la importancia de este residuo para su secuencia señal (ver figura 1.7). Además, se encontró que se acumulan los polipéptidos de HBcAg en el núcleo solamente cuando ambas NLSs (los grupos 1 y 4 de arginina) son simultáneamente deletadas o modificadas por mutación.

Como se ha podido observar, el fenómeno de las LCS no está presente sólo en el material genético de los sistemas celulares. Las LCS detectadas en las regiones codificantes, tanto en los sistemas celulares como en los virales, pueden proporcionar la materia prima para la generación de nuevas funciones en las proteínas de las que formen parte.

En comparación con los sistemas celulares, los sistemas virales presentan características peculiares en el almacenamiento y expresión de la información que pueden ser importantes para explicar la presencia o ausencia y posible función de las LCS en el material genético, tales como: a) diversidad en el tipo y estructuración del ácido nucleico que utilizan para delimitar su genoma, los cuales pueden ser genomas de cadena sencilla o doble DNA o de

RNA (ver figura 1.3); b) el tamaños de sus genomas es generalmente pequeño, siendo los genomas virales de DNA, tanto de cadena doble como sencilla, los que presentan los tamaños más grandes; c) las polimerasas que sintetizan los genomas de los virus de RNA presentan generalmente altas tasas de mutación y presentan pocos o carecen de sistemas de reparación; d) los genomas de algunos grupos de virus de RNA se encuentran fragmentados, y sólo pueden llevar a cabo su replicación cuando se encuentran todos los segmentos en la célula hospedera; e) sus proteomas codifican para pocas clases funcionales en comparación con los sistemas celulares; y f) son de origen polifilético.

Dada la amplia cobertura de las LCS en los sistemas celulares y la correlación positiva de su presencia con el tamaño de los genomas se puede pensar que las secuencias simples son un fenómeno que, por un lado, genera nuevo material genético y que podría dar lugar a nuevas funciones, pero por el otro, que no se esperaría detectar en genomas con tamaños muy pequeños, tales como los que poseen los virus. De esta forma, se toma en este trabajo a los proteomas virales como buenos modelos para evaluar la posible función evolutiva de las LCS en las regiones codificantes del material genético, para lo cual se analizarán características, tales como, la frecuencia de aparición de estas secuencias en las secuencias de aminoácidos de los virus, su distribución por el tipo de genoma viral y por familia taxonómica, la composición de aminoácidos que las conforman principalmente, su longitud y ubicación física dentro de las secuencias (ya sea en las regiones amino, medio o carboxilo terminal) y el tentativo papel que puedan aportar estas LCS a la función general de las proteínas en las que se encuentran. Así, podemos conocer si el aporte de las LCS al incremento del material genético conlleva el generar nuevas funciones.

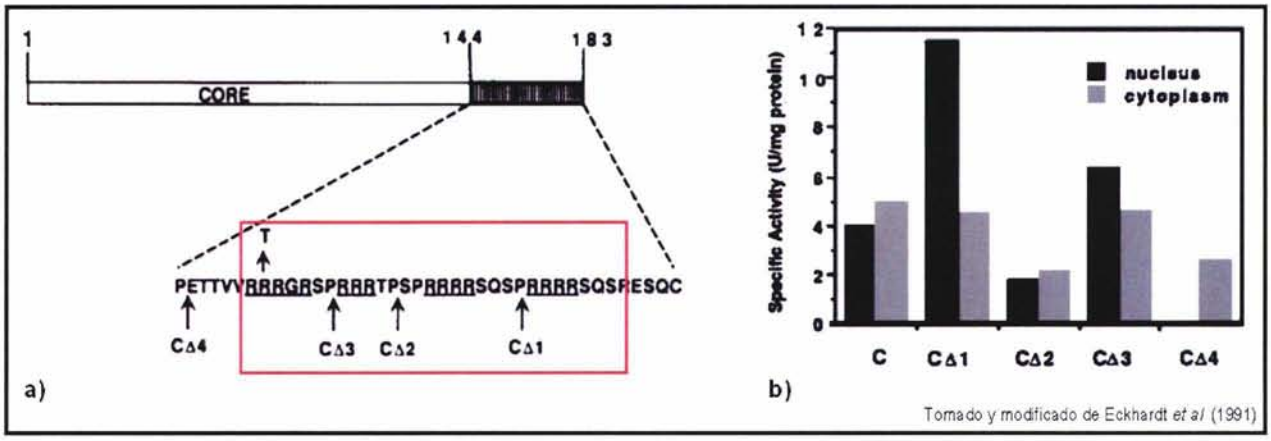


Fig. 1.7. Estructura del CDS de HBcAg que contiene a la LCS de arginina en el virus de la Hepatitis B. a) La región carboxilo terminal del CDS de 40 residuos de aminoácidos rica en arginina (LCS) es mostrada en el recuadro rojo. Los cuatro clusters ricos en arginina (numerados de derecha a izquierda) se encuentran subrayados. Las mutaciones sobre esta LCS están indicadas por CΔ1, CΔ2, CΔ3, CΔ4. **b)** Con análisis de fraccionamiento celular de antígenos relacionados a HBcAg expresados en líneas celulares transfectadas se puede observar que la actividad específica de esta LCS rica en arginina (que incluye a los cuatro grupos) se encuentra principalmente en el núcleo (CΔ1, CΔ2, CΔ3), mientras que la actividad de la región de CΔ4 sólo es detectada en el citoplasma.

2. OBJETIVOS.

GENERAL: Caracterizar la presencia, distribución y abundancia de las secuencias simples en los proteomas virales, y su posible papel biológico.

PARTICULARES:

A. Construcción de una Base de Proteomas Virales completamente secuenciados.

B. Cuantificar y caracterizar las secuencias simples (composición de aminoácidos, longitud y posición dentro de las secuencias) en los proteomas virales.

C. Determinar la posible contribución de las secuencias simples a la función de las proteínas que las contienen.

3. MATERIAL Y MÉTODOS.

I. Construcción de una Base de Proteomas Virales completamente secuenciados.

Se colectaron 2 304 proteomas virales completamente secuenciados hasta diciembre del 2003. A no encontrarse disponibles *vía* ftp, se colectaron de la versión http de las siguientes bases de datos: *National Center for Biotechnology Information* (**NCBI** http://www.ncbi.nlm.nih.gov/), *Kyoto Encyclopedia of Genes and Genomes* (**KEGG** http://www.genome.ad.jp/kegg/kegg2.html), *Virus Database at University College London* (**VIDA 2.0-UCL** http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html), *HIV Databases* (**HIVD** http://hiv-web.lanl.gov/content/index), *Universal Virus Database of the International Committee on Taxonomy of Viruses* (**ICTVDdB** http://life.bio2.edu/). Conformando dos bases de datos:

1) La primera contiene los 2 304 proteomas completamente secuenciados de virus representados por 34 253 secuencias. Se anexa en el apéndice III una tabla con los 2 304 nombres y números de acceso asignados en el NCBI de las especies virales y variantes de éstas que se colectaron en este trabajo.

   Hay varias formas de definir redundancia en una base de datos. Aquí se determinaron varios parámetros para tratar de evitar un alto grado de redundancia entre los genomas virales recopilados y curados, sin que ello signifique perder una representación más amplia de la diversidad característica de los genomas virales, incluso hasta el nivel de serotipos, cepas y aislados. Así pues se tomaron en cuenta, al menos dos parámetros:

   A) El primer filtro se basó en la nomenclatura de los nombres de los virus en las bases de datos empleadas, determinando la mayor cantidad de sinónimos y acrónimos posibles para una misma especie viral. Además se verificó que no se repitiera(n) números de acceso del NCBI de cada especie viral. Así, se dejó un representante de aquellos proteomas de virus completamente secuenciados con más de un sinónimo y acrónimo para la misma especie viral nombrada con el sinónimo más común, según la comparación hecha en todas las bases de datos analizadas.

   B) El segundo filtro se realizó con el programa llamado *Cluster Database at High Identity with Tolerante* (**CD-HIT**) (Weizhong *et al*, 2001) el cual se aplicó bajo el parámetro de 100% de identidad, donde sólo se eliminaron aquellos proteomas virales que fueran 100% idénticos a otro, dejando un sólo un representante.

2) La segunda base de datos contiene la información adicional de los proteomas virales completamente secuenciados e incluye: nombre de la familia taxonómica, nombres, acrónimos y sinónimos de las especies virales que la conforman, sus respectivos números de acceso asignados en el NCBI, hospedero por dominio y grupo taxonómico celular, tamaño del genoma en pares de bases, número de secuencias, lugar donde se colectó, autores y referencias. La estructura general de esta base bibliográfica toma como referencia a la clasificación viral de Baltimore y del **ICTV,** donde se encuentran los seis tipos de genomas virales, por tipo de hospedero eucarionte o procarionte y los 76 grupos taxonómicos que se analizaron en este trabajo, tales como familia y para algunos casos, género.

II. Detección de LCS en los proteomas.

Para la búsqueda de las LCS en las 34 253 secuencias de aminoácidos se utilizó el programa **SEG** (Wootton, 1993) (cuyas generalidades se pueden observar en el apéndice I) para calcular la frecuencia y distribución de las LCS en los proteomas virales y con las opciones –l y –x de este mismo programa se obtuvieron los datos para determinar la composición de aminoácidos y la localización de las LCS en la secuencia, en todos los casos se utilizaron los parámetros: w=12, k1=1.9 y k2=2.1 (ver apéndice I, V y VI). Con el objetivo de evitar falsos positivos se utilizaron valores en los parámetros del SEG más estrictos que los usados de forma estándar en otros trabajos y programas que filtran LCS (Wootton y Federhen, 1993; Wootton, 1994; Altschul *et al*, 1994; Huntley y Golding, 2002).

Por medio de programas de cómputo en lenguaje *perl* se contó el número de LCS detectadas para cada secuencia y el número de éstas que presentaran al menos una LCS en cada grupo taxonómico viral, así como, el número de proteomas que presentara al menos una secuencia con LCS (ver apéndice VI). Además, se analizó la distribución LCS en las regiones carboxilo, medio o amino de las secuencias que presentaron al menos un segmento de LCS. Estas tres regiones (carboxilo, medio o amino) se definieron a cada tercera parte en la secuencia, respectivamente en ese orden. Por último, se calculó la longitud de la LCS detectadas y el porcentaje de cobertura en las secuencias que las contienen con el objetivo de conocer su aportación al incremento en el tamaño del proteoma.

Los residuos de aminoácidos para analizar la composición, tanto de las LCS, las HCS (**HCS,** por *Hight Complexity Sequences*, las que no son LCS de aquellas secuencias que las contienen) como de los proteomas fueron obtenidas con programas en lenguaje *perl* y de los archivos de salida –l y –x del programa SEG. A partir de ello, la composición de los residuos de aminoácidos se determinó con el comando *aacomp* del paquete FASTA (Pearson y Lipman, 1988), modificado y compilado para ampliar el número de símbolos a analizar (aminoácidos, en

este caso) (ver apéndices VI y VII), tanto para las LCS, las HCS y los proteomas completos en cada uno de los 76 grupos taxonómicos virales.

Se calcularon, según los datos obtenidos, las medias poblacionales ($\mu = \Sigma\, x_i / n$), las desviaciones estándar ($\sigma = \sqrt{\Sigma\,(x_i - \mu)^2 / n}$) y el Z-score, a partir de los dos anteriores ($Z_{score} = (x_i - \mu) / \sigma$), para obtener los valores críticos con estas pruebas estadísticas se determinó un nivel de significancia de $p < 0.05$. Se calcularon también, el coeficiente de correlación lineal ($r = [1/n * \Sigma\,(x_i - \mu_x) * (y_i - \mu_y)] / \{ [1/n * \Sigma\,(x_i - \mu_x)^2] * [1/n * \Sigma\,(y_i - \mu_y)^2] \}^{1/2}$) y la prueba de Sperman ($r_s = 1 - [\, 6\, \Sigma\, d^2 / n\,(n-1)\,(n-1)]$), en las cuales, los valores calculados que pueden tomar r y r_s son: $-1 < r / r_s < 1$; donde si $r/r_s > 0$, la correlación lineal es positiva (si sube el valor de una variable sube la otra también) y la correlación es tanto más fuerte cuanto más se aproxime a 1; si $r/r_s < 0$, la correlación lineal es negativa (si sube el valor de una variable disminuye el de la otra) y la correlación negativa es tanto más fuerte cuanto más se aproxime a -1; si $r/r_s = 0$, no existe correlación entre las variables, al menos no lineal, porque podría haber otras del tipo parabólica o exponencial; los valores críticos se tomaron en cuenta a partir de un nivel de significancia de $p < 0.01$ para esta prueba estadística. En las fórmulas X_i y Y_i es la serie de los valores independientes y dependientes, respectivamente, según los resultados obtenidos, n es el total de los datos de la muestra estadística que se lleve a cabo y, d es la diferencia entre dos rangos de datos.

VI. Construcción de la clasificación de las proteínas virales para determinar la función de las LCS en los proteomas virales.

Las categorías funcionales de los virus no están del todo bien establecidas, tanto por la diversidad del tipo de virus que se han reportado y, en muchos casos, por la poca o ninguna similitud con las secuencias celulares de sus hospederos, además de los diversos problemas en la anotación de los CDS. De esta forma, la asignación de alguna de las tentativas funciones de las LCS detectadas en las secuencias de virus se llevó a cabo bajo la búsqueda de dominios conservados usando los modelos ocultos de Markov (**HMM** - *Hidden Markov Models*) de la base de familias de proteínas (**PFAM** - *Protein families database*) (Bateman *et al*, 2000), a la par de su detección en los grupos de ortólogos (**VOG's** - *Viral Orthologous Groups*) para los virus con tipo de genoma de DNA disponibles en el NCBI, los cuales se construyen bajo el mismo principio de los *Clusters of Orthologous Groups of proteins* (**COGs**) (Tatusov *et al*, 1997; Tatusov *et al*, 2001). Antes de detallar la metodología de estas bases de datos, explicaré brevemente el soporte teórico para los HMM, los COG's y la base de datos PFAM, importantes para entender la reconstrucción de las funciones en las secuencias virales y a partir de ellos, inferir en qué tipo de funciones pueden estar involucradas las LCS.

1. Perfiles de HMM.

El análisis de motivos conservados y la creación de bases de datos de motivos se han desarrollado debido a dos causas principales. Históricamente, el análisis de motivos se ha usado para la detección de secuencias cuya similitud con la secuencia problema es demasiado baja como para que puedan ser detectadas por métodos clásicos como BLAST o FASTA. El análisis de motivos descansa sobre la identificación de pequeñas regiones conservadas que pueden ser identificadas en homólogos remotos a pesar de un grado de similitud bajo a nivel global que impida su detección por métodos clásicos. Por otra parte, estos motivos o dominios existen debido a que ciertas regiones de las proteínas son funcional y/o estructuralmente importantes, de forma que pueden ser reconocibles mediante el uso de HMM's y permiten también tener una idea de cuál puede ser el papel funcional que una proteína lleva a cabo.

Existen diversas maneras de describir los motivos o dominios, pero casi todas están relacionadas con expresiones regulares, perfiles o HMMs. A continuación sólo describiré las que se utilizaron en este trabajo.

Primeramente definiremos el concepto de **dominio**, ya que tiene una relación estrecha con los perfiles, los cuales usualmente cubren una mayor parte de las secuencias que los motivos. El concepto de dominio se utiliza con cierta flexibilidad, pero generalmente define una unidad estructural independiente. Sin embargo, en estudios genéticos de deleción a veces se utiliza como sinónimo de la mínima secuencia capaz de realizar la función estudiada. En las bases de datos de dominios como PFAM, un dominio se corresponde con el núcleo del domino estructural, aquella zona más similar entre todas las proteínas de una familia, aunque no necesariamente coincida con los límites del dominio estructural (Bateman *et al*, 2000).

Un **perfil** es una tabla de pesos asociados a aminoácidos y costes asociados a deleciones o *gaps* (Durban *et al*, 1998). Estos valores (también llamados puntuaciones) se usan para calcular puntuaciones de los alineamientos del perfil con secuencias. Cuando el alineamiento de una secuencia con un perfil da una puntuación por encima de un cierto umbral se considera que en dicha secuencia existe el motivo correspondiente al perfil. Los perfiles se pueden construir de muchas maneras. El método más usado, de Gribskov *et al* (1990), requiere de un alineamiento múltiple de secuencias, a partir del cual, usando una matriz de puntuaciones para los cambios de aminoácidos, obtiene los perfiles a partir de las distribuciones de frecuencias de residuos. Los perfiles, al contrario de otras estrategias de búsqueda de motivos, no están confinados a pequeñas regiones con elevada similitud de secuencia, y se intenta más bien que caractericen familias de proteínas.

Los perfiles son más sensibles y robustos que otras aproximaciones, como los consensos, ya que a partir de los alineamientos se puede obtener una puntuación discriminatoria no sólo para los residuos presentes en una posición determinada, sino para aquellos que anteceden y suceden a esta posición particular. Los pesos para los

aminoácidos no encontrados se extrapolan de las matrices de puntuación usadas (PAM, BLOSUM, etc.) (Durban *et al*, 1998).

Los **perfiles de HMM** (*Hidden Harkov Models*) son modelos estadísticos de la estructura primaria consenso de una famillia de secuencias (Eddy, 1996; Durban *et al*, 1998). Este método se ha convertido en uno de los más populares ya que su uso se ha extendido gracias a que ha sido implementado en el paquete HMMER, utilizado en este trabajo; además, la construcción de la base de datos PFAM (Sonnhammer *et al*, 1997) se basa en gran medida en el uso de HMMER. La figura 3.1 muestra un HMM para un alineamiento de 4 secuencias con tres posiciones o estados (m_1, m_2, m_3) en la terminología de HMMs. Cada posición tiene los 20 valores de probabilidad de ser uno de los 20 posibles aminoácidos (barras), cuatro estados de inserción (i_0, i_1, i_2, i_3) y tres estados de deleción (d_1, d_2, d_3). Las flechas representan las probabilidades de transición entre estados. Todos o algunos de los parámetros se estiman del alineamiento.

La ventaja práctica de de usar HMMs es que pueden ser derivados con secuencias no alineadas, siendo el alineamiento uno de los resultados del proceso de obtención del perfil. El PSI-BLAST (Altschul *et al*, 1997) generaliza el algoritmo de BLAST para utilizar un perfil en vez de una secuencia, de forma que mediante una búsqueda iterativa va construyendo un perfil que utiliza para realizar una búsqueda más refinada. Sin embargo, en este trabajo se utilizaron los perfiles de HMMs, cuya ventaja sobre los perfiles que se construyen con PSI-BLAST es que éstos toman en cuenta el símbolo antecesor y sucesor (aminoácidos, nucleótidos o gaps) y no sólo una posición específica conservada de una determinada región (consenso) de la secuencia.

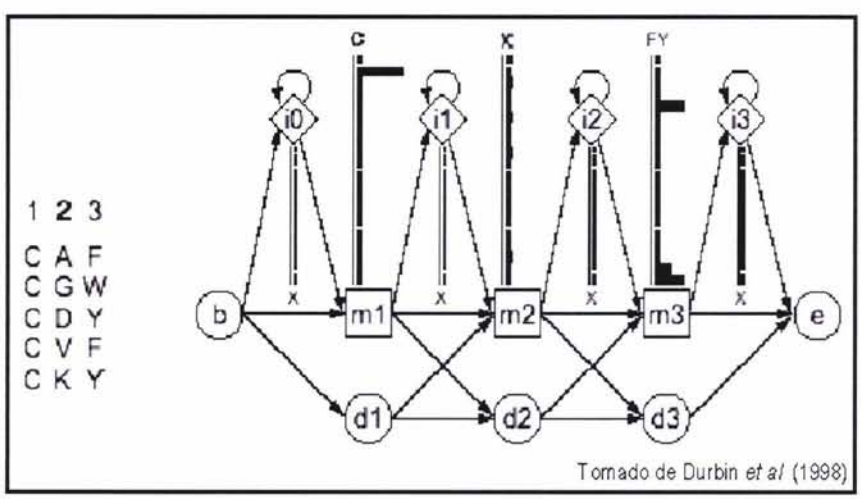


Fig. 3.1. HMM para un alineamiento de cuatro secuencias con tres posiciones. Las columnas del alineamiento son los tres estados (m_1, m_2, m_3) del sistema con 20 probabilidades de ser un residuo (barras correspondientes a las frecuencias observadas de los 20 posibles aminoácidos) cuatro estados de inserción (i_0, i_1, i_2, i_3) y tres estados de deleción (d_i, d_2, d_3). Las fechas representan las probabilidades de transición entre estados. Todos o algunos de los parámetros se pueden estimar del sistema estudiado.

2. Base de datos PFAM.

PFAM (http://www.sanger.ac.uk/Software/Pfam/index.shtml) es una base de datos de perfiles tipo HMM. Se divide en dos partes: **pfam-A** y **pfam-B**. La primera se construye semimanualmente: cada vez que se identifica una nueva familia de proteínas un experto elabora un HMM diagnóstico, un HMM capaz de detectar a las otras proteínas de la familia y sólo a éstas. Por otra parte, como pfam-A sólo cubre el 73% de Swiss-Prot y TrEMBL (http://us.expasy.org/sprot/), existe pfam-B. Ésta se genera automáticamente a partir de aquellos perfiles que existen en PRODOM (http://prodes.toulouse.inra.fr/prodom.htm) (también generados automáticamente) que no se corresponden con ningún pfam-A. Un 20% de las proteínas de Swiss-Prot y TrEMBL presentan al menos un pfam-B.

PFAM trabaja con *dominios*, es decir, cada perfil HMM se corresponde con un dominio, aunque no necesariamente cumplan la definición de dominio estructural independiente, sino más bien suelen ser regiones funcionales características de una determinada familia de proteínas.

Además de las ventajas que de por sí tiene esta clasificación, PFAM resulta útil para:

- Analizar los alineamientos múltiples que contiene.
- Estudiar la organización de dominios de las proteínas.

- Examinar la distribución filogenética de las proteínas que presentan el dominio.
- También permite ver la estructura tridimensional de los dominios, cuando ésta se conoce.
- Permite buscar con una secuencia de una proteína los dominios funcionales conservados empleando los métodos de HMM, los cuales son los más eficaces en el análisis de secuencias actualmente.

3. Grupos de ortólogos.

Por definición, dos proteínas, o genes, son **ortólogos** si han evolucionado a partir de un mismo ancestro del cual divergen a la par de la especiación (Fitch, 1970). Generalmente, genes o proteínas ortólogas tienen la misma función; no tiene por que ocurrir lo mismo entre genes o proteínas parálogas, que en vez de haber sido originadas por especiación (divergencia entre especies), han sido originadas por duplicación de un gen y posterior evolución hasta, a veces, desarrollar una nueva funcionalidad.

La base de datos **COGs (**http://www.ncbi.nlm.nih.gov/COG/**)** (*"clusters of orthologous groups"*) contiene grupos de genes ortólogos. Su objetivo es clasificar en tales grupos las proteínas de aquellos organismos de los que se conoce el genoma completo. Dado que cada COG incluye proteínas de organismos filogenéticamente muy diversos, este sitio es especialmente útil para asignar función a proteínas con función desconocida *a priori*.

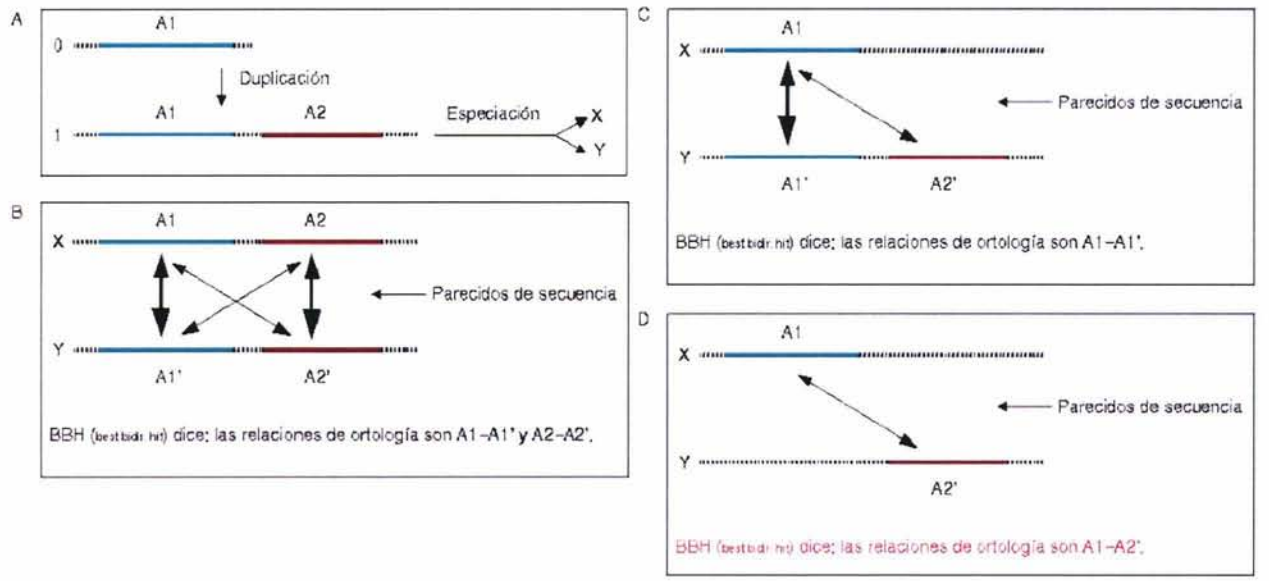


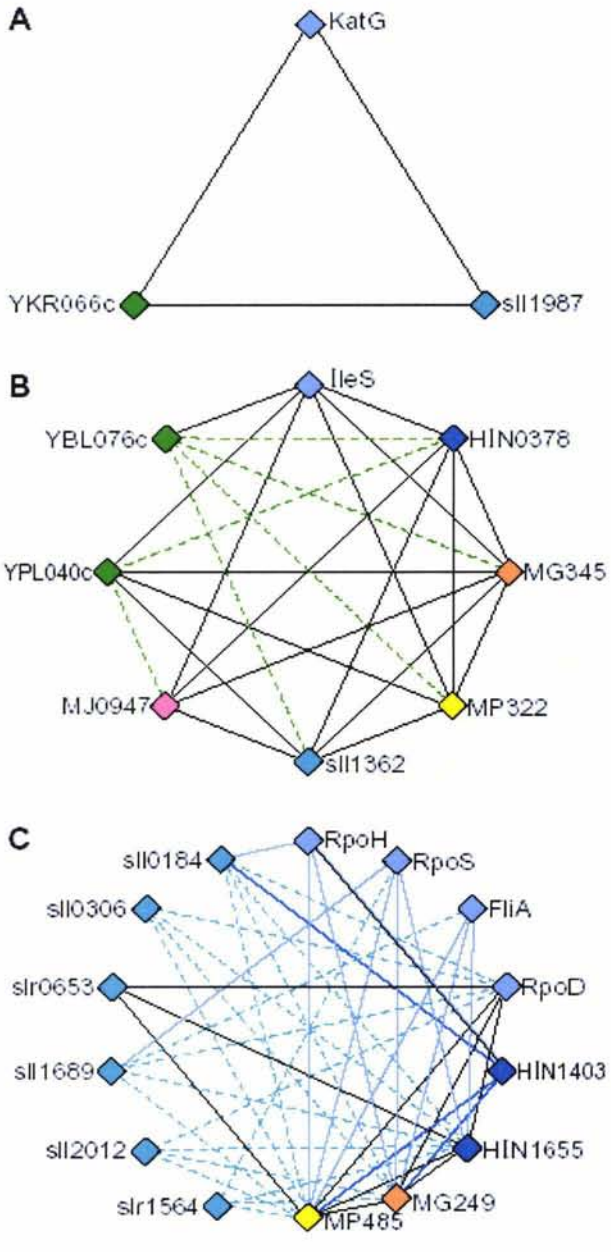
Fig. 3.2. Método de Best Bidirectional Hits para la detección de genes ortólogos. Este no es el método empleado por COGs, pero se parece un poco. La idea del método de ***"best bidirectional hits (BBHs)"*** o *mayores parecidos en las dos direcciones* es que si una proteína de un genoma es la más parecida de otra en otro genoma, y viceversa (bidireccional), entonces es muy probable que ambas sean ortólogos. Por ejemplo, supongamos que tenemos dos genomas X e Y, y en cada uno de ellos hay dos proteínas homólogas A1 y A2 que proceden de una duplicación ancestral en un genoma 0 (**A**). Si ninguno de los dos genomas X e Y sufre una deleción, el método funcionará correctamente (caso **B**). Si se produce una deleción dependiendo de a qué genes afecte el método funcionará bien (**C**) o mal (**D**).

El método de determinación de grupos de ortólogos de COGs.

Es un método **semiautomático**, ya que después de aplicar el método que a continuación se describe, se realiza una corrección de los resultados.

- Lo primero que se hace es determinar los BeTs (best hits, **pero no bidireccionales,** ver figura 3.2) para cada proteína se determina cuál es la más parecida en cada uno de los otros genomas.
- Fusión de *in-paralogs*: Como los *in-paralogs* (duplicaciones recientes dentro de una misma especie) pueden crear confusión (puede que no tenga sentido determinar con cuál de los in-paralogs se ha de establecer el BBH; puede que sea imposible determinarlo), lo que se hace es *fusionarlos*, tomándolos como si fueran uno solo. El criterio para la fusión es que su parecido sea muy elevado.

- Con las relaciones de BeTs se construye un grafo a partir del cual se buscan patrones consistentes de BeTs. El más sencillo de éstos es un triángulo de genes procedentes de tres linajes diferentes (ver figura 3.3).
- Los triángulos que comparten dos vértices (un lado) se unen.
- En muchos casos, dos o más grupos de ortólogos quedan unidos. En esos casos se construye un árbol filogenético y se separan.
- Finalmente, se asigna una función particular a cada COG y también una **clase funcional** general.



Tomado de Tatusov *et al* (1997)

Fig. 3.3. Ejemplos de COGs (Tatusov *et al*, 1997). Las líneas sólidas muestran los BeTs simétricos (ver figura 3.2). Las líneas discontinuas muestran los BeTs asimétricos, con colores correspondientes a las especies para el cual el BeT es observado. Genes de la misma especie están contiguos; de otra manera los nombres de los genes están posicionados arbitrariamente. Un único ID COG es indicado en la esquina superior izquierda. **A)** BeTs congruentes forman un triángulo, el **COG mínimo**. Origen de las proteínas: KatG, *Escherichia coli*; sll1987, *Synechocystis* sp.; y YKR066c, *Saccharomyces cerevisiae*. Nótese que todos los BeTs son simétricos. **B)** Un **simple COG con dos parálogos** de levadura. Origen de las proteínas: Iles, *E. coli*; HIN0378, H. influenza; MG345, *M. genitalium*; MP322, *M. pneumoniae*; MJ0947, *M. jannaschii*; y YBL076c y YPL040c, *S. cerevisiae*. Nótese los triángulos contiguos con un lado común, por ejemplo, IleS-MG345-MJ0947 y sll1362-MG345-MJ1362. YPL040c es la isoleucyl-tRNA sintetasa de la mitocondria de la levadura; los ortólogos bacterianos y aquellos de *M. jannaschii* son los BeTs para esta proteína de la levadura, pero el inverso es positivo sólo de las proteínas bacterianas (BeTs simétricos). A la inversa, para YBL076c, la cual es la isoleucyl-tRNA sintetasa citoplasmática de la levadura, el ortólogo *M. jannaschii* es un BeT simétrico, donde los BeTs bacterianos son asimétricos. **C)** Un **COG complejo con parálogos múltiples**. Origen de las proteínas: RpoH, RpoS, RpoD, y FliA, *E. coli*; HIN1403 y HIN1655, *H. influenzae*; MG249, *M. genitalium*; MP485, *M. pneumoniae*; sll0184, sll0306, slr0653, sll1689, sll2012, y slr1564, *Synechocystis* sp. RpoD, HIN1655, slr0653, y MG249 son los principales factores sigma (σ70), cuya función es universal en el dominio bacteria; nótese las relaciones completamente simétricas entre estas proteínas. Las otras proteínas son especializados factores sigma, cuya radiación de la familia ancestral, aparentemente, fue acompañada por la modificación de la función e implicada evolución acelerada; nótese los BeTs asimétricos.

4. Asignación de función a las LCS en las secuencias que las contienen.

La asignación de alguna de las tentativas funciones de las LCS detectadas en las secuencias se llevó a cabo bajo la aproximación de la búsqueda de dominios conservados usando **HMM** de **PFAM** y las clasificaciones funcionales que se utilizan en los *Viral Orthologous Groups* (**VOG's**). Se describe puntualmente la metodología empleada para esta parte del trabajo y el diagrama 3.1 esquematiza la estrategia empleada aquí:

1) Se compararon las 34 253 secuencias de aminoácidos que se colectaron contra los modelos de **HMM** de **PFAM** version 9 (Bateman *et al*, 2000), usando el programa HMMER 2.3.1 (Eddy, 1996) con un valor de expectancia menor al 0.01.

2) Para asignarle la función viral a los PFAMs detectados en las secuencias de virus se utilizaron los **VOGs** para los virus con tipo de genoma de DNA disponibles en el NCBI (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/vog.html) (Tatusov *et al*, 1997; Tatusov *et al*, 2001) , dentro de cada VOG hay una función general compartida para las secuencias virales analizadas derivada a partir de la anotación de las mismas y muchos de estos VOGs presentan PFAM caracterizados previamente también. Las categorías funcionales de cada VOG para los virus con tipo de genoma de DNA pueden incluir una o más de las siete jerarquías funcionales, tales como:

Símbolo	Funciones Virales
M	*Movement proteins.*
L	*DNA replication, Repair and Nucleotide Metabolism.*
T	*RNA replication, Transcription and Modification.*
S	*Structural proteins.*
A	*Auxiliary proteins.*
R	*Regulation of cellular metabolism.*
U	*Unknown proteins.*

Dado que los VOGs no se encuentran disponibles *vía* ftp en el NCBI, se colectaron de forma manual y con un programa en lenguaje *perl* (ver apéndice IX) se asignó la categoría funcional a todas las secuencias virales de nuestra base de datos filtradas en el inciso uno que presentaran el mismo PFAM asignado por HMMER y que se encontrara en los VOGs.

3) Por último, por medio de un programa en lenguaje *perl* (ver apéndice IX) se asignó tentativamente una o más de las categorías funcionales, arriba mencionadas, a aquellas LCS que se alinearan con un 25% de cobertura con la posición del PFAM detectado para esa secuencia viral por HMMER y, a la vez que también se encontrará registrado en los VCOGs.

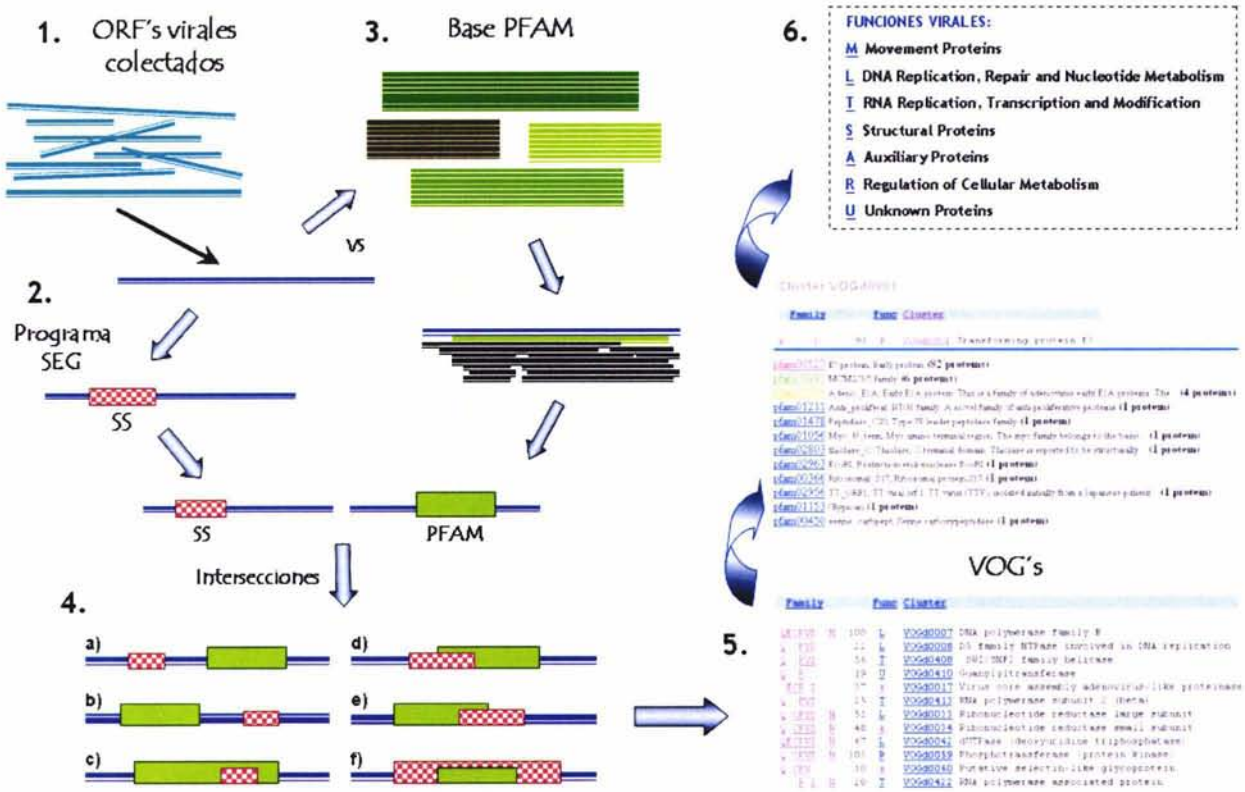


Diagrama 3.1. Se esquematiza la metodología llevada a cabo para asignar funciones a las LCS detectadas en las secuencias virales. El paso uno consistió en obtener la base de secuencias virales (de proteomas completamente secuenciados). En el paso dos se detectaron las LCS en las secuencias virales. En el paso 3 se compararon los secuencias virales contra los modelos *Hidden Markov Models* (**HMM**) de *Protein families database of alignments and HMMs* (**PFAM**) version 10 para detectar aquellas secuencias que presentaran similitud con un valor de e >0.01 a algún dominio PFAM. Los pasos dos y tres se llevaron a cabo simultáneamente y previos al cuatro. En el paso cuatro se compararon las regiones donde se encontraban tanto la LCS y el PFAM caracterizado para una secuencia determinada, sólo si y sólo si en una misma secuencia se encontraban ambos, de 4a a 4f se muestran todos los casos esperados para la localización de las LCS dentro de las secuencias que las contienen y que presentan al menos un dominio PFAM. Para los casos 4c a 4f se procedió con el paso cinco, en el cual, se detectaron los PFAM's que se caracterizaron para la construcción de los VOG's, y cada VOG tiene una categoría funcional previamente caracterizada en el NCBI. Entonces, aquellas secuencias que tuvieron, al menos, una LCS y que tuvieran el mismo PFAM (o más de uno) que alguno de los caracterizados en los VOG's se les asignaron la categoría funcional reportada, las cuales pueden presentar a más de una de las siete, mencionadas anteriormente (paso seis).

4. RESULTADOS.

I. Análisis de la base de proteomas virales construida.

La tabla 4.1 muestra cuatro características generales para los seis tipos de genomas virales (dsDNA, ssDNA, dsRNA, ssRNA-, ssRNA+ y Retrovirus), a saber: 1) el número de familias y géneros que los conforman, 2) el tipo de hospedero específico, 3) el intervalo del tamaño del genoma en nucleótidos y, 4) el intervalo del número de secuencias de aminoácidos que conforman los proteomas de cada grupo. Estas características se obtuvieron del análisis de la base de proteomas colectada y serán de importancia para la discusión. Es importante hacer notar que los proteomas de dsDNA poseen los tamaños de genomas más grandes, y por ende, se conforman por un mayor número de secuencias de aminoácidos; sin embargo, los grupos que se encuentran en una mayor diversidad de diferentes tipos de hospederos son los de RNA, a excepción de aquellos que se encuentran en el grupo de los Retrovirus. Es importante mencionar que no hay genomas virales completamente secuenciados reportados con hospederos procariontes (bacteriófagos) para los grupos de ssRNA- y Retrovirus. Un desglose de esta tabla por tipo de grupo viral se puede observar de las tablas I a VI del anexo 1.

			VIRUS CON HOSPEDEROS EUCARIONTES		VIRUS CON HOSPEDEROS PROCARIONTES	
Tipo de Genoma	**Grupo taxonómico**	**Hospederos específicos**	**Tamaño genoma (nts)**	**No. secuencias por proteoma**	**Tamaño genoma (nts)**	**No. secuencias por proteoma**
dsDNA	19 familias	V, Ins, H, A, B	4 669 – 335 593	4 – 698	10 079 – 280 334	11 – 381
ssDNA	6 familias	V, Pl, B	1 291 – 10 958	1 – 11	4 421 – 9 183	4 – 15
dsRNA	6 familias	V, Ins, Pl, Pr, H, B	3 090 – 29 210	1 – 13	13 173 – 14 984	13 – 16
ssRNA-	7 familias 2 géneros	V, I, Pl	3 694 – 25 159	3 – 12	NR	NR
ssRNA+	22 familias 11 géneros	V, Pl, H, Ins, B	2 343 – 31 357	1 – 14	3 466 – 4 276	3 – 4
Retrovirus	3 familias	V, Pl	1 928 – 13 246	1 – 11	NR	NR

Tabla 4.1. Datos generales de los proteomas que se agruparon por tipo de genoma a partir de la base de datos colectada. La primera columna muestra el número y tipo de grupos taxonómicos que conforman cada tipo de genoma viral. La segunda columna muestra a los hospederos específicos que parasitan los virus con ese determinado tipo de genoma. La simbología y categorías utilizadas (Regenmortel *et al,* 2000) para los nombres de los hospederos son: (V) vertebrados, (Pl) plantas, (Pr) protistas, (H) hongos, (Ins) insectos, (B) bacterias y, (A) arqueas. La tercera columna contiene el intervalo del tamaño de los genomas, en pares de bases, que conforman a cada uno de los seis grupos de genomas virales. La cuarta columna contiene el intervalo del número de secuencias de aminoácidos que contiene cada uno de los proteomas dentro de su grupo taxonómico viral. Estos datos nos permiten observar la heterogeneidad de los proteomas virales por su tamaño y contenido genómico, aún incluso dentro de cada grupo taxonómico, como la categoría de familia (ver tabla I a VI del anexo 1). NR significa que no hay genomas virales completamente secuenciados para esos grupos.

II. La presencia y distribución de las LCS en los proteomas virales.

Se recolectaron un total de 2 304 proteomas virales completamente secuenciados representado s por 34 253 secuencias de aminoácidos. En estas secuencias se encontró un total de 12 317 LCS formando parte de 7 617 secuencias (22% del total). La distribución de LCS en los seis grupos de virus clasificados por tipo de genoma se muestra en la tabla 4.2 y por tipo de hospedero celular en la tabla 4.3.

	Colectados		Detección de Secuencias Simples			
Tipo de Genoma	**Proteomas**	**Secuencias**	**Proteomes**	**Secuencias**	**% Sec**	**Segments**
dsDNA	393	26 009	386	5 136	20	8 734
ssDNA	405	2 273	200	283	12	329
dsRNA	77	390	51	104	27	125
ssRNA-	146	1 030	107	224	22	274
ssRNA+	914	2 646	615	912	34	1 590
Retrovirus	369	1 905	340	958	50	1 265
TOTALES	**2 304**	**34 253**	**1 699**	**7 617**	**22**	**12 317**

Tabla 4.2. Distribución de las LCS detectadas en los diferentes tipos de genoma. Las dos primeras columnas muestran la distribución de los 2 304 proteomas virales colectados y las 34 253 secuencias de aminoácidos que los conforman en los seis tipos de genoma. Las columnas restantes muestran: el número de proteomas que tuvieron al menos una secuencia con un segmento de LCS, el número total de secuencias que presentaron LCS, el porcentaje que esto representa para el grupo y el número de segmentos de LCS que se detectaron en todos los proteomas virales para cada uno de los seis grupos.

EUCARIONTES							PROCARIONTES					
	Colectados		Detección de Secuencias Simples				Colectados		Detección de Secuencias Simples			
Tipo de Genoma	**Proteomas**	**Sec**	**Proteomas**	**Sec**	**% Sec**	**Segmentos**	**Proteomas**	**Sec**	**Proteomas**	**Sec**	**% Sec**	**Segmentos**
dsDNA	263	16 236	260	4 195	26	7 471	130	9 773	126	941	10	1 263
ssDNA	370	1 909	165	219	11	253	35	364	35	64	18	76
dsRNA	73	333	47	91	27	107	4	57	4	13	23	18
ssRNA-	146	1 030	107	224	22	274	0	0	0	0	0	0
ssRNA+	904	2 607	613	910	35	1 583	10	39	2	2	5	2
Retrovirus	369	1 905	340	958	50	1 265	0	0	0	0	0	0
TOTALES	**2 125**	**24 020**	**1 532**	**6 597**	**27**	**10 953**	**179**	**10233**	**167**	**1020**	**10**	**1359**

Tabla 4.3. Distribución de las LCS detectadas en los diferentes tipos de genoma y tipo de hospedero celular. Se muestra la misma información que en la tabla 4.2, pero ahora se desglosa por tipo de hospedero celular, tanto eucarionte como procarionte.

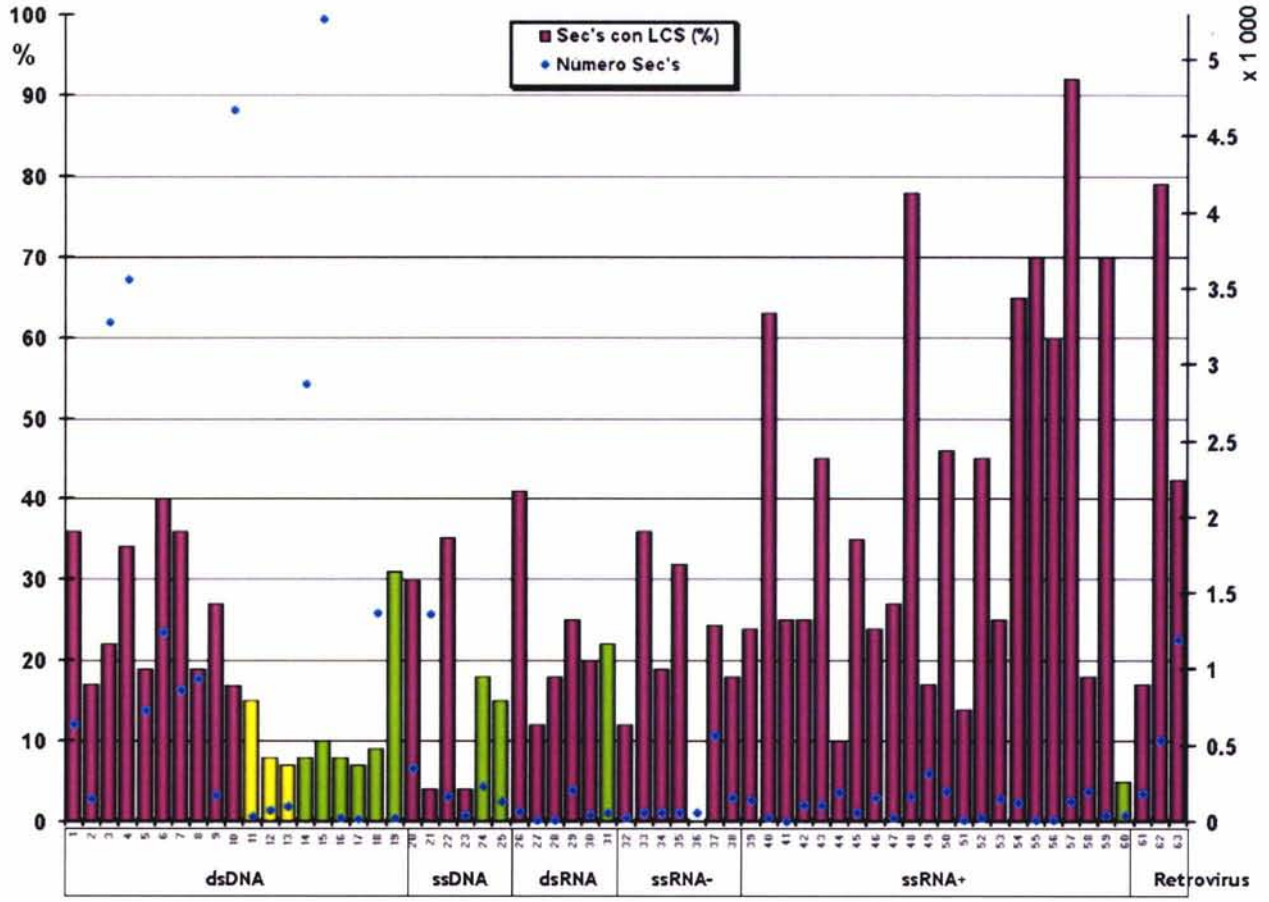
En la gráfica 4.1 se muestra la frecuencia de las LCS en los proteomas de las 63 familias virales, de las cuales sólo en la familia Orthomyxoviridae (ssRNA-) no se detectó ninguna secuencia con LCS. En esta gráfica se puede

observar que los proteomas virales con hospederos procariontes presentan porcentajes menores al 10% de LCS (barras en amarillo para virus con hospederos en arqueas y en verde para bacterias), al encontrarse por debajo de una desviación estándar (σ) de los porcentajes obtenidos para los 76 grupos taxonómicos, a excepción de las familias Fuselloviridae, Siphoviridae, Tectiviridae, Inoviridae, Microviridae y Cystoviridae (números 11, 15, 19, 24, 25 y 31 en la gráfica 4.1, respectivamente) los cuales presentan entre el 15 y el 30% de LCS en sus proteomas. Por otro lado, de forma general, los virus que claramente presentan una mayor cantidad de LCS en sus proteomas son los de ssRNA+ y los Retrovirus, así, dentro de los proteomas de ssRNA+ hay una sobre-representación de LCS por arriba de una σ en las familias Astroviridae (40), Flaviviridae (49), Potyviridae (54), Sequiviridae(55), Tetraviridae (56), Togaviridae (57) y Tymoviridae (59) y, las familias Retroviridae (63) y Hepadnaviridae(62) en los Retrovirus. De los 13 géneros virales analizados tampoco en Idaeovirus (F) (ssRNA+) se detectaron secuencias con LCS (ver gráfica 4.1).

La distribución de los proteomas colectados y de las LCS detectadas en ellos para cada uno de los seis tipos de genoma y los 76 grupos taxonómicos que los conforman se muestran en las tablas y gráficas I a VI en el anexo 1, las cuales incluyen: el tipo de hospedero, el número de proteomas colectados, el intervalo y tamaño promedio del genoma, el intervalo y promedio del número de secuencias que los conforman, el número de proteomas que presentaron al menos una secuencia con LCS y, el número y porcentaje de secuencias con LCS.

Se analizó si existe una correlación entre de la presencia de las LCS y el tamaño del proteoma, sea por el número de aminoácidos o por el número de secuencias que los conforman. Los resultados obtenidos y analizados bajo el coeficiente de correlación (**r**) y la prueba de Spearman ($\mathbf{r_s}$) con un $p < 0.01$ para los 6 grupos de genomas virales (ver material y métodos) nos muestran que, a excepción de los proteomas de aquellos virus con genomas de dsDNA y de dsRNA, no existe una clara correlación entre la presencia de las LCS y el tamaño del proteoma (ver figura 4.1). Tampoco parece existir una clara correlación entre la longitud de las LCS detectadas y el tamaño del proteoma dado el número de aminoácidos que lo conforman (ver I a VI en el anexo 1), sólo en los virus con genomas de dsDNA ($r = 0.725$) se detectó un incremento en la longitud de las LCS conforme iba incrementándose el proteoma, de hecho, en este tipo de genoma se encontraron las LCS de mayor longitud de nuestro estudio, mayores a 300 aminoácidos.

Al analizar el aporte en residuos de aminoácidos de las LCS al tamaño de los proteomas virales, se encontró que en la mayoría de los grupos virales las LCS contribuyen con menos del 1% al tamaño total del proteoma en residuos de aminoácidos, siendo las familias Herpesviridae (4), Nimaviridae (6), Phycodnaviridae (8), Poxviridae (10), Fuselloviridae (11), Tectiviridae (19), Circoviridae (20), Luteoviridae (50), Tymoviridae (59), Retroviridae (63) y Hepadnaviridae (62) cuyas LCS aportan poco más del 1% al tamaño total del proteoma, entre 3 y 6%, donde 11 y 19 son familias virales con hospederos procariontes (ver gráfico 4.2).



Gráfica 4.1. Frecuencia (%) de las LCS dentro de las 63 familias y tipo de genoma viral. Las barras del 1 al 63 representan las familias. Las barras rojas representan a los proteomas virales con hospederos eucariontes, las amarillas arqueanos y las verdes con hospederos bacterianos, agrupados todos ellos por tipos de genoma. En el eje izquierdo de las ordenadas se encuentran los porcentajes de las presencia de LCS para cada familia y, en el derecho el número de secuencias, en miles, que componen a los proteomas (puntos en azul para cada barra). Abajo se muestra qué número representa a cada familia viral y con qué letra a cada género.

TIPO DE GENOMA	FAMILIA GENERO	NUMERO LETRA
dsDNA	Adenoviridae	1
	Asfarviridae	2
	Baculoviridae	3
	Herpesviridae	4
	Iridoviridae	5
	Nimaviridae	6
	Papillomaviridae	7
	Phycodnaviridae	8
	Polyomaviridae	9
	Poxviridae	10
	Fuselloviridae	11
	Lipothrixviridae	12
	Rudiviridae	13
	Myoviridae	14
	Siphoviridae	15
	Corticoviridae	16
	Plasmaviridae	17
	Podoviridae	18
	Tectiviridae	19
ssDNA	Circoviridae	20
	Geminiviridae	21
	Parvoviridae	22
	Nanoviridae	23
	Inoviridae	24
	Microviridae	25
Retrovirus	Caulimoviridae	61
	Hepadnaviridae	62
	Retroviridae	63
dsRNA	Birnaviridae	26
	Hypoviridae	27
	Partitiviridae	28
	Reoviridae	29
	Totiviridae	30
	Cystoviridae	31
ssRNA-	Arenaviridae	32
	Bornaviridae	33
	Bunyaviridae	34
	Filoviridae	35
	Orthomyxoviridae	36
	Paramyxoviridae	37
	Rhabdoviridae	38
	Tenuivirus	A
	Ophiovirus	B
ssRNA+	Benyvirus	C
	Furovirus	D
	Hordeivirus	E
	Idaeovirus	F
	Iflavirus	G
	Pecluvirus	H
	Pomovirus	I
	Sobemovirus	J
	Tobamovirus	K
	Tobravirus	L
	Umbravirus	M
ssRNA+	Arteriviridae	39
	Astroviridae	40
	Barnaviridae	41
	Bromoviridae	42
	Caliciviridae	43
	Closteroviridae	44
	Comoviridae	45
	Coronaviridae	46
	Dicistroviridae	47
	Flaviviridae	48
	Flexiviridae	49
	Luteoviridae	50
	Narnaviridae	51
	Nodaviridae	52
	Picornaviridae	53
	Potyviridae	54
	Sequiviridae	55
	Tetraviridae	56
	Togaviridae	57
	Tombusviridae	58
	Tymoviridae	59
	Leviviridae	60

Los números corresponden a familias.

Las letras corresponden a géneros.

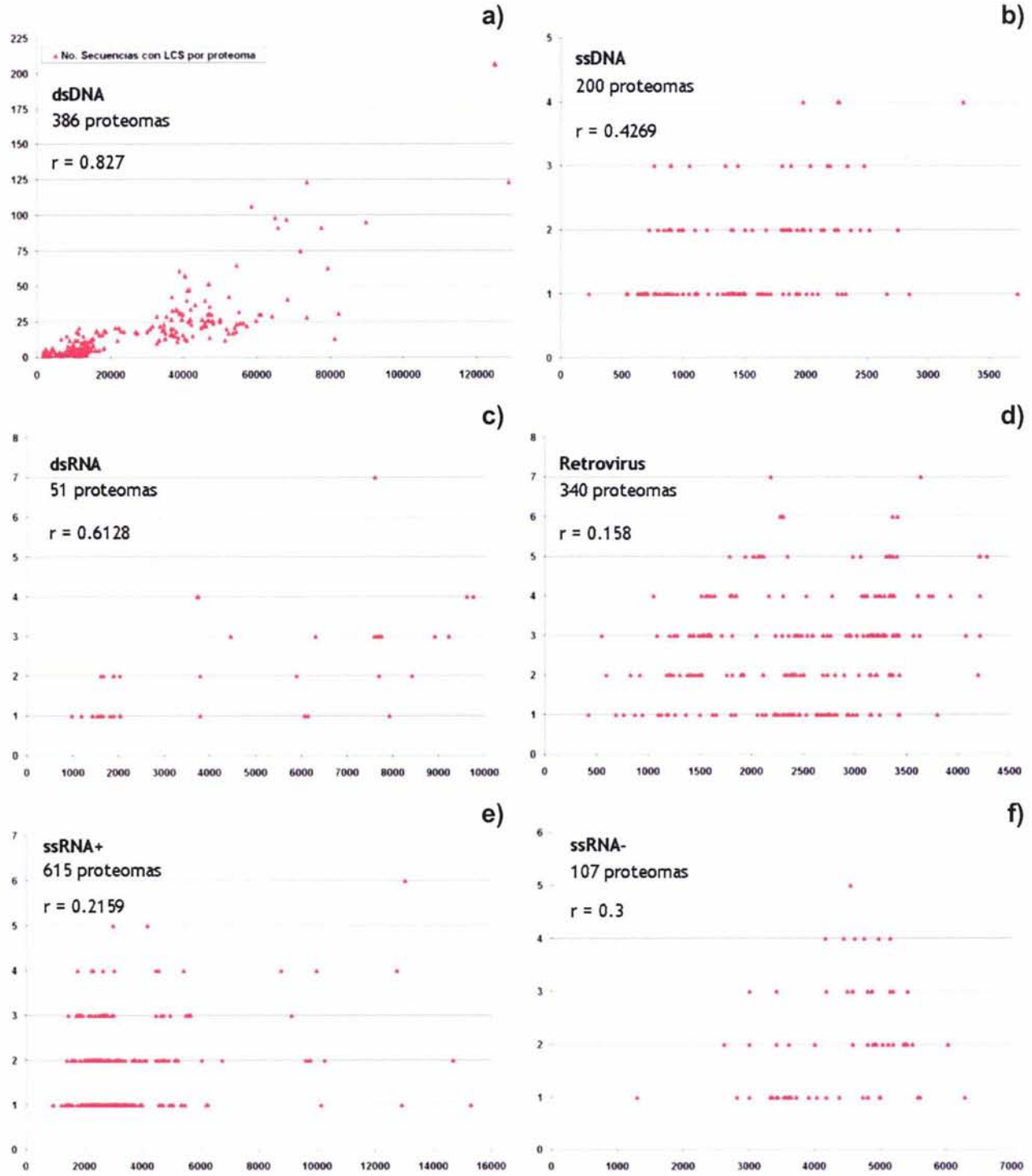
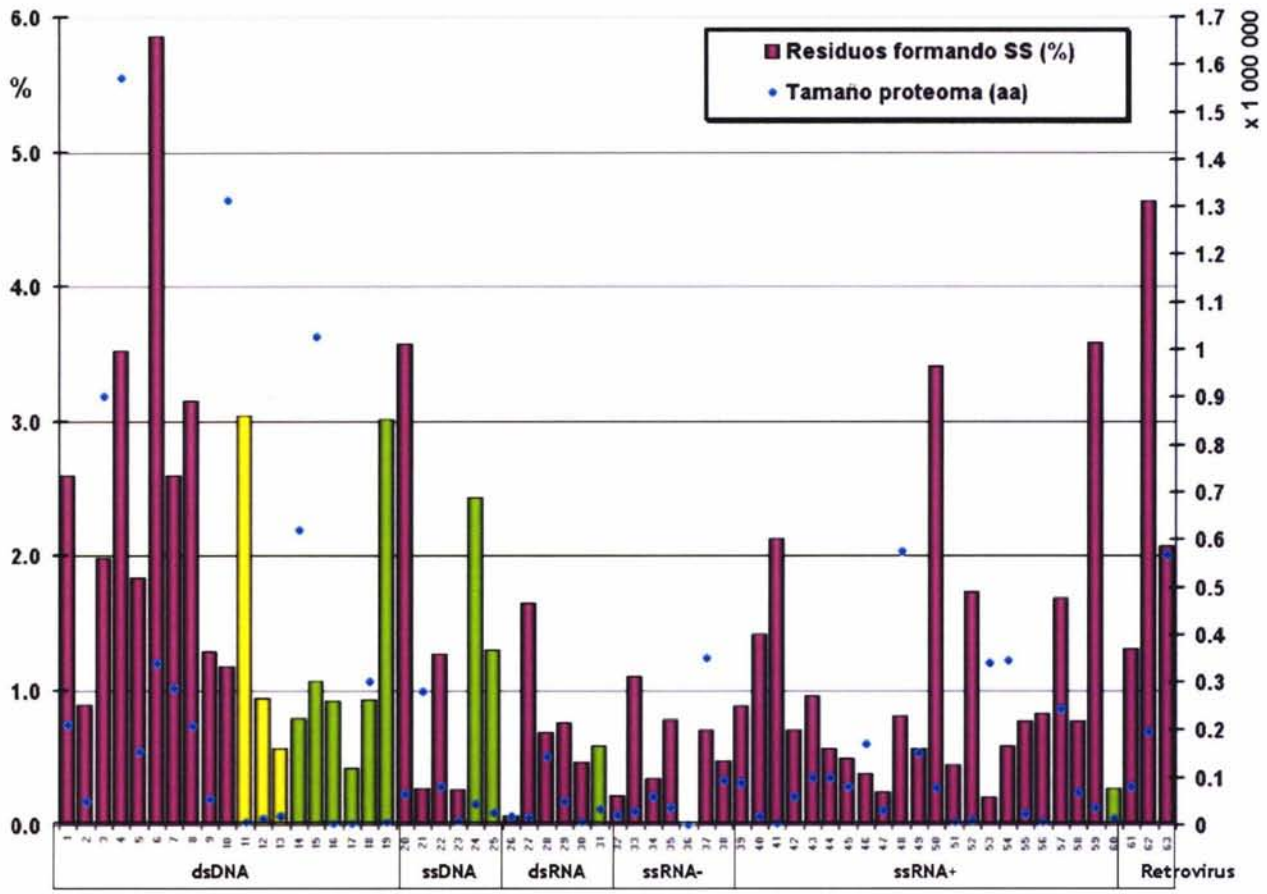


Fig. 4.1. Análisis de la correlación entre el tamaño del genoma dado el número de aminoácidos que lo conforman y la presencia de las LCS en los seis tipos de genomas virales. Sobre el eje de las abscisas se encuentran de forma creciente los tamaños de los proteomas virales dado el número de aminoácidos que los conforman en los seis tipos de genomas virales (a: dsDNA, b: ssDNA, c: dsRNA, d: Retrovirus, e: ssRNA+ y, f: ssRNA-); sobre el eje de las ordenadas se encuentran el número de secuencias con al menos un segmento de LCS por proteoma detectado. Se realizó un análisis de correlación y la prueba de Spearman con un valor de significancia de $p < 0.01$ para determinar esta correlación (ver material y métodos). Según este análisis parece que no existe una clara correlación entre la presencia de las LCS y el tamaño del proteoma, a excepción de los virus con genomas de DNA y de RNA de cadena doble, en los que esta correlación es clara.



Gráfica 4.2. Contribución relativa (%) de las LCS al tamaño de los proteomas virales en términos de los residuos de aminoácidos. Las barras del 1 al 63 representan las familias virales. Las rojas representan a los virus con hospederos eucariontes, las amarillas estrictamente con hospederos en arqueas y las verdes con hospederos en bacterias, agrupados todos ellos por tipos de genoma. En el eje izquierdo de las ordenadas se encuentran los porcentajes y en el derecho el número de total de residuos de aminoácidos (puntos en azul para cada barra) (la escala está en millares) que conforman los proteomas de un determinado grupo. El aporte de las LCS no sobrepasa el 1% para la mayoría de las familias virales.

III. La posición relativa dentro de las secuencias y composición de aminoácidos de las LCS en los proteomas virales.

La presencia del número de segmentos de las LCS no es unitaria en las secuencias virales que las contienen, ya que de forma general, existe una sobre-representación de tales segmentos en una sola secuencia según el tipo de genoma, en la tabla 4.4 y en I a VI del anexo 1 se puede observar que los genomas de dsDNA presentan una mayor cantidad de segmentos de LCS por secuencia analizada, de forma especial en las familias Herpesviridae (4) y Nirmaviridae (6).

Al analizar la posición relativa (%) de las LCS dentro de las secuencias bajo un análisis de Z-score (ver material y métodos), se pudo observar que de forma general las LCS se pueden encontrar con la misma frecuencia a lo largo de las secuencias, como se muestra en las figura 4.2a y 4.2d-f. Sin embargo, en los virus de ssDNA, se puede observar que un gran número de LCS se agrupan en las regiones amino terminal y, que la mayoría de las LCS de virus de dsRNA se ubican en la región carboxilo terminal preferencialmente (ver figura 4.2b y 4.2c). Para ambos casos, los valores de la σ son grandes (ver tabla 4.4, columna cinco). Un desglose más detallado de las posiciones en las que se ubican las LCS dentro de las secuencias para cada tipo de genoma de puede observar en I a VI del anexo 1.

			Posición relativa (%) de LCS en la Sec			
Tipo de Genoma	**Seg_LCS**	**LCS_dentro_Sec's**	**μ**	**σ**	**inicial (+/- una σ)**	**final**
dsDNA	8 734	1 - 28	48	30	18	78
ssDNA	329	1 - 4	30	31	-1	61
dsRNA	125	1 - 4	58	32	26	90
ssRNA-	274	1 - 3	51	34	17	84
ssRNA+	1 590	1 - 8	41	30	11	70
Retrovirus	1 265	1 - 5	54	29	25	83
TOTALES	12 317					

Tabla 4.4. Número y distribución relativa de las LCS dentro de las secuencias que las contienen por tipo de genoma viral. En la primera columna se encuentran los seis tipos de genoma virales, en la segunda el número total de segmentos de LCS detectados en las secuencias que las contienen, en la tercera columna se encuentra el intervalo de segmentos de LCS que se pueden encontrar dentro de una sola secuencia, según el tipo de genoma. Las columnas 4 a 7 muestran los datos estadísticos realizados para calcular el Z-score (ver material y métodos), donde se muestra el promedio de la posición relativa (en porcentaje para normalizar las longitudes de las secuencias) de las LCS dentro de las secuencias (μ), la desviación estándar de las mismas (σ), y las regiones relativas (%) por arriba y debajo de una σ en las que se agrupan principalmente las LCS dentro de las secuencias que las contienen.

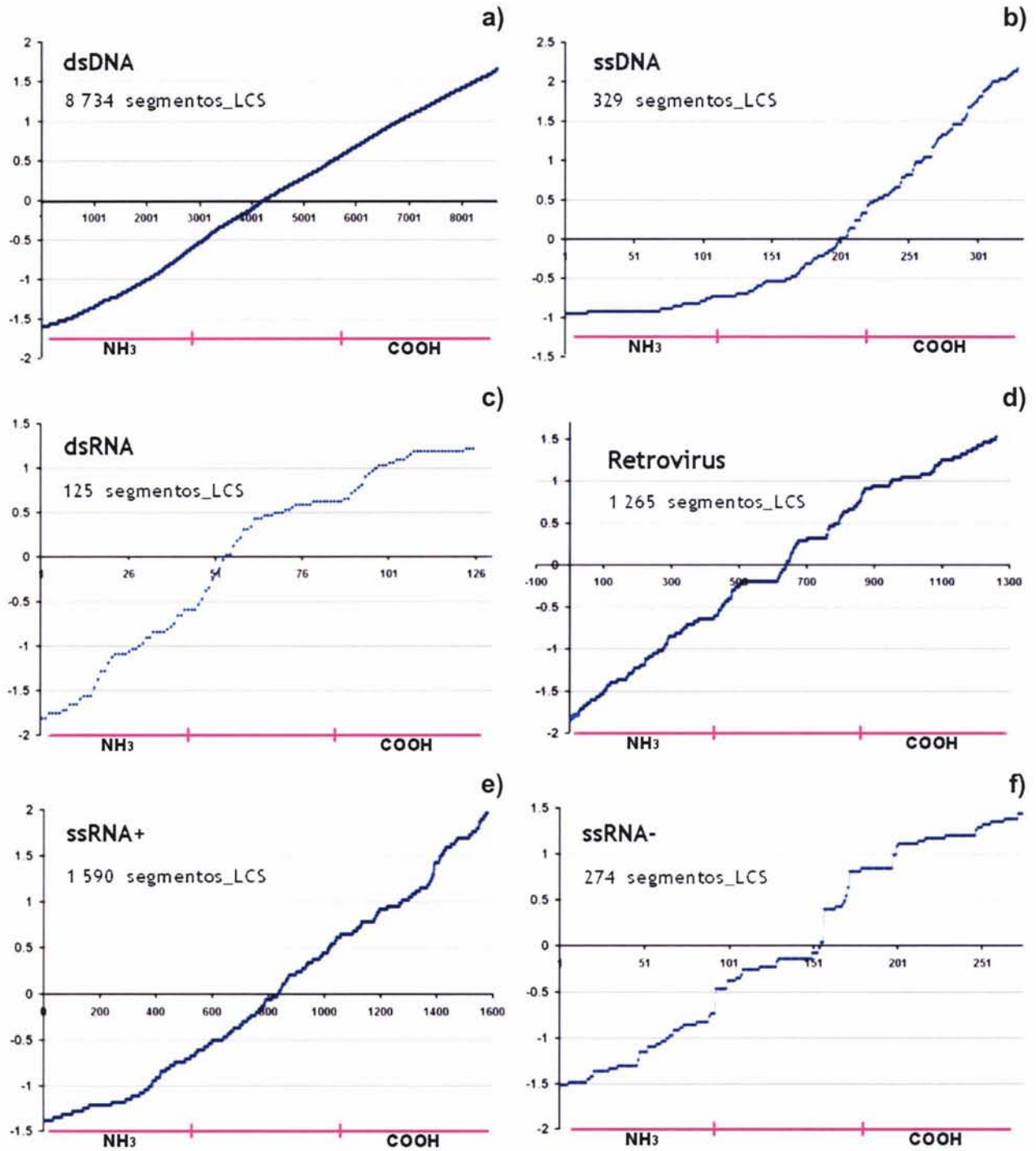


Fig. 4.2. Análisis de la posición relativa de las LCS dentro de las secuencias que las contiene, según el tipo de genoma en los que se agrupen. Se realizó el Z-score (ver material y métodos) para determinar la posición (%) de las LCS en las secuencias. Mientras menores o mayores sean los valores del Z-score calculados para cada posición nos dice que LCS se encuentran ubicadas hacia la parte amino o carboxilo terminal de la secuencia, respectivamente. Por otro lado, los intervalos del Z-score se encuentran entre 3 y -3, lo cual nos dice la tendencia de las LCS a ubicarse estadísticamente en las regiones que por arriba y debajo de una σ se obtuvieron bajo este análisis (ver tabla 4.4). De forma general, las LCS se localizan a lo largo de todo las secuencias, sólo en los virus con genomas de DNA y RNA de cadena doble (b y c) se observa que un gran número de LCS se agrupa preferencialmente en las regiones amino y carboxilo terminales de las secuencias.

Al llevar a cabo el análisis composicional de los proteomas virales según el tipo de genoma y el hospedero celular, se puede observar que existe una uniformidad dentro de los 20 aminoácidos que los conforman (menores al 7%); sin embargo, aminoácidos tales como glicina, alanina, valina, leucina y serina, en algunos casos más que en otros, se encuentran sobre-representados en la composición de los proteomas virales (ver figura 4.3a).

Con respecto a la composición de aminoácidos que componen a las LCS en los proteomas virales se encontró que están formadas principalmente por aquellos aminoácidos que se encuentran sobre-representados en la composición original de los proteomas, como los mencionados arriba, pero además se encontró que otros aminoácidos poco representados o sub-representados en la composición original de los proteomas también se encuentran, en algunos grupos, formando principalmente a las LCS, tales son isoleucina (en genomas de dsRNA con hospederos procariontes y en ssRNA-), prolina (dsDNA, dsRNA, Retrovirus, así como también en ssDNA y ssRNA+ con hospederos procariontes), treonina (ssRNA-), lisina (dsRNA con hospederos procariontes), arginina (dsDNA, ssDNA y Retrovirus), glutamato (dsDNA en ambos hospederos) y aspartato (ssRNA+ con hospederos procariontes). En las LCS hay una baja representación de aminoácidos con importantes restricciones a ciertas propiedades bioquímicas, tales como triptofano, cisteína e histidina (ver tabla 4.5 y figura 4.3b).

Aminoácidos	dsDNA	dsDNA_Pro	ssDNA	ssDNA_Pro	dsRNA	dsRNA_Pro	ssRNA+	ssRNA+_Pro	ssRNA-	Retrovirus
G	**8.94**	**13.1**	**12.72**	**23.56**	4.71	**8.88**	6.54	**14.29**	3.99	4.48
A	**11.09**	**20.53**	4.72	5.48	**16.7**	**21.62**	**7.31**	0	6.55	3.87
V	3.75	4.72	1.74	3.63	6.03	**18.15**	**7.62**	**11.43**	**7.65**	2.36
L	6.72	**7.09**	2.54	**12.17**	5.84	**12.74**	**9.02**	0	**13.94**	**15.89**
I	4.15	3.38	3.11	3.49	1.57	**8.88**	5.11	0	**14.67**	5.4
M	0.54	0.61	0.29	0.36	0.06	0.39	2.51	0	1.23	0.22
P	**11.38**	5.33	5.68	**7.76**	**19.15**	3.86	4.95	**45.71**	2.51	**11.69**
F	1.59	1.2	2.35	1.71	1.38	0.39	4.2	0	0.6	3.39
W	0.21	0.3	0.98	0.43	0	0.39	1.51	0	0.37	1
S	**12.6**	**8.1**	**11.25**	**11.89**	**11.11**	4.25	**7.12**	5.71	**10.49**	**10.29**
T	5.88	5.27	4.97	5.62	6.65	4.63	6.43	0	**11.15**	5.01
N	3.4	2.4	1.49	3.7	5.27	1.54	4.18	0	1.85	2.05
Q	2.87	3.05	2.3	1.92	2.2	0	3.37	0	1.72	3.88
D	5.98	4.68	1.3	3.13	4.33	2.7	5.2	**17.14**	4.93	1.68
E	**7.52**	**9.05**	2.94	5.2	1.88	0.77	5.69	2.86	6.55	6.06
C	0.82	0.26	0.07	1.35	0.06	0	2.17	0	0.37	1.19
Y	1.11	0.79	3.86	0.85	0.56	0	3.43	0	2.48	0.76
H	0.89	0.39	2.18	0.21	0.38	0	2.35	0	0.16	0.52
K	3.4	5.63	3.74	3.63	6.34	**8.49**	5.86	2.86	5.4	3.34
R	**7.15**	4.14	**31.75**	3.91	5.78	2.32	5.4	0	3.39	**16.84**
Total aminoácidos	132, 183	19, 199	4, 088	1, 405	1, 593	259	2, 782, 723	35	3, 831	21, 518
Total SS	7, 471	1, 263	253	76	107	18	1, 583	2	274	1, 265

Tabla 4.5. Se desglosan las frecuencias relativas de los aminoácidos que conforman a las LCS agrupadas por los seis tipos de genomas virales y hospederos celulares (hospederos eucariontes y procariontes). Las frecuencias se manejan en porcentajes de los aminoácidos acomodados en incremento por el peso molecular y por su tendencia a formar alfa-hélice (Creighton, 1984), G: glicina, A: alanina, V: valina, L: leucina, I: isoleucina, M: metionina, P: prolina, F: fenilalanina, W: triptofano, S: serina, T: treonina, N: asparagina, Q: glutamina, D: asparagina, E: glutamato, C: cisterna, Y: tirosina, H: histidina, K: lisina, R: arginina. Aquellos aminoácidos que se observan sobrerepresentados en la composición de las LCS se resaltan con números en negritas. El total de aminoácidos analizados con el comando AACOMP del paquete FASTA (ver material y métodos) se maneja en unidades enteras, así como también número total de segmentos de LCS que se encuentran constituidas por estos aminoácidos.

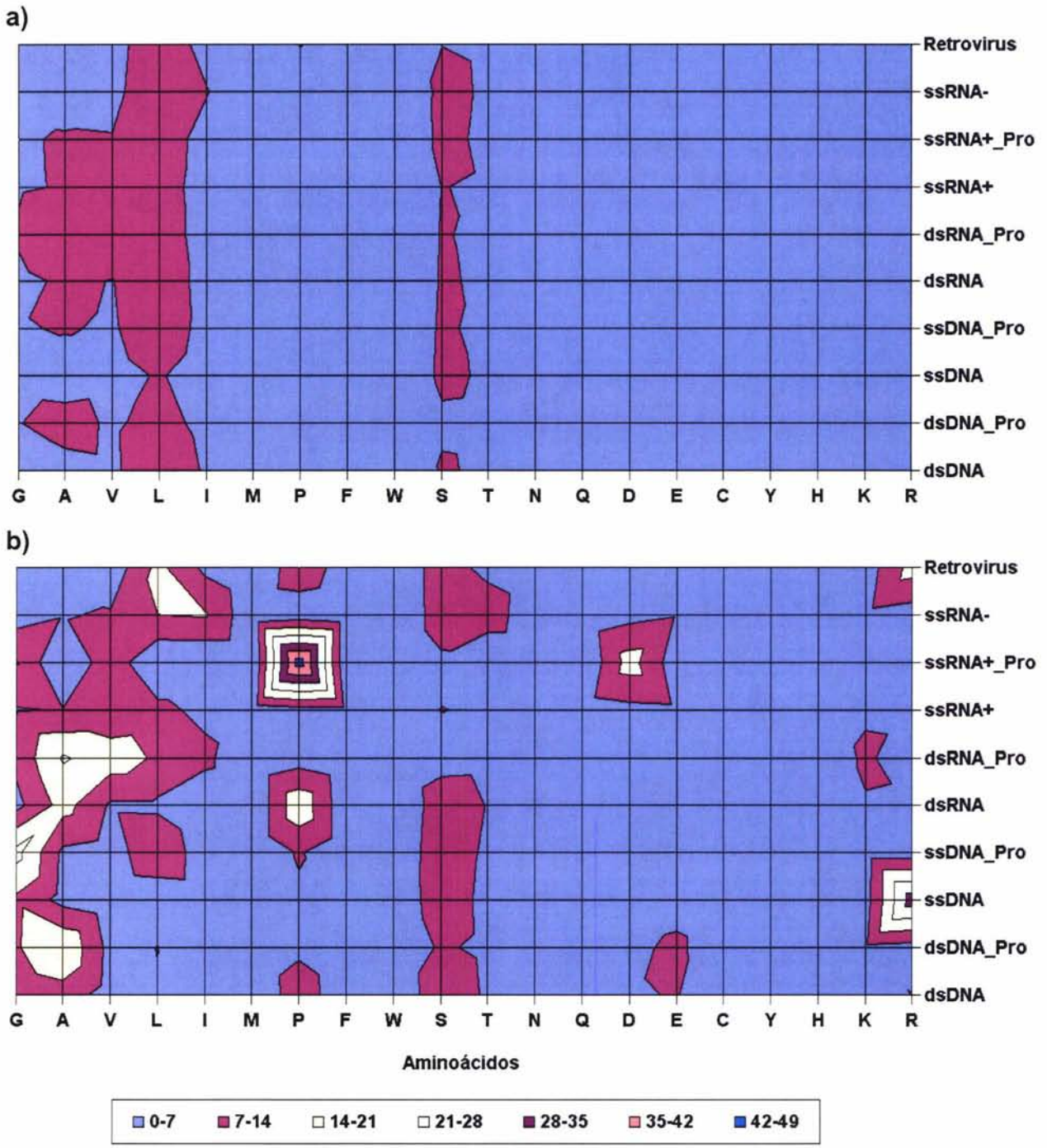


Fig. 4.3. Análisis de la composición de aminoácidos de los proteomas virales y de las LCS que se encontraron en ellos. En el inciso a) se encuentra la frecuencia de ocurrencia de los 20 aminoácidos dado su peso molecular y tendencia a formar alfa-hélice (de izquierda a derecha en el eje de las abscisas) de los proteomas por tipo los seis tipos de genoma viral, pero además, se incluye al tipo de hospedero celular (eucarionte y procarionte -PRO) (eje de las ordenadas); en el inciso b) se encuentra la frecuencia de ocurrencia de las LCS que se encuentran en los proteomas virales distribuidos bajo las mismas agrupaciones. Las frecuencias se toman de intervalos de 7%, los que están dentro del primer intervalo (0-7%) no se toman como representativos en la composición de aminoácidos, los aminoácidos que conforman los intervalos subsecuentes son los que se sugieren como representativos en la composición de aminoácidos, ya sea en el proteoma o en las LCS.

IV. Las posibles funciones de las LCS en los proteomas virales.

Al realizar una clasificación funcional de las secuencias virales de nuestra base de datos se encontró que de las 34 253 secuencias que la componen, más del 50% de ella (16 499 secuencias) presentan al menos una región similar a un dominio PFAM, a las cuales se le ha asignado una o más de una de las siete categorías funciones virales descritas con anterioridad en los VOG's del NCBI (ver material y métodos). La distribución de estas clases funcionales en los proteomas para cada uno de los seis tipos de genomas virales se puede observar en la gráfica 4.3a. En esta gráfica se muestra que las funciones virales están sobre-representadas por aquellas involucradas en auxiliares (A), estructurales (S), la de replicación y reparación del DNA y metabolismo de nucleótidos (L); mientras que en la misma proporción, pero menor que las anteriores, las de replicación del RNA, transcripción y modificación (T) y, regulación de metabolismo celular (R), la función que se encuentra poco representada en los proteomas virales es la que corresponde a proteínas del movimiento (M). La categoría de desconocidos (U) es muy baja también.

Al detectar aquellas secuencias que presentan LCS y una secuencia similar a algún PFAM, se identificaron todos los casos en los que las LCS podrían estar formando parte del dominio PFAM caracterizado en estas secuencias, así, de las 12 317 LCS detectadas, el número de casos en los que segmentos de LCS no se encontraban dentro de un dominio PFAM particular son 2 564 (ver diagrama 4.1a) y, por ende, a estos 2 564 segmentos de LCS no se les puede atribuir función alguna bajo nuestra aproximación. Una vez detectadas aquellas LCS formando parte de un dominio PFAM, en al menos un 25% de cobertura en el alineamiento (ver diagrama 4.1c), sólo un caso se encontró que un PFAM está constituido totalmente por una LCS, siendo esta última más grande que el PFAM. Finalmente, se buscaron los PFAM's intersectados con LCS en los VOG's para a las LCS las funciones virales clasificadas previamente. Si bien, no se puede dar una explicación definitiva para las funciones de una LCS determinada, se puede inferir bajo esta aproximación que muy tentativamente participan las LCS en alguna o más de las siete funciones virales asignadas a los dominios funcionales que las contienen. Se esquematizan estas aproximaciones en el diagrama 4.1.

La gráfica 4.3b esquematiza la distribución de las siete categorías funcionales para aquellas secuencias que presentan al menos una LCS. Cabe resaltar que algunas categorías funcionales son representadas uniformemente más en la mayoría de los grupos virales que otras, tales como las auxiliares y estructurales. Por otro lado, se puede observar que hay una baja representación de LCS que puedan estar involucradas en la función de proteínas del movimiento, la cual se encuentra totalmente ausente en los virus de ssRNA+ y en ssDNA. También se puede observar en la gráfica 4.3b que hay un gran número de familias a cuyas LCS no se pudo asignar a alguna categoría funcional, incluso ni la de desconocidos, esta observación se discutirá más adelante. De esta forma, se encontró que algunas LCS podrían estar involucradas en las funciones de las proteínas del tipo de replicación del DNA y del RNA, transcripción, estructurales y auxiliares, principalmente. Para un desglose más detallado ver de I al VI en el anexo 1.

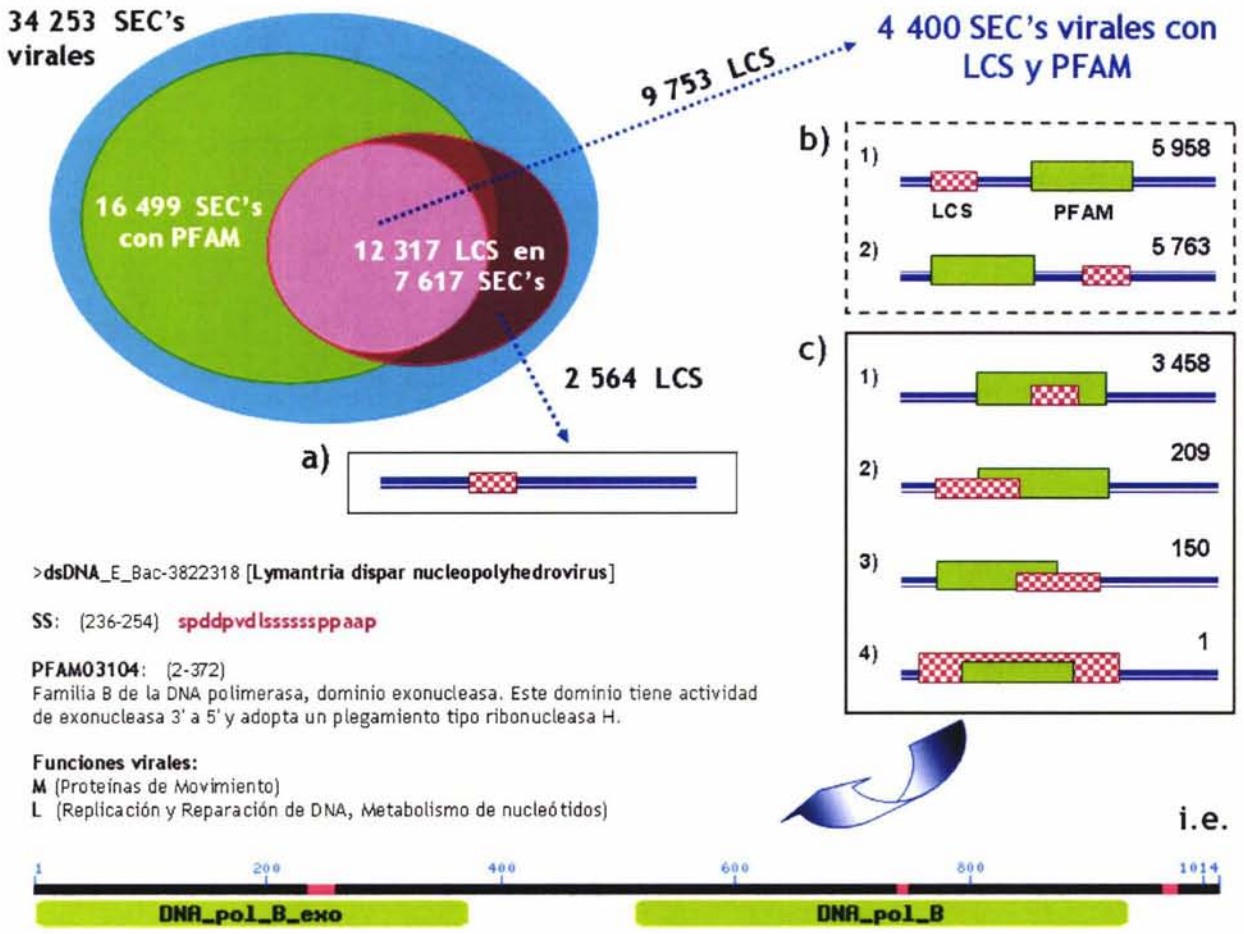
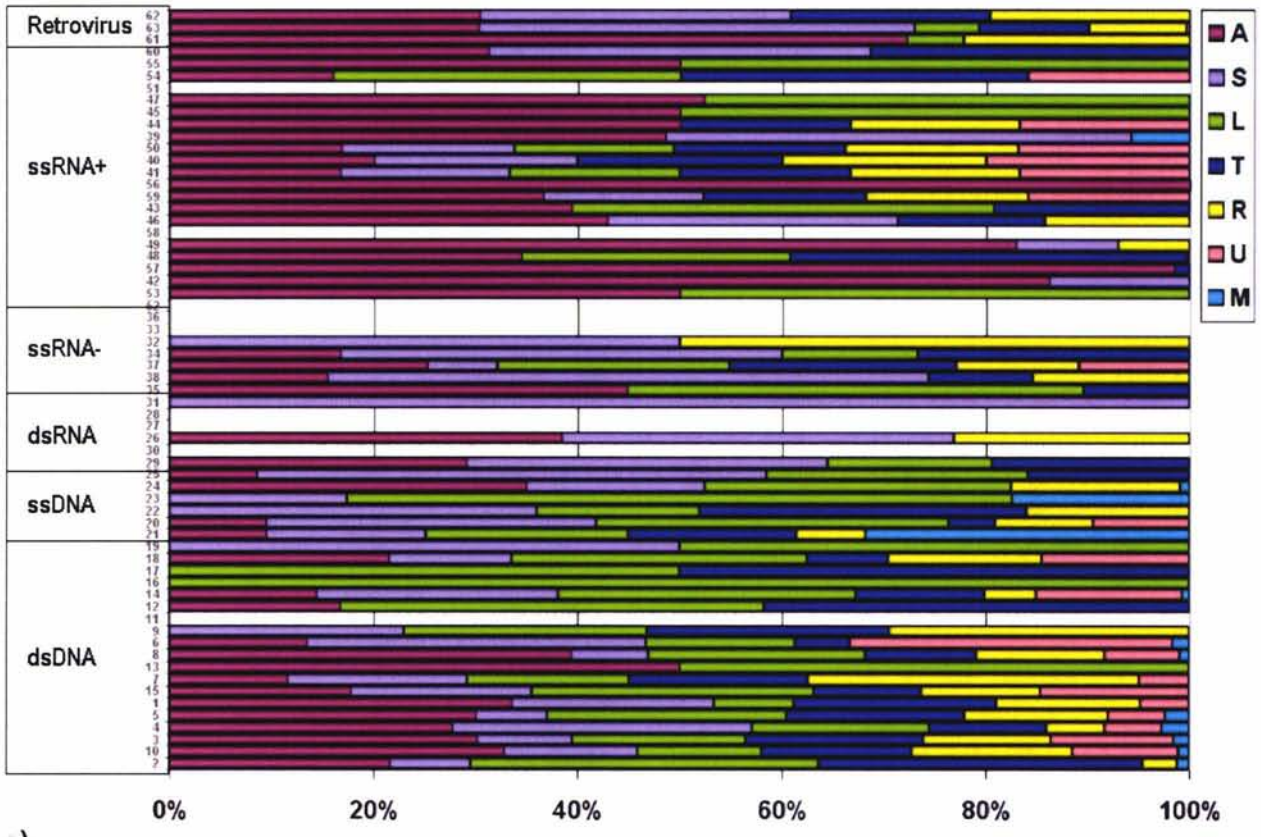
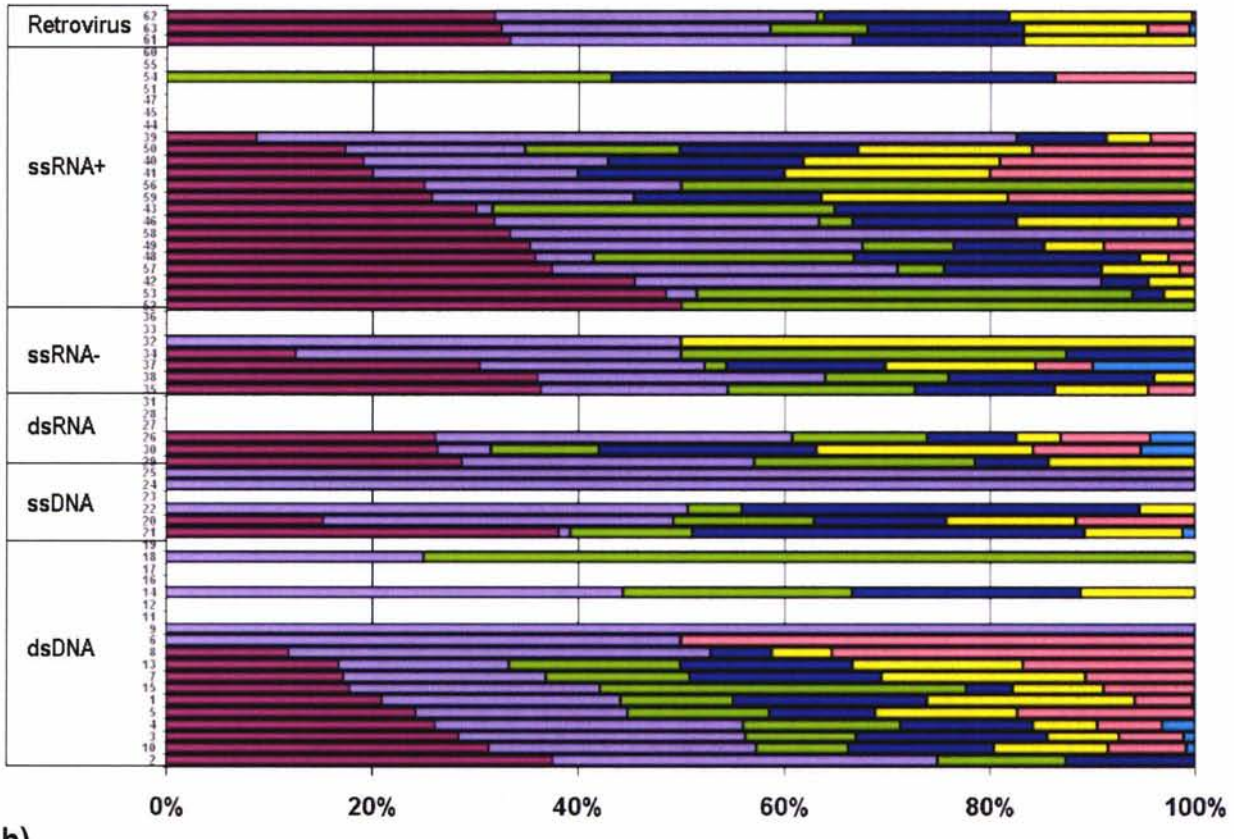


Diagrama 4.1. Se esquematiza la metodología empleada para detectar la posible función de las LCS en las proteínas virales. Se muestran las aproximaciones obtenidas en este trabajo y las categorías funcionales en los proteomas virales asignadas actualmente en los VOG's. Se compararon las 34 253 secuencias virales (óvalo azul): 1) contra PFAM (ver material y métodos), de los cuales 16 499 secuencias (óvalo verde) presentaron al menos un segmento de PFAM; 2) para detectar LCS por medio del programa SEG (ver material y métodos), de los cuales 7 617 secuencias presentaron al menos un segmento de LCS (óvalo en rojo), constituyendo así 12 317 segmentos de LCS. Al identificar aquellas secuencias que presentan LCS y una secuencia similar a algún dominio PFAM, se identificaron todos los casos en los que las LCS (bloques de cuadros rojos y blancos) podrían ser parte del dominio PFAM (bloques verdes) caracterizado en estos CDS (c) (casos agrupados en el recuadro negro), el número de casos en los que la LCS no se encontraba dentro de la secuencia del dominio PFAM particular se muestran en b) (casos agrupados en recuadro discontinuo). Una vez detectadas aquellas LCS formando parte del dominio PFAM, se buscaron estos PFAM's en los VOG's asumiendo que la secuencia que contiene a esa LCS pudiera tener las funciones virales descritas en esa base de datos. Si bien, no se puede hacer una propuesta contundente para las funciones de una LCS determinada, se puede inferir bajo esta aproximación que pudieran participar en alguna o más de las 7 funciones virales asignadas en VOGs. En a) se muestran los casos de las LCS restantes de las 12 317 que se detectaron inicialmente donde las secuencias que las contienen no presentaron alineamiento con alguna consenso de un dominio PFAM, y por ende, a estas 2 564 LCS no se les pudo atribuir función alguna. Finalmente, se muestra el ejemplo de una de los 3 465 casos de LCS que se encuentran dentro del dominio PFAM (c.1); esta secuencia de Lymantria dispar nucleopolyhedrovirus presenta una LCS rica en serina (en rojo), presenta el PFAM03104 (recuadro verde) que se encuentra en la familia B de la DNA polimerasa, el dominio exonucleasa y, se le asignaron las funciones virales: M (Proteínas de Movimiento) y L (involucrada en la replicación y reparación del DNA, así como en el metabolismo de nucleótidos) y así, la LCS que se encuentra en esta secuencia participaría tentativamente también en estas funciones virales.


Retrovirus
ssRNA+
ssRNA-
dsRNA
ssDNA
dsDNA
A
S
L
T
R
U
M
0%
20%
40%
60%
80%
100%

a)


Retrovirus
ssRNA+
ssRNA-
dsRNA
ssDNA
dsDNA
0%
20%
40%
60%
80%
100%

b)

Gráfica 4.3. Distribución de las siete categorías funcionales dentro de las 63 familias y seis tipos de genomas virales que se asignaron a las secuencias de los proteomas y a aquellas que presentaron, al menos, una LCS. Esta distribución está dada en porcentajes para normalizar los datos con respecto al número de dominios PFAM que presentan los proteomas según el tipo de genoma, así en el eje de las abscisas se encuentran los porcentajes de la frecuencia para cada una de las siete categorías funcionales en los 63 familias y, estos a su vez, agrupados por tipo de genoma viral. Estas categorías corresponden a: **(A)** proteínas auxiliares, **(S)** proteínas estructurales, **(L)** replicación y reparación del DNA, así como metabolismo de nucleótidos, **(T)** replicación del RNA, transcripción y modificación, **(R)** regulación de metabolismo celular, **(U)** desconocidas, y **(M)** proteínas de movimiento. En a) se muestra la distribución de las categorías funcionales para los proteomas virales a partir de las 16 499 secuencias con un PFAM, en comparación con b) la distribución de las categorías funcionales de aquellas secuencias con LCS que forman parte de un dominio PFAM (9 753 LCS en 4 400 secuencias virales; ver diagrama 4.1). Las familias que no presentan columnas de distribución representan, en **a),** a aquellas cuyas secuencias no presentan dominios PFAM y, en b) a aquellas cuyas LCS no presentan un PFAM y no se les pudo asignar alguna categoría funcional por medio de esta aproximación.

Por último, en la figura 4.4 se muestran los ejemplos de aquellas secuencias que tienen definida una estructura cristalizada y cuya LCS se localizó físicamente en el plegamiento.

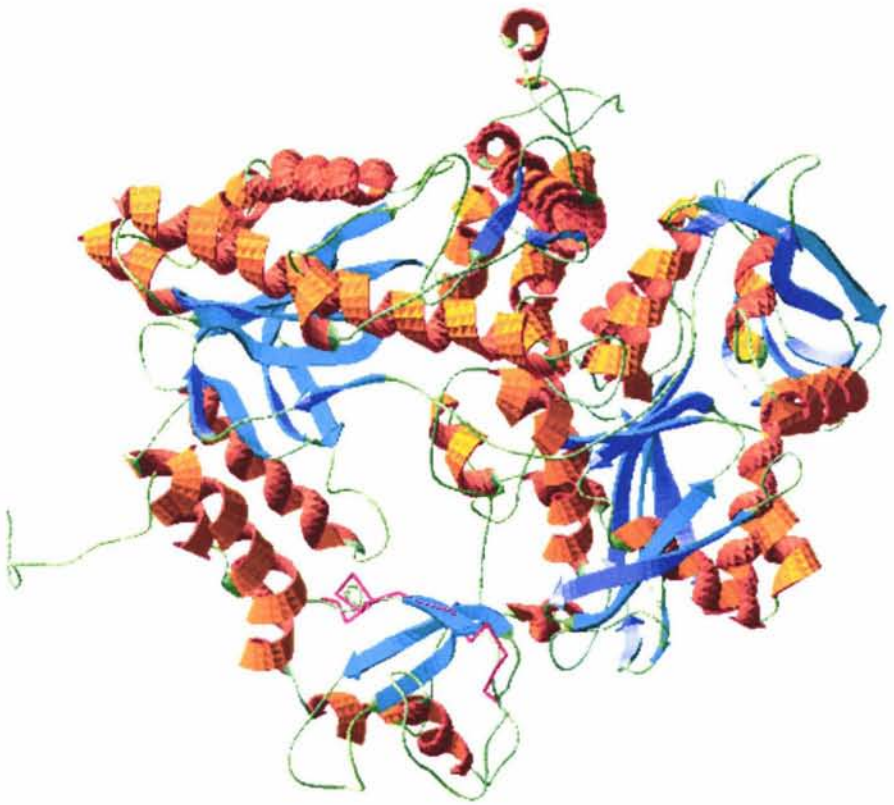


Fig. 4.4. Localización de una LCS en la estructura cristalizada de la cadena B de la DNA Polymerasa del Bacteriophage Rb69. Se obtuvo la estructura cristalizada [Wang et al, 1997] en el PDB, cuyo número de acceso es 1WAF. La secuencia primaria de la secuencia analizada se muestra debajo de la estructura cristalizada y la LCS detectada en esta secuencia se encuentra localizado del residuo número 236 al 254 (spddpvdlssssssppaap), formando una pequeña β-plegada y conectores (marcada en línea rosa mexicano; ver ejemplo de diagrama 4.1). La localización física se llevó a cabo por medio del paquete SPDBV del PDB versión 3.7 (http://spdbv.niehs.nih.gov).

>gi|3822318|gb|AAC70269.1|AAC70269 DNA polymerase [Lymantria dispar nucleopolyhedrovirus]

MSSVNLMEWSALKTQLQAGRDAGKARVSIGPADTARITRMTYADNHLIVFMNARLAKENHRLYQFYAEVRCDLYSYKSCYGTHASATCHRNCTSYKTFVMPGLRDVHTDKL
HVVKFKRSDEKRDKNCLDGYLADVNRVHMQTSLLEGQYVRFKNAHACRDYRLSHTAKDVHEFESMLERVQVSALSHEILPVVACYDIETHSDGQRFSAPDADFIISIAVVV
RRDAADTRICLFY**SPDDPVDLSSSSSSPPAAP**DTAAVHFRAERDMIAAFFQLLPLLNADVVLDFNGDKFDLPFLTGRANKLCGPAEAARATKIARYDLSPVNVVTQQSYDK
FSNKLHSHYLTYYIHIDLYQFLSTDSEHNDLENFQLNTVAEHYLKKSKVDLPIHDMLQMYGEKRLSRIVEYNVQDCVLPVELFLKLEIADYMYTQCMLLYLCTDDLLRNIS
HKITVAYFHLALTNTVAAADPTPDPYFFNKYDLSVTSGRKRAFNEPAPSANAIDLSQLKRTPVDAARIPPSAVKLCSTRQSCTYKGGKVLSPKPGFNRWVATLDFNALYPT
IMMWEGVCMSNVFIASDGNVYLDKNVNAVNPKLLKTLSEMRVRYKGLRDQCEYNSFYYKLYDKIQNALKRIANSIYGYYGIFFKPLANYITKMGRGKLKEVVGKVEAMSDD
PRILREFGLSKINFSVIYGDTDSCFIRVLFDEAELAPDRRQTAIRSIVQDHVCKTLNDSWCGYKMSLENIMLSLILLKKKKYCYLNNEQRTKYKGWLIKRDMPLFMRKAFR
ATVDSLLRGHSTACALGLLAEQMLRYYREFGAPRENLVDYCFSMSYNETSTTAKRRKEEDPARKPVITIAKHCRELLANPGVDFLPGNGDRIQYVLVDVKEKITQKAFPLK
LFDPDSPTLQISWLKHMNILCTFMNELIQVFGNRPEFEHYFGAIVDEYTSAQMYDVRYPVLVPTRRAKAGKSAKKNDSDSDSDSDDDDDPATTPVNYHSLFSMHLKKPKRQ
AVGEFEPCPQCVARA

5. DISCUSIÓN.

La primera observación que podemos hacer a partir de los datos obtenidos aquí es la detección de una alta cantidad de secuencias simples en los proteomas virales, especialmente en los virus que poseen genomas de RNA de cadena sencilla de sentido positivo (+) (ver gráfica 4.1). Se puede observar también que la presencia de estas secuencias en los sistemas virales esta ampliamente distribuida entre los diferentes tipos de genomas virales y tipos de hospederos, dado que tres cuartas partes de los proteomas analizados tuvieron al menos una secuencia simple, representando una cuarta parte del total de las secuencias virales (ver tabla 4.2 y 4.3).

Aún cuando la presencia de las secuencias simples se ha asociado al tamaño del genoma en los sistemas celulares, por lo que se han considerado importantes en la evolución del tamaño de los genomas (Tauzt *et al*, 1986; Hancock, 1995), en los sistemas virales no hemos detectado correlación entre la presencia de las secuencias simples y el tamaño del proteoma. La excepción a esta observación recae sobre los virus de DNA y RNA con genomas de cadena doble (ver figura 4.1), donde puede haber un efecto de selección, ya que los genomas de conformación en doble cadena además de ser una característica de los genomas en los sistemas celulares son de los genomas más grandes en los virus de DNA y en los de RNA (en estos últimos con genomas fragmentados). Esto podría sugerir que la presencia de las secuencias simples o el mecanismo que las origina (como el *slippage* descrito en la introducción) presentan, conjunta o independientemente, una tendencia a generarse y acumularse preferencialmente en genomas que se constituyen por una cadena doble del ácido nucleico. Incluso, en estos grupos se encontraron también las secuencias simples de mayor longitud y en mayor número por secuencia analizada (ver tabla 4.2). Por otro lado, el aporte de las secuencias simples al tamaño de los proteomas virales es minoritario, menor al 3% (ver gráfica 4.2) en la mayoría de los casos.

Al no encontrar una correlación generalizada entre la presencia de las secuencias simples y el tamaño del proteoma, uno podría suponer que las secuencias simples en los proteomas virales se mantienen por una presión favorable sobre la función de las secuencias que las contienen, tal como ha sido propuesto con anterioridad en los sistemas celulares (Levinson y Gutman, 1987; Tautz y Schlötterer, 1994; Wootton, 1994; Hancock, 1996; Albá *et al*, 1999; Huntley y Golding, 2000; Huntley y Golding, 2002; Tompa, 2003). Entonces, conocer las posibles funciones de las secuencias simples es prioridad para aseverar esta hipótesis.

Contrario a lo que se ha observado en los sistemas celulares (Becerra *et al*, enviado), las posiciones de las secuencias simples dentro de las secuencias que las contienen no presentan ninguna tendencia por alguna ubicación en particular, se les puede encontrar a lo largo de la secuencia viral, a excepción de los virus con genoma de DNA de cadena sencilla y RNA de cadena doble, en los cuales las secuencias simples se encuentran ubicadas

preferencialmente en las regiones amino y carboxilo terminal, respectivamente (ver figura 4.2). Una posible explicación sobre esta observación es que, un número no considerado de las secuencias virales puede codificar para más de una proteína a diferentes regiones de este polipéptido (Davis *et al*, 1985; Flint *et al*, 2000; Regenmortel *et al*, 2000;), entonces lo que nosotros delimitamos como una región amino, medio y carboxilo terminales de cada una de las secuencias analizadas aquí representa un primer nivel de delimitación; sin embargo, lo ideal sería conocer las posiciones inicial y final de cada uno de los productos codificantes de una secuencia viral, y sólo a partir de ello, delimitar estas tres regiones en cada uno de los productos codificantes, entonces, podría haber más de una de región amino y carboxilo terminal en una secuencia viral de las aquí estudiadas. A pesar de no contamos con estos datos, ya que para muchos virus no se conocen el número completo de los productos que codifican, nuestra aproximación nos permite suponer, que la ubicación de las secuencias simples no perturba la función de las proteínas en las que se encuentran.

Una observación importante sobre la composición de aminoácidos de las secuencias simples es que en comparación con los sistemas celulares (Hancock, 1996; Albá *et al*, 1999; Huntley y Golding, 2000; Huntley y Golding, 2002), los aminoácidos prolina y arginina se encuentran formando parte de forma significativa a las secuencias simples de algunos grupos virales, si bien la prolina comparte las características de ser de los aminoácidos de poco peso molecular y de tendencia a formar alfas hélices (Creighton, 1984) conjuntamente con serina, alanina y leucina (que también se encuentran formando parte de forma significativa a las secuencias simples), la diferencia de la prolina con los tres anteriores es que es de alta carga positiva, por lo que su función en las proteínas debe ser especial. Algunos casos de la función de secuencias simples con prolina se han reportado de forma aislada en los sistemas celulares y virales (Tompa, 2003; Perera *et al*, 2001). De forma contraria, arginina es de los aminoácidos con mayor peso molecular y cuya tendencia a formar alfas hélices es baja (Creighton, 1984), su función biológica en las secuencias simples no es clara; sin embargo, se han reportado casos donde se ha visto que las secuencias simples con esta composición en virus funciona como localización nuclear (Eckhardt *et al*, 1991). Por lo que podríamos sugerir que las secuencias simples compuestas de prolina y arginina son una característica particular de los proteomas virales.

Se puede observar que en la distribución de las funciones de los proteomas virales existe un sesgo por aquellas clases funcionales involucradas en proteínas auxiliares, estructurales, y de replicación y reparación del DNA y de metabolismo de nucleótidos (ver gráfica 4.3.a). Tal sesgo parece verse reflejado también en las funciones de aquellas proteínas que contienen secuencias simples, principalmente por aquellas involucradas en funciones auxiliares y estructurales.

Finalmente, en este trabajo se trató de conocer el papel de las secuencias simples en los proteomas virales y aunque esta base de proteomas virales presenta un sesgo por aquellos virus que infectan a especies de importancia médica y que no representa a toda la diversidad viral, los análisis hechos en este trabajo muestran que dada la distribución y frecuencia de las secuencias simples en los proteomas virales, así como su conservación y presencia en dominios funcionales, hacen posible sugerir que este tipo de secuencias tienen un papel relevante en la evolución de los proteomas virales.

CONCLUSIONES.

En este estudio, se detectó una frecuencia alta de secuencias simples en los proteomas virales analizados, especialmente en sobre los virus que poseen genomas de RNA de cadena sencilla de sentido positivo. Además la presencia de las secuencias simples se encuentra ampliamente distribuida en los diferentes grupos de virus, ya sea por el tipo de genoma o por el hospedero.

Con excepción de los virus con genomas de cadena doble de DNA y RNA no existe una clara correlación entre el tamaño del proteoma, ya sea por número de aminoácidos o por número de secuencias que lo conforman y la presencia de las secuencias simples.

Sólo en los virus con genomas de cadena sencilla de DNA y de cadena doble de RNA las secuencias simples se encuentran ubicadas principalmente en las regiones amino y carboxilo terminales de las secuencias que las contienen.

El sesgo composicional de las secuencias simples está formado principalmente por los aminoácidos G, A, S y L que presentan una sobre-representación en la composición de los proteomas virales y, sólo en algunos grupos de virus, las secuencias simples están formadas por P, R y D, los cuales no se encuentran sobre-representados en la composición original de los proteomas virales ni reportados en las secuencias simples de los sistemas celulares.

La mayoría de las secuencias simples fueron detectadas dentro de dominios de proteínas virales involucradas en funciones estructurales y auxiliares.

Dada la distribución y frecuencia de las secuencias simples en los proteomas virales, así como su conservación y presencia formando parte de en dominios funcionales, es posible que este tipo de secuencias tengan un papel relevante en la evolución de los virus.

REFERENCIAS.

Alba M.M., Santibáñez-Koref M.F. y Hancock J.M. (1999). **Amino Acid Reiterations in Yeast Are Overrepresented in Particular Classes of Proteins and Show Evidence of a Slippage-Like Mutational Process.** *J. Mol. Evol.* 49:789-797.

Alba M.M., Santibáñez-Koref M.F. y Hancock J.M. (1999). **Conservation of Polyglutamine Tract Size Between Mice and Humans Depends on Codon Interruption.** *J. Mol. Evol.* 16:1641-1644.

Alba M.M., Santibáñez-Koref M.F. y Hancock J.M. (2001). **The Comparative Genomics of Polyglutamine Repeats: Extrem Difference in the Codon Organization of Repeat-Encoding Regions Between Mammals and Drosophila.** *J. Mol. Evol.* 52:249-259.

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. y Lipman D.J. (1997). **"Gapped BLAST and PSI-BLAST: a new generation of protein database search programs".** *Nucleic Acids Res.* 25:3389-3402.

Andersson, JO y Andersson, SG (1999). **Insights into the evolutionary process of genome degradation.** *Curr. Opin. Genet. Dev.* **9:** 664–671.

Bateman, A., et al. (2000). **The Pfam contribution to the annual NAR database issue.** *Nucleic Acids Res.* 28, 263-266

Bayliss C.D., Dixon K.M., y Moxon R.E. (2004). **Simple sequence repeats (microsatellites): mutational mechanisms and contributions to bacterial pathogenesis.** *FEMS Immunology and Medical Microbiology* 40:11-19.

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J. y Wheeler D.L. (2003). **GenBank.** *Nucleic Acids Res.* 31:23-7.

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissing H., Shindyalov I.N., Bourne P.E. (2000). **The Protein Data Bank.** *Nucleic Acids Res.* 28:235-242.

Byrd C, Ohtsuka E, Moon MW, Khorana HG (1965). **Synthetic deoxyribo-oligonucleotides as templates for the DNA polymerase of Escherichia coli: new DNA-like polymers containing repeating nucleotide sequences.** *Proc. Natl. Acad. Sci* 53:79-86.

Bork P, Ouzounis C, Sander C, Scharf M, Schneider R, Sonnhammer E (1992). **What's in a genome?** *Nature* 358:287.

Bourque G, Pevzner PA (2002). **Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species** *Gen. Res.* 12:26–36.

Bremner K.H., Seymour L.W. y Pauton C.W. (2001). **Harnessing Nuclear Localization Pathways for Transgene Delivery.** *Curr. Opin. Mol. Theory.* 3(2):170-177.

Brown JR, Douady CJ, Italia MJ, Marshall WE y Stanhope MJ (2001). **Universal trees based on large combined protein sequence data sets.** *Nat. Gen.* **28**:281-285.

Bruccoleri RE, Dougherty TJ y Davison DB (1998). **Concordance analysis of microbial genomes.** *Nucleic Acids Res.***26:** 4482–4486.

Bzymek M., Lovertt S.T. (2001) **Instability of repetitive DNA sequences: The role of replication in multiple mechanisms.** *Proc. Natl. Acad. Sci.* 98:8319-8325.

Campbell A. (2001). **The origins and evolution of viruses.** *TRENDS in Microbiology* 9:61

Casari G, Andrade MA, Bork P, Boyle J, Daruvar A, Ouzounis C, Schneider R, Tamames J, Valencia A, Sander C (1995) **Challenging times for bioinformatics.** *Nature.* **367**:647-648.

Cavalier-Smith, T. 1985. **The evolution of genome size.** John Wiley y Sons, Chichester, UK.

Coates J.C. (2003). **Armadillo repeat proteins: beyond the animal kingdom.** *TRENDS in Cell Biology* 13:463-471.

Creighton T (1984). **PROTEINS: Structures and Molecular Principles.** (W.H. Freeman and Company Press, New York, 2000).

Dandekar T, Snel B, Huynen M y Bork P (1998). **Conservation of gene order: A fingerprint of proteins that physically interact.** *Trends Biochem. Sci.* **23:** 324–328.

Eddy, S. R. (1996). **Hiden Markov models.** *Curr. Opin. Struct. Biol.* 6: 361-365.

Eisen JA (2000). **Horizontal gene transfer among microbial genomes: New insights from complete genome analysis.** *Curr. Opin. Genet. Dev.* **10:** 606–611.

Fitz-Gibbon ST y House CH (1999). **Whole genome-based phylogenetic analysis of free-living microorganisms.** *Nucleic Acids Res.* **27:** 4218–4222.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. (1995). **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.** *Science.* 269: 496-512.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. (1995). **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.** *Science.* **269:** 496-512.

Flint S.J., et al. **Principles of Virology: Molecular Biology Pathogenesis and Control.** (Washington D.C. ASM Press, 2000).

Gaasterland T y Ragan MA (1998). **Microbial genescapes: Phyletic and functional patterns of ORF distribution among prokaryotes.** *Microb. Comp. Genomics* **3:** 199–217.

Gough J., Karplus K., Hughey R., Chothia C. (2001). **Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Know Structure.** *J. Mol. Biol.* 313: 903-919.

Hamada H. Petrino MG, Kakunaga T (1982). **A novel repeated element with Z-DNA-forming potencial is widely found in evolutionarily diverse eukaryotic genomes.** *Proc. Natl. Acad. Sci* 79: 6465-6469.

Hancock J.M. (1996). **Correspondence: Simple sequences in a 'minimal' genome.** *Nature Genetics* 14:14-15.

Hancock J.M. (1996). **Simple sequences and the expanding genome.** *BioEssays* 18: 421-425.

Hancock J.M. y Dover A. G. (1990). **'Compensatory slippage' in the evolution of ribosomal RNA genes.** *Nucleic Acids Res.* 18: 5949-5954

Hendrix R.W., Lawrence J.G., Hatfull G.F. y Casjens S. (2000). **The origins and ongoing evolution of viruses.** *TRENDS in Microbiology* 8:504-508.

Hunt G.M., Johnson D. y Tiemessen C.T. (2001). **Characterisation of the Long Terminal Repeat Regions of South African Human Immunodeficiency Virus Type 1 Isolates.** *Virus Genes* 23:27-34.

Huntley M. y Golding G.B. (2000). **Evolution of Simple Sequence in Proteins.** *J. Mol. Evol.* 51:131-140.

Huntley M.A. y Golding G.B. (2002). **Simple Sequences are Rare in the Protein Data Bank.** *PROTEINS: Structure, Function, and Genetics* 48:134-140.

Hutson Ms, Holzwarth G, Duke T, Viovy Jl (1995). **2-Dimensional Motion Of Dna Bands During 120-Degrees Pulsed-Field Gel-Electrophoresis .1. Effect Of Molecular-Weight. BIOPOLYMERS** 35: 297-306.

Huynen MA y Bork P (1998). **Measuring genome evolution.** *Proc. Natl. Acad. Sci.* **95:** 5849–5856.

J.C. Vega-Arreguín (1996). **El origen de la vida sobre la tierra.** *BEB* 15:184.

Jeanmougin F, et al (1998). **Multiple sequence alignment with Clustal X.** *Trends Biochem Sci* **23:**403-405.

Katti M., Ranjekar P.K. y Gupta V. (2001). **Differential Distribution of Simple Sequences Repeats in Eukaryotic Genome Sequences.** *Mol. Biol. Evol.* 18:1161-1167

Kunin V y Ouzounis CA (2003). **The Balance of Driving Forces During Genome Evolution in Prokaryotes.** *Gen. Res.* **13:**1589–1594.

Kornberg A, Bertsch LL, Jackson JF, Khorana HG (1964). **Enzymatic synthesis of deoxyribonucleic acid, XVI. Oligonucleotides as templates and the mechanism of their replication.** *Proc. Natl. Acad. Sci.* 51:315-323.

La Scola B., Audio S., Robert C., Jungang L., de Lamballerie X., Drancourt M., Birtles R., Claverie J.M. y Raoult Didier (2003). **A Giant Virus in Amoebae.** *Science* 299:2033.

Levinson G y Gutman G (1987). **Slipped-Strand Mispairing: A major Mechanism for DNA Sequence Evolution.** *Mol. Biol. Evol.* 3:203-221.

Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI y Koonin EV (1999). **Comparative genomics of the Archaea (Euryarchaeota): Evolution of conserved protein families, the stable core, and the variable shell.** *Genome Res.* **9:** 608–628.

Marcotte E.M. Pellegrini M, Yeates T.O. y Eisenberg D. (1998). **A Census of Proteins Repeats.** *J. Mol. Biol.* 293:151-160.

Matzke M, Mette M.F. y Matzke A.J.M. (2000). **Transgene Silencing by the Host Genome Defense: Implications for the Evolution of Epigenetic Control Mechanisms in Plants and Vertebrates.** *Plant Mol. Biol.* 43:401-415.

Matzke M., Matzke A.J.M. y Kooter J.M. (2001). **RNA: Guiding Gene Silencing.** *Science* 293:1080-1083.

Mayo M.A. y Pringle C.R. (1998). **Virus Taxonomy-1997.** *J. Gen. Virol.* 79:649-657.

McGeoch D.J. y Davison A.J. (1995). **Molecular Basis of Virus Evolution.** A. Gibbs, C. Calisher y F. Garcia-Arenal, eds.

Meeds T., Lockard E., y Livingston B.T. (2001). **Special Evolutionary Properties of Genes Encoding a Protein with a Simple Amino Acid Repeat.** *J. Mol. Evol* 53:180-190.

Mills R., Rozanov M., Lomsadze A., Tatusova T. y Borodovsky M. (2003). **Improving gene annotation of complete viral genomes.** *Nucleic Acids Res.* 31:7041-7055.

Mira A, Ochman H y Moran NA (2001). **Deletional bias and the evolution of bacterial genomes.** *Trends Genet.* **17:** 589–596.

Mira A, Klasson L y Andersson GE (2002). **Microbial genome evolution: sources of variability.** *Curr. Opin. Microbiol.* **5:** 506-512

O'Donovan C., Martin M.J. Gattiker A., Gasteiger E., Bairoch A., Apweiler R. (2002). **High-quality protein knowledge resource: SWISS-PROT and TrEMBL.** *Brief Bioinform.* 3:275-284.

Ohno S (1970). **Evolution by gene duplication.** Springer-Verlag, New York.

Pearson W.R. y Lipman D.J. (1988). **Improved Tools for Biological Sequence Comparison.** *Proc. Natl. Acad. Sci.* 85:2444-2448.

Pedulla M.L., Ford M.E., Houtz J.M. (2003). **Origins of Highly Mosaic Mycobacteriophage Genomes.** *Cell* 113:171-182.

Perera R., et al (2001). **Alphavirus Nucleocapsid Protein Contains a Putative Coiled Coil α–Helix Important for Core Assembly.** *J. Virol.* 75(1):1-10.

Pupko T. y Graur D. (1999). **Evolution of Microsatellites in the Yeast Sassharomyces cereviseae: Role of Length and Number of Repeated Units.** *J. Mol. Evol.* 48:313-316.

Rackovsky S. (1998). ***"Hidden" sequence periodicities and protein architecture.*** *Proc. Natl. Acad. Sci.* 95:8580-8584.

Reddy P.S., et al (1998). **Nucleotide sequence, genome organization, and transcription map of bovine adenovirus type 3.** *J. Virol.* 72:1394-1402.

Regenmortel M. H. V., *et al.* (2000). ***Virus Taxonomy: 7th Report of the International Committee on Taxonomy of Viruses.*** Academic Press, San Diego, CA.

Rohl C.A., Fiori W. y Baldwin R.L. (1999). **Alanine is helix-stabilizing in both template-nucleated and standard peptide helices.** *Proc. Natl. Acad. Sci.* 96:3682-3687.

Sandaa RA, Heldal M, Castberg T, Thyrhaug R, Bratbak G (2001). **Isolation and characterization of two viruses with large genome size infecting Chrysochromulina ericina (Prymnesiophyceae) and Pyramimonas orientalis (Prasinophyceae).** *Virology.* 290:272-80.

Schlötterer C. y Tautz Diethard (1992). **Slippage sintesis of simple sequence DNA.** *Nucleic Acids Res.* 20:211-215.

Snel B, Bork P y Huynen MA (2002). **Genomes in flux: The evolution of archaeal and proteobacterial gene content.** *Genome Res.* **12:** 17–25.

Snel B, Bork P, y Huynen MA (1999). **Genome phylogeny based on gene content.** *Nat. Genet.* **21:** 108–110.

Stark G.R., Wahl G.M. (1984). **Gene Amplification.** *Ann Rev Biochem* 53:447-491.

Tatusov R.L., et al (2001). **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res.* 29:22-28.

Tatusov R.L., Koonin E.V., Lipman D.J. A (1997). **Genomic Perspective on Protein Families.** *Science*, 278:631-637.

Tautz D, Renz M (1984). **Simple sequences are ubiquitous repetitive components of eukaryotic genomes.** *Nucleic Acids Res* 12:4127-4137.

Tautz D. y Schötterer (1994). **Simple sequences.** *Curr. Opin. Gen. Dev.* 4:832-837.

Tekaia F, Lazcano A y Dujon B (1999). **The genomic tree as revealed from whole proteome comparisons.** *Genome Res.* **9:** 550–557.

Tompa P. (2002). **Intrinsically unstructured proteins.** *TRENDS in Biochemical Sciences* 27:527-533.

Tompa P. (2003). **Intrinsically unstructured proteins evolve by repeat expansion.** *BioEssays* 25:847-855.

Vance V. y Vaucheret H. (2001). **RNA Silencing in Plants –Defense and Counterdefense.** *Science* 292:2277-2280.

Watanabe H, Mori H, Itoh T y Gojobori T (1997). **Genome plasticity as a paradigm of eubacteria evolution.** *J. Mol. Evol.* **44:** S57–64.

Weiner A.M. y Maizels N. (1999). **The genomic tag hypothesis: Modern virases as molecular fossils of ancient strategies for genomic replication, and clues regarding the origin of protein synthesis.** *The biological Bulletin* 196:327-330.

Weizhong Li, Lukasz Jaroszewski y Adam Godzik (2001). **"Clustering of highly homologous sequences to reduce the size of large protein database".** *Bioinformatics*, (2001) 17:282-283.

Wickstead B., Ersfeld K. y Gull K. (2003). **Repetitive Elements in Genomes of Parasitic Protozoa.** *Microbiol. Mol. Biol. Rev.* 67:360-375.

Wolf YI, Rogozin IB, Kondrashov AS, y Koonin EK (2001). **Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context.** *Gen. Res.* **11:**356–372.

Wooton J.C. (1994). **Non-globular domains in protein sequences: Automated segmentation using complexity measures.** *Comp. Chem.* 18(3):269-285.

Wooton J.C. y Federhen Scott (1993). **Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases.** *Comp. Chem.* 17:149-163.

Yamaguchi-Kabata Y. y Takashi Gojobori (2000). **Reevaluation of Amino Acid Variability of the Human Immunodeficiency Virus Type 1 gp120 Envelope Glycoprotein and Prediction of New Discontinuous Epitopes.** *J. Virol* 74:4335-4350.

ANEXO 1.

Desglose de los resultados obtenidos en este trabajo dentro de los 76 grupos taxónomicos y seis tipos de genomas virales.

Genomas de dsDNA.

TAXONOMÍA DE LOS GENOMAS VIRALES (dsDNA)			CARACTERÍSTICAS DE LOS GENOMAS			NO. ORF's			PRESENCIA DE SS		
Familia o grupo[x]	Género[x]	Hospedero[x]	No. genomas*	Tamaño promedio del genoma (nts)**	Intervalo del tamaño (nts) x grupo**	Total ORF's*	Promedio x genoma**	Intervalo de ORF's x grupo**	No. genomas con SS[E]	No. ORF's con SS[E]	% ORF'S con SS
Adenoviridae		V	23	33545	26163 _ 45063	638	27	11 _ 39	23	233	36
Asfarviridae		V	1	170101	170101	151	151	151	1	27	17
Baculoviridae		I	23	130532	99657 _ 178733	3274	142	109 _ 181	23	734	22
Hespesviridae	Alphaherpesvirinae	V	12	148659	124138 _ 177874	1013	84	72 _ 105	12	380	37
	Betaherpesvirinae	V	9	186376	144861 _ 241087	1230	136	86 _ 204	9	498	40
	Gammaherpesvirinae	V	15	138984	108409 _ 184427	1222	81	71 _ 95	15	323	26
	Ictalurid_Herpes_like	V	1	134226	134226	92	92	92	1	14	15
Iridoviridae		V, I	4	132888	102653 _ 212482	726	181	23 _ 468	4	138	19
Nimaviridae		I	3	301787	292967 _ 307287	1247	415	184 _ 532	3	508	40
Papillomaviridae		V	118	7716	5089 _ 8607	860	7	4 _ 15	117	314	36
Phycodnaviridae		Al	2	333168	330743 _ 335593	938	469	240 _ 698	2	186	19
Polyomaviridae		V	29	5137	4669 _ 5380	173	5	5 _ 7	26	49	28
Poxviridae	Chordopoxvirinae	V	21	183341	144575 _ 288539	4111	195	148 _ 273	21	592	14
	Entomopoxvirinae	I	2	234256	232392 _ 236120	561	280	267 _ 294	2	198	35
Fuselloviridae		A	1	15465	15465	32	32	32	1	5	15
Lipothrixviridae		A	1	40047	40047	72	72	72	1	6	8
Rudiviridae		A	2	33879	32308 _ 35450	99	49	45 _ 54	2	7	7
Myoviridae		B, A	22	91881	11624 _ 280334	2873	130	14 _ 381	21	245	8
Siphoviridae		B, A	73	45810	14510 _ 134416	5267	72	22 _ 237	72	531	10
Corticoviridae		B	1	10079	10079	23	23	23	1	2	8
Plasmaviridae		B	1	11965	11965	14	14	14	1	1	7
Podoviridae		B	28	34557	11660 _ 49534	1371	48	11 _ 86	26	137	9
Tectiviridae		B	1	14925	14925	22	22	22	1	7	31
TOTALES			393	2439324		26009	2727		385	5135	19.7

Tabla I. Información obtenida del análisis de los proteomas completos y las LCS para el tipo de genoma DNA de cadena doble (dsDNA).

Gráfica 1:

1.1 Distribución de los genomas colectados por grupo taxonómico en comparación con los que presentaron al menos una secuencia con LCS.

1.2 Distribución de las LCS por grupo taxonómico con respecto al contenido total de secuencias de cada grupo viral.

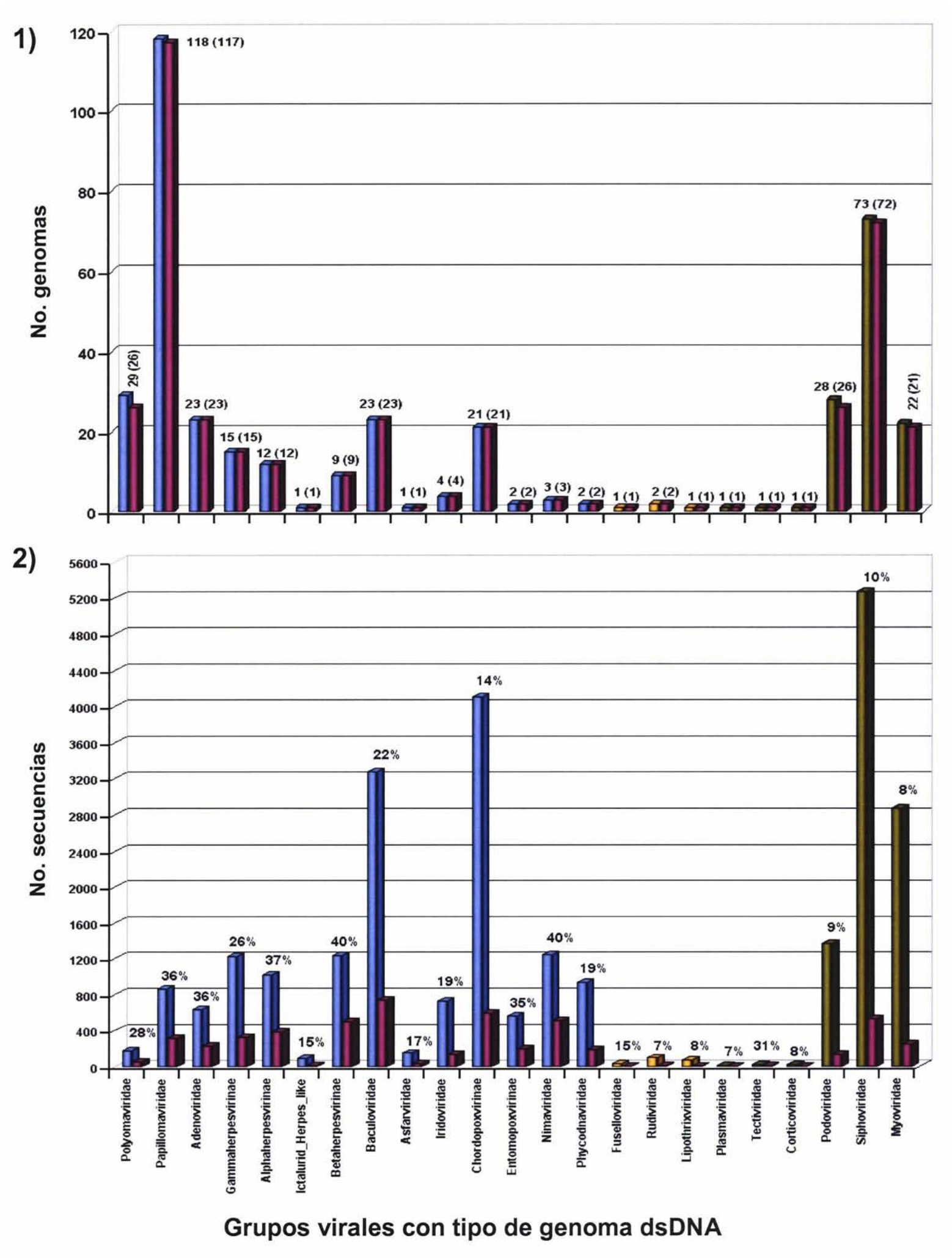
Las barras en azul, verde y amarillo son los datos colectados para virus con hospederos eucariontes, eubacterias y arqueas, respectivamente; en rojo, aquellos que presentan el fenómeno de LCS.

1.3 Correlación entre la presencia de las LCS y el tamaño del proteoma dado en número de aminoácidos y secuencias que los conforman.

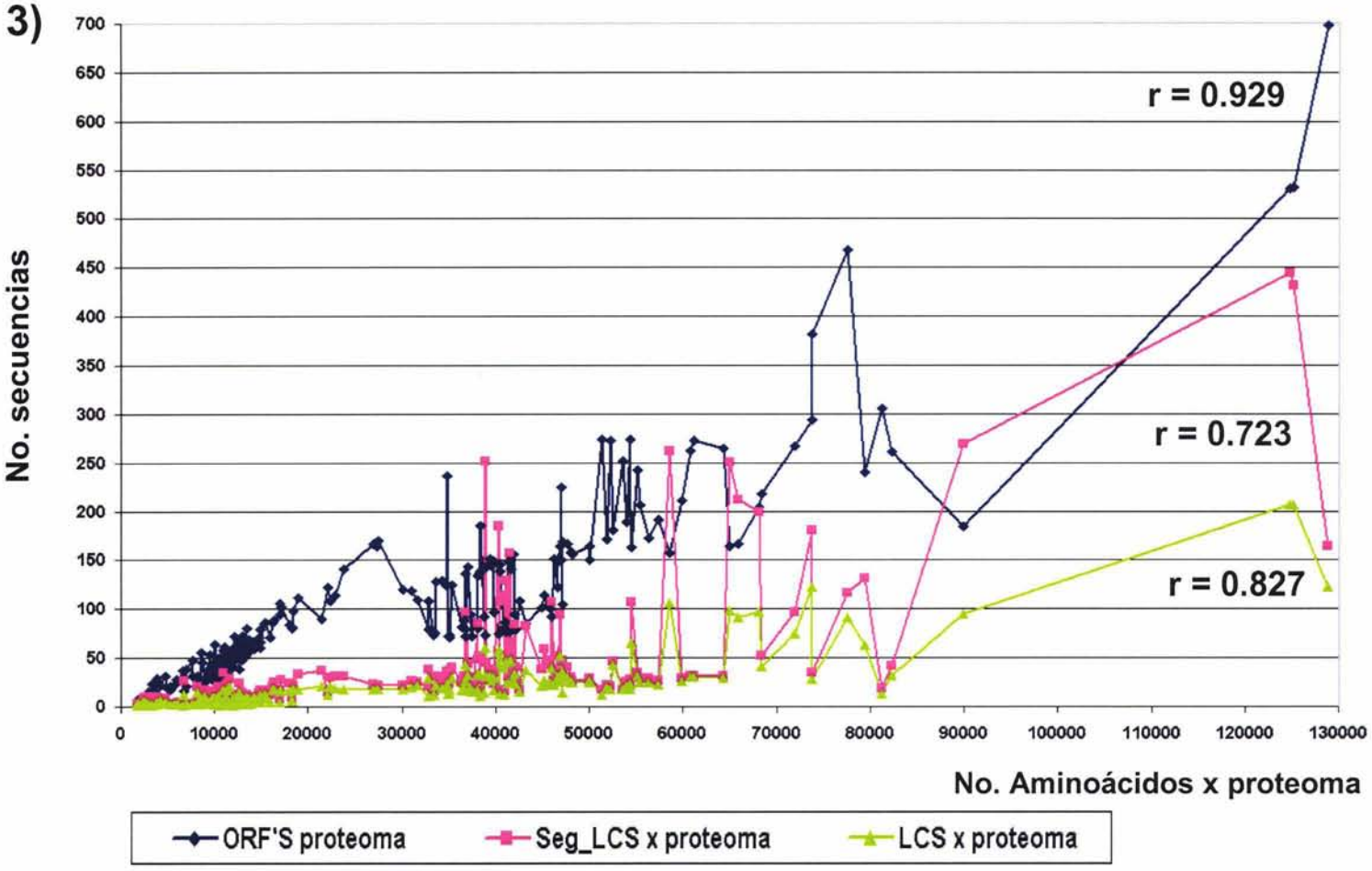
1.4 Análisis de la posición relativa de las LCS dentro de las secuencias que las contienen.

1.5 Análisis de la composición de los 20 aminoácidos tanto para el genoma completo como para las LCS que se encuentran dentro de ellos.

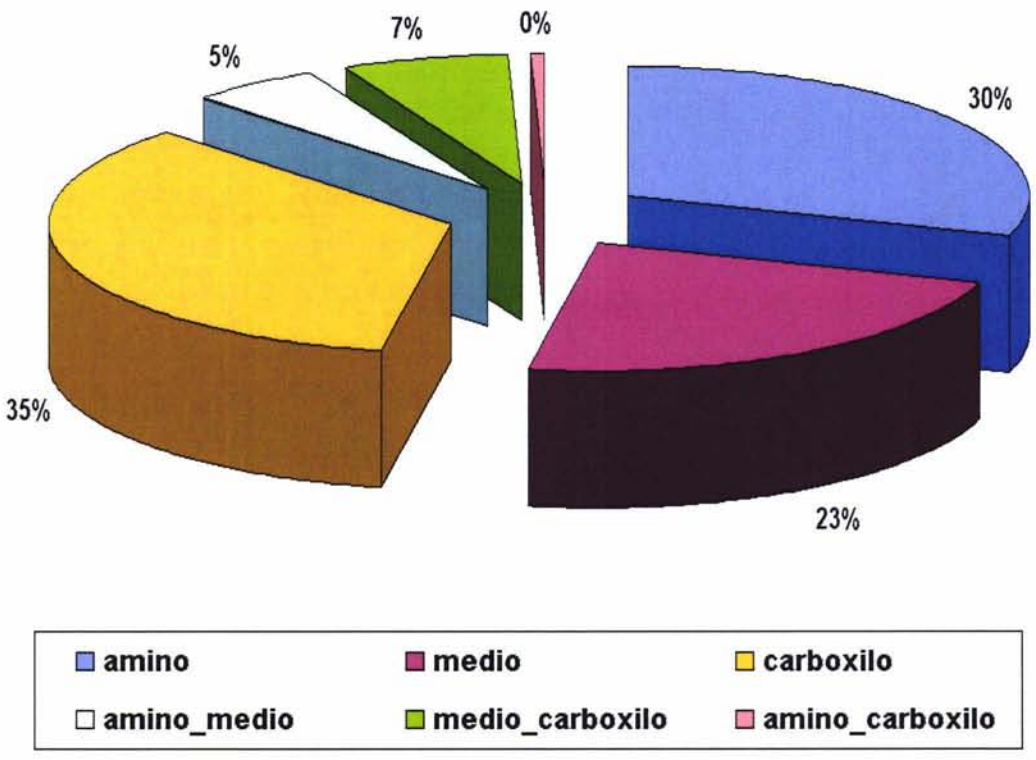
1.6 Distribución de las siete categorías funcionales de virus asignadas a las secuencias que contienen, al menos, una LCS, según el reporte previo del NCBI.


1)
No. genomas
120
100
80
60
40
20
0
118 (117)
73 (72)
29 (26)
23 (23)
15 (15)
12 (12)
1 (1)
9 (9)
23 (23)
1 (1)
4 (4)
21 (21)
2 (2)
3 (3)
2 (2)
1 (1)
2 (2)
1 (1)
1 (1)
1 (1)
1 (1)
28 (26)
22 (21)
2)
No. secuencias
5600
5200
4800
4400
4000
3600
3200
2800
2400
2000
1600
1200
800
400
0
28%
36%
36%
26%
37%
15%
40%
22%
17%
19%
14%
35%
40%
19%
15%
7%
8%
7%
31%
8%
9%
10%
8%
Polyomaviridae
Papillomaviridae
Adenoviridae
Gammaherpesvirinae
Alphaherpesvirinae
Ictalurid_Herpes_like
Betaherpesvirinae
Baculoviridae
Asfarviridae
Iridoviridae
Chordopoxvirinae
Entomopoxvirinae
Nimaviridae
Phycodnaviridae
Fuselloviridae
Rudiviridae
Lipothrixviridae
Plasmaviridae
Tectiviridae
Corticoviridae
Podoviridae
Siphoviridae
Myoviridae
Grupos virales con tipo de genoma dsDNA

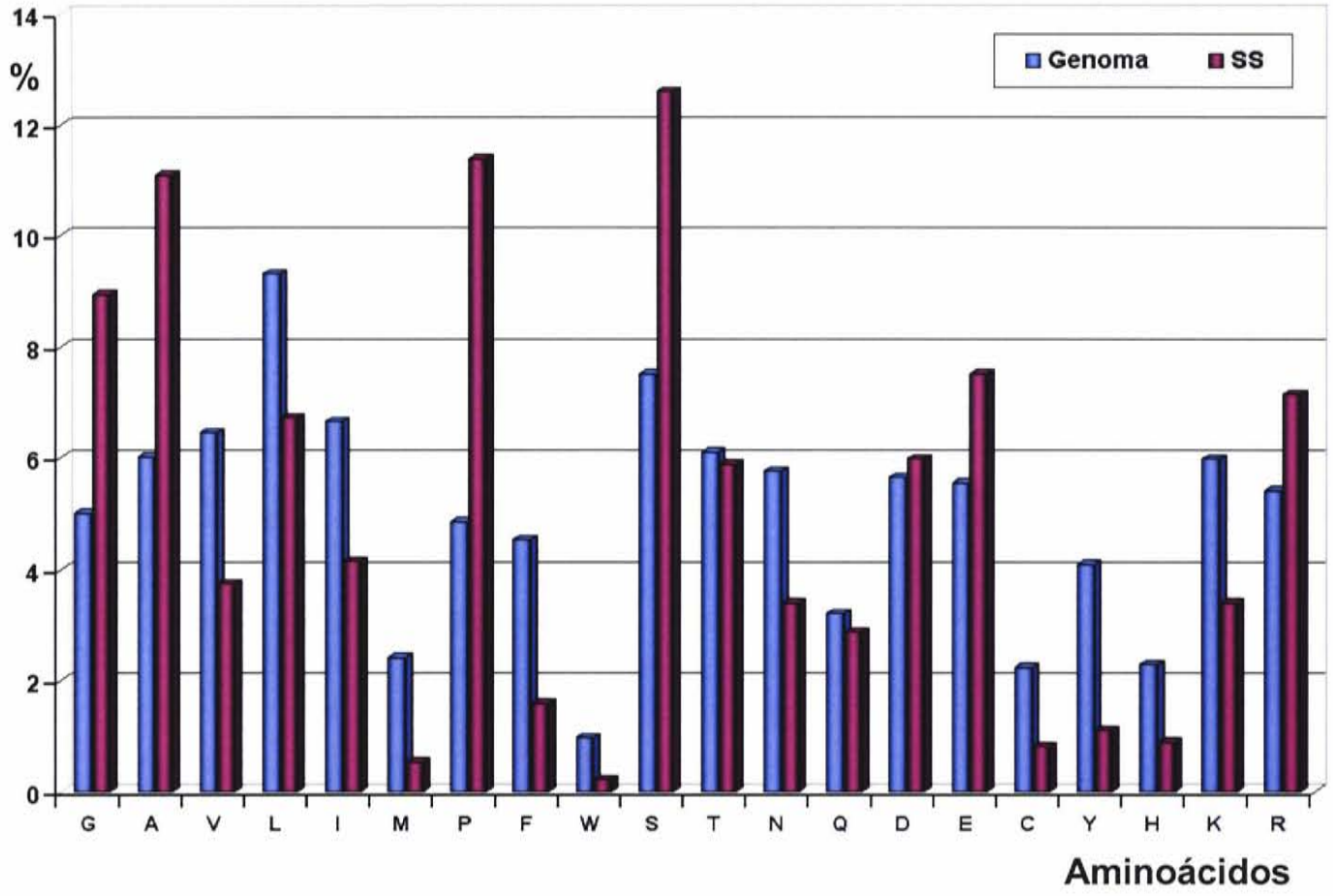
3)


No. secuencias
No. Aminoácidos x proteoma
r = 0.929
r = 0.723
r = 0.827
ORF'S proteoma
Seg_LCS x proteoma
LCS x proteoma

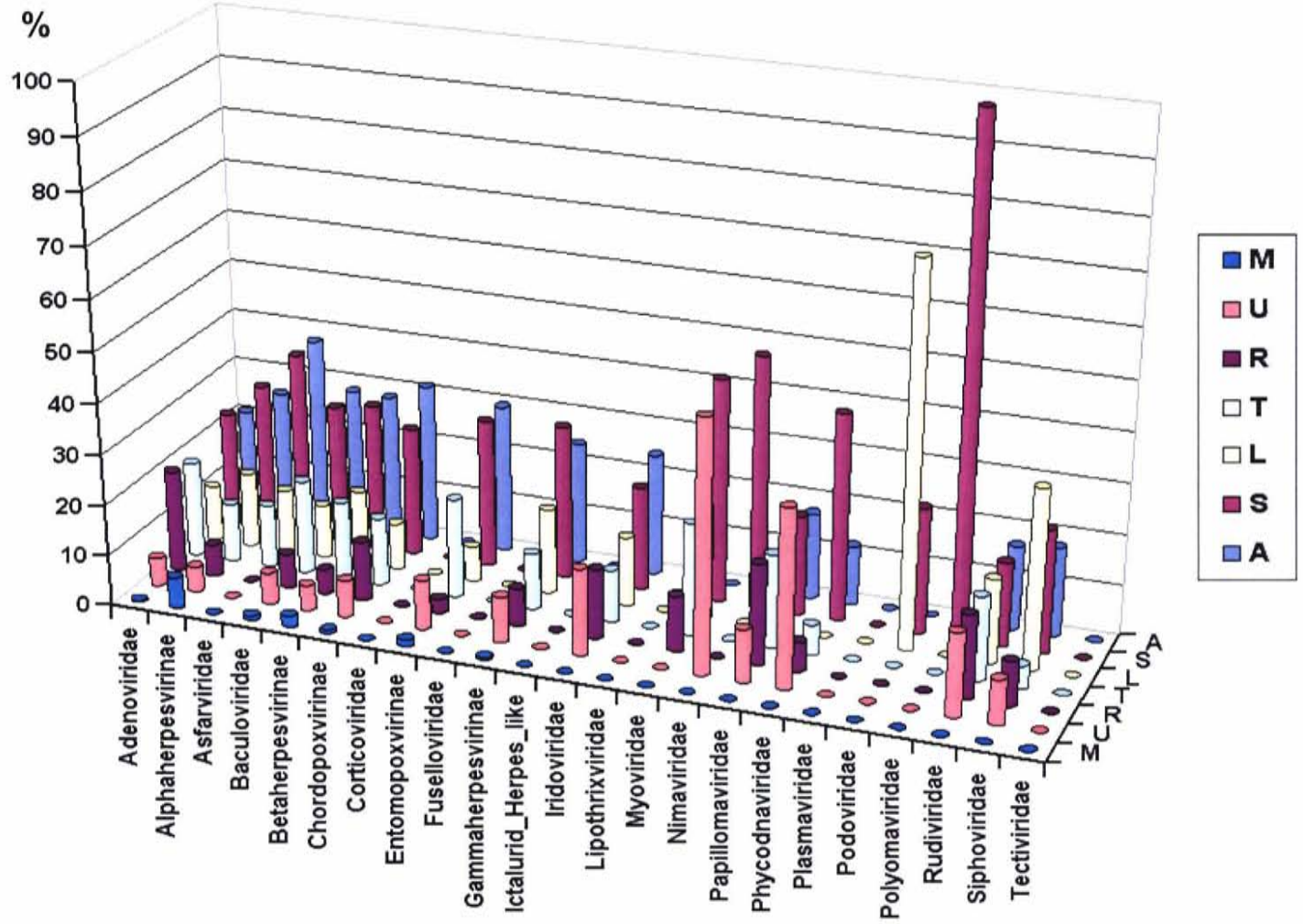
4)


5%
7%
0%
30%
35%
23%
amino
medio
carboxilo
amino_medio
medio_carboxilo
amino_carboxilo

5)


Genoma
SS
%
14
12
10
8
6
4
2
0
G A V L I M P F W S T N Q D E C Y H K R
Aminoácidos

6)


%
100
90
80
70
60
50
40
30
20
10
0
M
U
R
T
L
S
A
Adenoviridae
Alphaherpesvirinae
Asfarviridae
Baculoviridae
Betaherpesvirinae
Chordopoxvirinae
Corticoviridae
Entomopoxvirinae
Fuselloviridae
Gammaherpesvirinae
Ictalurid_Herpes_like
Iridoviridae
Lipothrixviridae
Myoviridae
Nimaviridae
Papillomaviridae
Phycodnaviridae
Plasmaviridae
Podoviridae
Polyomaviridae
Rudiviridae
Siphoviridae
Tectiviridae

Genomas de ssDNA.

TAXONOMÍA DE LOS GENOMAS VIRALES (ssDNA)			CARACTERÍSTICAS DE LOS GENOMAS			NO. ORF's			PRESENCIA DE SS		
Familia o grupo[x]	Género[x]	Hospedero[x]	No. genomas*	Tamaño promedio del genoma (nts)**	Intervalo del tamaño (nts) x grupo**	Total ORF's*	Promedio x genoma**	Intervalo de ORF's x grupo**	No. genomas con SS[E]	No. ORF's con SS[E]	% ORF'S con SS
Circoviridae		V	93	2029	1759 _ 3852	345	3	1 _ 11	86	105	30
Parvoviridae	Densovirinae	I	14	5003	3216 _ 6039	59	4	3 _ 8	9	24	40
	Parvovirinae	V	29	4975	4628 _ 5594	103	3	2 _ 7	21	33	32
Geminiviridae		Pl	230	3520	2561 _ 5576	1415	6	2 _ 10	49	59	4
Nanovirus		Pl	10	4352	1291 _ 10958	45	4	1 _ 11	2	2	4
Inoviridae		B	22	7093	4491 _ 9183	232	10	4 _ 15	22	43	18
Microviridae		B	13	5154	4421 _ 6089	132	10	7 _ 12	13	21	15
TOTALES			411	32126		2331	40		202	287	12.3

Tabla II. Información obtenida del análisis de los proteomas completos y las LCS para el tipo de genoma DNA de cadena sencilla (ssDNA).

Gráfica 2:

2.1 Distribución de los genomas colectados por grupo taxonómico en comparación con los que presentaron al menos una secuencia con LCS.

2.2 Distribución de las LCS por grupo taxonómico con respecto al contenido total de secuencias de cada grupo viral.

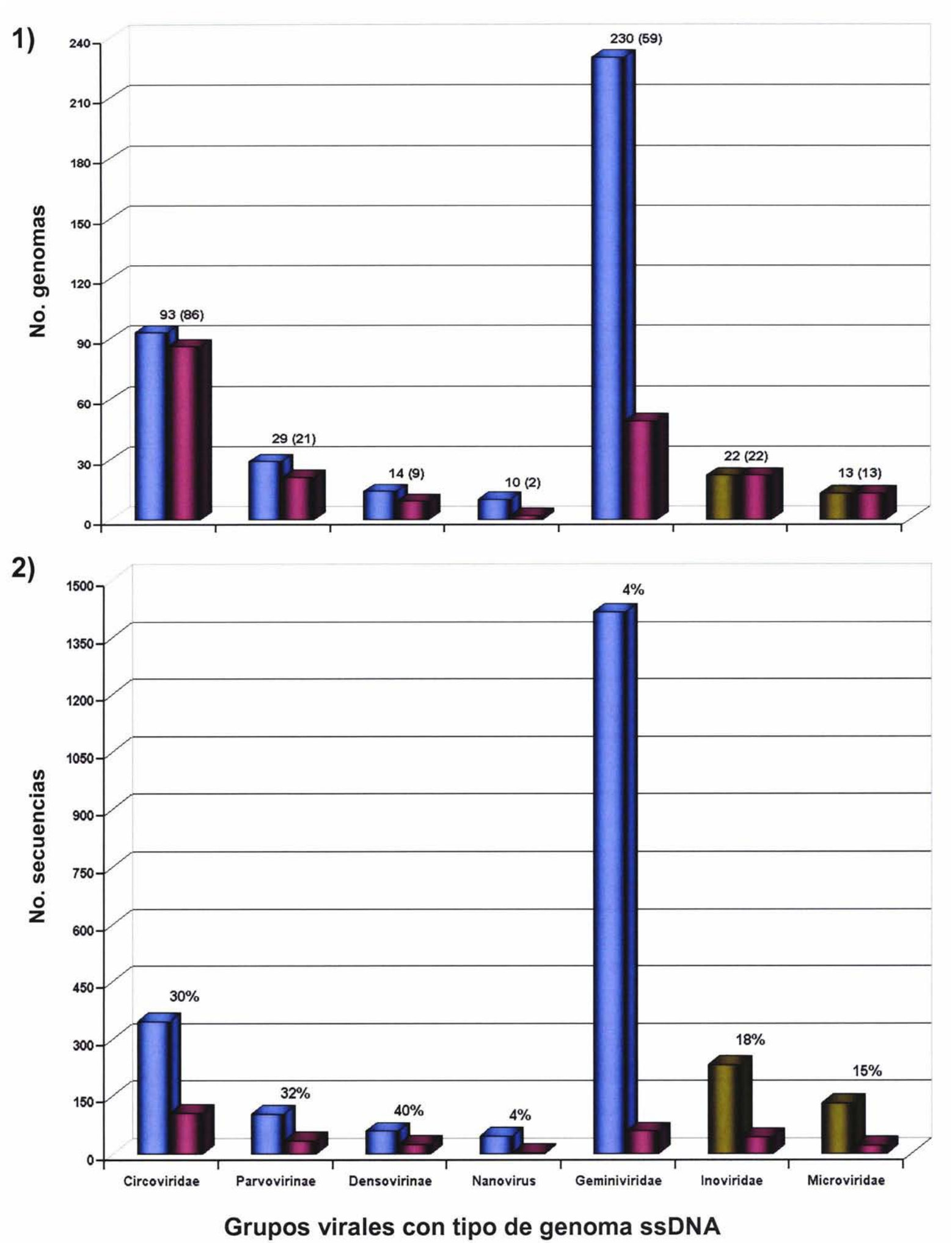
Las barras en azul, verde y amarillo son los datos colectados para virus con hospederos eucariontes, eubacterias y arqueas, respectivamente; en rojo, aquellos que presentan el fenómeno de LCS.

2.3 Correlación entre la presencia de las LCS y el tamaño del proteoma dado en número de aminoácidos y secuencias que los conforman.

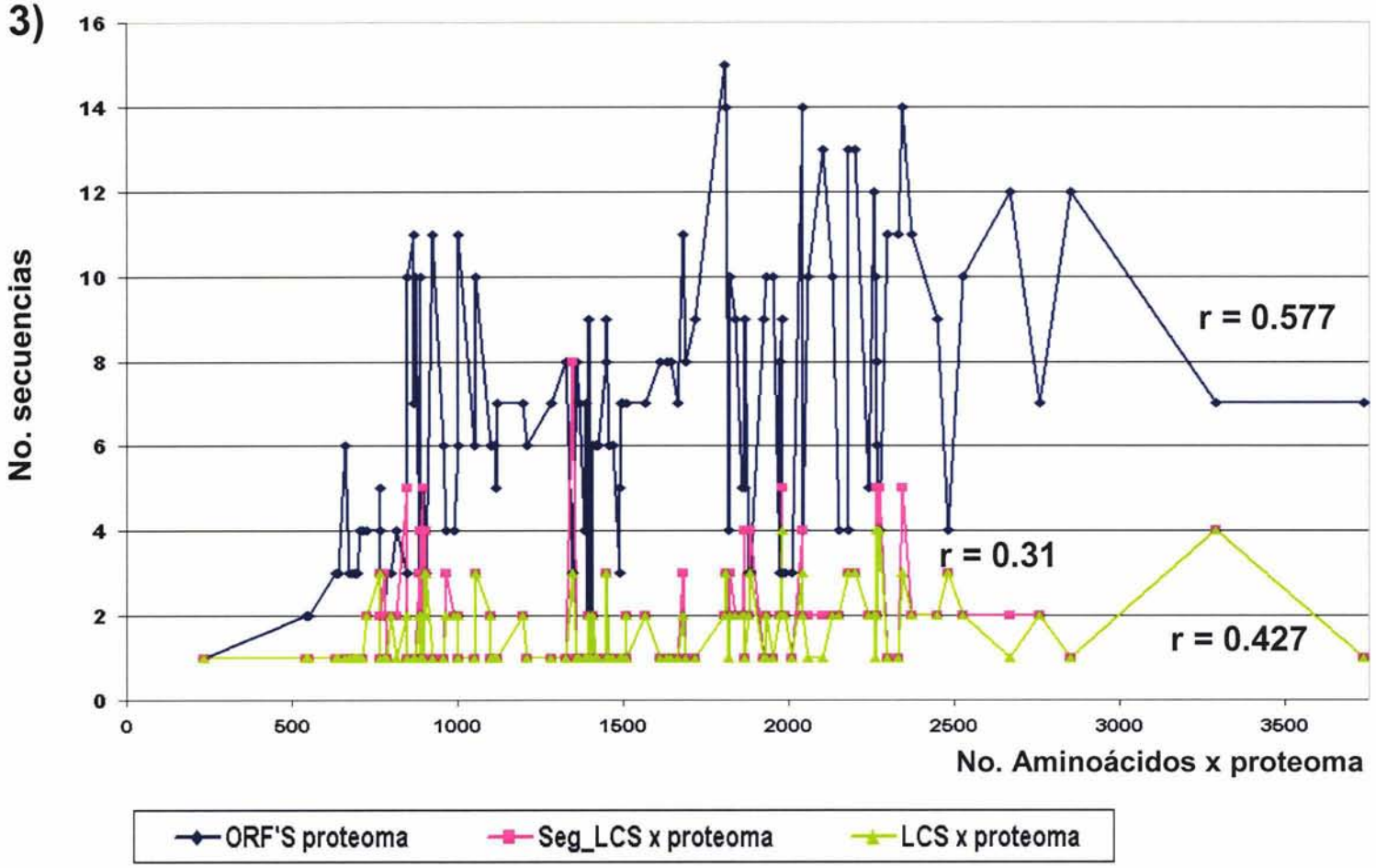
2.4 Análisis de la posición relativa de las LCS dentro de las secuencias que las contienen.

2.5 Análisis de la composición de los 20 aminoácidos tanto para el genoma completo como para las LCS que se encuentran dentro de ellos.

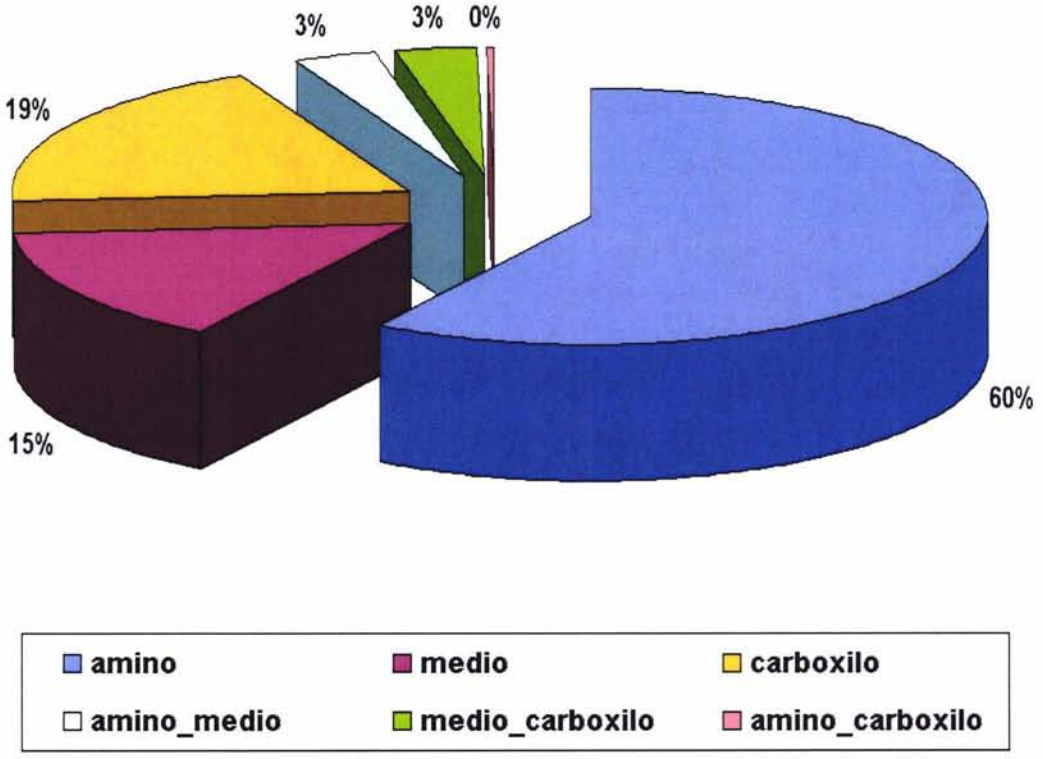
2.6 Distribución de las siete categorías funcionales de virus asignadas a las secuencias que contienen, al menos, una LCS, según el reporte previo del NCBI.


1)
No. genomas
240
210
180
150
120
90
60
30
0
93 (86)
29 (21)
14 (9)
10 (2)
230 (59)
22 (22)
13 (13)
2)
No. secuencias
1500
1350
1200
1050
900
750
600
450
300
150
0
30%
32%
40%
4%
4%
18%
15%
Circoviridae
Parvovirinae
Densovirinae
Nanovirus
Geminiviridae
Inoviridae
Microviridae
Grupos virales con tipo de genoma ssDNA

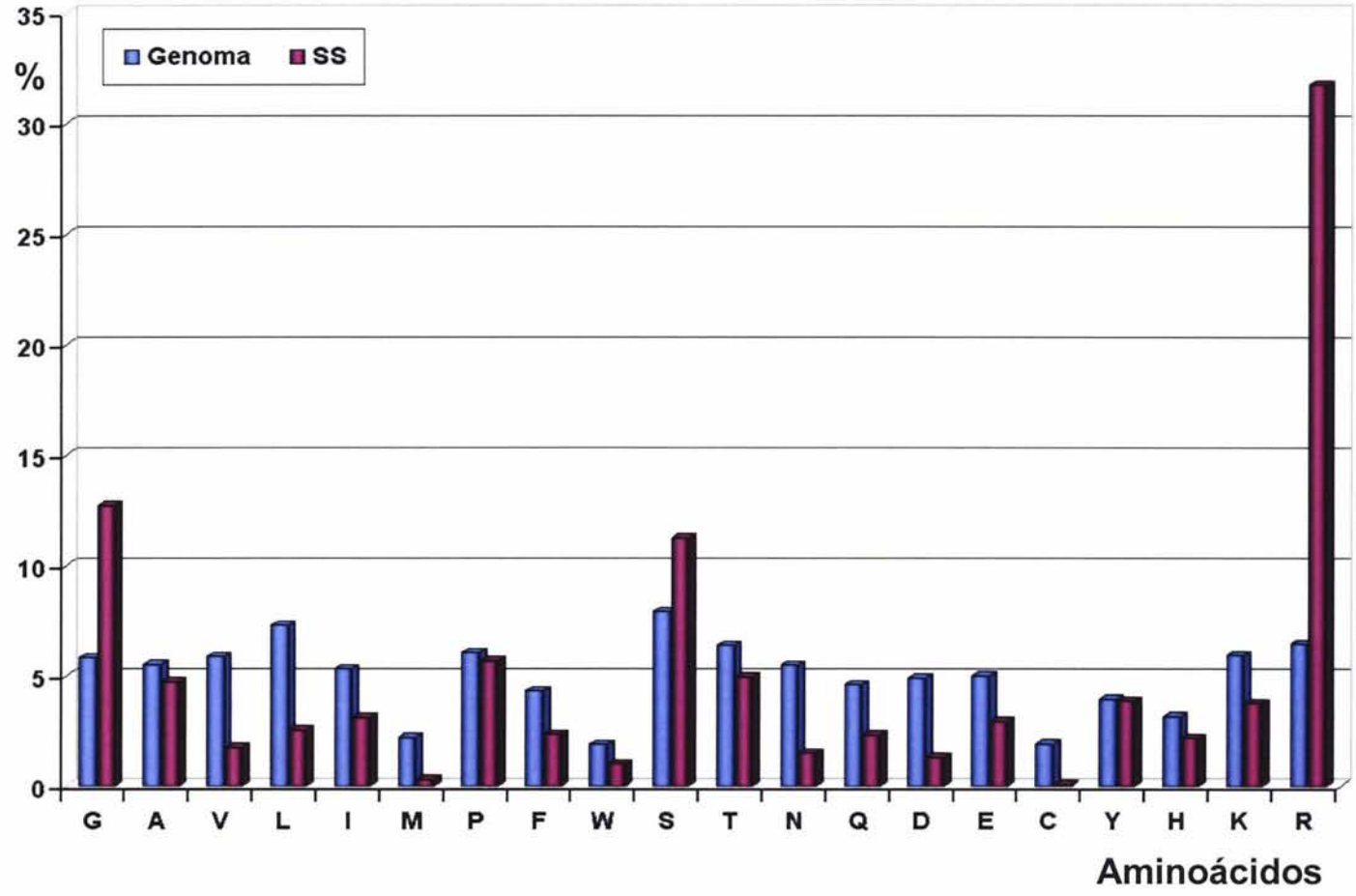
3)


No. secuencias
No. Aminoácidos x proteoma
r = 0.577
r = 0.31
r = 0.427
ORF'S proteoma
Seg_LCS x proteoma
LCS x proteoma

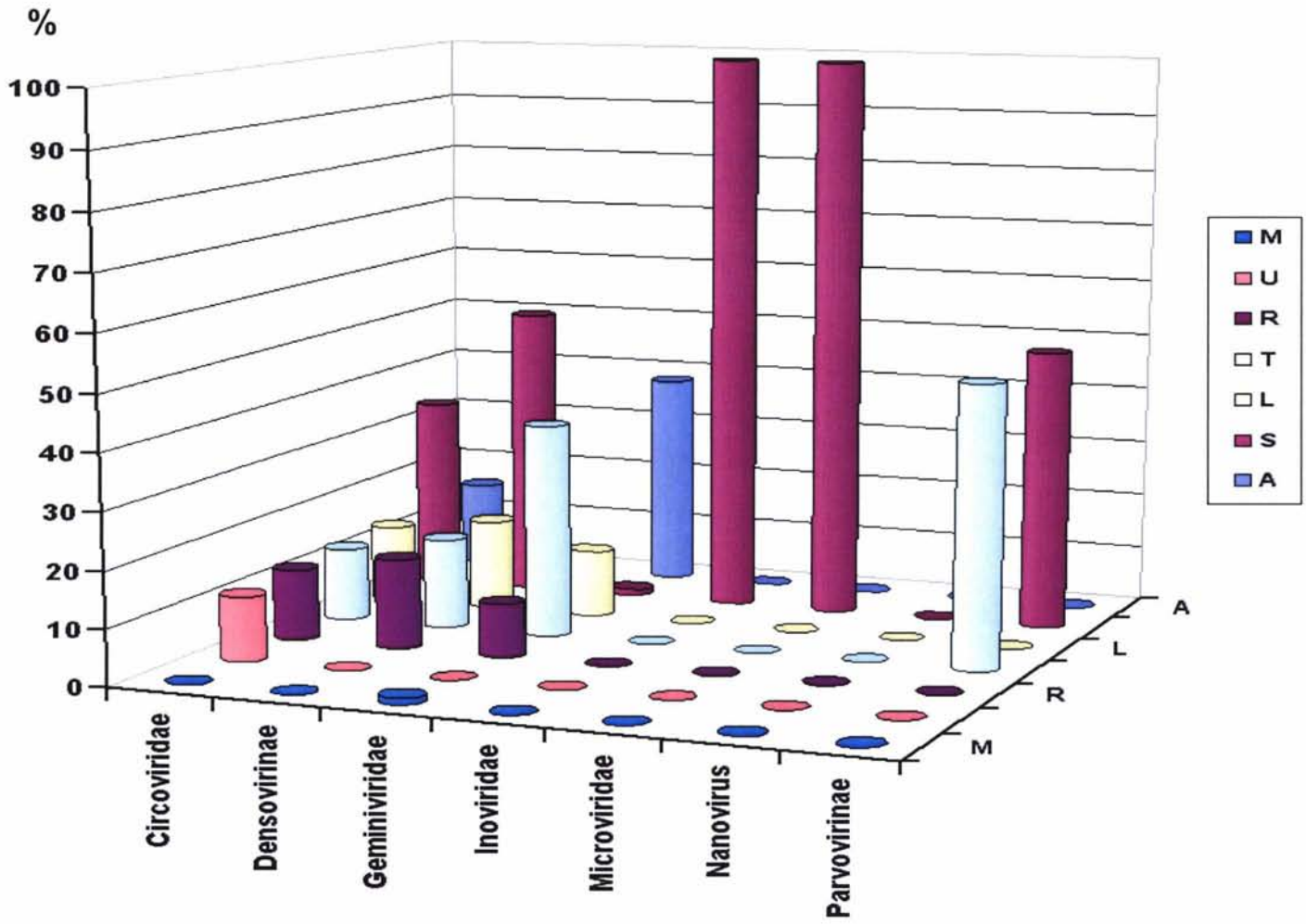
4)


3%
3%
0%
19%
60%
15%
amino
medio
carboxilo
amino_medio
medio_carboxilo
amino_carboxilo

5)


Genoma
SS
%
35
30
25
20
15
10
5
0
G A V L I M P F W S T N Q D E C Y H K R
Aminoácidos

6)


%
100
90
80
70
60
50
40
30
20
10
0
M
U
R
T
L
S
A
A
L
R
M
Circoviridae
Densovirinae
Geminiviridae
Inoviridae
Microviridae
Nanovirus
Parvovirinae

Genomas de dsRNA.

TAXONOMÍA DE LOS GENOMAS VIRALES (dsRNA)			CARACTERÍSTICAS DE LOS GENOMAS			NO. ORF's			PRESENCIA DE SS		
Familia o grupo[¥]	Género[¥]	Hospedero[¥]	No. genomas*	Tamaño promedio del genoma (nts)**	Intervalo del tamaño (nts) x grupo**	Total ORF's*	Promedio x genoma**	Intervalo de ORF's x grupo**	No. genomas con SS[£]	No. ORF's con SS[£]	% ORF'S con SS
Bimaviridae		V, I	25	5989	5592 – 6603	65	2	2 _ 3	20	27	41
Hypoviridae		H	5	11471	9591 – 12734	8	1	1 _ 2	1	1	12
Totiviridae		H, Pr	20	5176	3157 _ 7303	43	2	1 _ 3	7	9	20
Partitiviridae		H, Pl	5	3998	3090 – 4569	11	2	2 _ 3	2	2	18
Reoviridae		V, I, Pl	19	24481	18975 _ 29210	217	11	10 _ 13	18	53	24
Cystoviridae		B	4	13798	13173 _ 14984	57	14	13 _ 16	4	13	22
TOTALES			78	64913		401	32		52	105	26.18

Tabla III. Información obtenida del análisis de los proteomas completos y las LCS para el tipo de genoma RNA de cadena doble (dsRNA).

Gráfica 3:

3.1 Distribución de los genomas colectados por grupo taxonómico en comparación con los que presentaron al menos una secuencia con LCS.

3.2 Distribución de las LCS por grupo taxonómico con respecto al contenido total de secuencias de cada grupo viral.

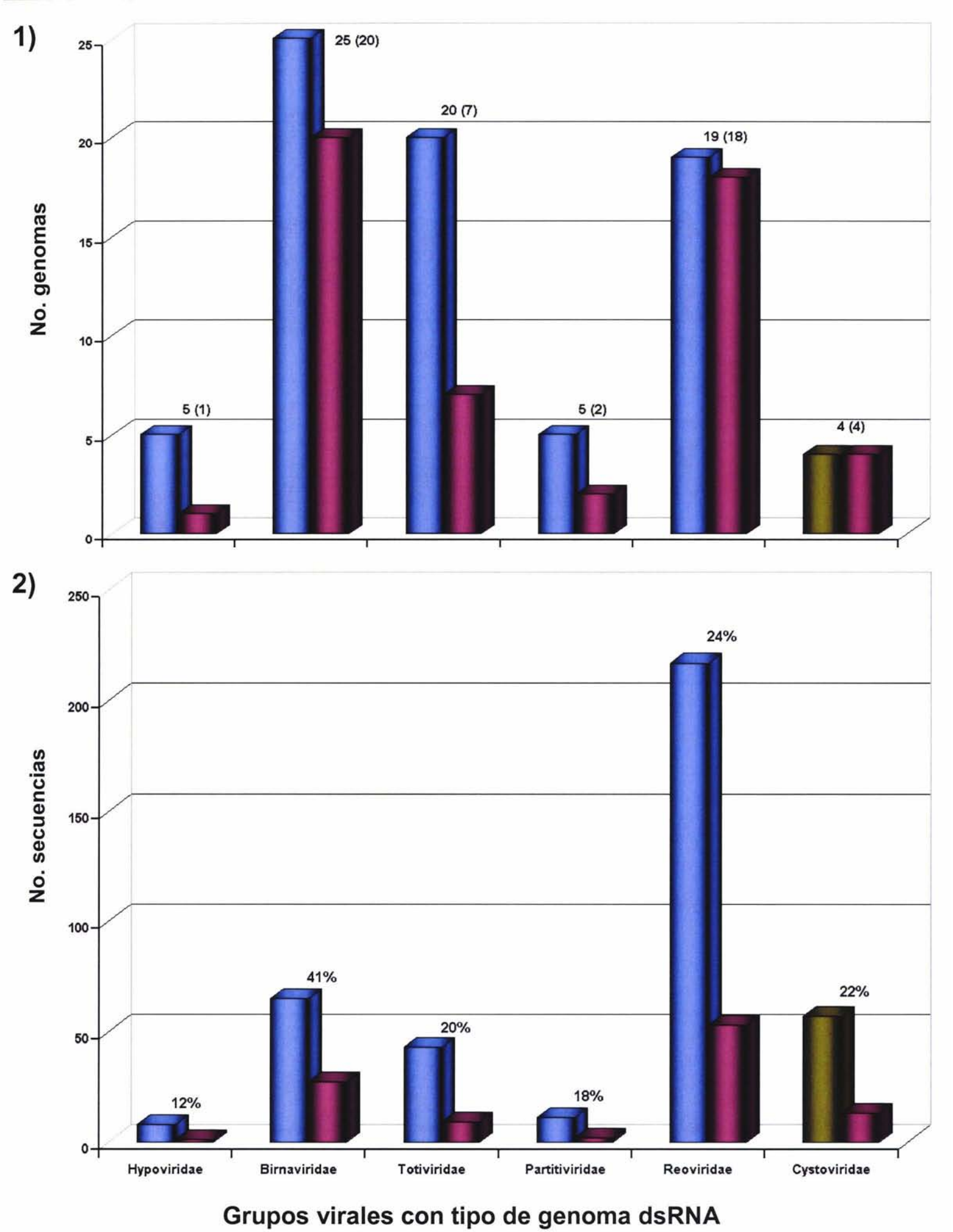
Las barras en azul, verde y amarillo son los datos colectados para virus con hospederos eucariontes, eubacterias y arqueas, respectivamente; en rojo, aquellos que presentan el fenómeno de LCS.

3.3 Correlación entre la presencia de las LCS y el tamaño del proteoma dado en número de aminoácidos y secuencias que los conforman.

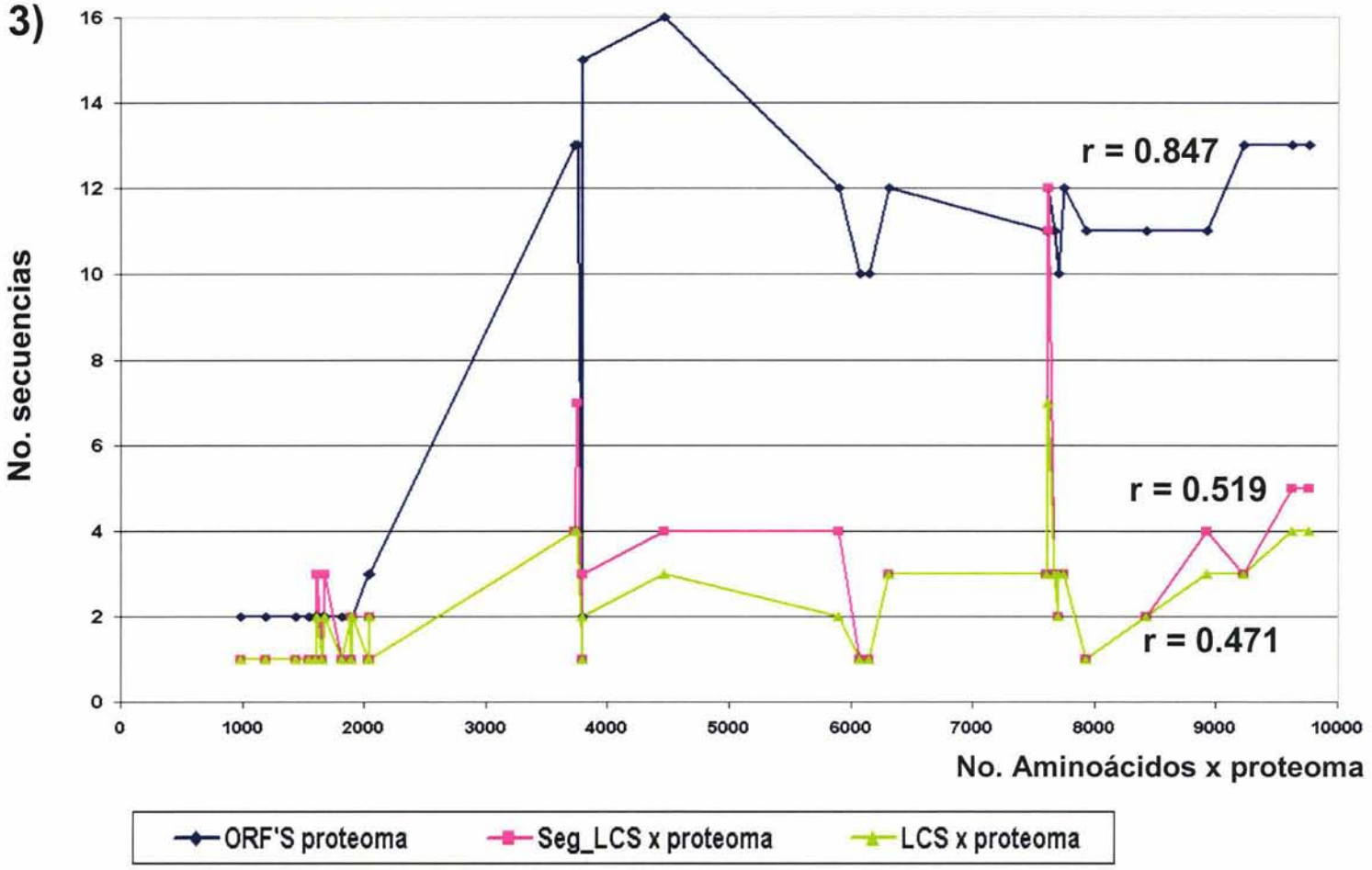
3.4 Análisis de la posición relativa de las LCS dentro de las secuencias que las contienen.

3.5 Análisis de la composición de los 20 aminoácidos tanto para el genoma completo como para las LCS que se encuentran dentro de ellos.

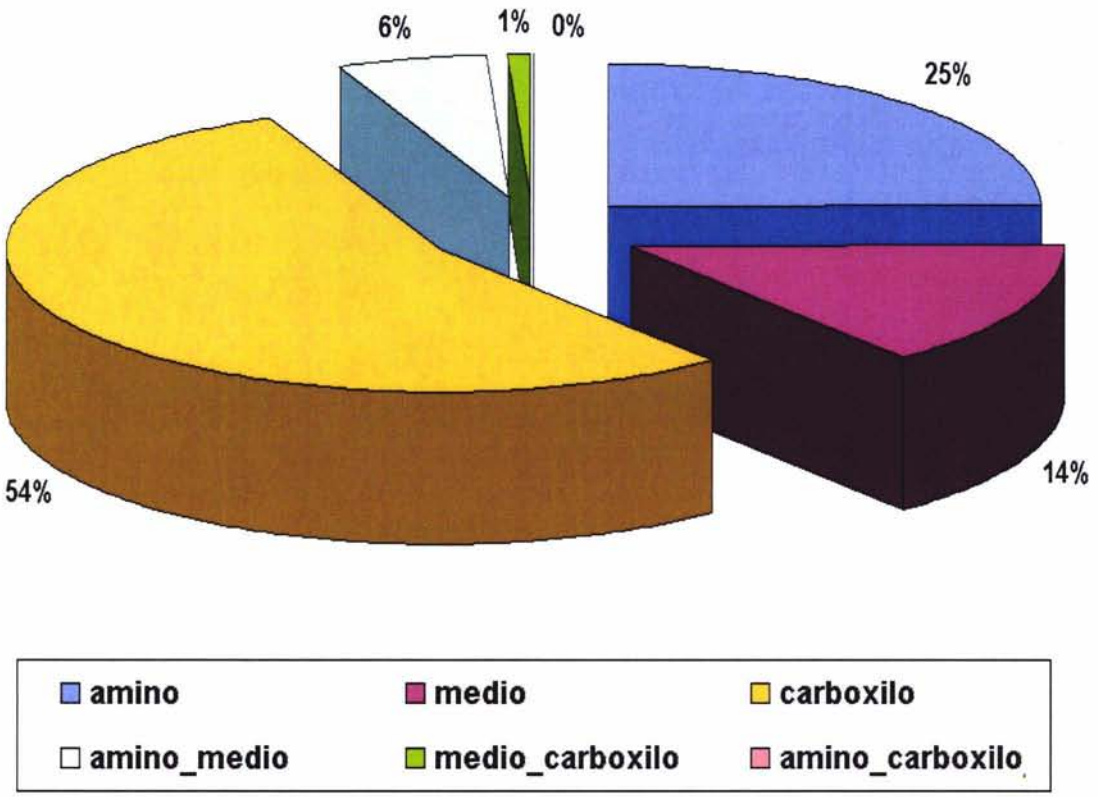
3.6 Distribución de las siete categorías funcionales de virus asignadas a las secuencias que contienen, al menos, una LCS, según el reporte previo del NCBI.


1)
No. genomas
25
20
15
10
5
0
5 (1)
25 (20)
20 (7)
5 (2)
19 (18)
4 (4)
2)
No. secuencias
250
200
150
100
50
0
12%
41%
20%
18%
24%
22%
Hypoviridae
Birnaviridae
Totiviridae
Partitiviridae
Reoviridae
Cystoviridae
Grupos virales con tipo de genoma dsRNA

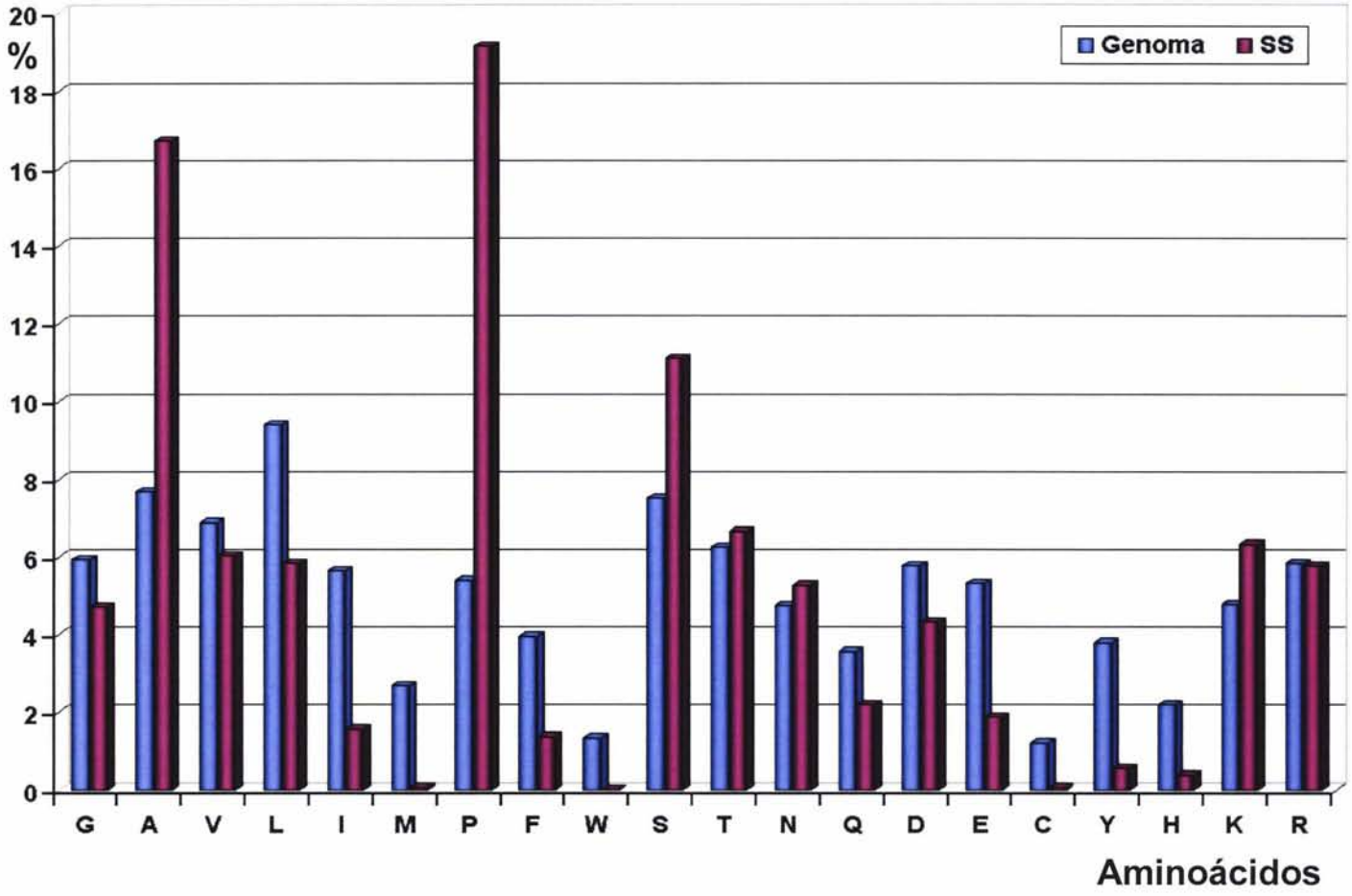
3)


No. secuencias
No. Aminoácidos x proteoma
r = 0.847
r = 0.519
r = 0.471
ORF'S proteoma
Seg_LCS x proteoma
LCS x proteoma

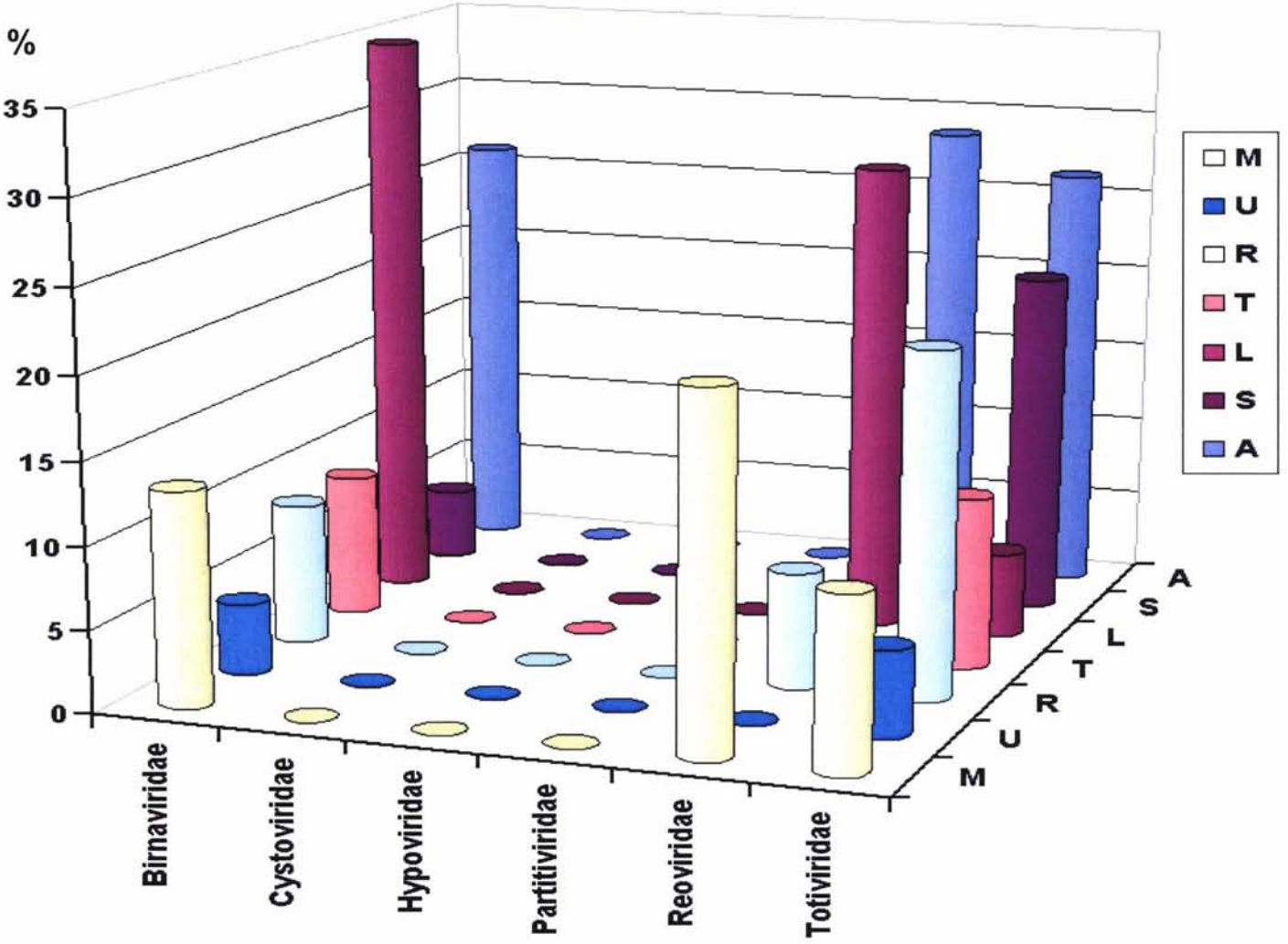
4)


6%
1%
0%
25%
54%
14%
amino
medio
carboxilo
amino_medio
medio_carboxilo
amino_carboxilo

5)


%
Genoma
SS
0
2
4
6
8
10
12
14
16
18
20
G A V L I M P F W S T N Q D E C Y H K R
Aminoácidos

6)


%
0
5
10
15
20
25
30
35
M
U
R
T
L
S
A
Birnaviridae
Cystoviridae
Hypoviridae
Partitiviridae
Reoviridae
Totiviridae

Genomas de ssRNA+.

TAXONOMÍA DE LOS GENOMAS VIRALES (ssRNA+)			CARACTERÍSTICAS DE LOS GENOMAS			NO. ORF's			PRESENCIA DE SS		
Familia o grupo*	Género*	Hospedero*	No. genomas*	Tamaño promedio del genoma (nts)**	Intervalo del tamaño (nts) x grupo**	Total ORF's*	Promedio x genoma**	Intervalo de ORF's x grupo**	No. genomas con SS[c]	No. ORF's con SS[c]	% ORF'S con SS
Arteriviridae		V	17	15108	12704 – 15717	140	8	7 _ 11	17	34	24
Astroviridae		V	9	6832	6440 _ 7355	22	2	2 _ 3	8	14	63
Barnaviridae		H	1	4009	4009	4	4	4	1	1	25
Benyvirus		Pl	2	15068	14222 – 15914	20	10	10 _ 10	2	9	45
Bromoviridae		Pl	25	8386	7830 – 8870	105	4	4 _ 5	21	27	25
Caliciviridae		V	39	7646	7320 _ 8513	105	2	2 _ 3	30	48	45
Closteroviridae		Pl	18	17216	15045 – 19302	189	10	8 _ 13	17	20	10
Comoviridae		Pl	24	10892	9219 _ 15487	56	2	2 _ 5	12	20	35
Coronaviridae		V	15	29938	27317 – 31357	156	10	6 _ 14	15	38	24
Dicistroviridae		I	11	9374	8550 – 10205	22	2	2 _ 2	4	6	27
Flaviviridae		V, I	168	10823	8930 _ 15521	169	1	1 _ 2	132	133	78
Flexiviridae		Pl	62	7310	5845 – 9306	317	5	2 _ 8	42	55	17
Furovirus		Pl	12	10629	10296 _ 10800	74	6	6 _ 7	11	23	31
Hordeivirus		Pl	4	8160	5894 _ 10221	28	7	7 _ 7	4	6	21
Idaeovirus		Pl	1	7680	7680	3	3	3	0	0	0
Ilavirus		Ins	5	9496	8832 – 10131	5	1	1 _ 1	1	1	20
Luteoviridae		Pl	34	5761	5179 _ 5987	201	6	5 _ 8	33	93	46
Namaviridae		H	7	2595	2343 – 2891	7	1	1	1	1	14
Nodaviridae		I	8	4464	4322 _ 4540	24	3	3 _ 5	7	11	45
Pecluvirus		Pl	2	10266	10131 – 10401	14	7	7 _ 7	1	3	21
Picornaviridae		V, I	149	7574	7055 _ 9476	153	1	1 _ 2	36	37	24
Pomovirus		Pl	4	11681	10848 – 12293	32	8	7 _ 9	3	7	21
Potyviridae		Pl	110	9868	9324 _ 11296	129	1	1 _ 9	78	85	65
Sequiviridae		Pl	7	11584	9871 – 12655	10	1	1 _ 2	5	7	70
Sobemovirus		Pl	14	4189	4037 _ 4451	56	4	3 _ 5	8	10	17
Tetraviridae		I	2	7207	6625 _ 7790	5	2	2 _ 3	2	3	60
Tobamovirus		Pl	40	6417	6297 _ 6618	157	3	3 _ 6	21	22	14
Tobravirus		Pl	4	9942	8627 _ 10646	25	6	5 _ 7	4	6	24
Togaviridae		V, I	62	11334	9743 – 11919	134	2	2 _ 3	62	124	92
Tombusviridae		Pl	38	4341	3641 _ 5346	201	5	4 _ 8	23	38	18
Tymoviridae		Pl	14	6409	6035 _ 7564	40	2	2 _ 4	13	28	70
Umbravirus		Pl	4	4156	4019 _ 4253	16	4	4 _ 4	1	2	12
Leviviridae		B	10	3890	3466 _ 4276	39	3	3 _ 4	2	2	5
TOTALES			922	300245		2658	136		617	914	34.4

Tabla IV. Información obtenida del análisis de los proteomas completos y las LCS para el tipo de genoma RNA de cadena sencilla sentido positivo (ssRNA+).

Gráfica 4:

4.1 Distribución de los genomas colectados por grupo taxonómico en comparación con los que presentaron al menos una secuencia con LCS.

4.2 Distribución de las LCS por grupo taxonómico con respecto al contenido total de secuencias de cada grupo viral.

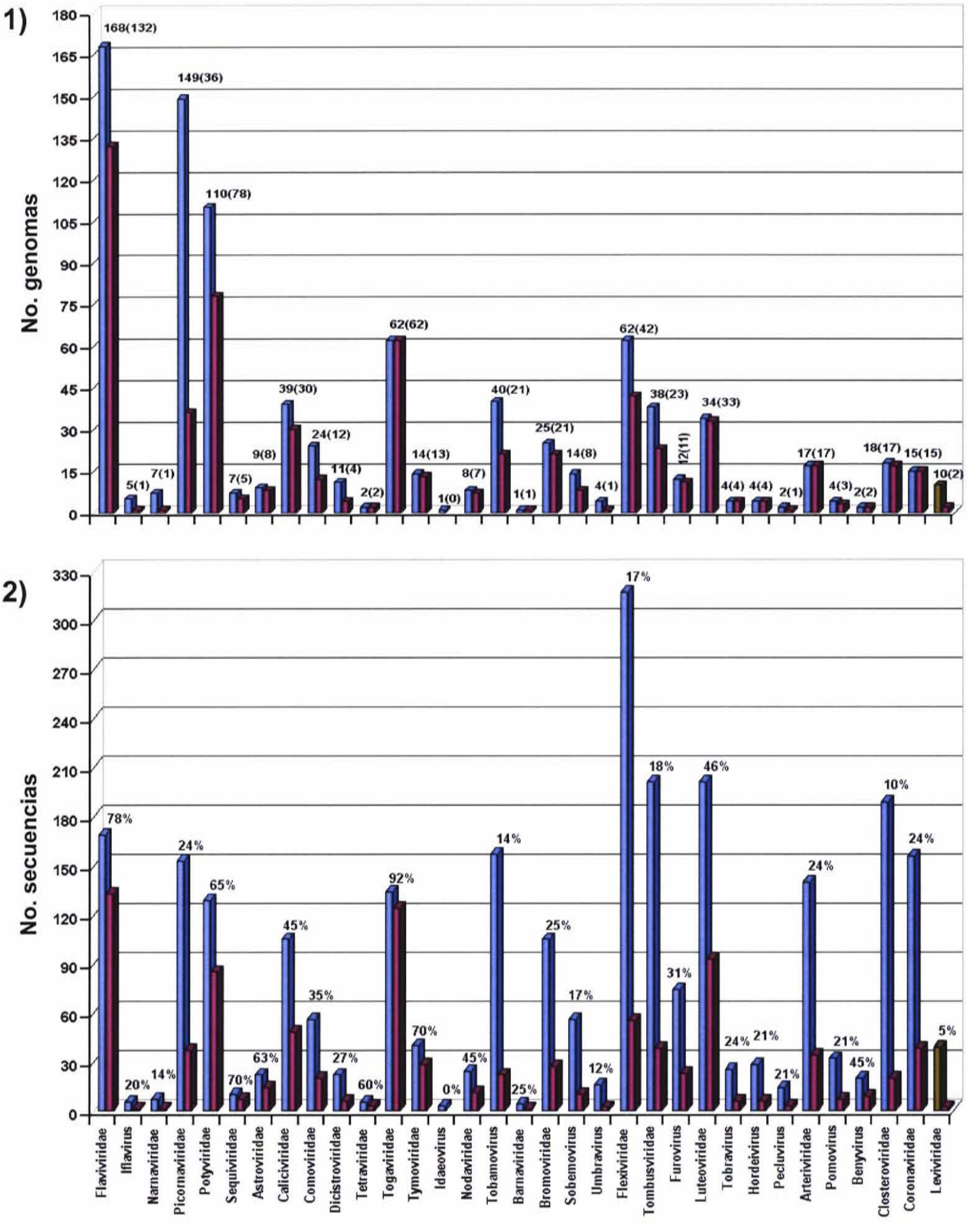
Las barras en azul, verde y amarillo son los datos colectados para virus con hospederos eucariontes, eubacterias y arqueas, respectivamente; en rojo, aquellos que presentan el fenómeno de LCS.

4.3 Correlación entre la presencia de las LCS y el tamaño del proteoma dado en número de aminoácidos y secuencias que los conforman.

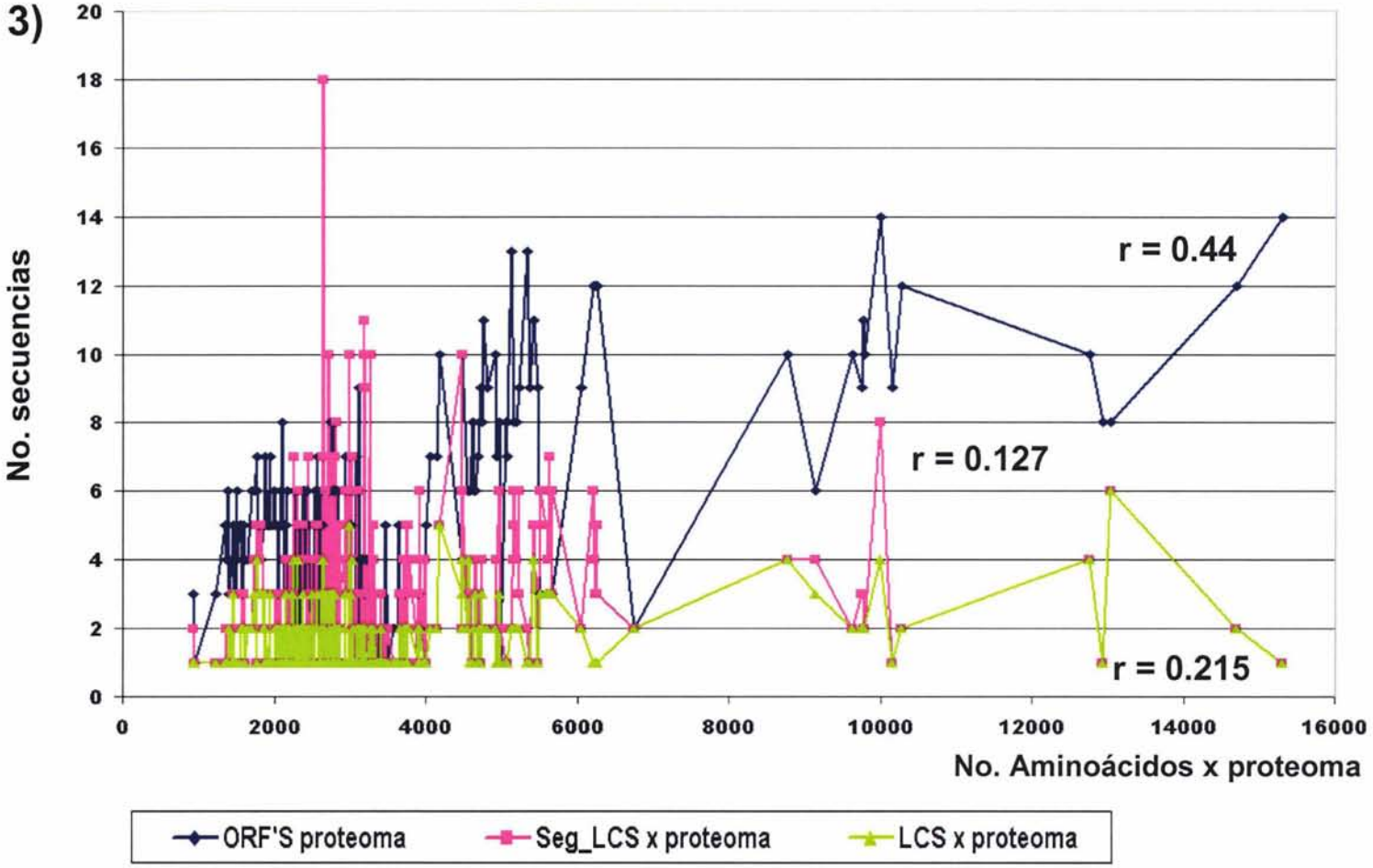
4.4 Análisis de la posición relativa de las LCS dentro de las secuencias que las contienen.

4.5 Análisis de la composición de los 20 aminoácidos tanto para el genoma completo como para las LCS que se encuentran dentro de ellos.

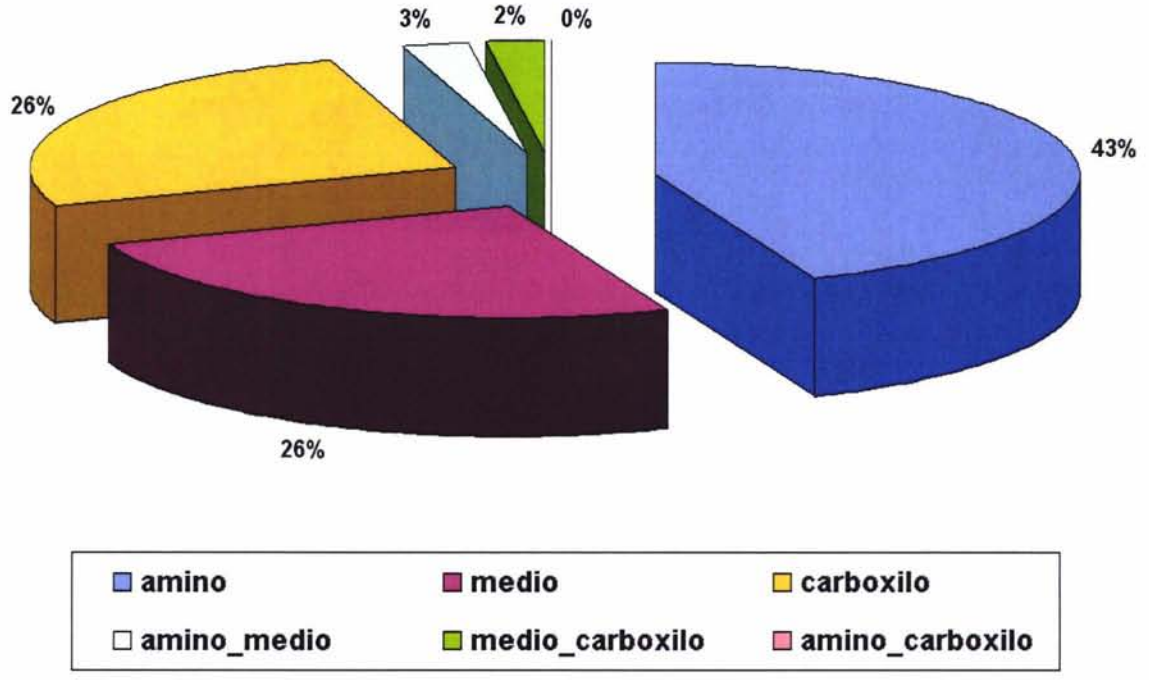
4.6 Distribución de las siete categorías funcionales de virus asignadas a las secuencias que contienen, al menos, una LCS, según el reporte previo del NCBI.


1)
No. genomas
168(132)
5(1)
7(1)
149(36)
110(78)
7(5)
9(8)
39(30)
24(12)
11(4)
2(2)
62(62)
14(13)
1(0)
8(7)
40(21)
1(1)
25(21)
14(8)
4(1)
62(42)
38(23)
12(11)
34(33)
4(4)
4(4)
2(1)
17(17)
4(3)
2(2)
18(17)
15(15)
10(2)
2)
No. secuencias
78%
20%
14%
24%
65%
70%
63%
45%
35%
27%
60%
92%
70%
0%
45%
14%
25%
25%
17%
12%
17%
18%
31%
46%
24%
21%
21%
24%
21%
45%
10%
24%
5%
Flaviviridae
Iflavirus
Narnaviridae
Picornaviridae
Potyviridae
Sequiviridae
Astroviridae
Caliciviridae
Comoviridae
Dicistroviridae
Tetraviridae
Togaviridae
Tymoviridae
Idaeovirus
Nodaviridae
Tobamovirus
Barnaviridae
Bromoviridae
Sobemovirus
Umbravirus
Flexiviridae
Tombusviridae
Furovirus
Luteoviridae
Tobravirus
Hordeivirus
Pecluvirus
Arteriviridae
Pomovirus
Benyvirus
Closteroviridae
Coronaviridae
Leviviridae
Grupos virales con tipo de genoma ssRNA positivo

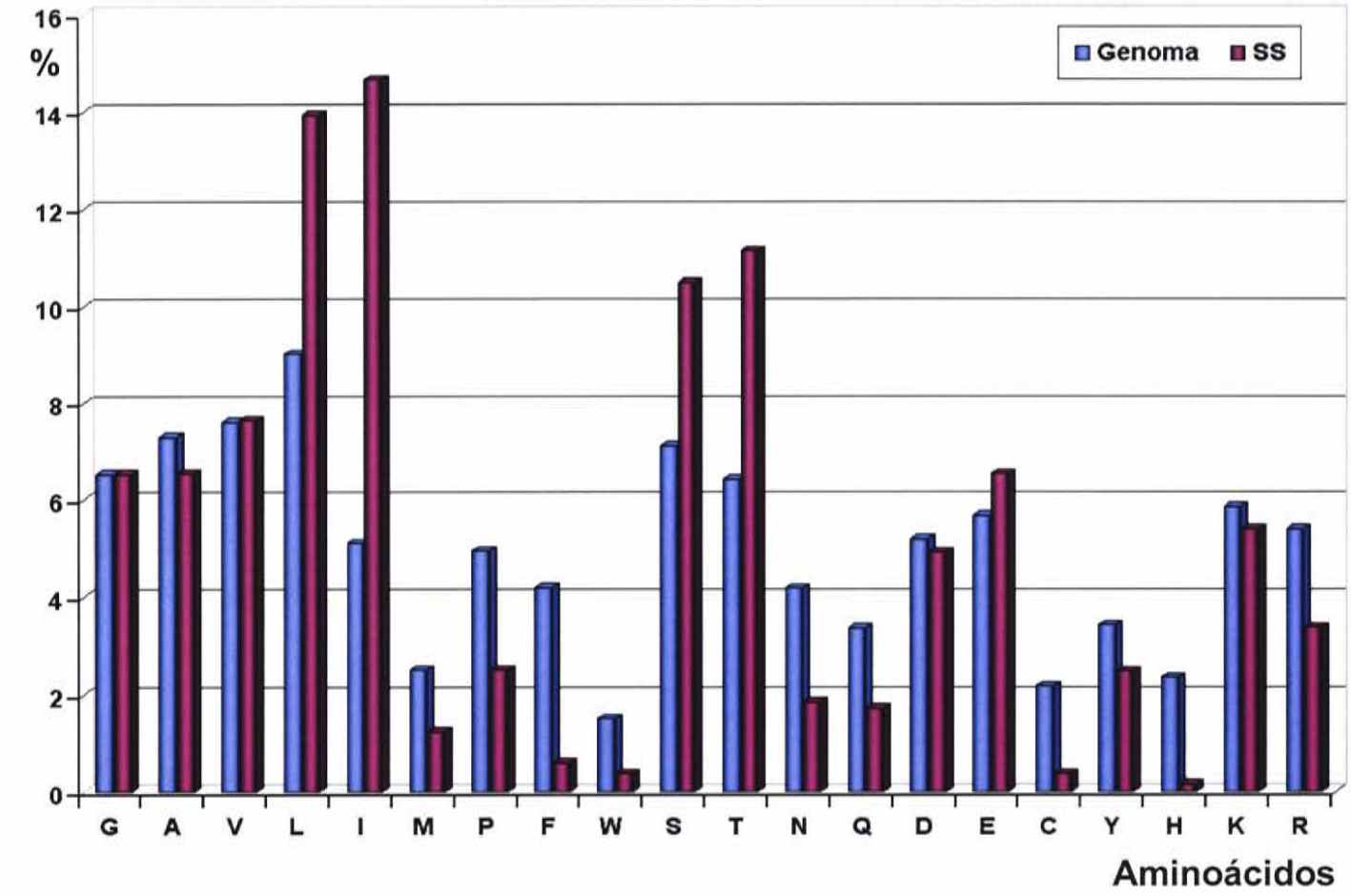
3)


No. secuencias
No. Aminoácidos x proteoma
r = 0.44
r = 0.127
r = 0.215
ORF'S proteoma
Seg_LCS x proteoma
LCS x proteoma

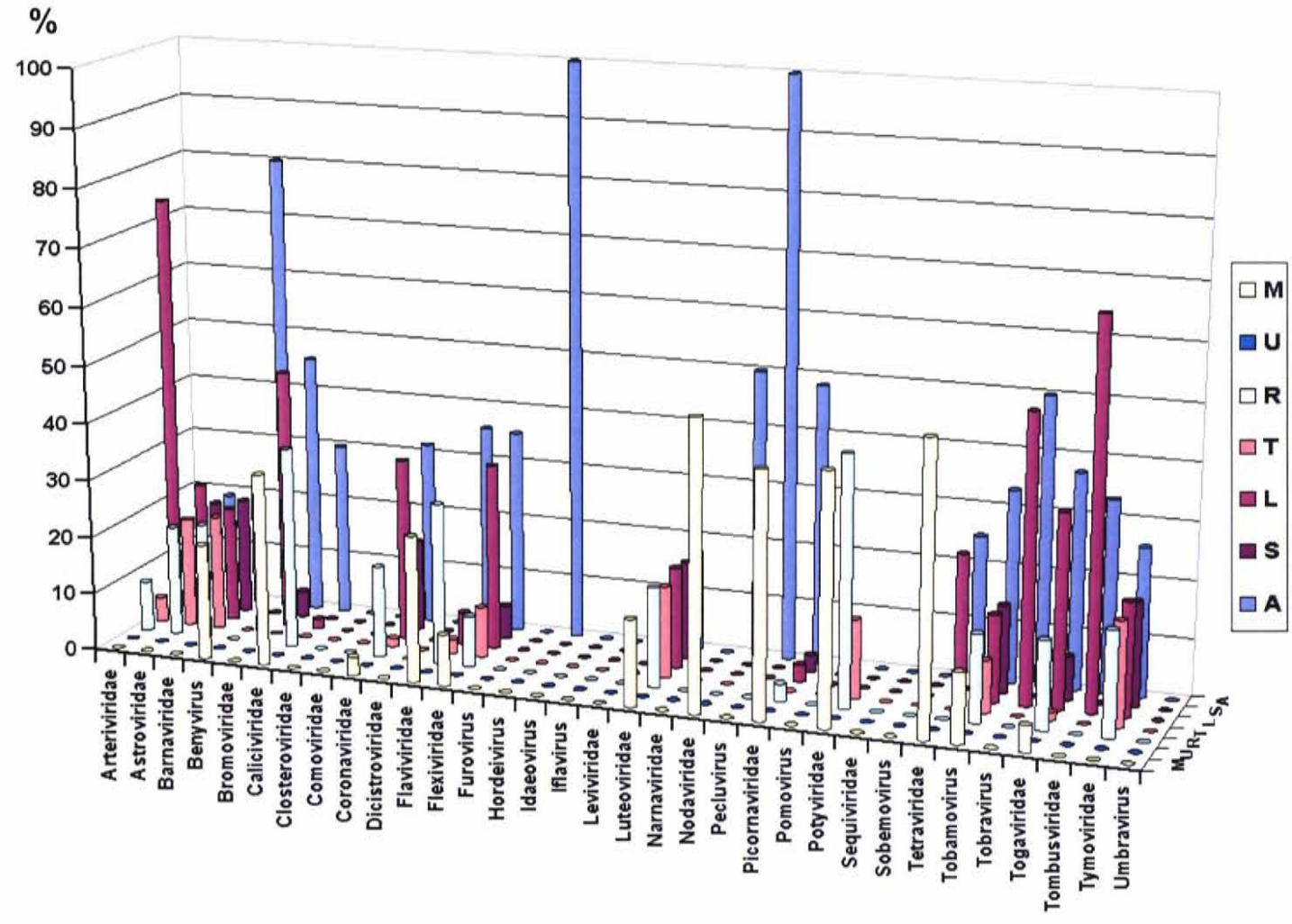
4)


3%
2%
0%
26%
43%
26%
amino
medio
carboxilo
amino_medio
medio_carboxilo
amino_carboxilo

5)


%
Genoma
SS
16
14
12
10
8
6
4
2
0
G A V L I M P F W S T N Q D E C Y H K R
Aminoácidos

6)


%
100
90
80
70
60
50
40
30
20
10
0
M
U
R
T
L
S
A
Arteriviridae
Astroviridae
Barnaviridae
Benyvirus
Bromoviridae
Caliciviridae
Closteroviridae
Comoviridae
Coronaviridae
Dicistroviridae
Flaviviridae
Flexiviridae
Furovirus
Hordeivirus
Idaeovirus
Iflavirus
Leviviridae
Luteoviridae
Narnaviridae
Nodaviridae
Pecluvirus
Picornaviridae
Pomovirus
Potyviridae
Sequiviridae
Sobemovirus
Tetraviridae
Tobamovirus
Tobravirus
Togaviridae
Tombusviridae
Tymoviridae
Umbravirus

Genomas de ssRNA-.

TAXONOMÍA DE LOS GENOMAS VIRALES (ssRNA-)			CARACTERÍSTICAS DE LOS GENOMAS			NO. ORF's			PRESENCIA DE SS		
Familia o grupo*	Género*	Hospedero*	No. genomas*	Tamaño promedio del genoma (nts)**	Intervalo del tamaño (nts) x grupo**	Total ORF's*	Promedio x genoma**	Intervalo de ORF's x grupo**	No. genomas con SS[c]	No. ORF's con SS[c]	% ORF'S con SS
Arenaviridae		V	6	10475	10056_10681	24	4	4_4	3	3	12
Bunyaviridae		V, Pl, Ro, Ar	14	13981	11372–18859	56	4	3_6	9	11	19
Orthomyxoviridae		V	6	13914	13506_14452	60	10	10_10	0	0	0
Bornaviridae		V	10	8387	3694_8910	57	5	5_6	10	21	36
Filoviridae		V	7	18982	18890_19112	55	7	7_9	7	18	32
Paramyxoviridae	Paramyxovirinae	V	61	15656	11200_18246	458	7	6_10	42	101	22
	Pneumovirinae	V	10	15070	13933_15225	110	11	9_12	10	38	34
Rhabdoviridae		V, I, Pl, Ar	26	11718	10845–14900	154	5	5_12	23	29	18
Ophiovirus		Pl	1	12499	12499	7	7	7	1	1	14
Tenuivirus		Pl	5	21948	17145–25159	49	9	7_12	2	2	4
TOTALES			146	142630		1030	69		107	224	21.7

Tabla V. Información obtenida del análisis de los proteomas completos y las LCS para el tipo de genoma RNA de cadena sencilla sentido negativo (ssRNA-).

Gráfica 5:

5.1 Distribución de los genomas colectados por grupo taxonómico en comparación con los que presentaron al menos una secuencia con LCS.

5.2 Distribución de las LCS por grupo taxonómico con respecto al contenido total de secuencias de cada grupo viral.

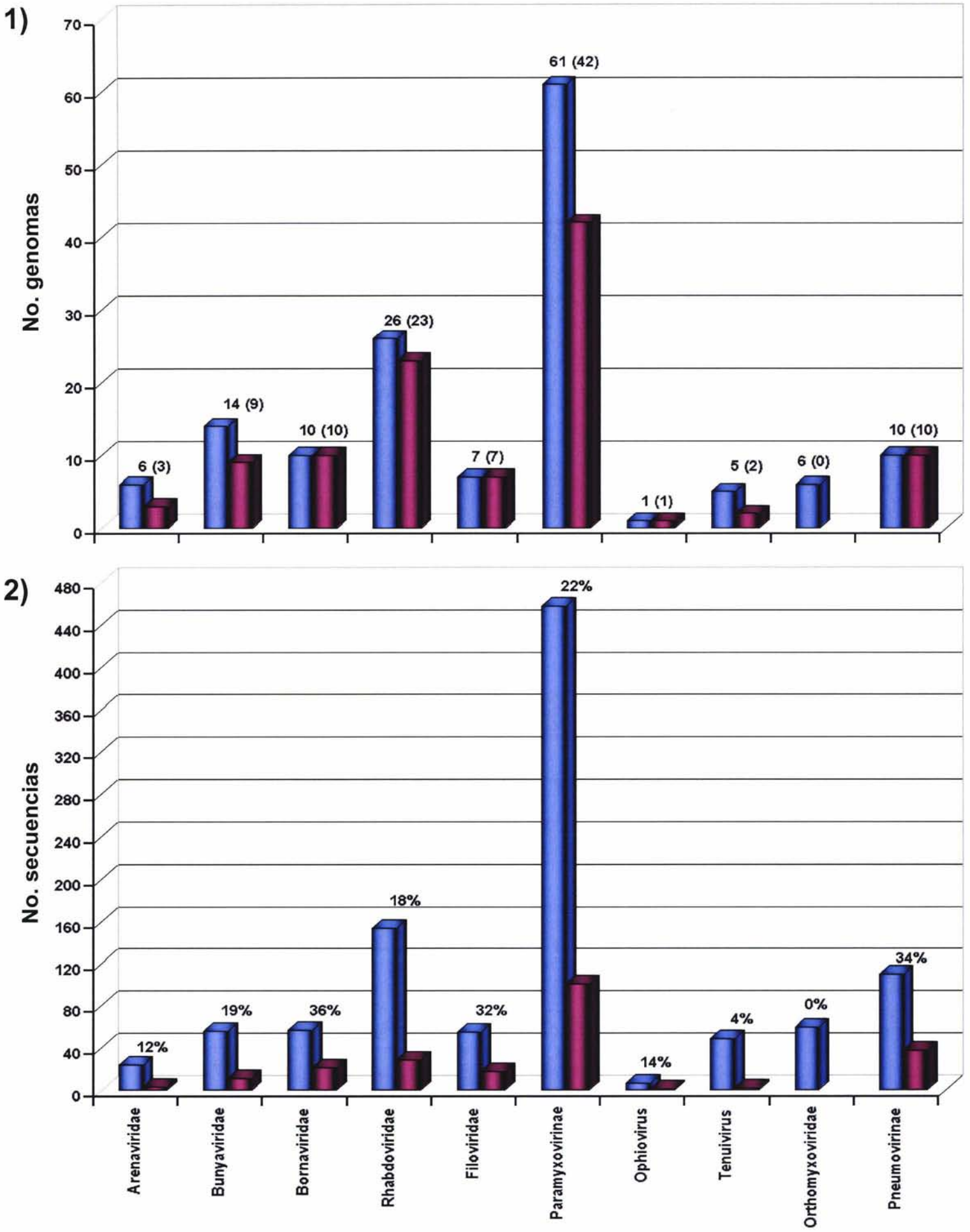
Las barras en azul, verde y amarillo son los datos colectados para virus con hospederos eucariontes, eubacterias y arqueas, respectivamente; en rojo, aquellos que presentan el fenómeno de LCS.

5.3 Correlación entre la presencia de las LCS y el tamaño del proteoma dado en número de aminoácidos y secuencias que los conforman.

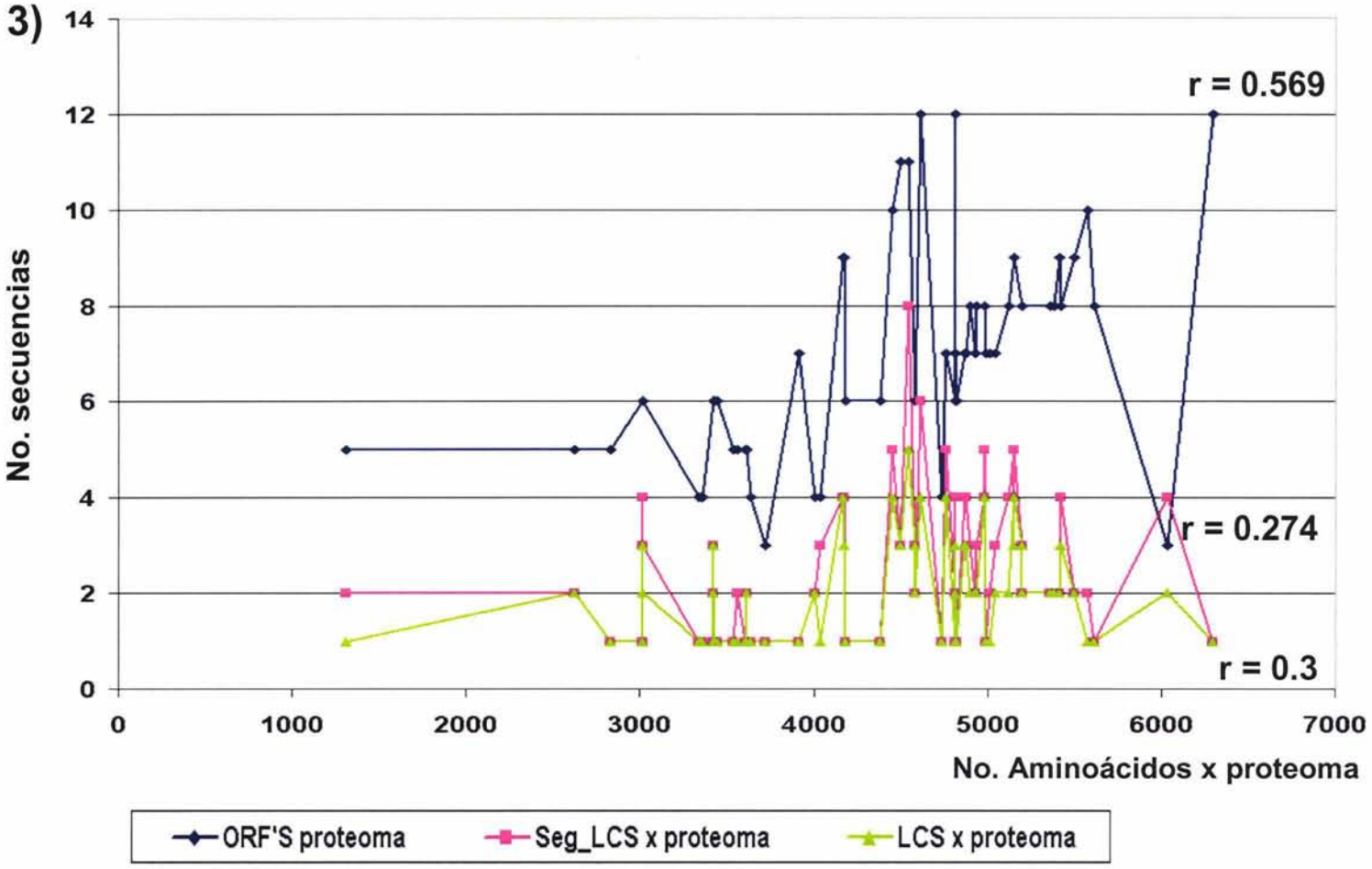
5.4 Análisis de la posición relativa de las LCS dentro de las secuencias que las contienen.

5.5 Análisis de la composición de los 20 aminoácidos tanto para el genoma completo como para las LCS que se encuentran dentro de ellos.

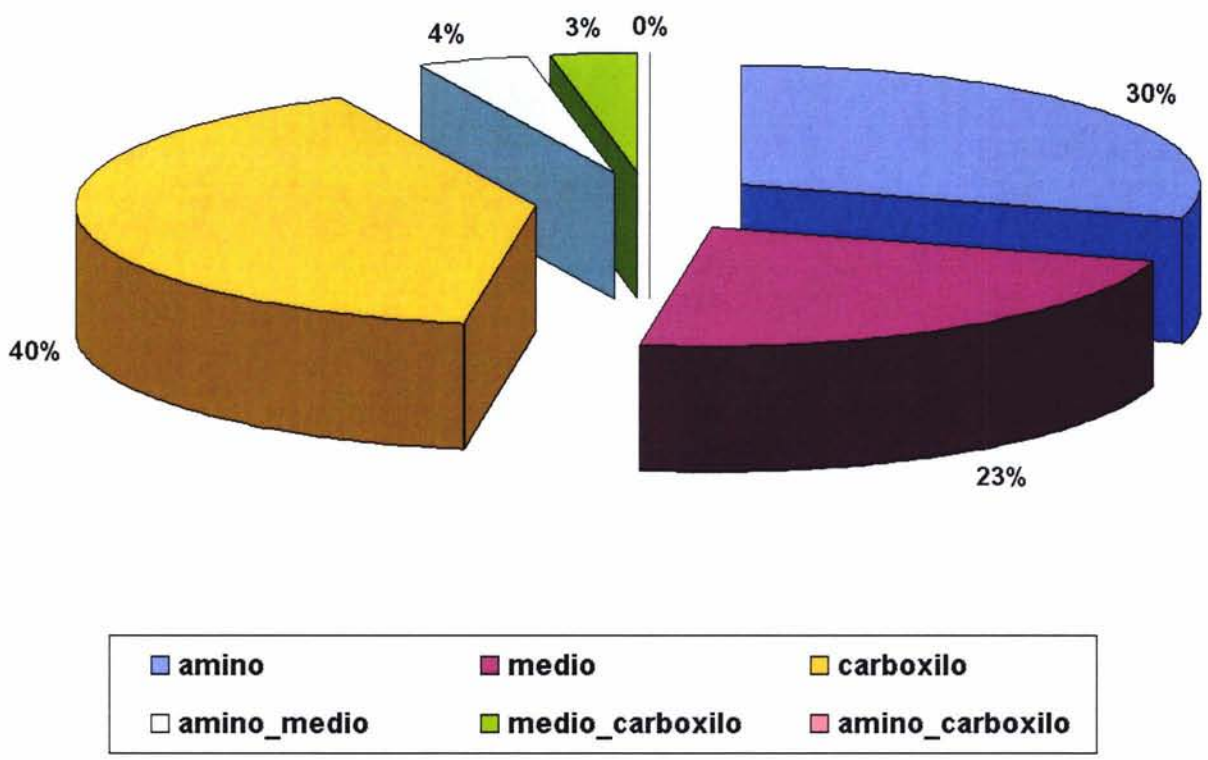
5.6 Distribución de las siete categorías funcionales de virus asignadas a las secuencias que contienen, al menos, una LCS, según el reporte previo del NCBI.


1)
No. genomas
70
60
50
40
30
20
10
0
6 (3)
14 (9)
10 (10)
26 (23)
7 (7)
61 (42)
1 (1)
5 (2)
6 (0)
10 (10)
2)
No. secuencias
480
440
400
360
320
280
240
200
160
120
80
40
0
12%
19%
36%
18%
32%
22%
14%
4%
0%
34%
Arenaviridae
Bunyaviridae
Bornaviridae
Rhabdoviridae
Filoviridae
Paramyxovirinae
Ophiovirus
Tenuivirus
Orthomyxoviridae
Pneumovirinae
Grupos virales con tipo de genoma ssRNA negativo

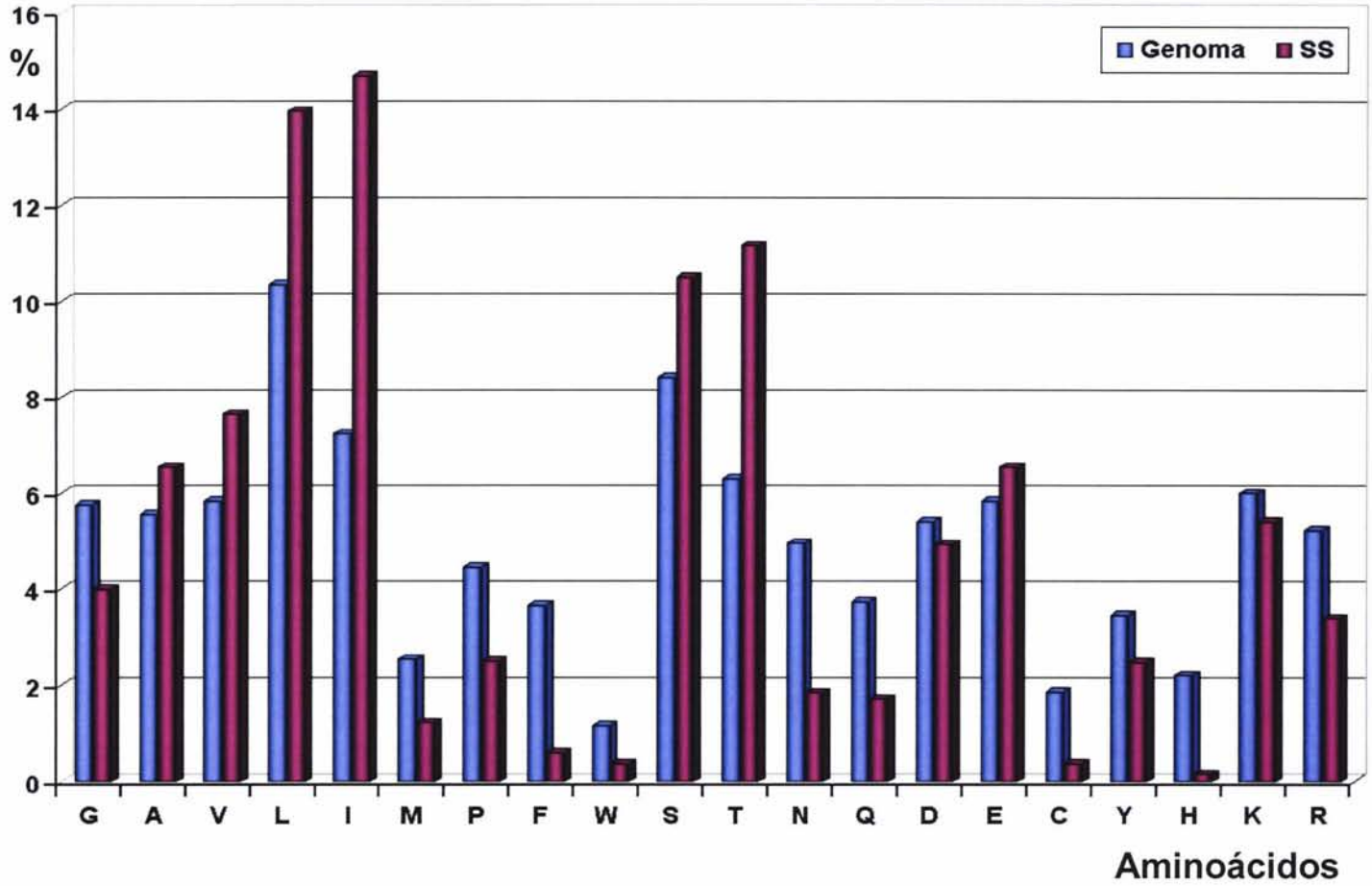
3)


14
12
10
8
6
4
2
0
No. secuencias
r = 0.569
r = 0.274
r = 0.3
0
1000
2000
3000
4000
5000
6000
7000
No. Aminoácidos x proteoma
ORF'S proteoma
Seg_LCS x proteoma
LCS x proteoma

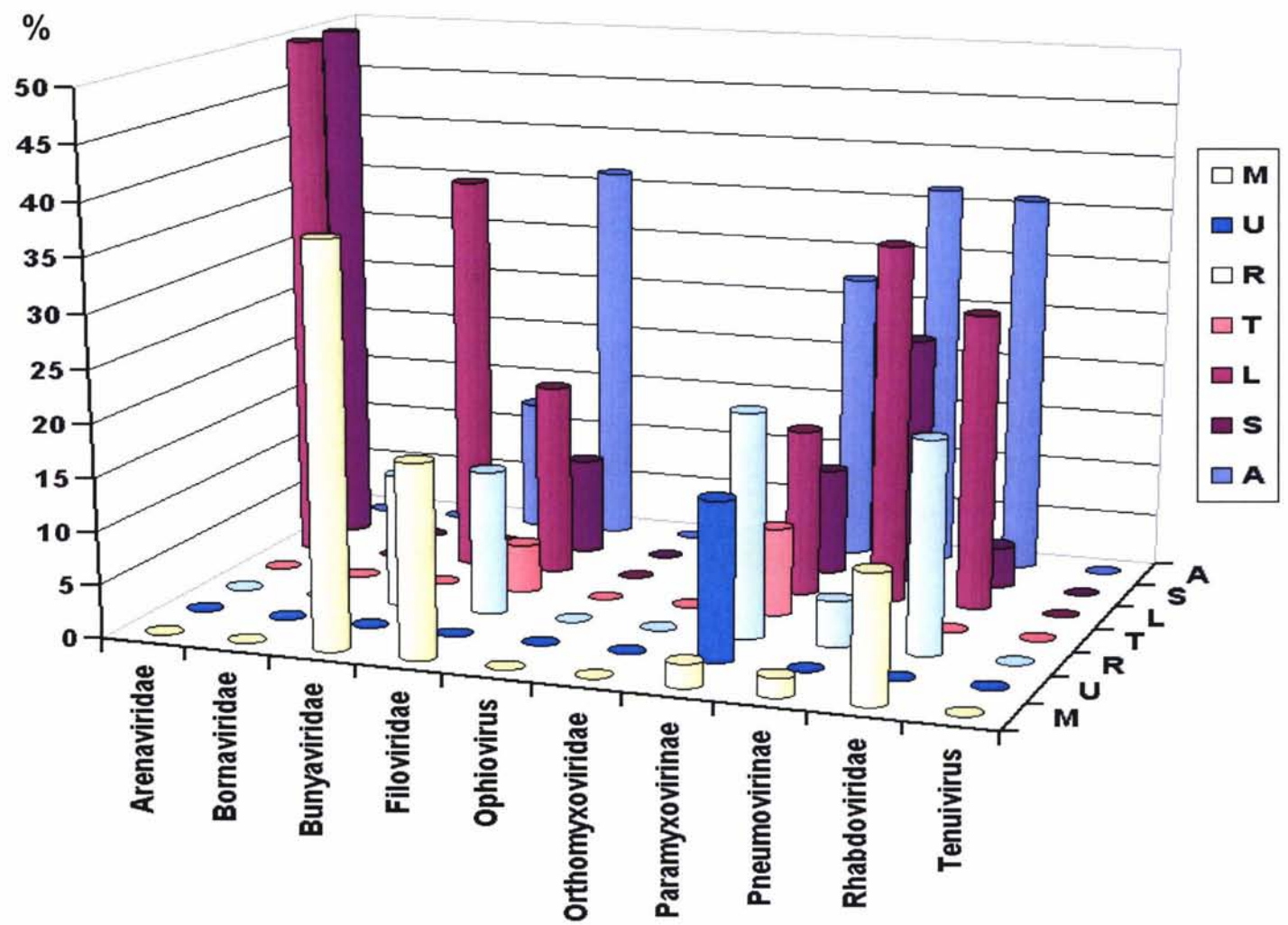
4)


4%
3%
0%
30%
40%
23%
amino
medio
carboxilo
amino_medio
medio_carboxilo
amino_carboxilo

5)


%
16
14
12
10
8
6
4
2
0
Genoma
SS
G
A
V
L
I
M
P
F
W
S
T
N
Q
D
E
C
Y
H
K
R
Aminoácidos

6)


%
50
45
40
35
30
25
20
15
10
5
0
M
U
R
T
L
S
A
Arenaviridae
Bornaviridae
Bunyaviridae
Filoviridae
Ophiovirus
Orthomyxoviridae
Paramyxovirinae
Pneumovirinae
Rhabdoviridae
Tenuivirus

Genomas de DNA o RNA que utilizan Reverso Transcriptasa.

(Retrovirus)

TAXONOMÍA DE LOS GENOMAS VIRALES (Retrovirus)			CARACTERÍSTICAS DE LOS GENOMAS			NO. ORF's			PRESENCIA DE SS		
Familia o grupo[¥]	Género[¥]	Hospedero[¥]	No. genomas*	Tamaño promedio del genoma (nts)**	Intervalo del tamaño (nts) x grupo**	Total ORF's*	Promedio x genoma**	Intervalo de ORF's x grupo**	No. genomas con SS[¢]	No. ORF's con SS[¢]	% ORF'S con SS
Caulimoviridae		PI	35	7890	7161 – 8303	180	5	2 _ 9	23	31	17
Hepadnaviridae		V	114	3183	2996 _ 3323	530	4	3 _ 7	114	420	79
Retroviridae	Alpharetrovirus	V	20	5901	1928 – 9625	43	2	1 _ 5	13	15	34
	Betaretrovirus	V	17	8696	7448 – 11791	76	4	3 _ 5	16	44	57
	Deltaretrovirus	V	25	8844	7933 _ 9160	116	4	3 _ 7	23	61	52
	Epsilonretrovirus	V	5	12445	10688 – 13125	30	6	5 _ 7	5	16	53
	Gammaretrovirus	V	41	7685	3811 _ 8918	110	2	1 _ 5	39	81	73
	Spumavirus	V	9	11352	5340 – 13246	44	4	2 _ 7	9	21	47
	Lentivirus	V	101	9482	7732 – 11443	777	7	2 _ 11	96	267	34
Inclasificados		V	2	4309	4302 – 4316	2	1	1 _ 1	1	1	50
TOTALES			369	79787		1908	39		339	957	50.2

Tabla VI. Información obtenida del análisis de los proteomas completos y las LCS para el tipo de genoma DNA o RNA que utilizan Reverso Transcriptasa (Retrovirus).

Gráfica 6:

6.1 Distribución de los genomas colectados por grupo taxonómico en comparación con los que presentaron al menos una secuencia con LCS.

6.2 Distribución de las LCS por grupo taxonómico con respecto al contenido total de secuencias de cada grupo viral.

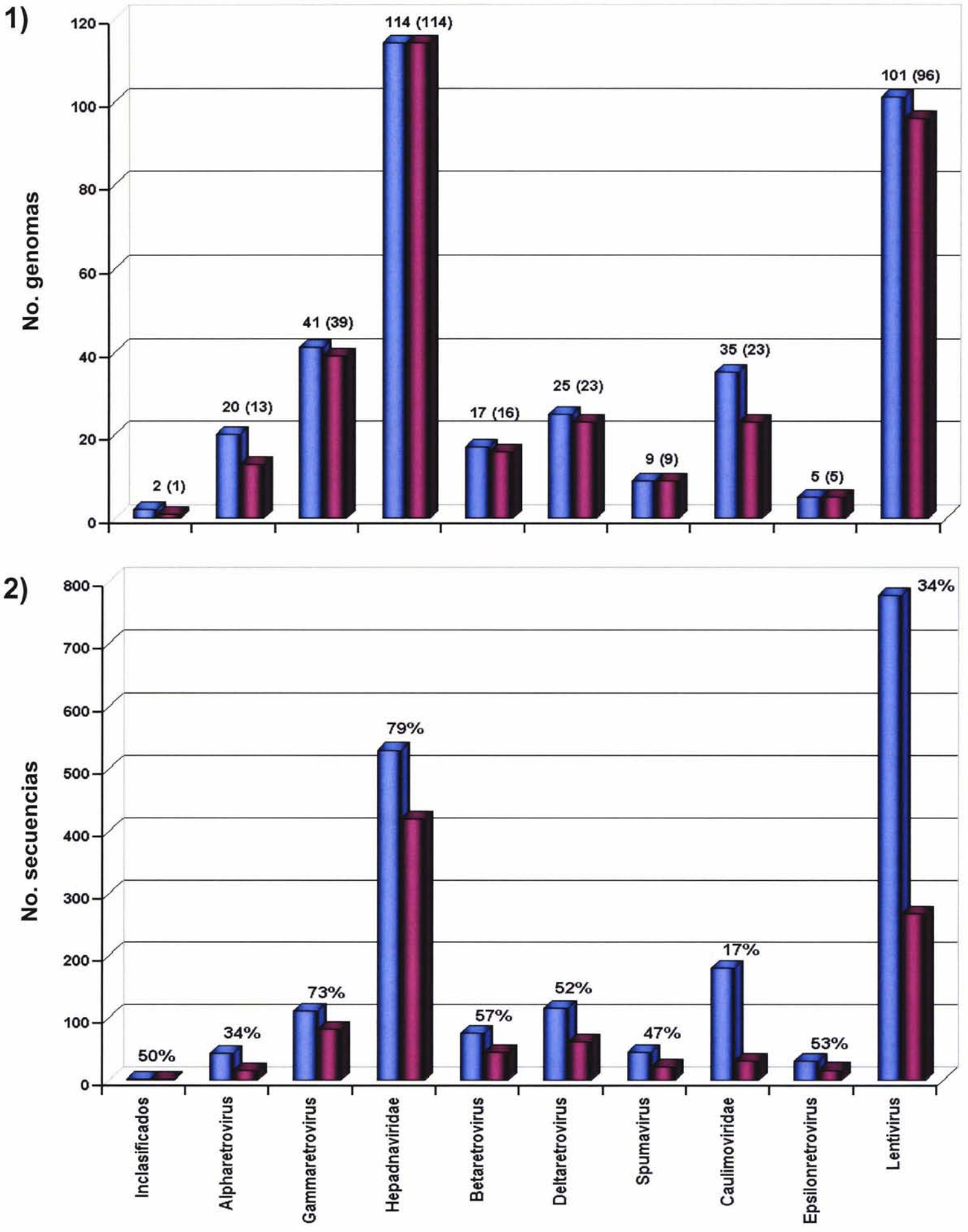
Las barras en azul, verde y amarillo son los datos colectados para virus con hospederos eucariontes, eubacterias y arqueas, respectivamente; en rojo, aquellos que presentan el fenómeno de LCS.

6.3 Correlación entre la presencia de las LCS y el tamaño del proteoma dado en número de aminoácidos y secuencias que los conforman.

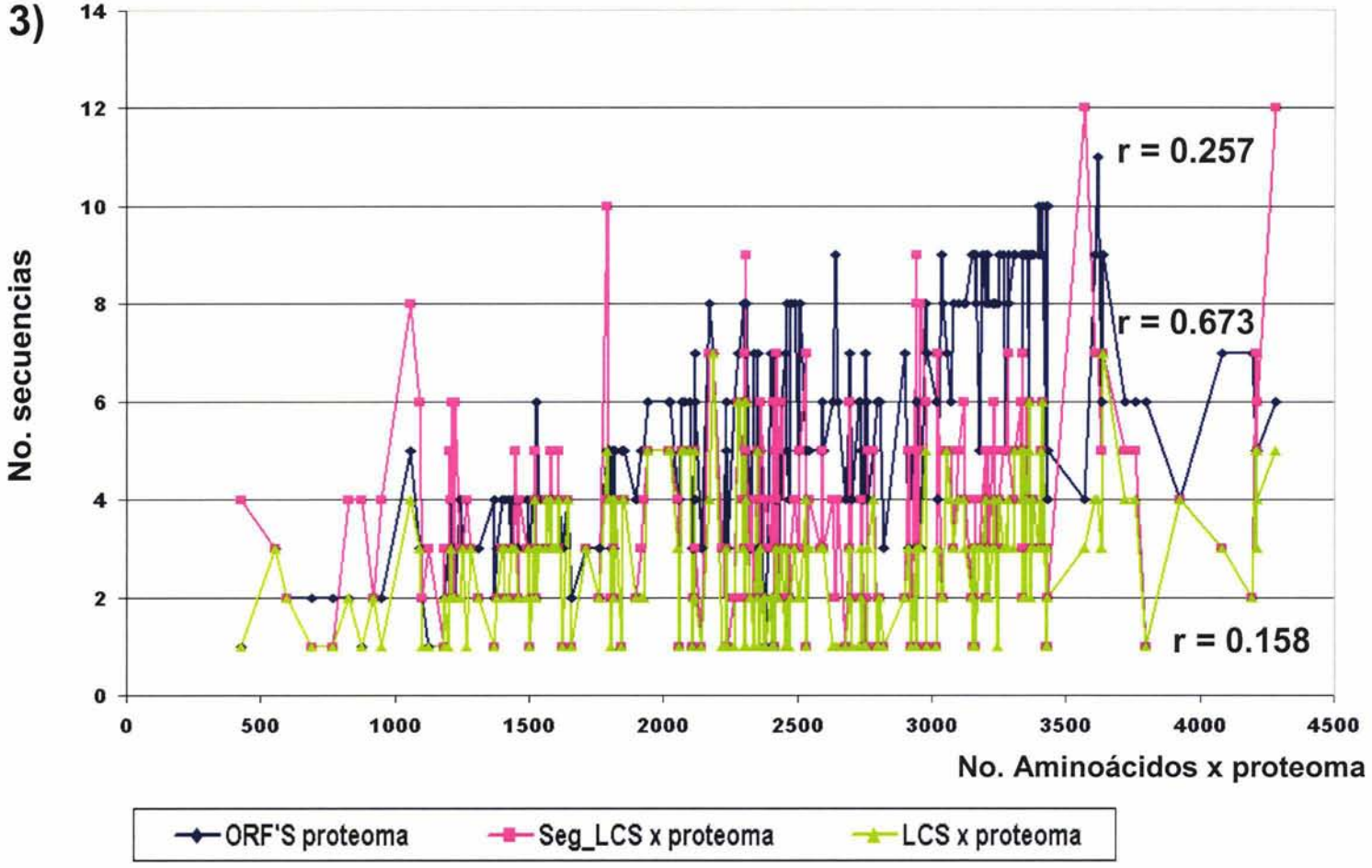
6.4 Análisis de la posición relativa de las LCS dentro de las secuencias que las contienen.

6.5 Análisis de la composición de los 20 aminoácidos tanto para el genoma completo como para las LCS que se encuentran dentro de ellos.

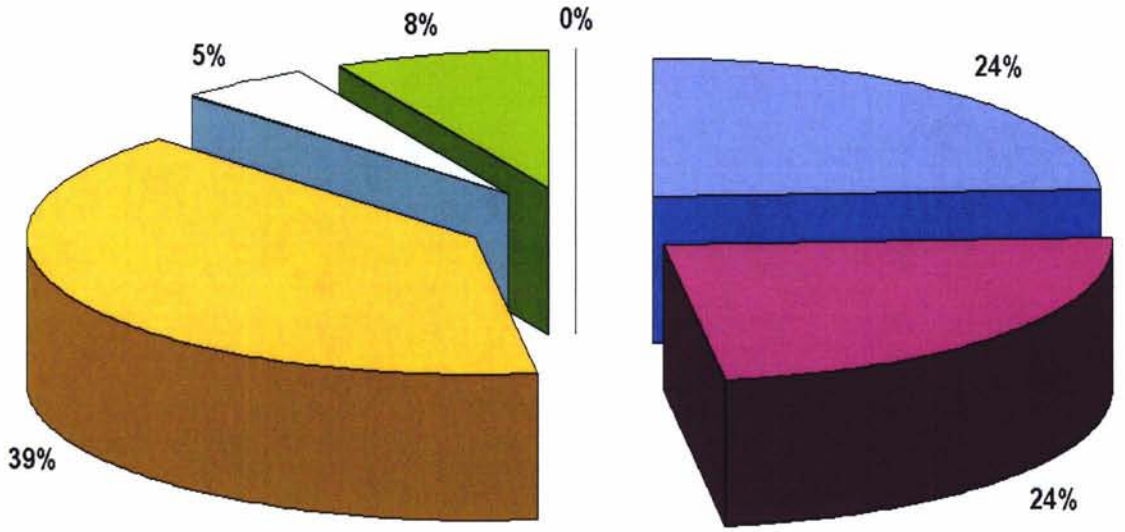
6.6 Distribución de las siete categorías funcionales de virus asignadas a las secuencias que contienen, al menos, una LCS, según el reporte previo del NCBI.


1)
No. genomas
120
100
80
60
40
20
0
2 (1)
20 (13)
41 (39)
114 (114)
17 (16)
25 (23)
9 (9)
35 (23)
5 (5)
101 (96)
2)
No. secuencias
800
700
600
500
400
300
200
100
0
50%
34%
73%
79%
57%
52%
47%
17%
53%
34%
Inclasificados
Alpharetrovirus
Gammaretrovirus
Hepadnaviridae
Betaretrovirus
Deltaretrovirus
Spumavirus
Caulimoviridae
Epsilonretrovirus
Lentivirus
Grupos virales con tipo de genoma Retrovirus

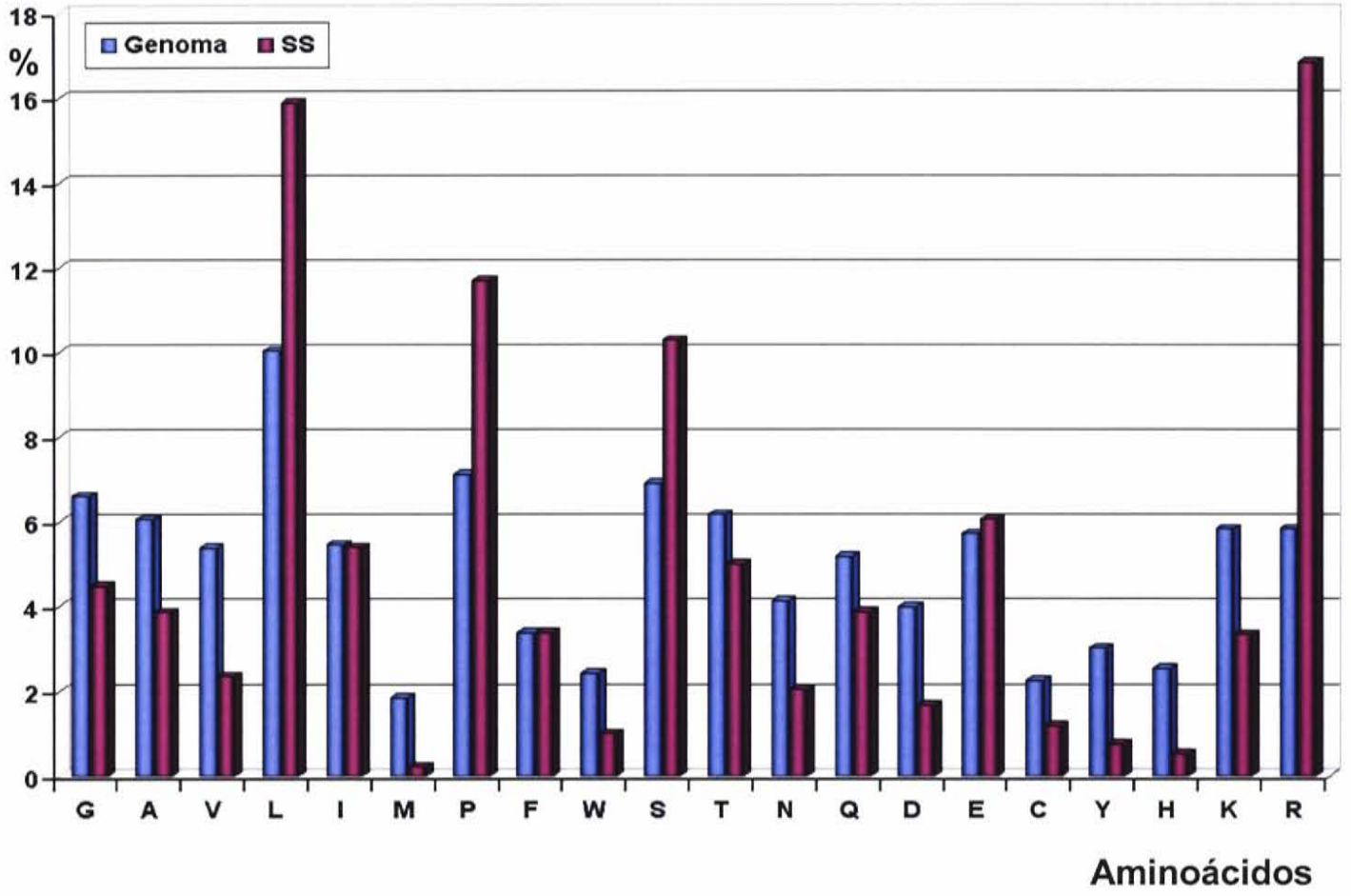
3)


No. secuencias
No. Aminoácidos x proteoma
r = 0.257
r = 0.673
r = 0.158
ORF'S proteoma
Seg_LCS x proteoma
LCS x proteoma

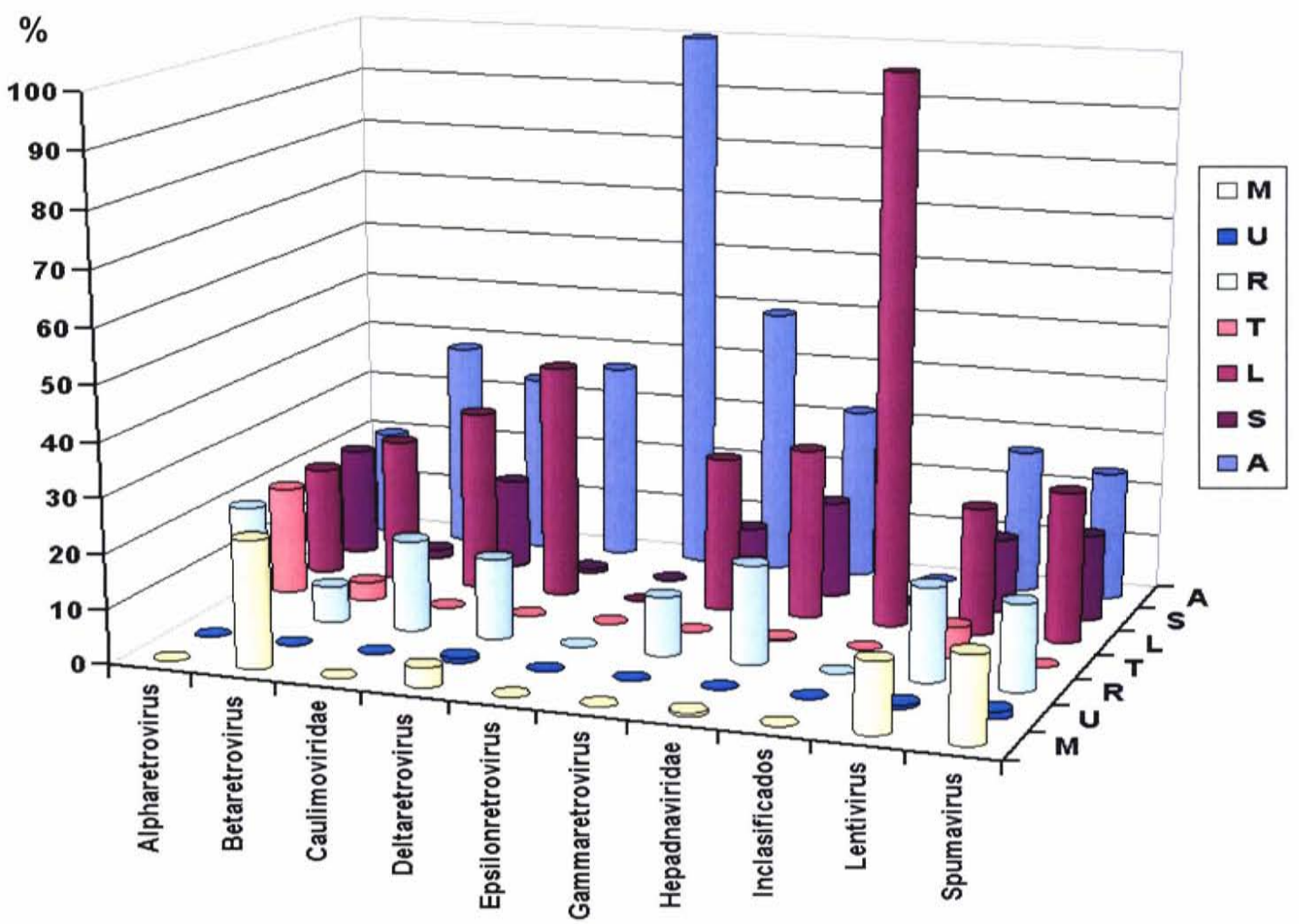
4)


8%
0%
5%
24%
39%
24%
amino
medio
carboxilo
amino_medio
medio_carboxilo
amino_carboxilo

5)


Genoma
SS
%
18
16
14
12
10
8
6
4
2
0
G A V L I M P F W S T N Q D E C Y H K R
Aminoácidos

6)


%
100
90
80
70
60
50
40
30
20
10
0
M
U
R
T
L
S
A
Alpharetrovirus
Betaretrovirus
Caulimoviridae
Deltaretrovirus
Epsilonretrovirus
Gammaretrovirus
Hepadnaviridae
Inclasificados
Lentivirus
Spumavirus

APÉNDICE I

Algoritmo SEG [Segment Sequence(s) by Local Complexity]

SINOPSIS: seq sequence [w] [K(1)] [K(2)] [-x] [opciones]

DESCRIPCIÓN.

Este algoritmo divide secuencias en segmentos contrastantes de baja y alta complejidad. Los segmentos definidos como de baja complejidad representan a las "**secuencias simples**" o "**regiones sesgadas composicionalmente**".

Los segmentos de baja complejidad optimizados localmente son producidos a niveles definidos de astringencia, basados sobre definiciones formales de complejidad composicional local (Wootton y Federhen, 1993). La longitud de los segmentos y el número de los mismos por secuencia están determinados automáticamente por el algoritmo.

La entrada es un archivo con una secuencia en formato FASTA, o un archivo con un banco de secuencias también en formato FASTA. SEG está calibrado para aminoácidos, pero hay opciones que proporcionan la búsqueda en nucleótidos.

La astringencia de la búsqueda para segmentos de baja complejidad es determinada por tres parámetros definidos por el usuario:

- **Longitud de la ventana del disparador [W].** Se refiere a la longitud de la ventana inicial de búsqueda y debe representar un entero más grande que cero. El valor por estándar es 12.
- **Complejidad del disparador [K1].** La complejidad máxima de una ventana inicial en unidades de bits. K1 debe ser igual o más grande que cero. El valor máximo es 4.322 (log [base2] 20) para secuencias de aminoácidos. El valor estándar es 2.2.
- **Complejidad de la extensión [K2].** La complejidad máxima de la extensión de una ventana en unidades de bits. Solamente valores más grandes que k1 son efectivos en extender ventanas adicionadas. El intervalo posible de valores es similar a k1. El valor estándar es 2.5.

ALGORITMO.

Este algoritmo tiene dos estados:

1) Identificación de segmentos crudos que se aproximen a la baja complejidad. Donde la astringencia y resolución de la búsqueda para segmentos de baja complejidad es determinada por los parámetros W, K(1) y K(2). Todas las ventanas disparadoras son definidas, incluyendo ventanas sobrelapantes, de longitud W y complejidad menor o igual a K(1).

"Complejidad" aquí es definida por la ecuación (3) de Wootton y Federhen (1993). Cada ventana disparadora es entonces extendida dentro de un contig en ambas direcciones por fusión con la extensión de las ventanas, los cuales son ventanas sobrelapadas de longitud W y complejidad menor o igual a K(2). Entonces cada contig es un segmento crudo de baja complejidad.

2) Cada segmento crudo es reducido a un simple segmento óptimo de baja complejidad, el cual, puede ser el segmento entero crudo, pero usualmente es una subsecuencia. Esta subsecuencia óptima tiene el valor más bajo de la probabilidad P (0) en la ecuación (5) de Wootton y Federhen (1993).

OPCIONES.

Las siguientes opciones pueden ser colocadas en cualquier orden en la línea de comando después de los parámetros W, K1 y k2:

Opción	Descripción
-a	Salida para los segmentos de alta y baja complejidad en un archivo con formato FASTA, como un sistema de entradas separadas con líneas principales.
-c	**[caracteres por línea]** Número de caracteres de secuencia por línea de salida. El valor estándar es 60. Otros caracteres, tales como el número de residuos, son adicionales.
-h	Salida solamente a los segmentos de alta complejidad en un archivo con formato FASTA, como un sistema de entradas separadas con líneas principales.
-l	Salida solamente a los segmentos de baja complejidad en un archivo con formato FASTA, como un sistema de entradas separadas con líneas principales.
-m	**[longitud]** Longitud mínima en residuos para un segmento de alta complejidad. El valor estándar es 0. Segmentos cortos son fusionados con segmentos de baja complejidad adyacentes.
-o	Muestra todos los sobrelapamientos, segmentos de baja complejidad disparados independientemente. Éstos son fusionados por estandarización.
-q	Produce un formato de salida con la secuencia en bloques numerados con marcas que ayudan a contar los residuos. Los segmentos de baja y alta complejidad están en minúsculas y mayúsculas, respectivamente.
-t	**[longitud]** Parámetro de la "longitud máxima del ajuste". El valor estándar es 100. Éste controla el espacio de la búsqueda (y el tiempo de la búsqueda) durante la optimización de segmentos crudos (véase el ALGORITMO arriba). De forma estándar, los subsequences 100 o más residuos más cortos que el segmento crudo son omitidos de la búsqueda. Este parámetro se puede aumentar para dar una búsqueda más extensa si los segmentos crudos son más largos de 100 residuos.
-x	Opción que enmascara a las secuencias de aminoácidos. Cada secuencia de entrada es representada por una sola secuencia de la salida en formato FASTA con las regiones de la baja complejidad substituidas por cadenas de caracteres con "x".

APÉNDICE II

Origen y Evolución de Virus.

Un pequeño desglose de las propuestas más argumentadas sobre el origen y evolución de los virus se describe a continuación.

1. **Los virus son parásitos intracelulares degenerados.** Ya que pudieron haber evolucionado por simplificación o degeneración de organismos unicelulares (Campbell, 2001). Ésta tiene muy poco apoyo actualmente, básicamente por la ausencia de formas intermedias y las extintas del proceso degenerativo, además, ni aun los virus conocidos más complejos y elaborados se asemejan en algún aspecto fenotípico o genotípico a la célula más sencilla.

2. **Los virus son relictos de vida pre-celular.** Desde el inicio, la hipótesis de la vida pre-celular presenta el dilema de que los virus requieren a los hospederos celulares y, por lo cual, no pudieron haber precedido a ellos. Una alternativa recurrente es que muchos de los virus de RNA se originaron en el mundo del RNA. Esta hipótesis aplica más bien al origen de los virus de RNA, es que éstos descienden de formas primitivas precelulares, y más directamente del "mundo del RNA", en el cual las moléculas de RNA eran capaces de catalizar todas las reacciones necesarias para sobrevivir y replicarse (Campbell, 2001). Se ha propuesto que las estructuras parecidas al RNA de transferencia y que se encuentran en el extremo 3' de algunos virus de RNA son fósiles moleculares que se han mantenido para alguna función determinada (Maizels y Weiner, 1993). De esta manera, los virus de RNA representan probablemente formas modificadas descendientes de un mundo de RNA prebiótico que llegaron a parasitar a las primeras células.

3. **Los virus son genes celulares que escaparon.** Muchos virólogos concuerdan en que una gran cantidad de virus de DNA son genes celulares escapados, mas bien se figuran como bloques de genes (posiblemente de diversos procesos celulares) que comprenden diversos módulos funcionales que se fusionaron unos con otros (llamados **morones**), dando al virus típico una ancestría quimérica (Campbell, 2001; Hendrix *et al*, 2000). En este sentido, las similitudes funcionales entre los virus de DNA y elementos genéticos celulares como los plásmidos y los transposones han dado lugar a una tercera hipótesis que propone que los virus se desarrollaron a partir de este tipo de elementos genéticos (Li, 1997). Este último punto de vista es actualmente uno de los más atractivos y con él se pretende explicar en forma más objetiva el origen de los virus.

APÉNDICE III.

Proteomas Virales de Cadena doble de DNA (dsDNA)

GRUPO VIRAL	VIRUS	CEPA	SEGM	SECs	NO. ACCESO DEL NCBI
Adenoviridae	Bovine adenovirus B	WBR-1	1	28	AF030154
	Bovine adenovirus D	THT/62	1	30	AF036092
	Canine adenovirus type 1	RI261	1	30	Y07760
		CLL	1	28	U55001
	Canine adenovirus type 2	Toronto A26/61	1	29	U77082
	Duck adenovirus 1	127	1	29	Y09566
	Fowl adenovirus A	Phelps (ATCC VR-432)	1	39	U46933
	Fowl adenovirus D	A-2A	1	29	AF083975
	Frog adenovirus A	CCL-50	1	23	AF224336
	Human adenovirus A		1	29	X73487
	Human adenovirus B	Ad11p Slobitski	1	38	AY163756
	Human adenovirus C		1	33	J01917 J01918 J01919 J01920 J01921 J01922 J01923 J01924 J01925 J01926 J01927 J01928 J01929 J01930 J01931 J01932 J01933 J01934 J01935 J01936 J01937 J01938 J01939 J01940 J01941 J01942 J01943 J01944 J01945 J01946 J01947 J01948 J01949 J01950 J01951 J01952 J01953 J01954 J01955 J01956 J01957 K00086 K00394 K00395 K02367 M13004 V00007 V00008 V00009 V00010 V00011 V00012 V00013 V00014 V00015 V00016 V00017 V00018 V00019 V00020 V00023 V00024
	Human adenovirus D		1	30	AF108105
	Human adenovirus E	Pan 9	1	32	AF394196
	Human adenovirus F	Dugan	1	37	L19443
	Murine adenovirus A	strain FL	1	26	M22245 J03353
	Ovine adenovirus 7	OAV287	1	30	U40839 U18755 U31557 U40837 U40838
	Ovine adenovirus A		1	21	"AF252654 AF035571 AF061853 L24204 S75673 X82686 VERSION AF252854.1 GI:9802268"
	Porcine adenovirus A	6618	1	11	AF083132 L43077 U10433 L43363
		6618	1	16	AB026117
		6618	1	16	AJ237815
	Porcine adenovirus C	5	1	30	AF289262
	Turkey adenovirus 3		1	24	AF074946
			23	**638**	
Asfarviridae	African swine fever virus	BA71V	**1**	**151**	U18466
Baculoviridae	Adoxophyes honmai nucleopolyhedrovirus	ADN001	1	125	AF006270
	Adoxophyes orana granulovirus		1	119	AF547984
	Autographa californica nucleopolyhedrovirus	C6	1	156	L22858
	Bombyx mori nuclear polyhedrosis virus	T3	1	143	L33180
	Choristoneura fumiferana defective nucleopolyhedrovirus		1	149	AY327402
	Choristoneura fumiferana MNPV		1	145	AF512031 AF512031 AF177329 S78506 S81690 U10441 U18677 U26876 U26734 U53854 U57401 U59008 U70432 U72240 X65395 S46001
	Cryptophlebia leucotreta granulovirus	CV3	1	128	AY229987 AY096241 AY096242 X77048 X79569
	Culex nigripalpus baculovirus	Florida1997*	1	109	AF403738
	Cydia pomonella granulovirus	Mexican 1*	1	143	U53466 AB010886 L05494 U44847 Y09478
	Epiphyas postvittana nucleopolyhedrovirus		1	136	AY043265
	Helicoverpa armigera nuclear polyhedrosis virus	C1	1	134	AF303045
	Helicoverpa zea single nucleocapsid nucleopolyhedrovirus		1	139	AF334030
	Helicoverpa armigera nucleopolyhedrovirus G4	HaSNPV-G4	1	136	AF271059
	Lymantria dispar nucleopolyhedrovirus		1	164	AF081810
	Mamestra configurata nucleopolyhedrovirus A	90/2	1	169	U59461 AF467808
	Mamestra configurata nucleopolyhedrovirus B		1	167	AY126275
	Orgyia pseudotsugata single capsid nuclear polyhedrosis virus		1	152	U75930
	Phthorimaea operculella granulovirus		1	130	AF499596
	Plutella xylostella granulovirus	K1	1	120	AF270937 AB029255 AB030018 AB030019 AB030174
	Rachiplusia ou multiple nucleopolyhedrovirus		1	149	AY145471
	Spodoptera exigua nucleopolyhedrovirus		1	139	AF169823 AF021837 AF054872 AF169824 S48199 X69615 X97577 X97578
	Spodoptera litura nucleopolyhedrovirus	G2	1	141	AF325155
	Xestia c-nigrum granulovirus		1	181	AF162221
			23	**3274**	
Herpesviridae Gammaherpesvirinae	Alcelaphine herpesvirus 1	C500	1	71	AF005370
	Ateline herpesvirus 3	73	1	73	AF083424
	Bovine herpesvirus 4		1	79	AF318573 AF271211
	Callitrichine herpesvirus 3	CJ0149	1	72	AF319782
	Equine herpesvirus 2	86/67	1	79	U20824 V01555 J02070 K01729 K01730 V01554 X00498 X00499
	Human herpesvirus 4	B95-8	1	94	X00784
		B95-8	1	92	AJ507799
	Cercopithicine herpesvirus 15	LCL8664	1	80	AY037858 AF148640 AF148641 AF159308 AF227125 AF159309 AF159310 AF227123 AF227124 U85993 U93160 U93909
	Human herpesvirus 8		1	82	U75698
			1	86	U93872
	Macaca mulatta rhadinovirus	17577	1	89	AF083501
		26-95	1	95	AF210726
	Murid herpesvirus 4	WUMS	1	81	U97553
		g2.4	1	73	AF105037
	Saimiriine herpesvirus 2		1	76	X64346
			15	**1222**	
Herpesviridae Alphaherpesvirinae	Bovine herpesvirus 1		1	73	AJ004801
	Cercopithecine herpesvirus 1	E2490	1	75	AF533758
	Cercopithecine herpesvirus 7		1	72	AF275348
	Equine herpesvirus 1	Ab4p	1	85	M86664
	Equine herpesvirus 4	NS80567	1	79	AF030027
	Gallid herpesvirus 2	GA	1	97	AF147806
		Md5	1	105	AF243438
	Gallid herpesvirus 3	HPRS24	1	100	AB049735
	Human herpesvirus 1	17	1	77	X14112 D00317 D00374 S40593
	Human herpesvirus 2	HG52	1	77	Z86099
	Human herpesvirus 3		1	72	X04370 M14891 M16612
	Meleagrid herpesvirus 1	FC126	1	101	AF291866
			12	**1013**	
Herpesviridae Betaherpesvirinae	Chimpanzee cytomegalovirus		1	165	AF480864
	Human herpesvirus 5	AD169	1	204	X17403
	Human herpesvirus 6	U1102	1	123	X83413
		HST	1	114	AB021506
	Human herpesvirus 6B	Z29	1	104	AF157706 L13152 L14772 L18947
	Human herpesvirus 7	JI	1	109	U43400
		RK	1	86	AF037218
	Rat cytomegalovirus	Maastricht	1	167	AF232689 AF046125 U50550 AF077758 U51788 AF133039 U57441 U57442 U87867 AF136519
	Tupaia herpesvirus	2	1	158	AF281817
			9	**1230**	
Ictalurid-Herpes-like	Ictalurid herpesvirus 1	auburn 1	**1**	**92**	M75136
Iridoviridae	Infectious spleen and kidney necrosis virus		1	125	AF371960
	Invertebrate iridescent virus 6		1	468	AF303741 AF003534 M99365 AF059606 M81387 M23627 AF083915 M23624 M23625 M81388 L22299 VERSION AF303741.1 GI:15042158
	Lymphocystis disease virus 1		1	110	L63545 J04360 L16683 M18368 M31137 M32131 M61115 VERSION L63545.1 GI:2276414
	Rana tigrina ranavirus		1	23	AF389451
			4	**726**	

Papillomaviridae	Bovine papillomavirus	307	1	8	X02346 J02044 M24622 X00473
	Bovine papillomavirus type 2		1	10	M20219 M19551
	Bovine papillomavirus type 3		1	8	AF486184
	Bovine papillomavirus type 4		1	11	X05817 D00146 X59063
	Bovine papillomavirus type 5		1	6	AF457465
	Canine oral papillomavirus	Y62	1	7	D55633
	Canine oral papillomavirus		1	7	L22695
	Common chimpanzee papillomavirus 1		1	8	AF020905
	Cottontail rabbit papillomavirus	Shope	1	10	K02708
	Cottontail rabbit papillomavirus	subtype b	1	10	AJ243287
	Cottontail rabbit papillomavirus	a4	1	10	AJ404003
	Deer papillomavirus		1	9	M11910
	Equinus papillomavirus		1	6	AF394740
	Equus caballus papillomavirus type 1		1	7	AF498323
	European elk papillomavirus		1	15	M15953
	Felis domesticus papillomavirus type 1		1	7	AF480454
	Fringilla coelebs papillomavirus		1	6	AY057109
	Hamster papovavirus		1	6	X02449
	Hamster papovavirus		1	6	M26281
	Human papillomavirus candHPV85		1	8	AF131950
	Human papillomavirus RTRX7	RTRX7	1	7	U85660
	Human papillomavirus type 10		1	7	X74465
	Human papillomavirus type 11		1	9	M14119
	Human papillomavirus type 12		1	6	X74466
	Human papillomavirus type 13		1	8	X62843 S43933
	Human papillomavirus type 14D		1	4	X74467
	Human papillomavirus type 15		1	6	X74468
	Human papillomavirus type 16		1	8	K02718
	Human papillomavirus type 16	18W12E	1	8	AF125673
	Human papillomavirus type 16		1	8	AF472508
	Human papillomavirus type 16		1	8	AF472509
	Human papillomavirus type 17		1	6	X74469
	Human papillomavirus type 18		1	8	X05015
	Human papillomavirus type 19		1	6	X74470
	Human papillomavirus type 1a		1	7	V01116 X03321
	Human papillomavirus type 20		1	7	U31778
	Human papillomavirus type 21		1	7	U31779
	Human papillomavirus type 22		1	7	U31780
	Human papillomavirus type 23		1	7	U31781
	Human papillomavirus type 24		1	7	U31782
	Human papillomavirus type 25		1	6	X74471
	Human papillomavirus type 26		1	6	X74472
	Human papillomavirus type 27		1	7	X74473
	Human papillomavirus type 28		1	7	U31783
	Human papillomavirus type 29		1	7	U31784
	Human papillomavirus type 2a		1	7	X55964
	Human papillomavirus type 3		1	7	X74462
	Human papillomavirus type 30		1	6	X74474
	Human papillomavirus type 31		1	8	J04353
	Human papillomavirus type 32		1	6	X74475
	Human papillomavirus type 33		1	8	M12732
	Human papillomavirus type 34		1	6	X74476
	Human papillomavirus type 35		1	8	M74117
	Human papillomavirus type 35		1	6	X74477
	Human papillomavirus type 36		1	7	U31785
	Human papillomavirus type 37		1	7	U31786
	Human papillomavirus type 38		1	7	U31787
	Human papillomavirus type 39		1	8	M62849 M38185
	Human papillomavirus type 4		1	7	X70827
	Human papillomavirus type 40		1	6	X74478
	Human papillomavirus type 41		1	11	X56147
	Human papillomavirus type 42		1	8	M73236
	Human papillomavirus type 44		1	7	U31788
	Human papillomavirus type 45		1	6	X74479
	Human papillomavirus type 47		1	8	M32305
	Human papillomavirus type 48		1	7	U31789
	Human papillomavirus type 49		1	6	X74480
	Human papillomavirus type 5		1	8	M17463
	Human papillomavirus type 50		1	7	U31790
	Human papillomavirus type 51		1	7	M62877
	Human papillomavirus type 52		1	6	X74481
	Human papillomavirus type 53		1	7	X74482
	Human papillomavirus type 54		1	7	U37488
	Human papillomavirus type 54	AE9	1	7	AF436129
	Human papillomavirus type 55		1	7	U31791
	Human papillomavirus type 56		1	6	X74483
	Human papillomavirus type 57		1	7	X55965
	Human papillomavirus type 58	GN479	1	8	D90400
	Human papillomavirus type 59		1	8	X77858
	Human papillomavirus type 5b		1	9	D90252
	Human papillomavirus type 6	6vc	1	9	AF092932
	Human papillomavirus type 60		1	7	U31792
	Human papillomavirus type 61		1	7	U31793
	Human papillomavirus type 63		1	7	X70828
	Human papillomavirus type 65		1	7	X70829
	Human papillomavirus type 66		1	7	U31794
	Human papillomavirus type 67		1	8	D21208
	Human papillomavirus type 69		1	8	AB027020
	Human papillomavirus type 6a		1	8	L41216
	Human papillomavirus type 6b	HPV6b	1	9	X00203
	Human papillomavirus type 7		1	6	X74463
	Human papillomavirus type 70		1	8	U21941
	Human papillomavirus type 71		1	5	AB040456
	Human papillomavirus type 74	AE10	1	8	AF436130
	Human papillomavirus type 8		1	6	M12737
	Human papillomavirus type 82		1	8	AB027021
	Human papillomavirus type 82	IS39/AE2	1	8	AF293961
	Human papillomavirus type 83		1	7	AF151983
	Human papillomavirus type 84		1	7	AF293960
	Human papillomavirus type 86	candHPV86	1	7	AF349909
	Human papillomavirus type 87		1	7	AJ400626
	Human papillomavirus type 89		1	7	AF436128
	Human papillomavirus type 9		1	7	X74464
	Human papillomavirus type 90		1	7	AY057438
	Human papillomavirus type 91		1	7	AF419318
	Human papillomavirus type 92		1	7	AF531420
	Human papillomavirus type 93		1	7	AY382778
	Human papillomavirus type 96		1	7	AY382779
	Monkey B-lymphotropic papovavirus		1	5	M30540
	Multimammate rat papillomavirus		1	6	U01834
	Ovine papillomavirus 1		1	5	U83594
	Ovine papillomavirus 2		1	5	U83595
	Phocoena spinipinnis papillomavirus	PsPV1	1	8	AJ238373
	Psittacus erithacus papillomavirus		1	6	AF502599
	Psittacus erithacus timneh papillomavirus		1	6	AF420235
	Rabbit oral papillomavirus		1	9	AF227240
	Reindeer papillomavirus		1	9	AF443292
	Rhesus monkey papillomavirus		1	8	M60184 M37718
			118	**860**	
Phycodnaviridae	Ectocarpus siliculosus virus	EsV-1	1	240	AF204951 AF204952 AF210454 U95206 X76296 VERSION AF204951.2 GI:13177282
	Paramecium bursaria Chlorella virus 1	NC64A	1	698	U42580 U17055 U32570 VERSION U42580.3 GI:1612644
			2	**938**	
Polyomaviridae	African green monkey polyomavirus	K38	1	5	K02562
	BK polyomavirus	Dunlop	1	6	J02038 K00058 V01108 VERSION J02038.1 GI:333378
		MM	1	6	V01109 J02039
	Bovine polyomavirus		1	6	D13942 D00755 M74843
	Budgerigar fledgling polyomavirus		1	6	M20775
	Goose hemorrhagic polyomavirus		1	6	AY140894
	JC virus	Mad1	1	6	J02226 J02227
	Polyomavirus muris	Kilham	1	5	M65904
		K	1	5	M57473
		BG	1	6	AF442959
	Polyomavirus sp.	A2-A3	1	7	J02288 J02290 J02291 J02292 K00932 K00997 K01041 K01071 K01072 V01117 VERSION J02288.1 GI:332752
		LID	1	6	U27813
		PTA	1	6	U27812
	Simian virus 40		1	7	J02400 J02402 J02403 J02406 J02407 J02408 J02409 J02410 J04139 M24874 M24914 M28728 V01380 VERSION J02400.1 GI:965480
		777-5	1	6	AF332699
		777-1	1	6	AF332562
		MC-026846B	1	6	AF180737
		GM00637H	1	6	AF345344

	GM00637H	1	6	AF345345
	N128-1	1	6	AY120890
	H328-1	1	6	AF316141
	Rh911-1	1	6	AF316140
	776	1	6	AF316139
	Baylor-1	1	6	AF155359
	Baylor-2	1	6	AF155358
	CPC/MEN	1	6	AF156108
	VA45-54-1	1	6	AF156107
	VA45-54-2	1	6	AF156105
	K661	1	6	AF038616
		29	**173**	
Poxviridae Entomopoxvirinae	Amsacta moorei entomopoxvirus		1	294	AF250284
	Melanoplus sanguinipes entomopoxvirus	Tucson	1	267	AF063866
			2	**561**	
Poxviridae Chordopoxvirinae	Camelpox virus	M-96	1	211	AF438165
		CMS	1	265	AY009089
	Cowpox virus	Brighton Red	1	218	AF482758 L08906 J02066
	Ectromelia virus	Moscow	1	173	AF012825 U19584 U67964
	Fowlpox virus		1	261	AF198100
	Lumpy skin disease virus	Neethling 2490	1	156	AF325528
		Neethling Warmbaths LW	1	158	AF409137
		Neethling vaccine LW 1959	1	159	AF409138
	Molluscum contagiosum virus		1	163	U60315
	Monkeypox virus	Zaire-96-I-16	1	191	AF380138
	Myxoma virus	Lausanne	1	171	AF170726
	Rabbit fibroma virus	Kasza	1	165	AF170722
	Sheeppox virus	TU-V02127	1	148	AY077832
	Swinepox virus	17077-99	1	150	AF410153
	Vaccinia virus	Copenhagen	1	273	M35027
		Ankara	1	262	U94848
		Tian Tan	1	243	AF095689
	Variola virus	India-1967-3	1	197	X69198
		Garcia-1966	1	206	Y16780
		Bangladesh-1975	1	189	L22579
	Yaba-like disease virus		1	152	AJ293568
			21	**4111**	
Nimaviridae	Shrimp white spot syndrome virus		1	531	AF332093
		Taiwan	1	532	AF440570
			1	184	AF369029
			3	**1247**	
Corticoviridae	Alteromonas phage PM2		**1**	**23**	AF155037
Fuselloviridae	Sulfolobus virus 1		**1**	**32**	X07234 S61065
Lipothrixviridae	Sulfolobus islandicus filamentous virus		**1**	**72**	AF440571
Plasmaviridae	Acholeplasma phage L2		**1**	**14**	L13696
Rudiviridae	Sulfolobus virus SIRV-1	KVE M10H3	1	45	AJ414696
	Sulfolobus virus SIRV-2	HVE 10/4	1	54	AJ344259
			2	**99**	
Tectiviridae	Enterobacteria phage PRD1		**1**	**22**	M69077
Caudovirales Myoviridae	Bacteriophage 44RR2.8t		1	252	AY375531
	Bacteriophage KVP40		1	381	AY283928
	Bacteriophage L-413C		1	40	AY251033
	Bacteriophage RM 378		1	146	AX059140
	Bacteriophage Wphi	ATCC9637	1	44	AY135739
	Enterobacteria phage RB49		1	274	AY343333 AB062453 AB062595 AF223003 AF303100 AF410869 AF410870 AJ508254 AY051145 AY051146 AY051147 AY051148 AY051149 AY051150 AY051151 AY051152 AY051153 AY051154 AY051155 AY051156 AY051157 AY051158 AY051159 AY051160 AY051161 AY051162 AY051163 AY051164 AY051165 AY051166 AY051167 AY051168 AY051169 AY051170 AY051171 AY051172 AY051173 AY051174 AY051175 AY051176 AY051177 Z78088 Z78090 AY051179 AY051180 AY051181 AY051182 AY051183 AY051184 AY051185 AY051186 AY051187 AY051188 AY051189 AY051190 AY051191 AY051192 AY051193 AY051194 AY051195 AY051196 AY249135 AY266306 Z78068 Z78069 Z78070 Z78071 Z78072 Z78073 Z78074 Z78075 Z78076 Z78077 Z78078 Z78079 Z78080 Z78081 Z78082 Z78083 Z78084 Z78085 Z78086 Z78087 Z78091 Z78092 Z78093 AY051178 AY051197
	Enterobacteria phage RB69		1	273	AY303349
	Mycobacteriophage Bxz1 virion		1	225	AY129337
	Pseudomonas phage phiKZ		1	306	AF399011
	Bacteriophage K139	MO10-O139	1	44	AF125163
	Bacteriophage P4		1	14	X51522
	Bacteriophage phi CTX	phiCTX-c-PA158	1	47	AB008550
	Burkholderia cepacia phage Bcep781		1	61	AF543311
	Enterobacteria phage 186		1	46	U32222 X53318 X04449 U51471
	Enterobacteria phage Mu	G(+) form	1	55	AF083977
	Enterobacteria phage P2		1	43	AF063097 J02474 X02300 X02301 M56023 M13202 L29304 X87173 M64677 X99627 M27836 M12772 X05655 Z11483 X99628 M27131 M34756 M59752 X61229 U02597
	Enterobacteria phage T4		1	274	AF158101
	Haemophilus phage HP1	HP1c1	1	42	U24159 U06847 M28366 M12911 M22941 M12910 M15313
	Haemophilus phage HP2		1	37	AY027935
	Halovirus HF2		1	114	AF222060 L26038
	Vibrio harveyi bacteriophage VHML		1	57	AY133112
	Virus PhiCh1		1	98	AF440695 AF172444 AF231018
			22	**932**	
Caudovirales	Acyrthosiphon pisum bacteriophage APSE-1		1	54	AF157835
Podoviridae	Bacillus phage GA-1		1	35	X96987 AF148209 AJ133524 AJ133525 AJ244026 AJ294726
	Bacteriophage B103		1	17	X99260
	Bacteriophage phi-29		1	27	M11813 M13904 M13905
	Bacteriophage Mx8		1	86	AF396866
	Bacteriophage phiKMV		1	48	AJ505558
	Bacteriophage phiYeO3-12		1	59	AJ251805
	Bacteriophage T3	Luria	1	55	AJ318471
	Cyanophage P60		1	80	AF338467
	Enterobacteria phage epsilon15		1	49	AY150271
	Enterobacteria phage P22		1	57	AF217253
			1	72	BK000583
			1	72	AF527606
	Enterobacteria phage SP6		1	52	AY288927
	Enterobacteria phage T7		1	60	V01146 J02518 X00411
	Mycoplasma virus P1		1	11	AF246223 L33717 L33718
	Pseudomonas aeruginosa phage PaP3		1	69	AY078382
	Pseudomonas phage gh-1		1	42	AF493143
	Roseophage SIO1		1	34	AF189021
	Salmonella typhimurium bacteriophage ST64T		1	65	AY052766
	Salmonella typhimurium phage ST64B		1	56	AY055382
	Shigella flexneri bacteriophage V		1	53	U82619 U82620 AF339141
	Staphylococcus aureus phage phiP68		1	22	AF513033
	Staphylococcus phage 44AHJD		1	21	AF513032
	Streptococcus phage C1 virion		1	20	AY212251
	Streptococcus phage Cp-1		1	28	Z47794
	Vibriophage VpV262		1	67	AY095314
	Yersinia pestis phage phiA1122		1	50	AY247822
			28	**1371**	
Caudovirales Siphoviridae	Bacteriophage 933W	EDL933	1	80	AF125520
	Bacteriophage A118		1	72	AJ242593
	Bacteriophage bIL170		1	64	AF009630
	Bacteriophage bIL285		1	62	AF323668
	Bacteriophage bIL286		1	61	AF323669
	Bacteriophage bIL309		1	56	AF323670
	Bacteriophage bIL310		1	29	AF323671
	Bacteriophage bIL311		1	22	AF323672
	Bacteriophage bIL312		1	27	AF323673
	Bacteriophage HK620	2158	1	58	AF335538

	Bacteriophage HK97		1	61	AF069529
	Bacteriophage lambda		1	71	J02459 M17233 M24325 V00636 X00906
	Bacteriophage N15		1	60	AF064539
	Bacteriophage P27	2771/97	1	58	AJ296298
	Bacteriophage phi ETA		1	66	AP001553
	Bacteriophage phi-105		1	51	AB016282
	Bacteriophage phi-BT1		1	55	AJ550940
	Bacteriophage phi-C31	Norwich stock	1	55	AJ006589
	Bacteriophage phi3626		1	50	AY082070
	Bacteriophage phiE125		1	71	AF447491
	Bacteriophage phig1e		1	50	X96106
	Bacteriophage PY54		1	67	AJ564013
	Bacteriophage r1t		1	50	U38906
	Bacteriophage sk1		1	56	AF011378
	Bacteriophage SPBc2		1	185	AF020713
	Bacteriophage SPP1		1	106	X97918
	Bacteriophage TP901-1		1	56	AF304433
	Bacteriophage Tuc2009		1	56	AF109874 L26219 L31348 L31364 L31365 L31366
	Bacteriophage VT2-Sa	RIMD0509894	1	82	AP000363
	Enterobacteria phage HK022		1	57	AF069306
	Lactobacillus bacteriophage phi adh		1	63	AJ131519 X78410 Z97974
	Lactobacillus casei bacteriophage A2	393	1	61	AJ251789 AJ251790 X97563 Y09454 Y12813
	Lactococcus lactis bacteriophage ul36		1	58	AF349457
	Lactococcus phage BK5-T		1	63	AF176025 L44593 M34487
	Lactococcus phage BK5-T		1	64	AJ245615
	Lactococcus phage c2	bIL67	1	39	L48605
	Lactococcus phage c2		1	37	L33769
	Lactococcus phage P335		1	49	AF489521
	Listeria phage 2389		1	57	AJ312240
	Methanobacterium phage psiM2		1	32	AF065411
	Methanothermobacter wolfeii prophage psiM100		1	37	AF301375
	Mycobacteriophage Barnyard		1	109	AY129339
	Mycobacteriophage Bxb1		1	86	AF271693
	Mycobacteriophage Bxz2		1	86	AY129332
	Mycobacteriophage Che8		1	112	AY129330
	Mycobacteriophage Che9c		1	84	AY129333
	Mycobacteriophage Che9d		1	103	AY129336
	Mycobacteriophage CJW1		1	141	AY129331
	Mycobacteriophage Corndog		1	122	AY129335
	Mycobacteriophage D29		1	79	AF022214
	Mycobacteriophage Omega		1	237	AY129338
	Mycobacteriophage Rosebush		1	90	AY129334
	Mycobacteriophage TM4		1	89	AF068845
	Mycobacterium phage L5		1	85	Z18946
	Phage phi 4795	O84 H4	1	48	AJ487680
	Pseudomonas phage D3		1	94	AF165214 L22692 U32623 U47623 AF077308 AF147978 AF165213
	Staphylococcus aureus bacteriophage PVL	49775	1	62	AB009866
	Staphylococcus aureus phage phi 11	8325	1	53	AF424781
	Staphylococcus aureus phage phi 12	8325	1	49	AF424782
	Staphylococcus aureus phage phi 13	8325	1	49	AF424783
	Staphylococcus aureus prophage phiPV83	P83	1	65	AB044554
	Staphylococcus aureus temperate phage phiSLT		1	62	AB045978
	Streptococcus pneumoniae bacteriophage MM1		1	53	AJ302074
	Streptococcus thermophilus bacteriophage 7201		1	46	AF145054 AF001793 AF118440 U89245
	Streptococcus thermophilus bacteriophage DT1	DT1	1	47	AF085222
	Streptococcus thermophilus bacteriophage Sfi11		1	53	AF158600 AF057033
	Streptococcus thermophilus bacteriophage Sfi19		1	45	AF115102 AF032122
	Streptococcus thermophilus bacteriophage Sfi21		1	50	AF115103 AF032121
	Streptococcus thermophilus temperate bacteriophage O1205	CNRZ1205	1	57	U88974
	Stx1 converting bacteriophage	Stx1 phage	1	167	AP005153
	Stx2 converting bacteriophage I	Stx2 phage-I	1	185	AP004402
	Stx2 converting bacteriophage II	Stx2 phage-II	1	170	AP005154
	Temperate phage PhiNIH1.1	NIH1	1	55	AY050245
			73	5267	

Proteomas virales de Cadena Sencilla de DNA (ssDNA)

Geminiviridae	Abutilon mosaic virus	REGEL-1	2	4	X15983
		REGEL-2		2	X15984
		Hawaii-1	2	5	U51137
		Hawaii-2		2	U51138
	African cassava mosaic virus	West Kenyan-844-1	2	7	J02057
		West Kenyan-844-2		2	J02058
		ACMV/CM:S1-1	2	6	AF112352
		ACMV/CM:S1-2		2	AF112353
		Nigerian-1	2	5	X17095
		Nigerian-2		2	X17096
		Ogoroco-1	2	6	AJ427910
		Ogoroco-2		2	AJ427911
	Ageratum enation virus	SB30-3-leaf	1	6	AJ437618
	Ageratum yellow vein China virus	Hn2	1	6	AJ495813
	Ageratum yellow vein virus		1	6	AF307861
			1	6	AF314144
		Tw-PD	1	6	AF327902
		pHN419	1	7	X74516
	Althea rosea enation virus		1	6	AF014881
	Bean calico mosaic virus		2	5	AF110189
				2	AF110190
	Bean dwarf mosaic virus		2	4	M88179
				2	M88180
	Bean golden mosaic virus-Brazil	Typel-Brazil-1	2	4	M88686
		Typel-Brazil-2		2	M88687
	Bean golden mosaic virus-Puerto Rico		2	2	D00200
				6	D00201
		Type II-Guatemala-1	2	4	M91604
		Type II-Guatemala-2		2	M91605
		Type II-BGMV-DR-1	2	3	L01635
		Type II-BGMV-DR-2		2	L01636
			2	4	M10070
				2	M10080
	Bean yellow dwarf virus		1	4	Y11023
	Beet curly top virus		1	5	M24597 X04144
	Beet severe curly top virus		1	8	U02311
		Logan	1	7	AF379637
	Beet mild curly top virus	Worland	1	7	U56975
	Bhendi yellow vein mosaic virus		1	7	AF241479
	Cabbage leaf curl virus		2	5	U65529
				2	U65530
	Chayote mosaic virus	Nigeria	1	5	AJ223191
	Chilli leaf curl virus		1	6	AF336806
	Chino del tomate virus		1	3	AF101476
			1	2	AF101478
	Chloris striate mosaic virus		1	4	M20021
	Cotton leaf curl virus	CLCuV-G	1	6	AF363011 AF329671 AF337533 AF337534 AF339845 AF363010
		CLCuV/SD okra-1	1	6	AY036008
		CLCuV/SD okra-2	1	6	AY036006
		CLCuV/SD Sida	1	6	AY036007
		33af	1	6	AJ496461
		31cf	1	6	AJ496287
		Faisalabad x2f	1	6	AJ496286
		Pakistan	1	6	AJ132430
		Pakistan-clc311	1	6	AJ002459
		Pakistan-clc26	1	6	AJ002458
		Pakistan-clc802a	1	6	AJ002455
		Pakistan-clc804a	1	6	AJ002452
		Pakistan-clc806b	1	6	AJ002449
		Pakistan-clc72b	1	6	AJ002448
		Pakistan-clc62	1	6	AJ002447
			1	6	AF155064
	Croton yellow vein mosaic virus		1	7	AJ507777
	Cucurbit leaf crumple geminivirus	A	2	4	AF224760
		B		2	AF224761
	Cucurbit leaf curl virus		2	4	AF256200

				2	AF327559
	Dicliptera yellow mottle virus		2	4	AF139168
				2	AF170101
	Digitaria streak virus		1	4	M23022
	East African cassava mosaic virus	Ug1-1	2	6	AF422174
		Ug1-2		2	AF230375
		EACMV/Ug2/Svr-1	2	2	AF126806
		EACMV/Ug3/Svr2		7	AF126807
		WACMV/CM-1	2	6	AF112354
		WACMV/CM-2		2	AF112355
		Ivory Coast-1	2	2	AF259696
		Ivory Coast-2		2	AF259697
	Egyptian sugarcane streak virus	Giza	1	4	AF037752
		Aswan	1	4	AF039528
		Beni-Suef	1	4	AF039529
		Mansoura	1	4	AF039530
		Naga Hammady	1	4	AF239159
	Eupatorium yellow vein virus		1	6	AB007990
	Havana tomato virus	Quivican-1	2	5	Y14874
		Quivican-2		2	Y14875
	Hollyhock leaf crumple virus	HLCrV hollyhock	1	6	AY036009
	Honeysuckle yellow vein mosaic virus		1	6	AB020781
	Horseradish curly top virus		1	6	U49907
	Indian cassava mosaic virus		2	8	Z24758
				2	Z24759
		ICMV-Mah-1	2	6	AJ314739
		ICMV-Mah-2		2	AJ314740
	Ipomoea yellow vein virus		1		AJ132548
	Macroptilium mosaic virus		2	5	AY044133
				2	AY044134
	Macroptilium yellow mosaic virus	Florida-1	2	5	AY044135
		Florida-2		2	AY044136
	Macroptilium yellow mosaic virus	Cuba	1	4	AJ344452
	Maize streak virus	South African	1	5	Y00514
		MSV-D-Raw	1	4	AF329889
		MSV-E-Pat	1	4	AF329888
		MSV-B-Jam	1	4	AF329887
		MSV-B-Mom	1	4	AF329886
		MSV-A-MtKA	1	4	AF329885
		MSV-A-MalD	1	4	AF329884
		MSV-A-MatC	1	4	AF329883
		MSV-A-MatB	1	4	AF329882
		MSV-A-MatA	1	4	AF329881
		MSV-A-Sag	1	4	AF329880
		MSV-A-Gat	1	4	AF329879
		MSV-A-Ama	1	4	AF329878
		MSV-Km	1	3	AF395891
		Komatipoort	1	4	AF003952
		Setaria	1	4	AF007881
		Tas	1	4	AF239962
		VM	1	4	AF239961
		VW	1	4	AF239960
		Kenyan-K	1	4	X01069
		SP2-Reunion Island	1	7	AJ225011
		SP2-R12	1	7	AJ225010
		SP2-R11	1	7	AJ225009
		SP2-R7	1	7	AJ225008
		SP1-R10	1	7	AJ225007
		N2A-R8	1	7	AJ225006
		N2A-Reunion-R2	1	7	AJ224504
		N2A-Reunion Island-R6	1	7	AJ224508
		N2A-Reunion Island-R5	1	7	AJ224507
		N2A-Reunion Island-R4	1	7	AJ224506
		N2A-Reunion Island-R3	1	7	AJ224505
		MSV-R	1	7	X94330
			1	6	X01633 K02026
	Malvastrum yellow vein virus [Y47]	Y47	1	6	AJ457824
	Melon chlorotic leaf curl virus		1	4	AF325497
	Miscanthus streak virus		1	5	D01030
			1	5	D00800

	Mungbean yellow mosaic virus		2	6	D14703
				4	D14704
		Akola-1	2	6	AY271894
		Akola-2		2	AY271893
	Mungbean yellow mosaic virus-Vigna		2	6	AJ132575
				2	AJ132574
	Mungbean yellow mosaic India virus	Cowpea-1	2	7	AF481865 AF513504
		Cowpea-2		2	AF503580
			2	2	AJ420331
				7	AJ416349
		mungbean-1	2	7	AF416742
		mungbean-2		2	AF416741
		MYMV-Sb-1	2	7	AY049772
		MYMV-Sb-2		2	AY049771
		bg3-1	2	7	AF126406
		bg3-2		2	AF142440
	Okra enation virus		1	6	AF155064
	Okra leaf curl virus	OkLCV/EG okra	1	6	AY036010
	Okra yellow vein mosaic virus	Pakistan-oyvm201	1	6	AJ002451
	Panicum streak virus	Karino-pKom100	1	3	L39638
		geminivirus-leaf	1	3	X60168
	Papaya leaf curl virus		1	7	Y15934
	Pepper huasteco virus		2	5	X70418
				2	X70419
		Sinola-1	2	5	AY044162
		Sinola-2		2	AY044163
	Pepper leaf curl virus		1	6	AF134484
			1	6	AF336806
			1	6	AF314531
	Potato yellow mosaic virus	pMAH2	2	4	D00940
		pMBH2		2	D00941
		Guadeloupe-1	2	5	AY120882
		Guadeloupe-2		2	AY120883
		PYMV/TT-1	2	5	AF039031
		PYMV/TT-2		2	AF039032
		Panama-1	2	5	Y15034
		Panama-2		2	Y15033
	Rhynchosia Golden Mosaic Virus		1	4	AF239671
		tobacco	1	4	AF406199
	Sida golden mosaic virus	Florida	2	5	AF049336
		Florida		2	AF039841
		DNA-AI	2	7	U77963
	Sida golden yellow vein virus	DNA-AII	2	6	U77964
		Honduras-DNA A	2	5	Y11097
		Honduras-DNA B		2	Y11098
		Costa Rica-1	2	5	X99550
		Costa Rica-2		2	X99551
	Sida yellow vein virus	Honduras-yellow vein-1	2	5	Y11099
		Honduras-yellow vein-2		2	Y11100
	Sida mottle virus		1	6	AY090555
	Sida yellow mosaic virus		1	6	AY090558
	South African cassava mosaic virus		2	6	AF155806
				2	AF155807
	Soybean crinkle leaf virus		1	6	AB050781
	Squash leaf curl virus		2	4	M38183
				2	M38182
	Squash mild leaf curl virus	SLCV-R-1	2	4	AF421552
		SLCV-R-2		2	AF421553
	Squash leaf curl China virus		1	6	AB027465
	Squash leaf curl Yunnan virus	Y23	1	6	AJ420319
	Squash yellow mottle virus	98-631-1	2	4	AY064391
		98-631-2		2	AF440790
	Sri Lankan cassava mosaic virus	SLCMV-Col-1	2	6	AJ314737
		SLCMV-Col-2		2	AJ314738
	Stachytarpheta leaf curl virus	Hn5	1	6	AJ495814
	Sugarcane streak virus	Natal	1	3	M82918
		1	1	4	AF072672
	Sweet potato leaf curl virus		1	6	AF104036
	Sweet potato leaf curl Georgia virus		1	6	AF326775
	Taino tomato mottle virus		2	4	AF012300
				2	AF012301
			2	4	U57457

				2	AF499442
	Tobacco curly shoot virus	Y41	1	6	AJ457986
	Tobacco leaf curl virus		1	6	AF350330
		Y64	1	6	AJ457823
		Kochi-TLCV-KK	1	6	AB055009
		Nara-TLCV-Jp2	1	6	AB055008
		Y36	1	6	AJ420316
		Y35	1	6	AJ420318
		Y38	1	6	AJ420317
		Y8	1	6	AJ319677
		Y11	1	6	AJ319676
		Y10	1	6	AJ319675
		Y5	1	6	AJ319674
		Y3	1	6	AF240674
			1	6	AB028604
	Tobacco leaf curl Yunnan virus	Y136	1	6	AJ512761
		Y143	1	6	AJ512762
	Tobacco leaf curl Yamaguchi virus	TLCV-Y	1	7	AB079765
	Tobacco yellow dwarf virus	tobacco	1	4	M81103
	Tomato chlorotic mottle virus	BA-Se1-1	2	5	AF490004
		BA-Se1-2		2	AF491306
	Tomato geminivirus	Y72	1	6	AJ496812
		Y25	1	6	AJ457985
	Tomato golden mosaic virus		2	5	K02029
				2	K02030
	Tomato golden mottle geminivirus	GT94-R2	1	2	AF132852
	Tomato leaf crumple virus		2	5	AF101475
				2	AF101478
	Tomato leaf curl virus - Australia	Australia Australian	1	6	S53251
	Tomato leaf curl virus - Bangalore I	ITmLCV-Bangalore I	1	6	Z48182
	Tomato leaf curl virus - Bangalore II	Bangalore II-2	1	6	U38239
		LCV-Ban4	1	6	AF165096
	Tomato leaf curl virus - Taiwan	Taiwan	1	6	U88692
	Tomato leaf curl Bangladesh virus	TLCV-BD2	1	6	AF188481
	Tomato leaf curl Gujarat virus	Varanasi-1	2	6	AY190290
		Varanasi-1		2	AY190291
			1	6	AF449999
			1	6	AF413671
	Tomato leaf curl Laos virus	TLCV-LA	1	6	AF195782
	Tomato leaf curl Malaysia virus		1	6	AF327436
		New Delhi-1	2	8	U15015
		New Delhi-2		2	U15017
			1	6	AF102276 AF061339
		New Delhi	1	8	U15016
			1	6	AF274349
		Viet Nam	1	6	AF264063
	Tomato mottle virus	Florida-1	2	4	L14460
		Florida-2		2	L14461
	Tomato pseudo-curly top virus	K77B-11	1	6	X84735
	Tomato rugose mosaic virus		2	5	AF291705
				2	AF291706
	Tomato yellow leaf curl virus	Spain_Almeria	1	6	AJ489258
			1	6	AJ132711
		TYLCV-CHI	1	6	AF311734
		Sp42199	1	6	AF271234
		Dominican Republic	1	6	AF024715
		Thailand	1	5	X63015
		Aichi-Gemini-A	1	6	AB014347
		Shizuoka-Gemini-S	1	6	AB014346
		Cuban	1	6	AJ223505
		TYLCV-Sicily	1	6	Z28390
			1	6	X76319
		TYLCV-M-Murcia	1	6	Z25751
			1	6	X15656
		Almeria	1	6	L27708
	Tomato yellow leaf curl virus - Israel	Portugal	1	6	AF105975
		Spain Sp7297	1	6	AF071228
	Tomato yellow leaf curl virus - Sardinia		1	6	X61153
	Tomato yellow leaf curl virus - Thailand	Thailand-1	2	7	AF141922
		Thailand-2		3	AF141897
	Watermelon chlorotic stunt virus	k22-11	2	6	AJ012081
		wcb		2	AJ012082
		WmCSV-SD-A	2	6	AJ245650
		WmCSV-SD-B		3	AJ245651
		WmCSV-IR-A	2	7	AJ245652
		WmCSV-IR-B		2	AJ245653
	Wheat dwarf virus	Enkoping1-Wheat	1	4	AJ311031
		French	1	5	X82104
			1	5	X02869
			305	**1423**	
Circoviridae	Beak and feather disease virus	PBMC	1	3	AF311302
		feather	1	3	AF311301
		Cacatua leadbeateri	1	3	AF311300
		Trichoglossus haematodus	1	3	AF311299
		Eolophus roseicapillus	1	3	AF311298
		Cacatua tenuirostris	1	3	AF311297
		Agapornis roseicollis	1	3	AF311296
		Psephotus haematogaster	1	3	AF311295
	Bovine circovirus	cattle	1	11	AF109397
	Canary circovirus	serinus canaria	1	2	AJ301633, AF346618
	Chicken anemia virus	3-1P60 MDCC-MSB1	1	3	AY040632
		SMSC-1P60	1	3	AF390102
		BD-3	1	3	AF395114
		Gg3-1	1	3	AF390038
			1	3	AF475908
		Cux-1 MDCC-MSB1	1	3	AJ297685
		Cux-1 MDCC-MSB1	1	3	AJ297684
		SMSC-1 MDCC-MSB1	1	3	AF285682
		TR20	1	3	AB027470
		Gg10	1	3	U66304
		26P4	1	3	D10068, D01218
		704	1	3	U65414
		Cuxhaven 1	1	3	M81223
	Columbid circovirus	Pigeon-7050	1	6	AJ298230
		Pigeon-9030	1	6	AJ298229
		Pigeon	1	5	AF252610
	Mulard duck circovirus	DuCV	1	2	AY228555
	Goose circovirus	Anser	1	4	AJ304456
	Porcine circovirus	CT-PCV-P6	1	2	AY099500
		PMWS-PCV-P1	1	2	AY099495
		CT-PCV-P7	1	2	AY099501
		PMWS-PCV-P4	1	2	AY099498
		PMWS-PCV-P3	1	2	AY099497
		CT-PCV-P5	1	2	AY099499
		PMWS-PCV-P2	1	2	AY099496
	Porcine circovirus type 2	JHP	1	2	AF520783
		SC	1	2	AF455211
		KSY-1	1	2	AF454546
		Pig-35	1	4	AB072303
		Pig-26	1	4	AB072302
		Pig-33	1	5	AB072301
		IAF2897	1	10	AF408635
		Pig-PMWS	1	10	AY035820
		Pig-BX	1	2	AF381177
		Pig-HR	1	2	AF381176
		Pig-BF	1	2	AF381175
		Chia-Yi	1	1	AF364094
		24657 NL	1	10	AF201897
		IAF-4370	1	1	AF118097
		IAF-614	1	1	AF118095
		34464	1	2	AF264043
		40895	1	2	AF264042
		40856	1	2	AF264041
		10489	1	2	AF264040
		26607	1	2	AF264039
		26606	1	2	AF264038
		FRA3	1	2	AF201311
		SPA3	1	2	AF201310
		SPA2	1	2	AF201309
		SPA1	1	2	AF201308
		GER3	1	2	AF201307
		GER2	1	2	AF201306

		GER1	1	2	AF201305
		ISU-31	1	1	AJ223185
		Tainan	1	1	AF166528
		MLTW98	1	1	AF154679
			1	1	AF012107
		412	1	7	AF085695
		M226	1	2	AF086836
		9741	1	2	AF086835
		B9	1	2	AF086834
			1	11	AF027217
		pPCV-PSTI	1	1	U49186
		PCVII	1	10	AF055394
		PCVII	1	10	AF055393
		PCVII	1	11	AF055392
		PCVII	1	11	AF055391
	Porcine circovirus type 2D	lung	1	11	AF117753
	Porcine circovirus type 2B	lung	1	11	AF112862
	Porcine circovirus type 2E	lung	1	11	AF109399
	Porcine circovirus type 2C	lung	1	10	AF109398
	TTV-like mini virus	PB4TL	1	2	AF291073
		TLMV-NLC030	1	3	AB038631
		TLMV-NLC026	1	3	AB038630
		TLMV-NLC023	1	3	AB038629
		TLMV-CLC205	1	3	AB038628
		TLMV-CLC156	1	2	AB038627
		TLMV-CLC138	1	3	AB038626
		TLMV-CLC062	1	3	AB038625
		TLMV-CBD279	1	3	AB026931
		TLMV-CBD231	1	3	AB026930
		TLMV-CBD203	1	3	AB026929
	TT virus	VT416	1	3	AB041007
			93	**345**	
Parvoviridae					
Parvovirinae	Adeno-associated virus 1		1	2	AF063497
	Adeno-associated virus 2	HeLa	1	5	J01901 M12405 M12468 M12469
			1	7	AF043303
	Adeno-associated virus 3	3H	1	2	U48704
	Adeno-associated virus 3B		1	2	AF028705
	Adeno-associated virus 4	ATCC VR-646	1	2	U89790
	Adeno-associated virus 6		1	2	AF028704
	Aleutian mink disease virus	ADV-G	1	4	M20036
	Avian adeno-associated virus ATCC VR-865	ATCC-VR-865	1	2	AY186198
	B19 virus	HV	1	3	AF182273
		Rm	1	3	AB030694
		M6	1	3	AB030693
		N6	1	3	AB030673
	Bovine parvovirus	Pat153	1	4	M14363 M21972 M21973 M21974 M21975
	Canine parvovirus	CPV-N	1	3	M19296
		Y1	1	2	D26079
		CPV-b-790312	1	3	M38245
	Goose parvovirus	Virulent B	1	4	U25749
	Human erythrovirus V9	V9	1	6	AJ249437
	Parvovirus LuIII	NBE324K	1	2	M81888
	Mice minute virus	MVM(p)	1	7	J02275 M12520 M12521 M14704
		MVM(i)	1	7	M12032
			1	5	V01115
	Minute virus of canines		1	4	AF495467
	Mouse parvovirus 1		1	2	U12469
	Parvovirus H1		1	3	X01457 J02198
	Porcine parvovirus	NADL-2	1	3	M38367
		Kresse	1	5	U44978
		NADL-2	1	3	D00623
			29	**103**	
Parvoviridae					
Densovirinae	Aedes albopictus Parvovirus	C6/36	1	3	X74945
	Blattella germanica densovirus		1	5	AY189948
	Bombyx mori densovirus 1		1	3	AY033435
	Bombyx mori densovirus 5	Shinshu	1	4	AB042597
	Casphalia extranea densovirus		1	3	AF375296
	Diatraea saccharalis densovirus		1	7	AF036333
	Galleria mellonella densovirus	GmDNV	1	4	L32896
	Infectious hypodermal and		1	3	AF218266

	hematopoietic necrosis virus				
		Thailand	1	3	AY102034
			1	3	AF273215
	Junonia coenia densovirus		1	4	S47266
	Myzus persicae densovirus		1	5	AY148187
	Periplaneta fuliginosa densovirus		1	8	AF192260
	Planococcus citri densovirus		1	4	AY032882
			14	**59**	
Nanoviridae	Ageratum yellow vein virus-associated	pGEM-AYVV2	1	1	AJ416153
	Banana bunchy top virus		6	1	S56276
					L41576
				1	L41574
				1	L41575
				1	L41578
				1	L41577
	Coconut foliar decay virus		1	6	M29963
	Faba bean necrotic yellows virus	EV1-93	10	1	AJ005968
		Egyptian EV1-93		1	AJ132180
		Egyptian EV1-94		1	AJ132181
		Egyptian EV1-95		1	AJ132182
		Egyptian EV1-96		1	AJ132183
		Egyptian EV1-97		1	AJ132184
		Egyptian EV1-98		1	AJ132185
		Egyptian EV1-99		1	AJ132186
		Egyptian EV1-100		1	AJ132187
		Egyptian EV1-101		1	AJ132179
	Milk vetch dwarf virus		11	1	AB000920
				1	AB000921
				1	AB000922
				1	AB000923
				1	AB000924
				1	AB000925
				1	AB000926
				1	AB000927
				1	AB009046
				1	AB009047
				1	AB027511
	Subterranean clover stunt virus		8	1	U16730
				1	U16731
				1	U16732
				1	U16733
				1	U16734
				1	U16735
				1	U16736
				1	AJ290434
	Tobacco curly shoot virus associated DNA1	Y35	1	1	AJ579345
	Tobacco leaf curl Yunnan virus associated DNA 1	Y143	1	1	AJ579361
	Tomato yellow leaf curl China virus associated DNA 1	Y8	1	1	AJ579353
	Tomato yellow leaf curl Thailand virus associated DNA 1	Y70	1	1	AJ579359
			41	**45**	
Microviridae	Chlamydia phage 2		1	8	AJ270057
	Chlamydia phage phiCPAR39		1	7	AE002163
	Chlamydia phage PhiCPG1	PhiCPG1	1	9	U41758
	Chlamydia psittaci bacteriophage chp1		1	12	D00624
	Coliphage alpha3	wildtype	1	10	X60322
	Coliphage phiK	wildtype	1	10	X60323
	Coliphage phiX174		1	11	J02482 M10348 M10379 M10714 M10749 M10750 M10866 M10867 M24859 VERSION J02482.1 GI:216019
	Enterobacteria phage G4		1	11	J02454 M10724 M11404 V00657 VERSION J02454.1 GI:215415
		Anc	1	11	AF454431
	Enterobacteria phage S13		1	12	M14428 M25197
		Anc	1	11	AF274751
	Phage phiMH2K		1	11	AF306496
	Spiroplasma phage 4		1	9	M17988
			13	**132**	
Inoviridae	Acholeplasma phage MV-L1	JA-1	1	4	X58839

	Bacteriophage VfO3K6	O3:K6	1	10	AB043678
	Bacteriophage VfO4K68	O4:K68	1	8	AB043679
	Bacteriophage VSKK	BO4	1	6	AF452449
	Colphage M13		1	10	V00604 J02461 M10377
		f1	1	10	J02448
		fd-478	1	10	J02451 M10731 M10767 M21666 M21667 M21668 M21669 M21670 M25196 VERSION J02451.1 GI:215394
		fd-tet	1	12	AF217317
	Enterobacteria phage f2-2	f2-2	1	9	X14336
	Enterobacteria phage If1		1	10	U02303
	Enterobacteria phage Ike		1	10	X02139 K02750
	Propionibacterium phage phiB5		1	10	AF426260
	Pseudomonas phage Pf1	ATCC 25102-B1	1	14	X52107
	Pseudomonas phage Pf3	Nijmegen-RP1	1	9	M11912
	Spiroplasma phage 1-C74	SpV1-C74	1	13	U28974
	Spiroplasma phage 1-R8A2B		1	13	X51344
	SVTS2 plectrovirus	pESV-1	1	14	AF133242
	Vibrio cholerae filamentous bacteriophage fs-2	MDO14	1	9	AB002632
	Vibrio cholerae O139 fs1 phage		1	15	D89074
	Vibrio cholerae phage VGJphi		1	13	AY242528
	Vibrio phage VSK	O139	1	14	AF453500
	Xanthomonas phage Cf1c		1	9	M57538
			22	232	

Proteomas Virales de Cadena Doble de RNA (dsRNA)

Birnaviridae	Drosophila x virus		2	1	U60650
				1	AF196645
	Infectious bursal disease virus	Very virulent-UK56-A	2	2	X92760
		Very virulent-UK56-B		1	X92761
		Australian 002-73-A	2	1	X03993 M24259 M27967 M64738
		Australian 002-73-B		1	M19336
		OH-A	2	2	U30818
		OH-B		1	U30819
		Harbin-1-A	2	2	AF454945
		Harbin-1-B		1	AF455136
		23/82-A	2	1	AF362773
		23/82-B		1	AF362774
		CT-A	2	2	AJ310185
		CT-B		1	AJ310186
		UPM94/273-A	2	2	AF527039
		UPM94/273-B		1	AF527038
		SH95-A	2	1	AY134874
		SH95-B		1	AY134875
		UPM97/61-A	2	2	AF247006
		UPM97/61-B		1	AF527040
		Irwin Moulthrop-A	2	2	AY029166
		Irwin Moulthrop-B		1	AY029165
		JD1-A	2	2	AF321055
		JD1-B		1	AY103464
		HZ2-A	2	2	AF321054
		HZ2-B		1	AF493979
		BD 3/99-A	2	1	AF362776
		BD 3/99-B		1	AF362770
		Cu-1wt-A	2	1	AF362747
		Cu-1wt-B		1	AF362748
		D6948-A	2	2	AF240686
		D6948-B		1	AF240687
		variant E-A	2	2	AF133904
		variant E-B		1	AF133905
		CEF94-A	2	2	AF194428
		CEF94-B		1	AF194429
		HK46-A	2	2	AF092943
		HK46-B		1	AF092944
		UK 661-A	2	1	AJ316896
		UK 661-B		1	AJ316897
	Infectious pancreatic necrosis virus	Jasper-A	2	2	M18049 X04124
		Jasper-B		1	M58756
		DRT-HL-1-A	2	1	D26526
		DRT-HL-1-B		1	D26527
		West Buxton-A	2	2	AF078668
		West Buxton-B		1	AF078669
		SP-A	2	1	U56907
		SP-B		1	M58757
	Marine birnavirus	Y-6-A	2	1	AB006783 AB001329
		Y-6-B		1	AY129662
			50	65	
Hypoviridae	Cryphonectria hypovirus	2-NB58	1	2	L29010
	Cryphonectria hypovirus 1	EP713	1	2	M57938
		Euro7	1	2	AF082191
	Cryphonectria hypovirus 3	CHV3-GH2 (WY)	1	1	AF188515
		CHV3-GH2 (CS)	1	1	AF188514
			5	8	
Totiviridae	Eimeria brunetti RNA virus 1	note="Eb-RV1"	1	2	AF356189
	Giardia lamblia virus	Portland 1	1	2	L13218
			1	2	AF525216
	Gremmeniella abietina RNA virus L1	type A, HR2	1	2	AF337175
	Helicobasidium mompa No.17 dsRNA virus		2	2	AB085814
				1	AB085815
	Helminthosporium victoriae virus 190S		1	2	U41345
	Leishmania RNA virus 1-1		1	3	M92355
	Leishmania RNA virus 1-4		1	2	U01899
	Leishmania RNA virus 2-1		1	3	U32106
	Saccharomyces cerevisiae virus L-A		1	3	J04692
		L014	1	2	M28353 X13426
	Saccharomyces cerevisiae virus L-BC (La)		1	2	U01060
	Sphaeropsis sapinea RNA virus 1		1	2	AF038665
	Sphaeropsis sapinea RNA virus 2		1	2	AF039080
	Trichomonas vaginalis virus	T1	1	2	U08999
		TVV-T5	1	2	U57898
	Trichomonas vaginalis virus 3		1	2	AF325840
	Trichomonas vaginalis virus II		1	2	AF127178
	Ustilago maydis virus H1	P1H1	1	1	U01059
	Zygosaccharomyces bailii virus Z		1	2	AF224490
			21	43	
Partitiviridae	Discula destructiva virus 2	331-1	2	1	AY033436
		331-2		1	AY033437
	Fusarium poae virus 1	A11_1	2	1	AF015924
		A11_2		1	AF047013
	Mycovirus FusoV (from Fusarium solani)	M1	2	1	D55668
		M2		1	D55669
	Gremmeniella abietina RNA virus MS1	C5-A-LTT-1	3	1	AY089993
		C5-A-LTT-2		1	AY089994
		C5-A-LTT-3		1	AY089995
	Rhizoctonia solani virus	Rhizoctonia solani 717 partitivirus-1	2	1	AF133290
		Rhizoctonia solani 717 partitivirus-2		1	AF133291
			11	11	
Reoviridae	Banna virus	JKT-6423-1	12	12	AF133430
		JKT-6423-2			AF134514
		JKT-6423-3			AF134515
		JKT-6423-4			AF134516
		JKT-6423-5			AF134517
		JKT-6423-6			AF134518
		JKT-6423-7			AF052018
		JKT-6423-8			AF052017
		JKT-6423-9			AF052016
		JKT-6423-10			AF052015
		JKT-6423-11			AF052014
		JKT-6423-12			AF019908
	Aquareovirus C		11	12	AF403398
					AF403399
					AF403400
					AF403401
					AF403402

				AF403403
				AF403404
				AF403405
				AF403406
				AF403407
				AF403408
Grass carp reovirus		11	12	AF260511
				AF260512
				AF260513
				AF403390
				AF403391
				AF403392
				AF403393
				AF403394
				AF403395
				AF403396
				AF403397
Bluetongue virus	serotype 10-L1	10	10	X12819
	serotype 10			M11787 M16566
	serotype 10			M22096
	serotype 10-4			Y00421
	serotype 10-M5			D12532 D01183
	serotype 10-6			Y00422
	serotype 10			X06463
	10-S8			D00500
	10-S9			D00509
	10			M28981
Bombyx mori cypovirus 1	I-1	10	10	AF323781
	I-2			AF323782
	I-3			AF323783
	I-4			AF323784
	I-5			AB035732
	I-6			AB030014
	I-7			AB030015
	I-8			AB016436
	I-9			AF061199
	I-10			M19112
Chuzan virus	Palyam-1	10	10	AB018086
	Palyam-2			AB014725
	Palyam-3			AB014726
	Palyam-4			AB018087
	Palyam-5			AB018089
	Palyam-6			AB014725
	Palyam-7			AB014727
	Palyam-8			AB018090
	Palyam-9			AB018088
	Palyam-10			AB018091
Colorado tick fever virus	Florio; N-7180-1	12	13	AF133428
	Florio; N-7180-2			AF139758
	Florio; N-7180-3			AF139759
	Florio; N-7180-4			AF139760
	Florio; N-7180-5			AF139761
	Florio; N-7180-6			AF139762
	Florio; N-7180-7			AF139763
	Florio; N-7180-8			AF139764
	Florio; N-7180-9			AF000720
	Florio; N-7180-10			AF139765
	Florio; N-7180-11			U72694
	Florio; N-7180-12			U53227
Eyach virus	Fr578-1	12	13	AF282467
	Fr578-2			AF282468
	Fr578-3			AF282469
	Fr578-4			AF282470
	Fr578-5			AF282471
	Fr578-6			AF282472
	Fr578-7			AF282473
	Fr578-8			AF282474
	Fr578-9			AF282475
	Fr578-10			AF282476
	Fr578-11			AF282477
	Fr578-12			AF282478
Kadipiro virus	JKT-7075-1	12	12	AF133429
	JKT-7075-2			AF134509
	JKT-7075-3			AF134510
	JKT-7075-4			AF134511
	JKT-7075-5			AF134512
	JKT-7075-6			AF134513
	JKT-7075-7			AF052023
	JKT-7075-8			AF052022
	JKT-7075-9			AF052021
	JKT-7075-10			AF052020
	JKT-7075-11			AF052019
	JKT-7075-12			AF019909
Lymantria dispar cypovirus 1	gypsy-1	10	10	AF389462
	gypsy-2			AF389463
	gypsy-3			AF389464
	gypsy-4			AF389465
	gypsy-5			AF389466
	gypsy-6			AF389467
	gypsy-7			AF389468
	gypsy-8			AF389469
	gypsy-9			AF389470
	gypsy-10			AF389471
Lymantria dispar cypovirus 14	gypsy-1	10	11	AF389452
	gypsy-2			AF389453
	gypsy-3			AF389454
	gypsy-4			AF389455
	gypsy-5			AF389456
	gypsy-6			AF389457
	gypsy-7			AF389458
	gypsy-8			AF389459
	gypsy-9			AF389460
	gypsy-10			AF389461
Mammalian orthoreovirus 1	Lang-Dearing-1	10	11	M24734
	Lang-Dearing			AF378003
	Lang-Dearing-L3			AF129820
	Lang-Dearing			AF461682
	Lang-Dearing			AF490617
	Lang-Dearing-M3			AF174382
	Lang-Dearing			M14779
	Lang-Dearing			M17598
	Lang-Dearing			M14325
	Lang-Dearing			M13139
Mammalian orthoreovirus 2	2-D/5 Jones	10	11	M31057
	2-D/5 Jones			AF378005
	2-D/5 Jones			AF129821
	2-D/5 Jones			AF124519
	2-D/5 Jones			M19355
	2-D/5 Jones			AF174383
	2-D/5 Jones			M10261
	2-D/5 Jones			L19775
	2-D/5 Jones			M18390
	2-D/5 Jones			X60066
Nilaparvata lugens reovirus	Izumo-S1	10	11	D49693
	Izumo-S2			D49694
	Izumo-S3			D49695
	Izumo-S4			D49696
	Izumo-S5			D49697
	Izumo-S6			D49698
	Izumo-S7			D49699
	Izumo-S8			D26127
	Izumo-S9			D49700
	Izumo-S10			D14691
Rice black streaked dwarf virus	S1	10	13	AJ294757
	S2			AJ409145
	S3			AJ293984
	S4			AJ409146
	S5			AJ409147
	S6			AJ409148
	S7			AJ297427
	S8			AJ297431
	S9			AJ297430
	S10			AJ297433
Rice dwarf virus	Chinese-1	12	12	U73201
	Chinese-2			U73202
	Chinese-3			U72757
	Chinese-4			U36562
	Chinese-5			U36563

		Chinese-6			U36564
		Chinese-7			U36402
		Chinese-8			U36565
		Chinese-9			U36566
		Chinese-10			U36567
		Chinese-11			U36568
		Chinese-12			U36569
	Rice ragged stunt virus	Thai-S1	10	11	AF020334
		Thai-S2			AF020335
		Thai-S3			AF020336
		Thai-S4			U66714
		Thai-S5			U33633
		Thai-S6			AF020337
		Thai-S7			U66713
		Thai-S8			L46682
		Thai-S9			L38899
		Thai-S10			U66712
	Simian rotavirus A	SA11-1	11	12	X16830
		SA11-2			X16831
		SA11-3			X16062
		SA11-4			X14204
		SA11-5			X14914
		SA11-6			X00421
		SA11-7			X00355
		SA11-8			J02353
		SA11-9			K02028
		SA11-10			K01138
		SA11-11			X07831
			203	**217**	
Inclasificados	Diaporthe ambigua RNA virus 1		**1**	**2**	AF142094
Cystoviridae	Bacteriophage phi-12	L	3	6	AF406636
		M		5	AY039807
		S		4	AY034425
	Bacteriophage phi-8	L	3	6	AF226851
		M		6	AF226852
		S		4	AF226853
	Pseudomonas bacteriophage phi-13	L	3	4	AF261668
		M		5	AF261667
		S		4	AF261666
	Pseudomonas phage phi-6	L	3	4	M17461
		M		4	M17462
		S		5	M12921
			12	**57**	

Proteomas Virales de Cadena Sencilla RNA sentido positivo (ssRNA+)

Astroviridae	Avian nephritis virus	G-4260	1	2	AB033998
	Human astrovirus		1	2	Z25771
		Berlin-3	1	3	AF141381
		Yuc-8	1	3	AF260508
		1-Oxford	1	3	L23513
	Mink astrovirus		1	3	AY179509
	Ovine astrovirus		1	3	Y15937
	Turkey astrovirus		1	3	Y15936
			1	3	AF206663
			9	**25**	
Barnaviridae	Mushroom bacilliform virus	AUS LF-1	**1**	**4**	U07551
Benyvirus	Beet necrotic yellow vein virus	S	5	10	D84410
		S			D84411
		S			D84412
		S			D84413
		S43			D63936
	Beet soil-borne mosaic virus	EA	4	10	AF280539
		EA			AF061869
		EA			AF280540
		EA			AF280541
			9	**20**	

Bromoviridae	Alfalfa mosaic virus	425-Leiden-1	3	1	L00163 J02000
		425-Leiden-2		1	K02702 J02002
		425-Madison-3		2	K02703
	American plum line pattern virus		3	1	AF235033
				1	AF235165
				2	AF235166
	Apple mosaic virus		3	1	AF174584
				1	AF174585
				2	U15608
	Broad bean mottle virus		3	1	M65138
				1	M64713
				2	M60291
	Brome mosaic virus		3	1	X02380 K02706
				1	X01678 K02707
				2	V00099 J02042 J02043
		KU1-1	3	1	X58456
		KU1-2		1	X58457
		KU1-3		2	X58458
	Citrus leaf rugose virus		3	1	U23715
				1	U17726
				2	U17390
	Cowpea chlorotic mottle virus		3	1	M65139
				1	M28817
				2	M28818
		T-1	3	1	AF325739
		T-2		1	AF325740
		T-3		2	AF325741
		R-1	3	1	AF325736
		R-2		1	AF325737
		R-3		2	AF325738
	Cucumber mosaic virus	Fny-1	3	1	D00356
		Fny-2		1	D00355
		Fny-3		2	D10538
	Elm mottle virus		3	1	U57047
				1	U34050
				2	U85399
	Olive latent virus 2		3	1	X94346
				1	X94347
				2	X76993
	Peanut stunt virus	ER-1	3	1	U15728
		ER-2		1	U15729
		ER-3		2	U15730
		J-1	3	1	D11126 D01123
		J-2		1	D11127 D01124
		J-3		2	D00668
		Western (W)-1	3	1	U33145
		Western (W)-2		1	U33146
		Western (W)-3		2	U31366
	Pelargonium zonate spot virus	tomato-1	3	1	AJ272327
		tomato-2		1	AJ272328
		tomato-3		2	AJ272329
	Prune dwarf virus	Salmo BC cherry	3	1	U57648
				1	AF277662
		ch 137		2	L26145
	Prunus necrotic ringspot virus		3	1	AF278534
				1	AF278535
		30/4 Prunus-3		2	U57046
	Spinach latent virus		3	1	U93192
				2	U93193
				2	U93194
	Spring beauty latent virus	KU1-1	3	1	AB080598
		KU1-2		1	AB080599
		KU1-3		2	AB080600
	Tobacco streak virus		3	1	U80934
		WC-2		2	U75538
				2	X00435
	Tomato aspermy virus	V-1	3	1	D10044 D01101
		V-2		2	D10663 D01102
		V-3		2	AJ277268
		KC-TAV-1	3	1	AJ320273
		KC-TAV-2		2	AJ320274
		KC-TAV-3		2	AJ237849

	Tulare apple mosaic virus		3	1	AF226160
				2	AF226161
				2	AF226162
			75	**105**	
Caliciviridae	Calicivirus	NB	1	2	AY082891
	Chiba virus	Hu/NLV/Chiba 407/1987/JP	1	3	AB042808
	Bovine enteric calicivirus	Jena	1	3	AJ011099
	European brown hare syndrome virus	EBHSV-GD	1	2	Z69620
	Feline calicivirus		1	2	M86379
		CFI/68	1	2	U13992
		FCV2024	1	3	AF479590
		F65	1	3	AF109465
		Urbana	1	3	L40021
	Norwalk virus	Hu/NLV/NV/1968/US	1	3	M87661
		Hu/NLV/GII/MD145-12/1987/US	1	3	AY032605
		Manchester	1	3	X86560
		Norwalk-like-BS5	1	3	AF093797
		Bristol/98/uk	1	2	AJ249939
		SzUG1	1	3	AB039774
		WUG1	1	3	AB081723
		Saitama U201	1	3	AB039782
		Saitama U18	1	3	AB039781
		Saitama U25	1	3	AB039780
		Saitama U17	1	3	AB039779
		Saitama U16	1	3	AB039778
		Saitama U4	1	3	AB039777
		Saitama U3	1	3	AB039776
		Saitama U1	1	3	AB039775
		Lorsdale virus Hu/NLV/LD/1993/UK	1	3	X86557
		Hu/NLV/Hawaii virus/1971/US	1	3	U07611
	Porcine enteric calicivirus	Cowden-1	1	2	AF182760
	Rabbit hemorrhagic disease virus	FRG-Ra/LV/RHDV/GH/1989/GE	1	2	M67473
		BS89	1	2	X87607
		SD	1	2	Z29514
		Czech strain V351	1	2	U54983
		Iowa2000	1	2	AF258618
		AST/89	1	2	Z49271
	Southampton virus		1	3	L07418
	Vesicular exanthema of swine virus	A48	1	3	U76874 AF181082
		Pan-1	1	3	AF091736
	San Miguel sea lion virus	serotype 1	1	3	U15301 AF181081
	Canine calicivirus		1	3	AB070225
	Walrus calicivirus	walrus	1	3	AF321298
			39	**105**	
Closteroviridae	Beet pseudo-yellows virus		2	2	AY330918
				8	AY330919 AY267369 AY268107
	Beet yellows virus	Ukrainian-BYV-U	1	8	X73476
		Californian	1	8	AF056575
		BYV-4Californian	1	8	AF190581
	Citrus tristeza virus	T36	1	12	U16304 U02547 L20760
		T385	1	12	Y18420
		SY568	1	12	AF001623
		vt	1	12	U56902
		T30	1	12	AF260651
		NUagA	1	12	AB046398
	Cucumber yellows virus		2	7	AB085612
				2	AB085613
	Cucurbit yellow stunting disorder virus	AILM-1	2	5	AY242077 AF389082
		AILM-2		8	AY242078
	Grapevine leafroll-associated virus 3	NY1	1	13	AF037268 U82937
	Grapevine rootstock stem lesion associated virusH		1	9	AF314061
	Lettuce infectious yellows virus	92-1	2	3	U15440
		92-2		5	U15441 U05242
	Little cherry virus	UW2	1	9	Y10237
	Little cherry virus 2	USA6b	1	11	AF531505
	Sweet potato chlorotic stunt virus	EA-1	2	4	AJ428554
		EA-2		7	AJ428555
			23	**189**	
Comoviridae	Apple latent spherical virus	S1	2	1	AB030940
		S2		1	AB030941
	Bean pod mottle virus	KY G-7-1	2	1	U70666
		KY G-7-2		1	M62738
		K-Hopkins1-1	2	1	AF394608
		K-Hopkins1-2		2	AF394609
		K-Hancock1-1	2	1	AF394606
		K-Hancock1-2		2	AF394607
	Beet ringspot virus	S-1	2	1	D00322
		S-2		1	X04062
	Blackcurrant reversion virus	S1	2	1	AF368272
		S2		1	AF020051 AF112118
	Broad bean wilt virus 2	ME-1	2	1	AF225953
		ME-2		1	AF225954
		MB7-1	2	1	AB013615
		MB7-2		1	AB013616
		IA-1	2	1	AB051386
		IA-2		1	AB032403
		IP-1	2	1	AB023484
		IP-2		1	AB018698
	Cowpea mosaic virus	B	2	1	X00206
		M		4	X00729
	Cowpea severe mosaic virus	Vigna unguiculata	2	1	M83830
				1	M83309
	Cycas necrotic stunt virus		2	1	AB073147
				1	AB073148
	Grapevine chrome mosaic virus		2	1	X15346
				1	X15163
	Grapevine fanleaf virus	F13-1	2	1	D00915
		F13-2		2	X16907
	Patchouli mild mosaic virus	Philippines-1	2	1	AB050782
		Philippines-2		1	AB011007
	Raspberry ringspot virus	cherry-1	2	1	AY303787
		cherry-2		1	AY303788
	Red clover mottle virus	S1	2	1	X64886 S46268
		S2		1	M14913
	Satsuma dwarf virus	S58-1	2	1	AB009958
		S58-2		1	AB009959
	Squash mosaic virus	Y-SqMV-1	2	1	AB054688
		Y-SqMV-2		1	AB054689
	Strawberry mottle virus		2	1	AJ311875
				1	AJ311876
	Tobacco ringspot virus	Bud Blight-1	2	1	U50869
		Bud Blight-2		1	AY363727
	Tomato black ring virus	MJ-1	2	1	NC_004439
		MJ-2		1	AY157994
	Tomato ringspot virus	raspberry-1	2	1	L19655
		raspberry-2		3	D12477 D01129
			48	**56**	
Dicistroviridae	Acute bee paralysis virus		1	2	AF150629
	Aphid lethal paralysis virus		1	2	AF536531
	Black queen cell virus	South African	1	2	AF183905
	Cricket paralysis virus		1	2	AF218039
	Drosophila C virus	EB	1	2	AF014388
	Himetobi P virus		1	2	AB017037
	Kashmir bee virus		1	2	AY275710
	Plautia stali intestine virus		1	2	AB006531
	Rhopalosiphum padi virus		1	2	AF022937
	Taura syndrome virus		1	2	AF277675
	Triatoma virus		1	2	AF178440
			11	**22**	
Flaviviridae	Aikhurma virus	1176	1	1	AF331718
	Apoi virus	ApMAR	1	1	AF160193
	Bovine viral diarrhea virus genotype 2	C413	1	1	AF002227
		New York'93	1	1	AF502399
		1373	1	1	AF145967

	Deer tick virus	db30-CT95	1	1	AF311056
	Dengue virus	S1-vaccine	1	1	M19197
		BR/01-MR	1	1	AF513110
		BR/97-233	1	1	AF311958
	Hepatitis C virus	1a-H77	1	1	AF009606
		M1LE-1b	1	1	AB080299
		chimera-HCV-CG1b	1	1	AF333324
	Hepatitis G virus	PNF2161	1	1	U44402
		R10291	1	1	U45966
		HGV-tw	1	1	D87255
		HGV-GD(Guangdong)	1	1	AF006500
		HGVCN	1	1	U94695
		HGV-GA128	1	1	AB013500
		HGV-tw-pHGVqz	1	1	AF081782
		HGV-IM71	1	1	AB008342
		HGV-BL230	1	1	AB013501
		HGV-MY14	1	1	AB021287
		HGV-VT48	1	1	AB018667
		lowan	1	1	AF121950
		PEI	1	1	AF309966
		HGV-T55875	1	1	AF031827
		HGV-1517	1	1	AF031828
		HGV-1539	1	1	AF031829
	Hepatitis GB virus A	Alab	1	1	U94421
		mx70047	1	1	AF023424
		tri1122	1	1	AF023425
			1	1	U22303
	Hepatitis GB virus B	unknown	1	1	AF179612
			1	1	U22304
		ACY/GBV-B/FL-3	1	1	AJ277947
	Hepatitis GB virus C		1	1	U36380
		CG12LC	1	1	AB003291
			1	1	AF070476
		K1775	1	1	D87715
		K1668	1	1	D87714
		K1916	1	1	D87712
		K1789	1	1	D87711
		K1741	1	1	D87710
		K1737	1	1	D87709
		K606	1	1	D87708
		G13HC	1	1	AB003293
		G05BD	1	1	AB003292
		CG07BD	1	1	AB003290
		CG01BD	1	1	AB003289
		BG1HC	1	1	AB003288
		GSI93	1	1	D87263
		GSI85	1	1	D87262
		GT230	1	1	D90601
		T110	1	1	D90600
		GBV-C(EA)	1	1	U63715
	Japanese encephalitis virus	JaOArS982	1	1	M18370
		T1P1	1	1	AF254453
		CH1392	1	1	AF254452
		YL-vaccine	1	1	AF486638
		Ling	1	1	L78128
		Vellore-P20778	1	1	AF080251
		attenuated SA14-12-1-7	1	1	AF416457
		Ishikawa	1	1	AB051292
		CH2195SA	1	1	AF221500
		CH2195LA	1	1	AF221499
		SA14-14-2-vaccine	1	1	AF315119
		FU	1	1	AF217620
		GP78	1	1	AF075723
		K94P05	1	1	AF045551
		JaGAr 01	1	1	AF069076
		TL	1	1	AF098737
		TC	1	1	AF098736
		HVI	1	1	AF098735
		RP-9	1	1	AF014161
		RP-2ms	1	1	AF014160
		SA(V)	1	1	D90194
		SA(A)	1	1	D90195

		p3	1	1	U47032
		SA-14-2-8	1	1	U15763
		SA-14	1	2	M55506
		SA14-wild-type	1	1	U14163
	Kamiti River virus	SR-82	1	1	AY149905
	Kunjin virus	MRM61C	1	1	D00246
	Langat virus	TP21	1	1	AF253419
		attenuated strain E5	1	1	AF253420
	Louping ill virus	369/T2	1	1	Y07863
	Modoc virus	M544	1	1	AJ242984
	Montana myotis leukoencephalitis virus		1	1	AJ299445
	Mosquito cell fusing agent		1	1	M91671
	Murray Valley encephalitis virus	MVE-1-51	1	1	AF161266
	Omsk hemorrhagic fever virus	Bogoluvovska	1	1	AY193805
	Pestivirus Giraffe-1	giraffe-1 H138	1	1	AF144617
	Pestivirus Reindeer-1	reindeer-1 V50-Krefeld	1	1	AF144618
	Bovine viral diarrhea virus	SD-1	1	1	M96751
		Osloss	1	1	M96687
		890	1	1	U18059
		ILLC	1	1	U86599
		ILLNC	1	1	U86600
		type 1-NADL	1	1	AJ133739
		NADL-type 1	1	1	AJ133738
		CP7-5A	1	1	AF220247
		Oregon	1	1	AF041040
		CP7	1	1	U63479
		Oregon C24V	1	1	AF091606
			1	1	AF268278
	Pestivirus type 2	Eystrup	1	1	AF326963
	Classical swine fever virus	CI114	1	1	AF333000
		LPC	1	1	AF352565
		39	1	1	AF407339
		Paderborn	1	1	AY072924
		HCLV	1	1	AF531433
	Hog cholera virus	CS-vaccine	1	1	AF099102 AF132116
			1	1	J04358
		Shimen	1	1	AF092448 AF121103 AF157835
		HCLV	1	1	AF091507
		ALD	1	1	D49532
		GPE	1	1	D49533
	Pestivirus type 2	Alfort A19	1	1	U90951
		Chinese-85b	1	1	Z46258
		Alfort/187	1	1	X87939
		CAP	1	1	X96550
		Glentorf	1	1	U45478
		Riems, C	1	1	U45477
		Brescia	1	1	M31768
	Pestivirus type 3	X818-Clover Lane	1	1	AF037405 U00892 U00114
	Border disease virus	BD31	1	1	U70263
	Powassan virus	LB	1	1	L06436
	Rio Bravo virus	RIMAR	1	1	AF144692
	Tamana bat virus		1	1	AF285080
	Tick-borne encephalitis virus	Neudoerfl	1	1	U27495 M27157 M21498 M33668
		263	1	1	U27491
		Vasilchenko	1	1	L40361
		Vasilchenko	1	1	AF069066
		Hypr	1	1	U39292
		Oshima 5-10	1	1	AB062063
		Sofjin-HO	1	1	AB062064
	West Nile virus	33/G8	1	1	M12294 M10103
		WN Italy 1998-equine	1	1	AF404757
		WN NY 2000-crow3356	1	1	AF404756
		WN NY 2000-grouse3282	1	1	AF404755
		WN NJ 2000 MQ5488	1	1	AF404754
		VLG-4	1	1	AF317203
		WN MD 2000-crow265	1	1	AF404753
		IS-98 STD1	1	1	AF481864
		NY99-flamingo382-	1	1	AF196835

		99			
		RO97-50	1	1	AF260969
		Eg101	1	1	AF260968
		NY99-eq/hs	1	1	AF260967
		Connecticut 1999-2741	1	1	AF206518
		HNY1999	1	1	AF202541
		17D vaccine-Flavivirus	1	1	X03700 K02749
	Yellow fever virus	85-82H Ivory Coast	1	1	U54798
		Trinidad 79A-788379	1	1	AF094612
		Pasteur 17D-204 yellow fever vaccine	1	1	X15062
		vaccine strain 17D-213	1	1	U17067
		vaccine strain 17DD	1	1	U17066
		17D-204-South Africa vaccine	1	1	AF052446
		17D-204-South Africa vaccine	1	1	AF052445
		17D-204-South Africa vaccine	1	1	AF052444
		17D-204-USA vaccine-5	1	1	AF052439
		17D-204-USA vaccine-4	1	1	AF052438
		17D-204-USA vaccine-1	1	1	AF052437
		French neurotropic virus	1	1	U21055
		French viscerotropic virus	1	1	U21056
	Yokose virus	Oita 36-bat	1	1	AB114858
			168	**189**	
Flexiviridae			1	6	AB051848
			1	3	M58152 M31714
			1	4	X99752
			1	3	D14996
			1	4	AJ243438
			1	2	D14995 S47260
			1	2	AB004063
			1	5	D21829
			1	6	D26017
			1	5	AF018156
			1	5	L77962
			1	5	AF314662
			1	6	L25658
			1	7	AF308158
			1	5	U23414
			1	3	AJ318061
			1	5	D29630 D01191
			1	5	U62963
			1	5	AF016914
			1	3	AF017780
			1	7	AJ291761
			1	4	AF170028
			1	7	AF237816
			1	2	X82547
			1	5	M62730
			1	5	AY121833
			1	6	AJ292226
			1	6	Z68502
			1	6	AB010300 D11157
			1	6	AB010302 D11159
			1	8	AJ292230
			1	6	U89243
			1	6	AJ292229
			1	6	AF026278
			1	5	X75433
			1	5	X75448 X75896
			1	6	AB032469
			1	6	AF406744
			1	6	AJ516059
			1	6	D13747 D00405
			1	5	D13957 D00580
			1	5	AF484251
			1	5	AJ436767
			1	5	Z21647
			1	6	S73580
			1	6	D14449 D00515 X53062
			1	5	X05198 M37458 M72416
			1	5	X72214
			1	5	AF111193
			1	5	AF172259
			1	5	AB056718
			1	5	D00344
			1	5	M31541 M28049 M63141
			1	5	M95516
			1	5	AF057136
			1	5	AJ316085
			1	6	M97264
			1	6	D12517 D01227
			1	5	AF315308
			1	5	AB066288
			1	5	X06728
			1	5	X16636
			62	**317**	
Furovirus	Chinese wheat mosaic virus	Yantai-China-1	2	3	AJ012005
		Yantai-China-2		4	AJ012006
		Rongcheng-1	2	3	AJ271838
		Rongcheng-2		3	AJ271839
	Oat golden stripe virus	Oat-1	2	3	AJ132578
		Oat-2		3	AJ132579
	Soil-borne cereal mosaic virus	wheat-1	2	3	AJ132576
	European wheat mosaic virus	wheat-2		3	AJ132577
		wheat-1	2	3	AJ252151
		wheat-2		3	AJ252152
		O-1	2	3	AF146280
		O-2		3	AF146283
		G-1	2	3	AF146278
		G-2		3	AF146282
		C-1	2	3	AF146279
		C-2		3	AF146281
	Soil-borne wheat mosaic virus	US-Nebraska-1	2	3	L07937
		US-Nebraska-2		3	L07938
		Japanese-JT1	2	3	AB033689
		Japanese-JT2		3	AB033690
		S1	2	3	AJ296068
		S2		3	AJ296069
	Sorghum chlorotic spot virus	S1	2	3	AB033691
		S2		4	AB033692
			24	**74**	
Hordeivirus	Barley stripe mosaic virus	alpha	3	1	J04342
		beta-ATCC-PV43		4	X03854
		gamma-ATCC-PV43		2	M16576
		ND18-1	3	1	U35767
		ND18-2		4	U35770
		ND18-3		2	M16577
	Lychnis ringspot virus		3	1	Z46630
				4	Z46351
				2	Z46353
	Poa semilatent virus		3	1	Z46352
				4	M61486
				2	M61487
			12	**28**	
Idaeovirus	Raspberry bushy dwarf virus		2	1	S51557
				2	S55890
			2	**3**	
Iflavirus	Deformed wing virus		1	1	AJ489744
	Ectropis obliqua picorna-like virus		1	1	AY365064
	Infectious flacherie virus		1	1	AB000906

	Perina nuda picorna-like virus		1	1	AF323747
	Sacbrood virus	Rothamstead	1	1	AF092924
			5	5	
Luteoviridae	Barley yellow dwarf virus - MAV	MAV-PS1	1	6	D11028 D01213
	Barley yellow dwarf virus - GAV		1	6	AY220739
	Barley yellow dwarf virus - PAV	PAV-129	1	6	AF218798
		PAV-III	1	6	AF235167
			1	6	X07653
		P-PAV	1	6	D11032 D01214
	Bean leafroll virus		1	5	AF441393
	Beet chlorosis virus	BChV-2a	1	5	AF352024
		BChV-CR	1	5	AF352025
	Beet mild yellowing virus	2ITB	1	6	X83110
	Beet western yellows virus	FL1	1	6	X13063
		USA	1	6	AF473561
	Cereal yellow dwarf virus-RPV	RPV Mex-1	1	6	AF235168
		NY	1	6	L25299
	Cucurbit aphid-borne yellows virus	N	1	6	X76931
	Pea enation mosaic virus	WSG	1	5	L04573
	Potato leafroll virus	1	1	6	D00530 X14600
		Zim13	1	8	AF453388
		OP	1	8	AF453389
		Noir	1	8	AF453390
		Fr1	1	8	AF453391
		CIP01	1	7	AF453392
		CU87	1	8	AF453393
		14.2	1	8	AF453394
			1	5	AY138970
		polish	1	6	X74789
		Wageningen	1	6	Y07496
	Soybean dwarf virus	YS-M93-1	1	5	AB038147
		YP-M94-1	1	5	AB038148
		DS-HS97-8	1	5	AB038149
		DP-M96-1	1	5	AB038150
		Tas-1	1	5	L24049
	Sugarcane yellow leaf virus	A-CP65-357	1	6	AF157029
		CP65-357	1	6	AJ249447
			34	207	
Narnaviridae	Cryphonectria parasitica mitovirus 1-NB631	NB631	1	1	L31849
	Ophiostoma mitovirus 3a		1	1	AJ004930
	Ophiostoma novo-ulmi mitovirus 4-Ld	Dutch Elm	1	1	AJ132754
	Ophiostoma novo-ulmi mitovirus 5-Ld	Dutch Elm	1	1	AJ132755
	Ophiostoma novo-ulmi mitovirus 6-Ld	Dutch Elm	1	1	AJ132756
	Saccharomyces cerevisiae narnavirus 20S RNA	37-4C-W	1	1	AF039063 M63893
	Saccharomyces cerevisiae narnavirus 23S RNA	37-4C-T	1	1	U90136 M86595
			7	7	
Nidovirales	Avian infectious bronchitis virus	Beaudette	1	10	M95169 M27569
Coronaviridae		Beaudette-CK	1	10	AJ311317
	Bovine coronavirus	BCoV-ENT	1	12	AF391541
		BCoV-LUN	1	12	AF391542
		Quebec	1	12	AF220295
	Human coronavirus 229E	229E	1	8	AF304460
	Human coronavirus OC43	VR-759-OC43	1	9	AY391777
	Murine hepatitis virus	MHV-A59-C12	1	14	AF029248
		Penn 97-1	1	10	AF208066
		MHV-2	1	11	AF201929
		ML-11	1	9	AF207902
		ML-10	1	10	AF208067
	Porcine epidemic diarrhea virus	CV777	1	6	AF353511
	SARS coronavirus TOR2	Tor2-SARS	1	14	AY274119
	Transmissible gastroenteritis virus	PUR46-MAD-TGEV Purdue	1	9	AJ271965
			15	156	
Nidovirales	Equine arteritis virus	Bucyrus	1	9	X53459
Arteriviridae	Lactate dehydrogenase-elevating virus	neuro-virulent-C	1	8	L13298
		Plagemann	1	8	U15146
	Porcine reproductive and respiratory syndrome virus	16244B-2/18/97(Nebraska)passage 3	1	8	AF046869
		MLV RespPRRS/Repro	1	8	AF159149
		SP	1	8	AF184212
		VR-2332	1	8	U87392 AF030244 U00153
		BJ-4	1	8	AF331831
		NVSL 97-7985 IA 1-4-2	1	8	AF325691
		RespPRRS vaccine	1	8	AF066183
		CH-1a	1	8	AY032626 AF059352 AF132118
		P129-A	1	9	AF494042
		PA8	1	9	AF176348
		HB-1(sh)/2002	1	7	AY150312
	Lelystad virus		1	8	M96262
			1	7	A26843
	Simian hemorrhagic fever virus	LVR 42-0/M6941	1	11	AF180391 L39091 U63121 U20522
			17	140	
Nodaviridae	Black beetle virus		2	3	K02560
				2	X00956
	Boolarra virus		2	1	X15960
				2	AF329080
	Epinephelus tauvina nervous necrosis virus		2	2	AF319555
				1	AF318942
	Flock house virus		2	1	X15959
				3	X77156
	Macrobrachium rosenbergii nodavirus		2	2	AY222839
				1	AY222840
	Nodamura virus		2	2	AF174533
				1	AF174534
	Pariacoto virus		2	2	AF171942
				1	AF171943
	Striped Jack nervous necrosis virus		2	2	AB056571
				1	AB056572
			16	27	
Pecluvirus	Peanut clump virus		2	2	X78602
		87/TGTA2		5	L07269
	Indian peanut clump virus	Hyderabad serotype	2	2	X99149
		L-2		5	AF239729
			4	14	
Picornaviridae	A-2 plaque virus		1	1	AF201894
	Aichi virus	A846/88	1	1	AB010145
			1	1	AB040749
	Avian encephalomyelitis virus	Calnek	1	1	AJ225173
	Bovine enterovirus	VG-5-27	1	1	D00214
		SL305	1	1	AF123433
		K2577	1	1	AF123432
	Bovine kobuvirus	U-1	1	1	AB084788
	Encephalomyocarditis virus	Ruckert	1	1	M81861
		PV2	1	1	X87335
		BEL-2887A/91	1	1	AF356822
			1	1	X00463
		M-D-ifp	1	1	M37588
		EMC-D	1	1	M22458 J04335
		EMC-B	1	1	M22457 J04335
		emcv-pv21	1	1	X74312
	Equine rhinitis A virus	PERV	1	2	X96870
			1	1	L43052
	Equine rhinitis B virus	P1436 /71	1	1	X96871
	Equine rhinovirus 3	P313/75	1	1	AF361253
	Foot-and-mouth disease virus Asia 1 IND 63/72	vaccine IND 63/72	1	1	AY304994
	Foot-and-mouth disease virus C	C-S8	1	1	AF274010
		C-rp146	1	1	AJ133359
		C-rp99	1	1	AJ133358
		C-C-s8c1	1	1	AJ133357
	Foot-and-mouth disease virus O	O	1	1	AF308157
		China/1/99(Tibet)O	1	2	AF506822
		Akesu/58-O	1	1	AF511039

	O1Campos	1	1	AJ320488
	O/SKR/2000	1	1	AF377945
Foot-and-mouth disease virus SAT 2	KEN/5/57	1	1	AJ251473
Hepatitis A virus	wild-type	1	2	M14707
	HM-175	1	1	M16632
	Angeles	1	1	K02990
	isolate MBB	1	1	M20273
	F.G.Frp/3	1	1	X83302
	AH1	1	1	AB020564
	AH2	1	1	AB020565
	FH1	1	1	AB020567
	FH2	1	1	AB020568
	LY5	1	1	AF485328
	SLF88	1	1	AY032861
	LU38/WT	1	1	AF357222
	L-A-1-23	1	1	AF314208
	GBM/FRhK	1	1	X75214
	GBM/HFS	1	1	X75216
	GBM/WT	1	1	X75215
	HM175-43c	1	1	M59809
	HM175-24a	1	1	M59810
	HM175-18f	1	1	M59808
Human echovirus 1	Farouk-VR-1038	1	1	AF029859
Human enterovirus A	G-10	1	1	U05876
	Tainan/5079/98	1	1	AF177911
Enterovirus 5666/sin/002209	Enterovirus 5666/sin/002209	1	1	AF352027
Enterovirus 5865/sin/000009	5865/sin/000009	1	1	AF316321
Human enterovirus 71	MS/7423/87	1	1	U22522
	BrCr	1	1	U22521
	Tainan/6092/98	1	1	AF304459
	Tainan/4643/98	1	1	AF304458
	Tainan/5746/98	1	1	AF304457
	SHZH98	1	1	AF302996
	TW/2066/98	1	1	AF119796
	TW/2272/98	1	1	AF119795
	NCKU/9622	1	1	AF136379
	1245a/98/tw	1	1	AF176044
Coxsackievirus A9	Griggs	1	1	D00627
Enterovirus CA55-1988	CA55-1988	1	1	AF241359 AF241360
Human enterovirus B		1	1	M16560
Coxsackievirus B2	Ohio	1	1	AF081485
Coxsackievirus B3	PD	1	1	AF231765
	P	1	1	AF231764
	Cox31-1-93	1	1	AF231763
	B3	1	1	U57056
		1	1	M33854
		1	1	M68483
	Nancy	1	1	M16572
Coxsackievirus B4	J.V.B. Benschoten	1	1	X05690 D00149
	E2 variant	1	1	AF311939
	JBV	1	1	S76772
Coxsackievirus B5	Faulkner	1	1	AF114383
	1954/65/UK	1	1	X67706
Coxsackievirus B6	Schmitt	1	1	AF114384
	Schmitt-2	1	1	AF105342
	Schmitt (1-15-21)	1	1	AF039205
Human echovirus 5	Noyce	1	1	AF083069
Human echovirus 6	lytic-Charles	1	1	U16283 U05851
Human echovirus 7	Wallace-UMMC	1	1	AY036579
	UMMC	1	1	AY036578
Human echovirus 9	prototype Hill	1	1	X84981
	Barty	1	1	X92886
Human echovirus 11	wv207	1	1	AJ276224
		1	1	X80059
Human echovirus 12		1	1	X77708
	wildtype-Travis	1	1	X79047
Human echovirus 30	Bastianni	1	1	AF311938
	Bastianni-2	1	1	AF162711
Human enterovirus C	Coe	1	1	D00538
Human enterovirus D	J670/71	1	1	D00820
Human parechovirus 2	Gregory	1	1	AJ005695
	CT86-6760	1	1	AF055846
Human rhinovirus 89		1	1	A10937

	Human rhinovirus A 1B	pUB512	1	1	D00239
	Human rhinovirus A 2	type 2	1	1	X02316
	Human rhinovirus A 16	type 16	1	1	L24917
	Human rhinovirus B	type 14	1	1	K02121
			1	1	M16248
	Ljungan virus	87-012	1	1	AF327920
		174F	1	1	AF327921
		145SL	1	1	AF327922
	Mengo virus	medium plague	1	1	L22089
	Poliovirus	Mahoney	1	1	V01149 J02281
	Type 1	RUS-1161-96-001	1	1	AF462419
		99/056-252-14	1	1	AF462418
		Cox	1	1	AJ430385
		CHAT 10A-11	1	1	AJ416942
		DOR00013	1	1	AF405690
		DOR00041C1	1	1	AF405682
		HAI00003	1	1	AF405669
		HAI01007	1	1	AF405666
		Sabin 1	1	1	V01150 J02282 J02285 J02286 V01133
			1	1	AJ132961
			1	1	AJ132960
	Type 3	USOL-D-bac-lol 2/66.01.12	1	1	AJ293918
		Sabin 3 (Leon 12a1b)	1	1	X00596
		23127	1	1	X04468
		P3/Leon/37 (type 3)	1	1	K01392
	Type 2	W-2-p1308	1	1	D00625
		Lansing	1	1	M12197
	Perina nuda picorna-like virus	PnPV	1	1	AF323747
	Porcine enterovirus A	V13-8	1	1	AF406813
	Porcine enterovirus B	UKG/410/73	1	1	Y14459
		UKG/410/73-2	1	1	AF363453
		LP 54	1	1	AF363455
	Porcine teschovirus 1	1-F65	1	1	AJ011380
		Talfan-1	1	1	AB038528
	Sacbrood virus	Rothamstead	1	1	AF092924
			1	1	AF469603
	Simian hepatitis A virus	AGM-27	1	1	D00924
	Simian picornavirus 1	2383	1	1	AY064708
	Swine vesicular disease virus		1	1	X54521
		J173	1	1	D16364
		NET/1/92	1	1	AF268065
		H/3`76	1	1	D00435
	Theilovirus	GDVII	1	1	X56019
		BeAn 8386	1	1	M16020
		GDVII	1	1	M20562
		DA-TO	1	1	M20301
			149	**153**	
Pomovirus	Beet soil-borne virus	Ahlum-1	3	2	Z97873
		Ahlum-2		2	U64512
		Ahlum-3		3	Z66493
	Beet virus Q		3	2	AJ223596
				4	AJ223597
				3	AJ223598
	Broad bean necrosis virus		3	2	D86636
				2	D86637
				4	D86638
	Potato mop-top virus	Swedish (Sw)-1	3	2	AJ238607
		Swedish (Sw)-2		4	AJ277556
		Swedish (Sw)-3		2	AJ243719
			12	**32**	
Potyviridae	Barley mild mosaic virus	common-UK-F-1	2	1	Y10973
		common-UK-F-2		1	X90904
		Na1-1	2	1	D83408
		Na1-2		1	D83409
		Reims-M-1	2	1	L49381
		Reims-M-2		1	X82625
	Barley yellow mosaic virus	Yancheng-1	2	1	AJ132268
		Yancheng-2		1	AJ132269
		BaYMV-G-1	2	1	X69757

		German-2		1	D01099
		II-1-1	2	1	D01091
		II-1-2		1	D01092
	Bean common mosaic necrosis virus	NL-3 (necrotic)-1	1	1	U19287
		NL-3 (necrotic)-2	1	1	AY138897
	Bean common mosaic virus	blackeye-R	1	1	AJ312437
		blackeye-Y	1	1	AJ312438
	Bean yellow mosaic virus	MB4	1	1	D83749
		BYMV-S	1	1	U47033
	Brome streak mosaic virus		1	1	Z48506
	Clover yellow vein virus	No.30	1	1	AB011819
	Cocksfoot streak virus		1	1	AF499738
	Cowpea aphid-borne mosaic virus	CABMV-Z	1	1	AF348210
	Dasheen mosaic virus	M13	1	1	AJ296033
	Johnsongrass mosaic virus		1	1	Z26920
	Leek yellow stripe potyvirus	Yuhang GYH	1	1	AJ307057
	Lettuce mosaic virus	E	1	1	X97705
		0	1	1	X97704
		AF199	1	1	AJ278854
	Lily mottle virus	Sb	1	1	AJ564636
	Maize dwarf mosaic virus	Bulgaria	1	1	AJ001691
		China: Henan	1	1	AF494510
		Beijing	1	1	AY042184
	Oat mosaic virus	Cranbrook laboratory-1	2	1	AJ306718
		Cranbrook laboratory-2		1	AJ306719
	Oat necrotic mottle virus	Type-NE	1	1	AY377938
	Onion yellow dwarf virus	Yuhang	1	1	AJ510223
	Papaya leaf-distortion mosaic potyvirus		1	1	BD171712
	Papaya ringspot virus	HA	1	1	X67673
		CI	1	1	AY027810
		W	1	1	AY010722
		YK-Taiwan	1	1	X97251
		HA-Hawaii	1	1	S46722
	Pea seed-borne mosaic virus	DPD1	1	1	D10930 D01152
		Pathotype P-2-L1	1	1	AJ252242
		pathotype P-4-NY	1	1	X89997
	Peanut mottle virus	M	1	1	AF023848
	Peanut stripe virus	Infectious Clone of the Blotch	1	1	U34972
		blotch	1	1	U05771
	Pepper mottle virus	California	1	1	M96425
		Florida	1	1	AF501591
	Peru tomato mosaic virus	PPK13	1	1	AJ437280
	Plum pox virus	PPV-NAT	1	1	D13751 D00424
		M-PS	1	1	AJ243957
		D	1	1	X16415
		SK 68	1	1	M92280 X56759
		PVV-SC	1	1	X81083
	Potato virus A	B11	1	1	AJ296311
		U-2	1	1	AF543709
		U-1	1	1	AF543212
			1	1	Z21670
		TamMV	1	1	AJ131403
		U	1	1	AJ131402
		Ali	1	1	AJ131401
		Her	1	1	AJ131400
	Potato virus V	DV-42	1	1	AJ243766
	Potato virus Y	N	1	1	X12456
		LYE84.2	1	1	AJ439545
		SON41	1	1	AJ439544
		N-Egypt	1	1	AF522296
		N-605	1	1	X97895
		Hungarian	1	1	M95491
		common	1	1	U09509
		Patente	1	1	A08776
	Ryegrass mosaic virus	Danish	1	1	Y09854
		RGMV-AV (Australia-Victoria)	1	1	AF035818
	Scallion mosaic virus		1	1	AJ316084
	Sorghum mosaic virus	Xiaoshan	1	1	AJ310197
		Yuhang	1	1	AJ310198
	Soybean mosaic virus	N	1	1	D00507

		G2	1	1	S42280
		G7	1	1	AF241739
	Sugarcane mosaic virus	maize	1	1	AJ297628
		Lingpin	1	1	AJ310102
		isolate="Xiangshan" specific_host="sugar cane"	1	1	AJ310103
		Yuhang	1	1	AJ310104
		Guangdong	1	1	AJ310105
		SCMV-SD	1	1	AY149118
	Sweet potato feathery mottle virus	S	1	1	D86371
	Sweet potato mild mottle virus		1	1	Z73124
	Tobacco etch virus	HAT	1	1	M11458
			1	1	M15239
		Non-wilting	1	1	L38714
	Tobacco vein mottling virus		1	1	X04083
		S	1	1	U38621
	Turnip mosaic virus	UK1	1	1	AF169561
		japanese	1	1	D83184
		C1	1	1	AF394601
		TW	1	1	AF394602
		Tu7-CHN 12-Jing 2	1	1	AY090660
	Wheat streak mosaic virus	Sidney 81	1	1	AF057533
		type	1	1	AF285169
		El Batan 3	1	1	AF285170
		Czech	1	1	AF454454
		Turkey 1	1	1	AF454455
	Wheat yellow mosaic virus		2	1	D86634
				1	D86635
		Yangzhou	2	1	AJ131981
		Yangzhou		1	AJ131982
		C-HC-1	2	1	AF067124
		C-HC-1		1	AF041041
		Yaan	2	1	AJ239039
				1	AJ242490
	Wild potato mosaic virus	Type Isolate	1	1	AJ437279
	Yam mosaic virus	mild	1	1	AB027007
		Japanese yam 1	1	1	AB016500
		Ivory Coast	1	9	U42596
	Zucchini yellow mosaic virus	TW-TN3	1	1	AF127929 AF343979
		Singapore	1	1	AF014811
		California	1	1	L31350
		Reunion Island	1	1	L29569
			121	129	
Sequiviridae	Apple latent spherical virus		2	2	AB030940 AB030941
	Maize chlorotic dwarf virus	Tennessee (TN)	1	1	U67839
	Parsnip yellow fleck virus	P121	1	1	D14066
	Rice tungro spherical virus		1	1	M95497
		Vt6	1	1	AB064963
	Satsuma dwarf virus	S-58-1	2	2	AB009958
		S-58-2			AB009959
	Strawberry mottle virus		2	2	AJ311875 AJ311876
			10	10	
Sobemovirus	Cocksfoot mottle virus		1	4	AB040447
			1	3	Z48630
		Russia	1	4	L40905
	Lucerne transient streak virus	New Zealand	1	5	U31286
	Rice yellow mottle virus		1	4	L20893
		Nigeria	1	4	U23142
	Ryegrass mottle virus		1	4	AB040445
	Sesbania mosaic virus		1	4	AY004291
	Southern bean mosaic virus	SBMV-B(ARK)	1	4	AF055887
		SBMV-S-Arkansas	1	4	AF055888
		Bean	1	3	L34672
	Southern cowpea mosaic virus		1	4	M23021
	Subterranean clover mottle virus	p23	1	4	AF208001
	Turnip rosette virus		1	5	AY177608
			14	56	

Tetraviridae	Helicoverpa armigera stunt virus		2	1	U18246
		black mountain		2	L37299
	Nudaurelia capensis beta virus		1	2	AF102884
			3	5	
Tobamovirus	Chinese Rape Mosaic Virus		1	3	U30944
	Crucifer tobamovirus	wasabi-Shizuoka	1	4	AB017503
		wasabi-Tochigi	1	4	AB017504
	Cucumber fruit mottle mosaic virus		1	3	AF321057
	Cucumber green mottle mosaic virus	SH	1	4	D12505 D01188
		Yodo	1	4	AB015145
		watermelon	1	4	AB015146
		KW	1	4	AF417242
		KOM	1	4	AF417243
	Kyuri green mottle mosaic virus	KGMMV-C1	1	4	AJ295948
	Obuda pepper virus	Ob	1	4	D13438
	Odontoglossum ringspot virus	18kDa	1	5	X82130
		Cy-1	1	4	S83257
			1	3	U89894
		Singapore 1	1	3	U34586
	Paprika mild mottle virus	Japanese	1	4	AB089381
	Pepper mild mottle virus	S	1	4	M81413
		PMMoV-Japan	1	4	AB000709
		Ia	1	4	AJ308228
	Ribgrass mosaic virus	Shanghai	1	4	AF254924
	Tobacco mild green mosaic virus	U2-TMV	1	4	M34077 M22483
		Japanese	1	4	AB078435
	Tobacco mosaic virus	variant 1	1	6	V01408 J02415
		Rakkyo	1	4	D63809
		OM-L	1	4	X02144
		K2-Kazakhstan	1	4	Z92909
			1	4	AF165190
		Kazakh-K1	1	4	AJ243571
		tomato	1	5	AF155507
		BS35A	1	4	AJ011933
			1	5	AF273221
		Fujian	1	3	AF395127
		TMV-017	1	3	AF395128
		TMV-152	1	3	AF395129
	Tomato mosaic virus	Queensland	1	4	AF332668
		camellia	1	4	AJ417701
		S-1	1	4	AJ132845
	Turnip vein-clearing virus	OSU	1	4	U03387 L22518
	Zucchini green mottle mosaic virus	Type-ZGMMV-K	1	4	AJ295949
			40	157	
Tobravirus	Pea early browning virus	British-SP5-1	2	4	X14006
		British-SP5-2		3	X51828
	Pepper ringspot virus	CAM-1	2	4	L23972
		CAM-2		1	X03241
	Tobacco rattle virus	PPK20-1	2	4	AF166084
		PPK20-2		3	Z36974
		ORY-1	2	3	AF034622
		ORY-2		3	AF034621
			8	25	
Togaviridae	Aura virus	New World	1	2	AF126284 S78478
	Barmah Forest virus	BH2193	1	3	U73745
	Chikungunya virus	S27-African	1	2	AF369024
	Eastern equine encephalitis virus	ssp. North American	1	3	X63135 X67111
		North American antigenic variety	1	2	U01034
	Igbo Ora virus	IBH10964	1	2	AF079457
	Mayaro virus		1	3	AF237947
	O'nyong-nyong virus	Gulu	1	2	M20303 M33999
		SG650	1	2	AF079456
	Ross River virus	NB5092	1	2	M20162
		Sagiyama	1	3	AB032553
	Rubella virus		1	2	M15240 M18901 M32735
		Surkova	1	2	AF435866
		Ulrike	1	2	AF435865
		TO-336 wild progenitor	1	2	AB047330
		TO-336 vaccine	1	2	AB047329
		Cendehill	1	2	AF188704
		RA 27/3	1	2	L78917
	Salmon pancreas disease virus	F93125	1	2	AJ316244
	Semliki forest virus		1	2	X04129 J02361 J02362 L00018 V01400 V01401
		L10	1	2	AY112987
	Sindbis virus	hrsp-wild-type	1	3	J02363 J02364 J02365 J02366 J02367 VERSION J02363.1 GI:334100
		Girdwood S.A.	1	2	U38304
		S.A.AR86	1	2	U38305
		XJ-160	1	2	AF103728
		YN87448	1	3	AF103734
		SW6562	1	3	AF429428
	Ockelbo virus	Edsbyn 82-5	1	2	M69205
	Sleeping disease virus		1	2	AJ316245
	Venezuelan equine encephalitis virus	P676	1	3	L04653 L00931
		P676	1	2	AF375051
		V198-IC	1	2	U55342
		CPA152-IE	1	2	AF448535
		OAX131-IE	1	2	AF448536
		CPA201-IE	1	2	AF448537
		OAX142-IE	1	2	AF448538
		80U76-IE	1	2	AF448539
		PMCHo5-IC	1	2	U55345
		6119-IC	1	2	U55347
		3908-IC	1	2	U55350
		83U434	1	2	U55362
		SH3	1	2	U55360
		243937	1	2	AF004459
		66457	1	2	AF004472
		66637	1	2	AF004458
		donkey/Trinidad/1943	1	3	L01442
		ZPC738-ID	1	2	AF100566
		Cabassou CaAr 508	1	2	AF075259
		AG80-663	1	2	AF075258
		78V-3531	1	2	AF075257
		Pixuna BeAr 35645	1	2	AF075256
		71D-1252	1	2	AF075255
		Tonate CaAn 410d	1	2	AF075254
		Mucambo BeAn 8	1	2	AF075253
		Mena II	1	2	AF075252
		Everglades Fe3-7c	1	2	AF075251
		equine/Texas/1971	1	2	AF069903
		68U201-IE	1	2	U34999
		TC-83	1	2	L01443
		Trinidad	1	2	J04332
		3880	1	3	L00930
	Western equine encephalomyelitis virus	71V-1658	1	2	AF214040 AF143811
			62	134	
Tombusviridae	Artichoke mottled crinkle virus	AMCV-Bari Dr.Gallitelli	1	5	X62493
	Beet black scorch virus		1	6	AF452684
	Cardamine chlorotic fleck virus	CCFV-CL	1	4	L16015
	Carnation Italian ringspot virus		1	5	X85215
	Carnation mottle virus	shanghai	1	4	AF192772
			1	6	X02986
	Carnation ringspot virus		2	4	L18870
				1	M88589
	Cowpea mottle virus		1	6	U20976 U07227
	Cucumber Bulgarian virus		1	5	AY163842
	Cucumber necrosis virus		1	5	M25270
	Cymbidium ringspot virus		1	5	X15511
	Dianthovirus RVX1		1	5	AB033715
	Galinsoga mosaic virus		1	5	Y13463
	Hibiscus chlorotic ringspot virus		1	7	X86448
	Japanese iris necrotic ring virus		1	6	D86123
	Johnsongrass chlorotic stripe mosaic virus		1	5	AJ557804
	Leek white stripe virus		1	5	X94560
	Maize chlorotic mottle virus		1	6	X14736
	Melon necrotic spot virus		1	6	M29671

	Oat chlorotic stunt virus		1	4	X83964
	Olive latent virus 1	citrus	1	5	X85989
	Panicum mosaic virus	Kansas	1	6	U55002
	Pea stem necrosis virus	specific_host="Pisum sativum"	1	5	AB066951
	Pear latent virus		1	5	AY100482
	Pelargonium flower break virus	MZ10	1	6	AJ514833
	Pelargonium necrotic spot virus		1	5	AJ607402
	Pothos latent virus	pigeonpea	1	5	AJ243370
	Red clover necrotic mosaic virus	Australia-1	2	4	J04357 M24621
		Australia-2		1	X08021
		Canada-1	2	4	AB034916
		Canada-2		1	AB034917
	Saguaro cactus virus		1	8	U72332
	Sweet clover necrotic mosaic virus	59-2	2	3	L07884
		59-2		1	S46028
	Tobacco necrosis virus A	TNV-A-FM1B	1	5	M33002
	Tobacco necrosis virus D	D-Hungarian-1	1	6	U62546
		D-Hungarian-2	1	6	D00942
	Tomato bushy stunt virus	cherry	1	5	M21958 M31019
		statice	1	5	AJ249740
		pepper	1	5	U80935
	Turnip crinkle virus		1	5	M22445
			42	201	
Tymovirus	Chayote mosaic tymovirus	Chayote (Sechium edule)	1	3	AF195000
	Eggplant mosaic virus		1	3	J04374
	Erysimum latent virus		1	3	AF098523
	Grapevine fleck virus	MT48	1	4	AJ309022
	Kennedya yellow mosaic virus	Jervis Bay	1	3	D00637
	Maize rayado fino virus	Costa Rican	1	2	AF265566
	Oat blue dwarf virus		1	2	U87832
	Ononis yellow mosaic virus	OYMTV-Tin	1	3	J04375
	Physalis mottle virus		1	3	Y16104
	Poinsettia mosaic virus		1	2	AJ271595
	Turnip yellow mosaic virus		1	3	X07441
		Blue Lake	1	3	AF035403
			1	3	X16378
		Club Lake	1	3	J04373
			3	40	
Umbravirus	Carrot mottle mimic virus	Australian	1	4	U57305
	Groundnut rosette virus	MC1	1	4	Z69910
	Pea enation mosaic virus-2		1	4	U03563 S53233
	Tobacco bushy top virus	Baoshan	1	4	AF402620 AF431890
			4	16	
Leviviridae	Bacteriophage AP205		1	4	AF334111
	Bacteriophage M11		1	4	AF052431
	Bacteriophage SP		1	4	X07489
	Enterobacteria phage fr		1	4	X15031
	Enterobacteria phage GA		1	3	D10027 D00046 X03869
	Enterobacteria phage KU1		1	4	AF227250
	Enterobacteria phage MX1		1	4	AF059242
	Enterobacteria phage NL95		1	4	AF059243
	Enterobacterio phage MS2		1	4	J02467 M24961 V00642
	Pseudomonas phage PP7		1	4	X80191
			10	39	

Proteomas Virales de Cadena Sencilla de RNA sentido negativo (ssRNA-)

Bunyaviridae	Andes virus	Chile-9717869-L	3	1	AF291704
		Chile-9717869-M		1	AF291703
		Chile-9717869-S		1	AF291702
	Bunyamwera virus		3	1	X14383
				1	M11852
				2	D00353
	Dugbe virus	ArD44313-L	3	1	U15018
		ArD44313-M		1	M94133
		ArD44313-S		1	AF434161
	Hantaan virus	76-118-L	3	1	X55901
		76-118-M		1	M14627
		76-118-S		1	M14626
	Impatiens necrotic spot virus	NL-07-L	3	1	X93218
		NL-07-M		2	M74904
		NL-07-S		2	X66972 S40057
	La Crosse virus	Human/78-L	3	1	AF528165
		Human/78-M		1	AF528166
		Human/78-S		2	AF528167
		La Crosse/original-L	3	3	U12396
		74-32813		1	D10370 D00202
				2	K00610
	Peanut bud necrosis virus		3	1	AF025538
				2	U42555
				2	U27809
	Rift Valley fever virus	ZH548-M12	3	1	X56464
		ZH-501		1	M11157
		MM12		2	X53771
	Sin Nombre virus	NMH10-L	3	1	L37901
		NMH10-M		1	L25783
		NMH10-S		1	L25784
		NM R11-L	3	1	L37902
		NM R11-M		1	L37903
		NM R11-S		1	L37904
	Tomato spotted wilt virus	BR-01 (CNPH1)	3	1	D10066 D01230
				1	S48091
		CPNH9		2	D00645
	Uukuniemi virus	S23	3	1	D10759
		S23		1	M17417
		S23		2	M33551
	Watermelon spotted wilt virus	Taiwan-L	3	1	AF133128
		Taiwan-M		2	U75379
		Taiwan-Tospo-W		2	U78734
			42	56	
Orthomyxoviridae	Influenza A virus	A/PR/8/34	8	1	V00603 J02153
		A/Puerto Rico/8/34		1	J02151
		A/PR/8/38		1	V01106 J02152
		A/PR/8/34(HON1)		1	V01088 J02143
		A/Puerto Rico/8/34		1	J02147
		A/Puerto Rico/8/34		1	J02146
		A/PR/8/34		2	V01099 J02145
		A/Puerto Rico/8/34		2	J02150
		A/Puerto Rico/8/34/Mount Sinai(H1N1)	8	1	AF389115
		A/Puerto Rico/8/34/Mount Sinai(H1N1)		1	AF389116
		A/Puerto Rico/8/34/Mount Sinai(H1N1)		1	AF389117
		A/Puerto Rico/8/34/Mount Sinai(H1N1)		1	AF389118
		A/Puerto Rico/8/34/Mount Sinai(H1N1)		1	AF389119
		A/Puerto Rico/8/34/Mount Sinai(H1N1)		1	AF389120
		A/Puerto Rico/8/34/Mount Sinai(H1N1)		2	AF389121
		A/Puerto Rico/8/34/Mount Sinai(H1N1)		2	AF389122
		A/Hong Kong/1073/99-H9N2	8	1	AJ404630
		A/Hong Kong/1073/99-H9N2		1	AJ404634
		A/Hong Kong/1073/99-H9N2		1	AJ404637
		A/Hong Kong/1073/99-H9N2		1	AJ404626

		A/Hong Kong/1073/99-H9N2		1	AF255742
		A/Hong Kong/1073/99-H9N2		1	AJ404629
		A/Hong Kong/1073/99-H9N2		2	AJ278646
		A/Hong Kong/1073/99-H9N2		2	AF256176
	Influenza B virus	B/Lee/40	8	1	M14880
		B/Lee/40		1	AF101982
		B/Lee/40		1	AF102017
		B/Lee/40		1	K00423
		B/Lee/40		1	K01395
		B/Lee/40		2	J02095
		B/Lee/40		1	J02094
		B/Lee/40		2	J02096
		B/Memphis/12/97	8	1	AY260942
		B/Memphis/12/97		1	AY260943
		B/Memphis/12/97		1	AY260944
		B/Memphis/12/97		1	AY260945
		B/Memphis/12/97		1	AY260946
		B/Memphis/12/97		2	AY260947
		B/Memphis/12/97		2	AY260941
		B/Memphis/12/97		1	AY260948
		B/Memphis/12/97-MA	8	1	AY260949
		B/Memphis/12/97-MA		1	AY260950
		B/Memphis/12/97-MA		1	AY260951
		B/Memphis/12/97-MA		1	AY260952
		B/Memphis/12/97-MA		1	AY260953
		B/Memphis/12/97-MA		2	AY260954
		B/Memphis/12/97-MA		2	AY260955
		B/Memphis/12/97-MA		1	AY260956
			48	**60**	
Arenaviridae	Lassa virus	1-LCMV-LASV	2	2	U73034
		Josiah		2	J04324
	Lymphocytic choriomeningitis virus	Armstrong 53b-L	2	2	J04331
		Armstrong 53b-S		2	M20869
	Tacaribe virus	T.RVL.II 573	2	2	J04340 M33513
				2	M20304 M65834
	Guanarito virus	INH-95551-L	2	2	AY358024
		INH 95551-S		2	AY129247
	Junin virus	XJ13-L	2	2	AY358022
		XJ13-S		2	AY358023
	Machupo virus	Carvallo-L	2	2	AY358021
		Carvallo-S		2	AY129248
			12	**24**	
Paramyxoviridae					
	Avian paramyxovirus 6	APMV-6/duck/Taiwan/Y1/98	1	7	AY029299
Paramyxovirinae	Bovine parainfluenza virus 3	Kansas/15626/84	1	6	AF178654
		Shipping Fever	1	6	AF178655
	Canine distemper virus	Onderstpoort	1	7	AF014953 L13194 L13195
		Onderstepoort-2	1	6	AF305419
		Onderstepoort (OS)	1	6	AF378705
		A75/17	1	6	AF164967
	Fer-de-lance virus	ATCC VR-895	1	8	AY141760
	Goose paramyxovirus SF02	SF02	1	6	AF473851
	Hendra virus		1	8	AF017149
	Human metapneumovirus	00-1	1	9	AF371337 AF371365 AF371366 AF371367 AF371345 AF371355 AF371364
	Human parainfluenza virus 1 strain Washington/1964	Washington 1964	1	10	AF457102
	Human parainfluenza virus 2		1	7	X57559
	Human parainfluenza virus 3	GP	1	8	AB012132
		910N	1	7	D84095
		JS	1	7	U51116
	Measles virus	Ichinose-B95a	1	7	AB016162
		Edmonston B	1	8	Z66517
		Edmonston-AIK-C	1	7	AB046218
		Edmonston (Schwarz vaccine)	1	8	AF266291
		Edmonston (Zagreb vaccine)	1	8	AF266290
		Edmonston (Rubeovax vaccine)	1	8	AF266289
		Edmonston-wild-type	1	8	AF266288
		Edmonston (Moraten vaccine)	1	8	AF266287
		Edmonston (AIK-C vaccine)	1	8	AF266286
		Ichinose-Vero	1	7	AB032157
		wild-type-9301B	1	7	AB012948
		9301V	1	7	AB012949
		Edmonston	1	8	K01711 X16565
		AIK-C	1	7	S58435
	Mumps virus	Miyahara	1	8	AB040874
		Jeryl Lynn	1	9	AF201473
		Jeryl-Lynn	1	9	AF345290
		88-1961-H	1	7	AF467767
		87 1005-clinical	1	9	AF314562
		Biken	1	9	AF314561
		87 1004-clinical	1	9	AF314560
		Smith-Kline Beecham live-attenuated vaccine*	1	9	AF314559
			1	9	AF314558
		Jeryl-Lynn	1	9	AF338106
		Glouc1/UK96	1	8	AF280799
	Nipah virus		1	8	AF212302
		UMMC2	1	8	AY029768
		UMMC1-CSF	1	8	AY029767
			1	7	AF376747
	Newcastle disease virus	B1	1	6	AF309418
		B1-Takaaki	1	6	AF375823
	-	ZJ1	1	6	AF431744
		LaSota	1	6	AF077761
			1	6	Y18898
	Sendai virus	Ohita-M1	1	9	AB005795
		Ohita-MVC11	1	9	AB005796
		Hamamatsu-E0	1	8	AB039658
		Hamamatsu	1	8	AB065187
		mutant T-5 revertant	1	6	M69046
		F1-R	1	6	M30203
		Z	1	6	M30202 M19661 M76992
		mutant ts-f1	1	6	M30204
	Simian parainfluenza virus 5	W3A	1	8	AF052755
	Tioman virus		1	8	AF298895
	Tupaia paramyxovirus		1	8	AF079780
			61	**458**	
Paramyxoviridae	Bovine respiratory syncytial virus	ATue51908	1	11	AF092942
Pneumovirinae		ATCC51908-BRSV A51908	1	11	AF295543
	Human respiratory syncytial virus	B1-wildtype	1	11	AF013254
		S2	1	10	U39662
		A2	1	12	U63644
		A2	1	12	AF035006
		A2	1	12	U50363
		A2	1	12	U50362
		B1	1	9	AF013255
	Respiratory syncytial virus	S2 ts1C	1	10	U39661
			10	**110**	
Rhabdoviridae	Australian bat lyssavirus	insectivorous	1	5	AF081020
			1	5	AF418014
	Bovine ephemeral fever virus	BB7721	1	12	AF234533
	Hirame rhabdovirus	CA 9703	1	6	AF104985
	Infectious hematopoietic necrosis virus	WRAC	1	6	L40883

			1	6	X89213
	Northern cereal mosaic virus		1	9	AB030277
	Rabies virus	PV	1	5	M13215 M21634
		SAD B19	1	5	M31046
		RC-HL	1	5	AB009663
		Nishigahara	1	5	AB044824
	Rice yellow stunt virus		1	7	AB011257 D89654 D87843 D87844 AB002822 AB010258 AB003092
	Snakehead rhabdovirus		1	6	AF147498
	Sonchus yellow net virus		1	6	L32603
	Spring viremia of carp virus	ATCC VR-1390	1	5	U18101
		Fijan Reference-VR-1390	1	5	AJ318079
	Vesicular stomatitis Indiana virus	Indiana-wild type	1	5	J02428 J02430 J02431 J02432 J02434 J02435 J02436 J02437 J02438 K00519 K00520 K00625 K01068 K01069 K01070 K01638 K01639
		94GUB-1994	1	5	AF473866
		85CLB-Indiana1	1	5	AF473865
		98COE-Indiana 1-1998	1	5	AF473864
	Viral hemorrhagic septicemia virus	Fl3	1	6	Y18263
		96-43-Atlantic herring	1	6	AF143862
		14-58-3-French	1	6	AF143863
		Hededam	1	6	Z93412
		Cod Ulcus	1	6	Z93414
		07_71	1	6	AJ233396
			26	154	
Bornaviridae	Borna disease virus	V	1	5	U04606
			1	5	L27077
		V/FR	1	6	AJ311521
		He/80/FR	1	6	AJ311522
		H1766	1	6	AJ311523
		No/98	1	6	AJ311524
		CRP3A	1	6	AY114161
		CRP3B	1	6	AY114162
		CRNP5	1	6	AY114163
		brain S589	1	5	AY066023
			10	57	
Filoviridae	Marburg virus	Popp	1	7	Z29337
		Musoke	1	7	Z12132 S55429
	Reston Ebola virus	Pennsylvania	1	8	AF522874
		Reston	1	8	AB050936
	Zaire Ebola virus	Mayinga	1	9	AF086833
		Mayinga-Zaire	1	8	AF272001
		Mayinga-Zaire	1	8	AF499101
			7	55	
Tenuivirus	Rice grassy stunt virus	IRRI-1	6	2	AB009656
		IRRI-2		2	AB010376
		IRRI-3		2	AB010377
		IRRI-4		2	AB010378
		IRRI-5		2	AB000403
		IRRI-6		2	AB000404
		South Cotabato-1	6	2	AB032180
		South Cotabato-2		2	AB023777
		South Cotabato-3		2	AB029694
		South Cotabato-4		2	AB023778
		South Cotabato-5		2	AB023779
		South Cotabato-6		2	AB023780
		shaxian-1	6	1	AF509470
		shaxian-2		2	AF511072
		shaxian-3		2	AF397468
		shaxian-4		2	AF290946
		shaxian-5		2	AF290947
		shaxian-6		2	AF287949
	Rice stripe virus	T-1	4	1	D31879
		T-2		2	D13176
		T-3		2	X53563
		T-4		2	D10979 D01164
		HZ-1	4	1	AY186788
		HZ-2		2	AY186789
		HZ-3		2	AF508865
		HZ-4		2	AF513505
			26	49	
Ophiovirus	Mirafiori lettuce virus	LS301-O-lettuce-1	4	2	AF525933
		LS301-O-lettuce-2		2	AF525934
		LS301-O-lettuce-3		1	AF525935
		LS301-O-lettuce-4		2	AF525936
			4	7	

Proteomas Virales de DNA y RNA con Reverso Transcriptasa (Retrovirus)

Caulimoviridae	Banana streak virus		1	3	AJ002234
	Blueberry red ringspot virus		1	8	AF404509
	Cacao swollen shoot virus		1	5	L14546
	Carnation etched ring virus		1	6	X04658
	Cauliflower mosaic virus	M13mp7	1	7	V00140 J02046
		Xinjiang	1	7	AF140604
		B29	1	6	X79465
			1	6	V00141 J02048
		BBC	1	6	M90542
		NY8153	1	6	M90541
		Cabb-D/H	1	8	M10376 J02047
		CMV-1	1	6	M90543
	Cassava vein mosaic virus		1	5	U20341
			1	5	U59751
	Cestrum yellow leaf curling virus		1	7	AF364175
	Citrus yellow mosaic virus		1	6	AF347695
	Commelina yellow mottle virus		1	3	X52938
	Figwort mosaic virus		1	7	X06166
	Kalanchoe top-spotting virus		1	3	AY180137
	Mirabilis mosaic virus		1	7	AF454635
	Peanut chlorotic streak caulimovirus	K1	1	4	U13988
	Petunia vein clearing virus		1	2	U95208
	Rice tungro bacilliform virus	Chainat	1	4	AF220561
			1	4	D10774 D01028 D01149 D10236
		West Bengal	1	4	AJ314596
		Serdang	1	4	AF076470
		RTBV-Ic	1	4	AF113832
		RTBV-G2	1	4	AF113831
		RTBV-G1	1	4	AF113830
		Philippines	1	4	X57924
	Soybean chlorotic mottle virus		1	9	X15828
	Strawberry vein banding virus		1	6	X97304
	Sugarcane bacilliform virus	Ireng Maleng	1	3	AJ277091 M89923
	Taro bacilliform virus		1	3	AF357836
	Tobacco vein-clearing virus		1	4	AF190123
			35	180	
Hepadnaviridae	Arctic ground squirrel hepatitis B virus		1	7	U29144
	Duck hepatitis B virus	DHBVQCA34	1	7	X60213
		Indiana	1	5	AF493986
			1	5	M60677
			1	3	M32990
			1	3	M32991
			1	3	M21953
			1	3	K01834
		Indian-IDHBV	1	3	X74623
		Australian DHBV	1	5	AJ006350
			1	3	AF404406
		Alberta-ALTA-16	1	4	AF047045
		RGHBV	1	3	M95589
	Ground squirrel hepatitis virus	27	1	4	K02715
	Hepatitis B virus		1	7	AF536524
		FG	1	5	AF498266 AF258595 AF258596
			1	7	AF537372
			1	6	AF537371
			1	6	AF533983
			1	7	AF536524
			1	4	AF286594
			1	7	AY128092
		1116Sal-adw4-F	1	4	AY090461

		LAS2523-adw4-H	1	4	AY090460
		7768H-adw4-F	1	4	AY090459
		70H-adw4-F	1	4	AY090458
		2928Nic-adw4-H	1	4	AY090457
		14-94-adw3-D	1	4	AY090453
		ZZ9-adw3-D	1	4	AY090452
		HBV-Th55	1	6	AB074756
			1	5	AB064316
		adw2-Wong-B	1	4	U87747 AF373066
			1	5	AB064313
			1	5	AB064312
			1	5	AB064315
			1	5	AB064314
			1	5	AB064310
		HBV AN29	1	4	AB049610
		HBV AN28	1	4	AB049609
		adyw-7782-A' (A40)	1	4	U87746 AF373065
		adw-7983-A' (A20)	1	4	U87742 AF364333
		235/01-G	1	6	AF405706
		G25	1	4	AY077735 AF274497
		G26	1	4	AY077736 AF274498 AF275379
		3c	1	7	AF479684
		2e	1	7	AF473543
			1	7	AJ344117
		HBsAg-X27-	1	6	AJ344116
		HBsAg-X27-	1	7	AJ344115
		adw	1	4	AF462041
			1	5	AF461043
		adw-FH10-c	1	4	AF461362
		adr-FH3Q-c	1	4	AF461361
		adw-FH14-2-b	1	4	AF461360
		adr-c16-HBsAg-c+	1	4	AF461359
		adrFH8-c	1	4	AF461358
		adr-FH5-c	1	4	AF461357
		adr-C733-HBsAg-c+*	1	4	AF458665
		adr-C84-HBsAg-c+	1	4	AF458664
		CB0376	1	5	AF305327
		adw-FH4-C	1	4	AY066028
		ayw1-B	1	4	AY033073
		G376-AS	1	4	AF384372
		Tibet127-C/D	1	4	AY057948
		Tibet705	1	4	AY057947
		black African-adw2-84-PM-HBsAg+ HBeAg+ A	1	4	AF297625
		adw2-80-656_HBsAg+ HBeAg+ A	1	4	AF297624
		adw2-83-KR-HBsAg+ - A	1	4	AF297623
		adw2-79-1714_HBsAg+ HBeAg+ A	1	4	AF297622
		adw2-76-DL-HBsAg+ HBeAg- A	1	4	AF297621
		C-adw-SK619	1	4	AB033553
		C-adw-SK494	1	4	AB033552
		C-adw-SK300	1	4	AB033551
		FMC#15-adr-C	1	4	AF411412
		FMC#97-adr-C	1	3	AF411411
		FMC#20-adr-C	1	4	AF411410
		FMC#16-adr-C	1	4	AF411409
		FMU#14-adr-C	1	3	AF411408
		WLAID	1	4	AF182805
		WLAP10	1	4	AF182804
		wenlinA1	1	4	AF182803
		wenlin61	1	4	AF182802
		adw1-57-A/D	1	4	AF297620
		adw2-41-A/D	1	4	AF297619
		T-1858-975484-Thailand-C	1	7	AF330110
		C-1858-ea8-C	1	7	AF223961
		C-1858-ea7-C	1	7	AF223960
		C-1858-ea6-C	1	7	AF223959
		C-1858-ea5-C	1	7	AF223958
		C-1858-ea4-C	1	7	AF223957
		C-1858-ea3-C	1	7	AF223956
		C-1858-ea2-C	1	7	AF223955
		C-1858-ea1-C	1	7	AF223954
		X104-114	1	6	AJ309371
		X104-110	1	7	AJ309370
		X104-108	1	6	AJ309369
		Chinese-C	1	6	AY040627
		HBV VietF-6	1	4	AB031267
		HBV VietF-3	1	4	AB031266
		HBV VietF-2	1	4	AB031265
		HBV VietA-3	1	4	AB031262
	Heron hepatitis B virus		1	4	M22056
	Orangutan hepadnavirus	Somad	1	4	AF193863
		Papa	1	4	AF193864
	Stork hepatitis B virus	STHBV-1	1	3	AJ251934
			1	3	AJ251935
		STHBV-16	1	3	AJ251936
		STHBV-21	1	3	AJ251937
	Woodchuck hepatitis B virus		1	4	M18752
			1	4	M19183
			1	3	J04514
			1	3	J02442
	Woodchuck hepatitis virus 2		1	4	M11082
	Woolly monkey hepatitis B Virus		1	5	AF046996
			114	**530**	
Retroviridae	Avian carcinoma virus		1	1	M14008
Alpharetrovirus		MH2	1	2	K02082
	Avian leukosis virus	RSA	1	3	M37980
		HPRS-103 (subgroup J)	1	2	Z46390
		Endogenous-1	1	3	AY013303
		Endogenous-3	1	2	AY013304
		ADOL-7501-J	1	3	AY027920
	Avian myelocytomatosis virus		1	1	AF033809
			1	2	J02013 V00871
	Avian sarcoma virus		1	2	M10455
		PR2257T	1	1	X51863
		CT10	1	1	Y00302
	Fujinami sarcoma virus		1	1	AF033810
			1	1	J02194 K01827 K01828
	Rous sarcoma virus		1	4	AF033808
		Schmidt-Ruppin-D	1	2	D10652
		Schmidt-Ruppin-D	1	4	AF052428
		Prague	1	5	J02342 J02021 J02343
		Prague C (Pr-C)	1	1	V01197
	Y73 sarcoma virus		1	2	J02027
			20	**43**	
Retroviridae	Mason-Pfizer monkey virus		1	4	AF033815
Betaretrovirus			1	3	M12349
	Simian retrovirus 2	D2/RHE/OR	1	4	AF126467
			1	3	M16605
		D2/RHE/OR/V1	1	4	AF126468
	Simian type D virus 1		1	5	U85505
			1	5	M11841
	Mouse mammary tumor virus		1	5	AF033807
		BRS	1	5	M15122
	Jaagsiekte sheep retrovirus Ovine pulmonary adenocarcinoma virus		1	5	M80216
			1	4	A27950
		JSRV21	1	5	AF105220
		JSRV	1	5	AF357971
	Exogenous mouse mammary tumor virus	C3H/HeN	1	5	AF228552
		C3H/HeJ	1	5	AF228551
	Endogenous mouse mammary tumor virus	C3H/HeN	1	5	AF228550
	Enzootic nasal tumour virus of goats		1	4	AY197548
			16	**72**	

Retroviridae	Bovine leukemia virus		1	6	AF033818
Deltaretrovirus		B19-Holstein-Argentina	1	5	AF257515
		Australian	1	3	D00647
			1	3	K02120
	Human T-lymphotropic virus 1		1	6	AF033817
		Caribbean	1	4	D13784 D00294
		WHP	1	4	AF259264
			1	3	J02029 M33896
		EBV-HTLV-VII	1	6	AF139170
		RKI3-Ger	1	3	AF042071
		ATL-YS	1	6	U19949
		BOI	1	5	L36905
	Human T-lymphotropic virus 2		1	6	M10060
			1	3	AF412314
		Ib96-a	1	5	AF326584
		RP329-a	1	5	AF326583
		SP-WV	1	5	AF139382
		G2-Guahibo Indian	1	6	AF074965
		Male-Gab-b	1	4	Y13051
		G12-Guaymi Indian	1	4	L11456
	Simian T-lymphotropic virus 1	Tan90-Tantalus	1	4	AF074966
	Simian T-lymphotropic virus 2	PP1664-STLV-2PP1564	1	7	Y14570
			1	3	U90557
	Simian T-lymphotropic virus 3		1	5	Y07616
		CTO-604	1	5	AF391797
			25	**116**	
Retroviridae	Bovine syncytial virus		1	5	U96514 L26493 L10921
Spumavirus	Equine foamy virus		1	5	AF201902
	Feline foamy virus	FUV	1	4	AJ223851
			1	5	Y06851
	Human foamy virus		1	5	Y07725
	Human spumaretrovirus		1	7	U21247
			1	2	AF033815
	Simian foamy virus	SFVcpz	1	6	U04327
		HU6	1	5	AF525404
			9	44	
Retroviridae	Abelson murine leukemia virus		1	3	AF033812
Gammaretrovirus			1	3	J01998 J01999 K00016 K00017 K00018 K01394
	Feline leukemia virus	Rickard-A	1	2	AF052723
		FeLV-FAIDS-61E	1	2	M18247 M19392
	Friend murine leukemia virus	FB29	1	3	Z11128
		FrCS-A8F5	1	3	D88386
		PVC-211	1	3	M93134 M81185
			1	2	X02794 J02192 M12528 M19209
	Friend spleen focus-forming virus		1	2	K00021
	Gibbon ape leukemia virus	SEATO-SF	1	3	M26927
		X-HUT78	1	3	U60065
	Moloney murine sarcoma virus		1	4	AF033813
			1	2	J02266
	Moloney murine leukemia virus		1	3	AF033811
			1	4	AF462057
			1	2	J02255 J02256 J02257 M76668
	Murine leukemia virus	SL3-3	1	3	AF169256
		SRS-19-6	1	3	AF019230
			1	2	AF221065
		Duplan MuLV	1	5	X14576
		MCF1233	1	3	U13766
		RadLV/VL3(T+L+)	1	2	K03363 M18449
	Murine osteosarcoma virus		1	2	AF033814
		FBR	1	1	X03347
			1	1	K02712
	Murine sarcoma virus	124	1	4	V01185 J02263
	Murine type C retrovirus		1	2	X94150
	Mus dunni endogenous virus		1	4	AF053745
	Porcine endogenous retrovirus	C	1	1	AJ133818
		PERV-17	1	2	AY099324
		PERV-60	1	2	AY099323
			1	3	AY056035
			1	3	AJ279057
			1	3	AJ279056
		Bac-PERV-A(151810)	1	3	AF435967
		Bac_PERV-A(463H12)	1	3	AF435966
		C-33	1	1	AJ133816
		C-42	1	1	AJ133817
	Rauscher murine leukemia virus	RV-1	1	3	U94692
	Woolly monkey sarcoma virus	D-SMRV-HLB	1	5	M23385
			1	4	V01201 J02394 J02396 J02397
			41	**110**	
Retroviridae	Bovine immunodeficiency virus	HXB3	1	5	M32690
Lentivirus	Caprine arthritis-encephalitis virus	Clements	1	6	M33677
		1GA	1	6	AF322109
	Equine infectious anemia virus		1	4	AF033820
		vaccine	1	6	AF327878
		Liaoning	1	6	AF327877
			1	6	AF247394
		V26	1	6	AB008197
		V70	1	6	AB008196
			1	6	AF028232
			1	6	AF028231
			1	6	AF016316
			1	4	M16575 K03334 M11337 M14855
		Malmquist of Wyoming	1	4	M87581
	Feline immunodeficiency virus	Petaluma-34TF10	1	7	M25381 M25729
		FIV-Oma	1	4	U56928
		USIL2489_78	1	4	U11820
		PPR-San Diego	1	7	M36968
	Human immunodeficiency virus 1	reference	1	9	AF033819
		X558-G	1	9	AF423760
		X477-CRF14_BG	1	9	AF423759
		X475-CRF14_BG	1	9	AF423758
		X421-CRF14_BG	1	9	AF423757
		X397-CRF14_BG	1	9	AF423756
		X254-Portugal-CRF14_BG	1	9	AF423755
		X138-CRF14_G	1	9	AF450098
	Human immunodeficiency virus 2	BEN-MK[2.6]	1	9	M30502
		7312a-JK	1	9	L36874
		Ab96	1	9	AF208027 U67390 U67391 U67392 U67393 U67394
		D194	1	9	J04542
		HIV-2ALI-ALI	1	9	AF082339
		EHO-AIDS	1	9	U27200 L14545
		CAM 2	1	9	D00835
		ALT-D205	1	2	X61240 X16109 SOLO TIENE DOS PROTEÍNAS
		HIV2BEN	1	8	U38293
		ROD isolate	1	7	X05291
		HIV-2/ST	1	9	M31113
		D194	1	8	X52223
		ROD	1	9	M15390
		FG	1	9	J03654
		SBLISY	1	9	J04498
		MDS	1	8	Z48731
		HIV-2 KR	1	9	U22047
		2UC1	1	9	L07625
		GH-1	1	9	M30895 D00477
	Jembrana disease virus	Tabanan/B7	1	9	U21603
	Ovine lentivirus	SA-OMVV	1	5	M34193
			1	7	M31646
	Primate T-lymphotropic virus 3	CTO-604	1	5	AF391797
			1	5	Y07616
	Simian immunodeficiency virus	677-gri-1	1	9	M58410
		SIVgsn-99CM166	1	9	AF468659
		SIVgsn-99CM71	1	9	AF468658
		STLV-III(MAC)	1	8	Y00277 M16403
		GB1	1	8	M27470 X15781
		SIVsmSL92b	1	9	AF334679
		SIVmnd14cg	1	8	AF328295
		SIV/17E-Br_SIVmac239	1	9	AY033233

		SIVmac239	1	9	AY033146
			1	8	M66437
		SIVmac	1	7	M76764
		Mm251	1	8	M19499 M15897 M16125 M24614 Y00294
		SIVcpzant	1	9	U42720
		syk173/1.2	1	8	L06042
		SIVCPZ-CAM5	1	9	AJ271369
		SIVmac32H	1	11	D01065
		SIVcpz-Cam3	1	8	AF115393
		SIVsun	1	8	AF131870
		TY0-1	1	8	X07805
		US-Marilyn	1	6	AF103818
		SIVlhoest	1	8	AF075269
		smmPGm	1	9	AF077017
		tantalus-1	1	8	U58991
		SIVsmE543	1	9	U72748
		SIVMne027	1	9	U79412
		PBj14-mangabey	1	8	L03295
		F236	1	6	X14307
			1	9	M80193
		155	1	8	M29975
		STM	1	9	M83293
			1	9	M32741
		AGM3	1	8	M30931
		smmPBj/bc13	1	9	L09211
		SMMPBj14	1	9	M31325
		SMM9	1	9	M80194
		smmPBj/bc13	1	9	L09213
		smmPBj/bc13	1	9	L09212
		SIVagmSAB-1/MU8	1	8	U04005
	Chimpanzee immunodeficieny virus	SIV(cpz)	1	8	X52154
	Simian-Human immunodeficiency virus	SHIV-89.6	1	10	AF038398
		SHIV-4; HXBc2	1	10	AF038399
		SHIV-89.6P	1	10	U89134
			1	10	AF041850
		SHIV-C2/1	1	10	AF217181
		1B3-1.4	1	9	AF465242
	Visna virus	kv1772	1	6	L06906
		Icelandic-LV1-1-1514	1	4	M51543
		KV1772	1	6	S55323
			1	4	A15114
		Icelandic 1514	1	5	M60610 M37977 M37978 M60609
		Icelandic 1514-LV1-1	1	3	M10608 M18039
			101	777	
Retroviridae	Snakehead retrovirus		1	7	U26458
Epsilonretrovirus	Walleye dermal sarcoma virus		1	6	AF033822
			1	5	L41838
	Walleye epidermal hyperplasia virus type 1		1	6	AF133051
	Walleye epidermal hyperplasia virus type 2		1	6	AF133052
			5	30	
Retroviridae	Avian endogenous retrovirus EAV-HP	EAV-HP1	1	1	AJ238124
Inclasificados		EAV-HP2	1	1	AJ238125
			2	2	
TOTALES			2 304	34 253	

Tabla con los 2 304 proteomas completamente secuenciados virales colectados manualmente del National Center for Biotechnology Information (NCBI: http://www.ncbi.nlm.nih.gov/ **).**

Simbología:

Grupo viral: Categorías taxonómicas, incluye clase, familia, género o grupo según el caso (datos obtenidos del **ICTV**: http://www.ncbi.nlm.nih.gov/ICTV/).

Virus: Nombre científico del virus, se tomó el más comúnmente reportado y usado (datos obtenidos del **ICTV**: http://www.ncbi.nlm.nih.gov/ICTV/).

Cepa: Variante de la especie, dadas las características de los autores para delimitar las condiciones de crecimiento o secuenciación del virus (datos obtenidos del **NCBI:** http://www.ncbi.nlm.nih.gov/).

Segmentos: Número de segmentos que conforman al genoma viral.

SECs: Número de Secuencias detectados en el proteoma viral.

No. de acceso del NCBI: Para obtener las secuencias en nucleótidos o aminoácidos, en muchas ocasiones el mismo virus se encuentra reportado con más de un número de acceso (Datos obtenidos del NCBI: http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html).

Totales: Se calculan los totales del número de segmentos, secuencias y longitudes de los genomas virales de ese grupo taxonómico.

APÉNDICE IV

Muestra de la base de datos bibliográfica derivada de los 2 304 proteomas virales colectados manualmente del NCBI (ver apéndice III).

Se tendrán disponibles en una página web la información bibliográfica de los 76 grupos taxonómicos de virus colectados en este trabajo como se muestra, a continuación, para la familia Adenoviridae con tipo de genoma de cadena doble de DNA (dsDNA) y hospederos del dominio Eukarya.

VIRUS	SINÓNIMOS	CEPA	HOSPEDEROS	PAÍS	ABREV	SEG	LONG (pbs)	SECs	NO. ACCESO DEL NCBI	FECHA
Bovine adenovirus B	Bovine Adenovirus 3 Mastadenovirus bos3 Bovine adenovirus type 3 BAd3	WBR-1	**sh:**bovino **lh:**Madin-Darby bovine kidney cells	Britain	BAdV-WBR-1	1	34446	28	AF030154	03/06/98
Bovine adenovirus D		THT/62-4	**sh:***Bos Taurus* **lh:**primary testicular cell culture	Hungary	BAdV-D4-THT62	1	31300	30	AF036092	11/27/97
Canine adenovirus type 1	Dog adenovirus 1 Mastadenovirus can1 Canine adenovirus type 1 Cav-1	RI261	**sh:**dog		CAdV-1-RI261dog	1	30536	30	Y07760	09/05/96
		CLL	**lh:**dog kidney	Toronto	CAdV-1-CLLdog	1	30288	28	U55001	01/08/92
Canine adenovirus type 2	Dog adenovirus 2 Mastadenovirus can2 Canine adenovirus type 2 Cav-2	Toronto A26/61	**sh:***Canis familiaris*	Toronto	CAdV-2TorA26-61	1	31323	29	U77082	14/12/1996
Duck adenovirus 1	Avian adenovirus EDS Egg drop syndrome -1976 virus	127			DAdV1-EDSV	1	33213	29	Y09598	5/31/97
Fowl adenovirus A	Avian adenovirus CELO Aviadenovirus gal1 Fowl adenovirus 1 (CELO, 112, Phelps) Adenovirus a1	Phelps (ATCC VR-432)	**sh:** chicken embryo lethal orphan	Berlin	FAdV-A1-CELO	1	43804	39	U46933	5/16/96
Fowl adenovirus D	Fowl adenovirus E Fowl adenovirus 8	A-2A			FAdV-E8	1	45063	29	AF083975	03/08/02
Frog adenovirus A		VR-896	**lh:**TH-1 cell line		FrAdV-A1	1	26163	23	AF224336	11/10/00
Human adenovirus A	Adenovirus type 12				HAdV-A12	1	34125	29	X73487	07/06/93
Human adenovirus B		Ad11p Slobitski			HAdV-B11-Slobitski	1	34794	38	AY163756	06/11/02
Human adenovirus C	Adenovirus type 2				HAdV-C2-some	1	35937	33	J01917 J01918 J01919 J01920 J01921 J01922 J01923 J01924 J01925 J01926 J01927 J01928 J01929 J01930 J01931 J01932 J01933 J01934 J01935 J01936 J01937 J01938 J01939 J01940	4/27/93
									J01941 J01942 J01943 J01944 J01945 J01946 J01947 J01948 J01949 J01950 J01951 J01952 J01953 J01954 J01955 J01956 J01957 K00086 K00394 K00395 K02367 M13004 V00007 V00008 V00009 V00010 V00011 V00012 V00013 V00014 V00015 V00016 V00017 V00018 V00019 V00020 V00023 V00024	
Human adenovirus D	Human adenovirus type 17				HAdV-D17	1	35100	30	AF108105	03/11/02
Human adenovirus E	Simian adenovirus 25	Pan 9-CV68-25			HAdV-E-SAdV25	1	36521	32	AF394196	11/27/01
Human adenovirus F	Human adenovirus type 40	Dugan-VR-931			HAdV-F40	1	34214	37	L19443	7/26/93
Murine adenovirus A	Mouse adenovirus type 1	strain FL			MAdV1-FL	1	30944	26	M22245 J03353	03/07/02
Ovine adenovirus 7	Ovine adenovirus D Ovine adenovirus OAV287	287-OAV287	**sh:**sheep	Western Australia	OAV-D287	1	29574	30	U40839 U18755 U31557 U40837 U40838	12/28/96
Ovine adenovirus A	Bovine adenovirus type 2				OAdV-A-BAdV-2	1	33034	21	AF252854 AF035571 AF061653 L24204 S75673 X82686	03/11/02
Porcine adenovirus A	Porcine adenovirus 3	P6618	**sh:***Sus scrofa*		PAdV-A3-6618	1	34094	11	AF083132 L43077 U10433 L43363	12/31/94
	Porcine adenovirus 3	IAF-6618	**lh:**PKA-pig kidney		PAdV-A3	1	34094	16	AB026117	22/04/1999
		IAF-6618	**lh:**PKA-pig kidney		PAdV-A3-IAF	1	34094	16	AJ237815	
Porcine adenovirus C		5			PAdV-C5	1	32621	30	AF289262	03/11/02
Turkey adenovirus 3					TAdV-B3-HEV	1	26263	24	AF074946	10/19/98
TOTALES						23	771545	638		
PROMEDIO							33545			

FUENTE DE COLECCIÓN DE LOS PROTEOMAS COMPLETOS:

National Center of Biotechnology Information (NCBI: http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html)

SIMBOLOGÌA:

La información fue obtenida de los reportes escritos originales de los genomas completos y de la bibliografía disponible en el servidor del NCBI.

Virus: Nombre más común de la especie viral.
Sinónimos: Nombres alternos con los que se ha reportado en trabajos previos y también autorizados por el International Committe on Taxonomy of Viruses (ICTV).
Cepa: Variante de la especie, dadas las características de los autores para delimitar las condiciones de crecimiento o secuenciación del virus.
Hospederos: **sh-specific host** (hospedero específico del virus y de donde originalmente se colectó) **lh-lab host** (hospedero que se utiliza para cultivarlo en el laboratorio).
País: Lugar donde fue colectado originalmente el virus.
Abreviación: Es el acrónimo que se emplea para identificar a ese virus en la base de proteomas completos en la base de datos construida.
Segmentos: Número de segmentos que conforman al genoma viral.
Longitud: Tamaño en residuos de nucleótidos del genoma viral.
SEC's: Número de Secuencias de aminoácidos que conforman el proteoma viral.
No. de acceso del NCBI: Para obtener las secuencias en nucleótidos o aminoácidos, en muchas ocasiones el mismo virus se encuentra reportado con más de un número de acceso.
Totales: Se calculan los totales del número de segmentos, secuencias y longitudes de los genomas virales de ese grupo taxonómico.
Promedio: Se calculan los promedios de los datos para los que se calcularon los totales de los genomas virales de ese grupo taxonómico.

APÉNDICE V

Sobre el formato y manejo de la base de proteomas virales completos por grupos taxonómicos.

---

1. Programa chequeo.pl

```
#!/usr/bin/perl

## PROGRAMA QUE FILTRA LA REDUNDANCIA DE LOS NUMEROS DE ACCESO DE LAS SECUENCIAS EN LA BASE DE
DATOS DE LOS PROTEOMAS VIRALES.

## Elaborado por Irma Lozada (ilozada@ ibt.unam.mx) y escrito en leguaje de programacion PERL.

## Necesita como archivo de entrada:
## 1) El archivo chequeo.txt (se crea en: /home/irmine/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo
taxonómico]/LCS/LCS_proteomes/)
##    Este archivo contiene todas las anotaciones de los ORFs de los proteomas virales de ese grupo taxonómico,
##    en la cual se encuentra el numero de acceso o identificador del NCBI de esa secuencia.
##    la anotación se obtiene con el comando grep de UNIX (grep ">" cat >chequeo.txt)

## 2) El archivo cat que tiene concatenado todos los proteomas en formato FASTA de un determinado grupo taxonómico viral.
##    Sobre este archivo se hace la búsqueda de los números de acceso repetidos con la lista del archivo chequeo.txt.

## Genera como archivo de salida:
## 1) redundance.txt  Archivo que contiene los números de acceso o identificadores que se repitieron en esa base de datos
##    en particular.


system "rm redundance.txt";

open EL, "chequeo.txt" or die "no pude abrir chequeo.txt";
open SAL, ">>redundance.txt" or die "no pude abrir redundance.txt";

@EL= <EL>;
chomp @EL;

foreach (@EL) {
        print SAL "$_\n";
        system "grep \"$_\" cat >>redundance.txt";
}
```

---

2. Programa Formato_seg.pl

```
#! /usr/bin/perl

## PROGRAMA QUE CAMBIA LA ANOTACION DE LAS SECUENCIAS EN FORMATO FASTA DE LA BASE DE DATOS
ORIGINAL PARA ANALISIS POSTERIORES.

##  Elaborado por Irma Lozada (ilozada@ibt.unam.mx) y escrito en leguaje de programación PERL.

##  Ubicación de trabajo de este programa, donde se encuentran los archivos de entrada y salida:
##  /home/irmine/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]

##  Necesita como archivos de entrada:
##  1) Los proteomas en archivos diferentes y con extensión (.faa)

##  Genera como archivos de salida:
##  1) Los mismos proteomas, se toman patrones de la anotación original de las secuencias para generar una nueva anotación,
##     pero mas corta y con palabras claves que permiten su mejor manejo.
##     Una vez creados se hace una copia concatenada (ver programa cat) en
##     /home/irmine/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_proteomes
##     a todos los subdirectorios en los que se realizaron los análisis restantes de este trabajo.


my @directorios = qw(/home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus);

foreach my $i (@directorios){
    chdir $i;
    opendir DIR, ".";
    my @archivos= readdir DIR;
    &baja(@archivos);
    closedir DIR;
}

sub baja{

   my @archivos= @_;
   if($#archivos > 1){
          @archivos= @archivos[2..$#archivos];
          foreach(@archivos){
             if($_ =~ /\~/){
                   system "rm  $_";
             }
          }

          my @cadenas;
          @archivos= sort @archivos;

          foreach(@archivos){
             @palabras= ();
             $i= 0;

             if ($_ =~ /^.*\.faa$/) {
                   open ARCHIVO, $_;
                   open SALIDA, ">spRE_Pom_$_" || die "No pude abrir";
                   @lineas= <ARCHIVO>;
                   chomp @lineas;

                   $i=0;
                   foreach $todo (@lineas){
                      chomp $todo;
                      $todo=~ s/\[(.*)\]//gs;
                      $new= $_;
                      $new=~ s/\.faa//;
                      $new=~ s/\/_/gs;
                      $todo=~ s/(^>.*?\()(.*?\|)(.*?\|)*/>spRE_Pom\|$2\|$new\|/g;
                      $todo=~ s/\n\s*/\|/g;
                      $todo=~ s/(\(.*\))//g;
                      $todo=~ s/(AF\d*_\d*\s*)//g;
                      $todo=~ s/\s*proteins?\s*/_/gi;
```

```
$todo=~ s/-/_/gis;
$todo=~ s/\s+/_/g;
$todo=~ s/[_+]/g;
$todo=~ s/_*$//gis;
#print "$todo\n";
$todo=~ s/pol.merase/pol/gis;
$todo=~ s/pol.protein/polyprot/gis;
$todo=~ s/dependent/dep/gis;
$todo=~ s/directed/dir/gis;
$todo=~ s/(non)structural(non.?structural)/nstruc/gis;
$todo=~ s/structural/struc/gis;
$todo=~ s/putative/put/gis;
$todo=~ s/double.?stranded/ds/gis;
$todo=~ s/single.?stranded/ss/gis;
$todo=~ s/pol.peptide/polypep/gis;
$todo=~ s/capsid/cap/gis;
$todo=~ s/minor/mi/gis;
$todo=~ s/major/ma/gis;
$todo=~ s/binding/bind/gis;
$todo=~ s/unidentified.?reading.?frame/URF/gis;
$todo=~
s/open.?reading.?frame|reading.?frame|complete.?transcript|complete.?genomic.?coding.?region|unnamed.?protein.?product)/ORF/gis;
$todo=~ s/hypothetical/hyp/gis;
$todo=~ s/replicase/repa/gis;
$todo=~ s/replicat(ion|ive)/replic/gis;
$todo=~ s/tran.?cription(al)?/transcr/gis;
$todo=~ s/transactiva(tion|tor)/transac/gis;
$todo=~ s/associated/asso/gis;
$todo=~ s/movement/mov/gis;
$todo=~ s/activat(or|ion)/act/gis;
$todo=~ s/(regulator|regulation|regulated)/reg/gis;
$todo=~ s/function/func/gis;
$todo=~ s/cell.?to.?cell_c/gis;
$todo=~ s/nuclear/nuc/gis;
$todo=~ s/shuttle/shut/gis;
$todo=~ s/membrane/memb/gis;
$todo=~ s/coding.?(sequence|region)/cs/gis;
$todo=~ s/involved.?in.?a?.?virus/invol_virus/gis;
$todo=~ s/replication.?enhan(cement|cer)/RcN/gis;
$todo=~ s/transcription.?activa(tion|tor)/TrAp/gis;
$todo=~ s/replication.?initia(tion|tor)/Rep/gis;
$todo=~ s/enhancer/enh/gis;
$todo=~ s/\s*and\s+/_/gis;
$todo=~ s/enhanc(er|ement|ing)/enha/gis;
$todo=~ s/pathogenicity/path/gis;
$todo=~ s/determinant/deter/gis;
$todo=~ s/composite/comp/gis;
$todo=~ s/gl.coprotein.?precursor/GCP/gis;
$todo=~ s/gl.coprotein/glyprot/gis;
$todo=~ s/envelope/env/gis;
$todo=~ s/nucleo.?protein/NuP/gis;
$todo=~ s/surface/surf/gis;
$todo=~ s/precursor/prec/gis;
$todo=~ s/haemagglutinin/HGT/gis;
$todo=~ s/neuraminidase/NRMD/gis;
$todo=~ s/phosphoprotein/PPP/gis;
$todo=~ s/secreted/secr/gis;
$todo=~ s/hydrop.?obic/hydrop/gis;
$todo=~ s/attachment/attach/gis;
$todo=~ s/non.?virion/nonv/gis;
$todo=~ s/reverse.?transcriptase/revTrans/gis;
$todo=~ s/inclusion/incl/gis;
$todo=~ s/transmission/transm/gis;
$todo=~ s/subunit/subt/gis;
$todo=~ s/meth.?transferase/MT/gis;
$todo=~ s/helicase/HEL/gis;
$todo=~ s/heat.?shock/HSP/gis;
$todo=~ s/gl.cosylated/glycos/gis;
$todo=~ s/transforming/transf/gis;
$todo=~ s/middle/mid/gis;
$todo=~ s/complete(d)?/comp/gis;
$todo=~ s/genomic/geno/gis;
$todo=~ s/Helper.?component.?proteinase/HC-Pro/gis;
$todo=~ s/Cytoplasmic.?inclusion/CI/gis;
$todo=~ s/Nuclear.?inclusion/NI/gis;
$todo=~ s/\s*of\s*/_/gis;
$todo=~ s/uncharacterized/unknow/gis;
$todo=~ s/conserved/conserv/gis;
$todo=~ s/family/fam/gis;
$todo=~ s/interacting/interac/gis;
$todo=~ s/aminoacyl.?tRNA.?synthetase/aaRS/gis;
$todo=~ s/signature(s)?/sfirm/gis;
$todo=~ s/(encoded.?by|corresponds.?(to)?.?)Swiss.?Prot.?Accession.?Number.+//gis;
$todo=~ s/(encoded.?by|corresponds.?(to)?.?)GenBank.?Accession.?Number.+//gis;
$todo=~ s/contains//gis;
$todo=~ s/similar.?to/sim/gis;
$todo=~ s/region/rg/gis;
$todo=~ s/transmembrane/transmemb/gis;
$todo=~ s/ribonucleotide/ribonuc/gis;
$todo=~ s/transrepressor/transrepres/gis;
$todo=~ s/packaging/pack/gis;
$todo=~ s/transrepressor/transrepres/gis;
$todo=~ s/suppressor/suppres/gis;
$todo=~ s/antigenic/Ag-mo/gis;
$todo=~ s/antigen/B\/Ag/gis;
$todo=~ s/precursor/precur/gis;
$todo=~ s/product.?/av/gis;
$todo=~ s/catalytic/cataly/gis;
$todo=~ s/alternative(ly)?/alterna/gis;
$todo=~ s/interleukin/IL/gis;
$todo=~ s/G.?(protein)?.?coupled.?receptor/GCR/gis;
$todo=~ s/origin.?(binding)?.?(protein)?/OBP/gis;
$todo=~ s/dihydrofolate.?reductase/DHFR/gis;
$todo=~ s/thymidine.?kinase/TK/gis;
$todo=~ s/thymidylate.?synthase/TS/gis;
$todo=~ s/interferon.?regulatory.?factor/v-IRF/gis;
$todo=~ s/macrophage.?inflammatory.?(protein)?/v-MIP-I/gis;
$todo=~ s/adhesion.?molecule/v-Adh/gis;
$todo=~ s/tegument/teg/gis;
$todo=~ s/accessory/access/gis;
$todo=~ s/prolifera(ting|tion).?(cell|cellular).?(nuclear|nuc).?(antigen|-like)?(Ag)?/PCNA/gis;
$todo=~ s/early.?transcription.?factor/VETF/gis;
$todo=~ s/natural.?killer.?cell.?receptor/NKcr/gis;
$todo=~ s/tumor.?necrosis.?factor.?receptor/NTFr/gis;
$todo=~ s/prokaryotic/prokar/gis;
$todo=~ s/lipoprotein/lipoprot/gis;
$todo=~ s/restriction/restric/gis;
$todo=~ s/specificity/specific/gis;
$todo=~
s/(amino.?acid.?repeat.?gene.?(family|fam)|.?(protein)?|repeat.?gene.?(family|fam)|.?(protein)|rich.?repeat.?gene.?(family|fam)|.?(protein)?)/aaRGF/gis;
$todo=~ s/late.?expression.?factor/LEF/gis;
$todo=~ s/very.?late.?factor/VLF/gis;
$todo=~ s/fibroblast.?growth.?factor/FGF/gis;
$todo=~ s/interferon.?gamma/IFN/gis;
$todo=~ s/interferon.?gamma.?receptor/IFNr/gis;
$todo=~ s/morphogenesis/morphogn/gis;
$todo=~ s/transformation/transf/gis;
$todo=~ s/rearrangement/rearg/gis;
$todo=~ s/present.?in.?the//gis;
$todo=~ s/Bacteriophage.*T6.*complete.*sequence.*translated.*products/trans_prod/gis;
$todo=~ s/Genome.*bacteriophage.*T7.*translated.*products/trans_prod/gis;
$todo=~ s/Bacteriophage.*Cp.*1.*DNA.*complete.*genome.*translated.*products/trans_prod/gis;
$todo=~ s/Bacteriophage.*SPBc2.*translated.*products/trans_prod/gis;
$todo=~
s/Vibrio.*cholerae.*filamentous.*bacteriophage.*fs.*2.*DNA.*sequence.*translated.*products/trans_prod/gis;
$todo=~ s/required_for_*/_/gis;
$todo=~ s/Bacteriophage_*IK_*translated_*products/trans_prod/gis;
$todo=~ s/Bacteriophage_*PI1_*//gis;
$todo=~ s/Bacteriophage_*alpha3_*//gis;
$todo=~ s/Bacteriophage_PP7_mRNA_for_*//gis;
$todo=~ s/site.*translation.*initiation.*has.*not.*been.*determined/STI_nond/gis;
```

```
                $todo=~ s/translated_*products/trans_prod/gis;
                $todo=~ s/(complete)?.?(sequence)?.?genome//gis;
                $todo=~ s/similarity_*with_*possible_*/swp_/gis;
                $todo=~ s/Author_*given_*sequence_*is_*in_*//gis;
                $todo=~ s/Bacteriophage_*fr_*RN_*//gis;
                $todo=~ s/Group_*II_*RNA_*coliphage_*G_*//gis;
                $todo=~ s/with_*homology_*to/wht/gis;
                $todo=~ s/This_*is_*the_*largest_*//gis;
                $todo=~ s/_*is_*thought_*to//gis;
                $todo=~ s/_*at_*nucleotide_*position//gis;
                $todo=~ s/\~+/_/gis;
                $todo=~ s/\*//gis;
                $todo=~ s/\\//gis;
                $todo=~ s/\./_/gs;
                $todo=~ s/\=/_/gs;
                $todo=~ s/:+_?/_/gis;
                $todo=~ s/\'+//gis;
                $todo=~ s/\+/_/gis;
                $todo=~ s/(V)/_/gs;
                $todo=~ s/(\;)//gs;
                $todo=~ s/(\,)//gs;
                $todo=~ s/_+/_/gs;
        $todo=~ s/\]_+/\]/g;
        $todo=~ s/_*$//gs;
                #$todo=~ s/\[/-/gs;
                #$todo=~ s/\]/-/gs;
                #$todo=~ s/\)/-/gs;
        $todo=~
s/_*which_*expressed_*in_*bacteria_*unstable_*when_*expressed_*in_*baculovirus_*start_*codon_*in_*poor_*context_*so_*is_*poor
ly_*expressed_*if_*at_*all_*put//gis;
                print "$todo\n";
                $palabras[$i]= $todo;
                print SALIDA "$palabras[-1]\n";
                $i++;
              }
             system "mv spRE_Pom_$_ /home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_proteomes";
             #print "$brevia[-1]\n";
        }
        close ARCHIVO;
        close SALIDA;
    }
  }
}
```

---

3. Programa cat.pl

```
#! /usr/bin/perl

##  PROGRAMA QUE HACE UN CAT DESDE EL DIRECTORIO RAIZ PARA REDIRECCIONARLO A LOS SUBDIRECTORIOS
SIGUIENTES PARA ANALISIS POSTERIORES.

##  Elaborado por Irma Lozada (ilozada@ibt.unam.mx) y escrito en leguaje de programación PERL.
.
##  Ubicación de trabajo de este programa, donde se encuentran los archivos de entrada y salida:
##  /home/irmine/Virus-[Eucariontes/Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_proteomes

##  Necesita como archivos de entrada:
##  1) Los proteomas en archivos diferentes, dado que hace una concatenación de los mismos

##  Genera como archivos de salida:
##  1) Cat que contiene todos los proteomas concatenados (previo se hizo un cambio en el formato de la anotación de los mismos)
##  Una vez creado se hace una copia a todos los subdirectorios en los que se realizaron los análisis restantes de este trabajo


my @directorios = qw(/home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_proteomes);

foreach my $i (@directorios){
  chdir $i;
  opendir DIR, ".";
  my @archivos= readdir DIR;
  &baja(@archivos);
  closedir DIR;
}

sub baja{

  my @archivos= @_;
  if($#archivos > 1){
    @archivos= @archivos[2..$#archivos];
    foreach(@archivos){
      if($_ =~ /\~/){
        system "rm $_";
      }
    }

        system "cat * >cat";
        system "cp cat /home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_seg";
        system "cp cat /home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_aa";
        system "cp cat /home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_func";
        system "cp cat /home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_ergo";
        print "Cree y transferi cat\n";
  }
}
```

4. Programa STA_total.pl

```
#! /usr/bin/perl

## PROGRAMA QUE PERMITE HACER EL CALCULO DE LOS TOTALES DE LOS PROTEOMAS DE LA BASE DE DATOS DE VIRUS PARA CADA FAMILIA.

## Elaborado por Irma Lozada ([illegible]) y escrito en leguaje de programación PERL.

## Necesita como archivo de entrada:
## 1) El archivo data (se crea en: /home/irmlor/Virus/[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_proteomes/)
##    Este archivo contiene todos los datos de los proteomas virales de ese grupo taxonómico, tales como:
##    1) numero de segmentos que los conforman, 2) el tamaño del proteoma en nucleótidos y, 3) el numero de ORFs que conforman a cada uno de estos.

## Genera como archivo de salida:
## 1) data.sta:  Archivo que contiene el análisis de los datos de los proteomas completos, tales como:
## a) Frecuencia de proteomas por tipo de hospedero específico.
## b) Calculo del total de proteomas y del numero de segmentos que los conforman.
## c) Tamaño total, promedio e intervalo en pares de bases de los proteomas virales de ese grupo taxonómico.
## d) Tamaño total, promedio e intervalo de ORFs de los proteomas virales de ese grupo taxonómico.


open FH, "data" or die "cannot open data\n";
open OUT, ">data.sta" or die "cannot open data.sta\n";

@all_data= <FH>;

$i= 0;
$g= 0;
$f= 0;
foreach (@all_data) {
   if ($_ =~ /[illegible]/){
            $i++;
            $virus_seg{$2}++;
            $virus_bp{$1}+= $3;
            $virus_prot{$1}+= $4;
            $virus_spe{$1}++;
            @name= split (" ", $1);
            $sp= $name[0];
            $names{$sp}++;
            $bps[$g++]= $3;
            $prots[$f++]= $4;
            $sc_bps+= $3;
            $sc_prots+= $4;
   }
   $total= $i;
}

print OUT "\n**********************************\nHOSTS\n\n";
sub byk_names {$a cmp $b}
foreach $na (sort byk_names keys %names) {
   $tot_gen+= $names{$na};
   print OUT "$na: $names{$na}\n";
}
print OUT "\nTotal Genomes: $tot_gen\n";

print OUT "\n**********************************\nSEGMENTS\n\n";
sub byk_seg {$a cmp $b}
foreach $seg (sort byk_seg keys %virus_seg) {
   print OUT "$seg\tFrecuency: $virus_seg{$seg}\n";
   $suma_seg+= $seg;
   $pre_total= $seg * $virus_seg{$seg};
   $suma_gen+= $pre_total;
}
print OUT "Total segmentos: $suma_gen\n";
print OUT "Total genomas: $suma_seg\n";

print OUT "\n**********************************\nGENOME SIZE\n\n";
$p_bp= int($sc_bps/$total);
@bps2= sort {$a <=> $b} @bps;
$first_bp= shift (@bps2);
$last_bp= pop (@bps2);
print OUT "Total en bps: $sc_bps\n";
print OUT "Size_x: $p_bp\n";
print OUT "$first_bp - $last_bp\n";

print OUT "\n**********************************\nPROTEINAS\n\n";
$p_prot= int($sc_prots/$total);
@prots2= sort {$a <=> $b} @prots;
$first_prot= shift (@prots2);
$last_prot= pop (@prots2);
print OUT "Total proteins: $sc_prots\n";
print OUT "Size_x: $p_prot\n";
print OUT "$first_prot - $last_prot\n";
```

APÉNDICE VI

Sobre el análisis de las LCS en los proteomas virales.

1. Programa seg.pl

```
#!usr/bin/perl

## PROGRAMA QUE HACE TODO EL ANALISIS DE SEG EN BASE A CAT DESDE LA RAIZ AL SUBDIRECTORIO
CORRESPONDIENTE

## Elaborado por Irma Lozada (ilozada@fcfm.buap.mx) y escrito en leguaje de programación PERL.

## Necesita como archivos de entrada:
## 1) El programa "Fre_LCS.pl" (se encuentra en: /home/irmine/LCS_programs/ y
##    redirigido a: /home/irmine/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
##    (ver programa Fre_LCS.pl).

## 2) El programa "Locprot.pl" (se encuentra en: /home/irmine/LCS_programs/ y
##    redirigido a: /home/irmine/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
##    (ver programa Locprot.pl).

## 3) El programa "seg-l.pl" (se encuentra en: /home/irmine/LCS_programs/ y
##    redirigido a: /home/irmine/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
##    (ver programa seg-l.pl).

## 4) El programa "seg-x.pl" (se encuentra en: /home/irmine/LCS_programs/ y
##    redirigido a: /home/irmine/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
##    (ver programa seg-x.pl).

## 3) El archivo cat que es el que tiene concatenado todos los proteomas de un determinado grupo taxonómico viral.
##    (ver programa cat.pl).
##    Sobre este se aplica el algoritmo SEG sin opciones (seg 12.0 1.9 2.1 cat >seg) para obtener las secuencias simples.

## Ejecuta el los programas anteriores a donde son redirigidos y genera varios archivos de salida (ver cada uno de los programas).

## Genera como archivos de salida:

## 1) seg (se queda en: /home/irmine/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
##    Divide, según el algoritmo SEG, a las secuencias en dos estados de complejidad, de baja y de alta complejidad, donde el primero
##    presenta un formato en minúsculas y el segundo con mayúsculas, en ambos se indica la localización de los dos tipos
##    de secuencias dentro de cada ORF.



system "cp Fre_LCS.pl /home/irmine/Virus-Eucariontes/dsDNA/Polyomaviridae/LCS/LCS_seg";
system "cp Locprot.pl /home/irmine/Virus-Eucariontes/dsDNA/Polyomaviridae/LCS/LCS_seg";
system "cp seg-l.pl /home/irmine/Virus-Eucariontes/dsDNA/Polyomaviridae/LCS/LCS_seg";
system "cp seg-x.pl /home/irmine/Virus-Eucariontes/dsDNA/Polyomaviridae/LCS/LCS_seg";

my @directorios= qw(/home/irmine/Virus-Eucariontes/dsDNA/Polyomaviridae/LCS/LCS_seg);

foreach my $i (@directorios){
   chdir $i;
   opendir DIR, ".";
   my @archivos= readdir DIR;
   &baja(@archivos);
   closedir DIR;
}

sub baja{
   my @archivos= @_;
   if($#archivos > 1){
      @archivos= @archivos[2..$#archivos];
   foreach(@archivos){
      if($_=~ /\~/){
         system "rm $_";
      }
   }
      system "seg cat 12.0 1.9 2.1 >seg";
      system "perl /home/irmine/Virus-Eucariontes/dsDNA/Polyomaviridae/LCS/LCS_seg/seg-l.pl";
      system "perl /home/irmine/Virus-Eucariontes/dsDNA/Polyomaviridae/LCS/LCS_seg/seg-x.pl";
      system "perl /home/irmine/Virus-Eucariontes/dsDNA/Polyomaviridae/LCS/LCS_seg/Fre_LCS.pl";
      system "perl /home/irmine/Virus-Eucariontes/dsDNA/Polyomaviridae/LCS/LCS_seg/Locprot.pl";
   }
}
```

2. Fre_LCS.pl

```perl
#!/usr/bin/perl

## PROGRAMA QUE CALCULA EL NUMERO DE ORFS CON AL MENOS UNA SS, EL NUMERO DE SEGMENTOS DE SS POR ORF, EL NUMERO DE GENOMAS QUE TIENEN
## AL MENOS UN ORF CON SS, LA LONGITUD Y POSICION DE LAS SS DENTRO DE CADA ORF, LA FRECUENCIA DEL TIPO DE SS (PATRON) POR CADA GRUPO
## TAXONOMICO Y ASIGNA A LA SS LA PALABRA CLAVE DE LA ANOTACION QUE TIENE EL ORF PARA VER LA FUNCION DE ESTE.
## TAMBIEN HACE ALGUNAS MODIFICACIONES SOBRE EL FORMATO DEL ARCHIVO DE SALIDA DE SEG-L PARA EL ANALISIS DE LA COMPOSICION DE AAs.

###   Elaborado por Irma Lozada ([illegible]) y escrito en lenguaje de programación PERL.

## Necesita como archivos de entrada:
## 1) El archivo cat que es el que tiene concatenado todos los proteomas de un determinado grupo taxonómico viral.
## 2) El archivo cat.seg-l que tiene al archivo cat analizado previamente con seg y la opción -l (seg 12 0 1.9 2.1 cat -l).

## Genera como archivos de salida:

## 1) ClasPro_LCS.txt (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_func)
## Número y tipo de patrones en los segmentos de SS que se repiten en un mismo ORF.

## 2) Clusters_LCS.txt (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_func)
## Numero y tipo de patrones en los segmentos de SS en todos los ORF's que constituyen al conjunto de proteomas de ese grupo viral.

## 3) LCS_seg-l.txt (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
## Presenta un formato similar al cat.seg-l, pero en este se encuentra el segmento de SS unido a la información del ORF dada por la
## opción -l del seg previamente aplicado.

## 4) Proteins_LCS.txt (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_func)
## Numero de ORF's que presentan SS y la misma anotación funcional según la aplicación previa del programa Formato_seg.pl

## 5) Complexity_LCS.txt (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
## Numero de ORF's en cada especie viral que presentan en su SS el mismo valor de complejidad.

## 6) Frecuency_LCS.txt (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
## Numero de segmentos de SS que se encuentran en cada ORF por grupo viral.

## 7) Frecuency_LCS.sta (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
## Distribución del numero de segmentos de SS en el total de los ORF's de ese grupo viral.
## También presenta el total de segmentos de SS en los proteomas de ese grupo viral.
## Numero de ORF's totales que contienen a estos segmentos de SS.
## Numero de proteomas que tienen al menos un ORF con una SS.

## 8) Prot_SP_LCS.txt (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_func)
## Numero de ORF's que presentan al menos una SS y la misma anotación funcional por especie viral para todo el grupo taxonómico.

## 9) aa_LCS.txt (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_aa)
## Aquí se encuentran todos los segmentos de SS de todos los proteomas de ese grupo viral para el análisis de aminoácidos.

## 10) Tot_seg-l.txt (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_func)
## Presenta un formato similar a cat.seg-l, pero en este se encuentran todos los segmentos de SS de un mismo ORF juntos
## y con la ubicación física en residuos de aminoácidos dentro del mismo.

## 11) seg.num (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
## Contiene el numero de ORF's totales de todos los proteomas del grupo viral,
## el numero de ORF's que presenta al menos una SS y el porcentaje que ello representa para este grupo.

## 12) Genomas_LCS.txt (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
## Numero de ORF's por proteoma viral de ese grupo taxonómico que tuvieron una SS.

## 13) LCS_genomas.num (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
## Numero de proteomas con al menos un ORF que presenta, al menos, una SS.

## 14) Proteins_GENOME.txt (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_func)
## Numero de ORF's de todos los proteomas de ese grupo taxonómico que presentan la misma anotación funcional según la aplicación
## previa del programa Formato_seg.pl

## 15) Prot_SP_GENOME.txt (redirigido a: /home/[illegible]/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_func)
## Numero de ORF's por especie viral de todos los proteomas de ese grupo taxonómico que presentan la misma anotación funcional
## según la aplicación previa del programa Formato_seg.pl


open SALE, ">ClasPro_LCS.txt" || die "No pude abrir ClasPro_LCS.txt\n";
open SAL, ">Clusters_LCS.txt" || die "No pude abrir Clusters_LCS.txt\n";
open SALE1, ">LCS_seg-l.txt" || die "No pude abrir LCS_seg-l.txt\n";
open SALE2, ">Proteins_LCS.txt" || die "No pude abrir Proteins_LCS.txt\n";
open SALE3, ">Complexity_LCS.txt" || die "No pude abrir Complexity_LCS.txt\n";
open SALE4, ">Frecuency_LCS.txt" || die "No pude abrir Frecuency_LCS.txt\n";
open SALE4A, ">Frecuency_LCS.sta" || die "No pude abrir Frecuency_LCS.sta\n";
open SALE5, ">Prot_SP_LCS.txt" || die "No pude abrir Prot_SP_LCS.txt\n";
open SALE6, ">aa_LCS.txt" || die "No pude abrir aa_LCS.txt\n";
open SALE7, ">Tot_seg-l.txt" || die "No pude abrir Tot_seg-l.txt\n";
open SALE8, ">seg.num" || die "No pude abrir seg.num\n";
open SALE9, ">Genomas_LCS.txt" || die "No pude abrir Genomas_LCS.txt\n";
open SALE10, ">LCS_genomas.num" || die "No pude abrir LCS_genomas.num\n";
open FIN, ">Proteins_GENOME.txt" || die "No pude abrir Proteins_GENOME.txt\n";
open FIN2, ">Prot_SP_GENOME.txt" || die "No pude abrir Prot_SP_GENOME.txt\n";
open ENTRA, "cat.seg-l" || die "No pude abrir cat.seg-l\n";
open ENTRA2, "cat" || die "No pude abrir cat\n";

@cat_date= <ENTRA2>;
chomp @cat_date;
$all_cat= join "", @cat_date;
@new_cat= split ">", $all_cat;

foreach $dog (@new_cat) {
   $dog=~ s/^/>/;
   if ($dog=~ [illegible]) {
            #print "$1\n";
            #print "*$2\n";
            $almost= $2;
            $protein_cat= $3;
            $protein_cat=~ s/[illegible]//;
            $f_cat{$protein_cat}++;
            $sp_cat= $almost.$protein_cat;
            $f_sp_cat{$sp_cat}++;
   }
}
print FIN "\n";
sub byvCAT {$f_cat{$a} cmp $f_cat{$b}}
foreach $prot_CAT (sort byvCAT keys %f_cat) {
   print FIN "$prot_CAT\t$f_cat{$prot_CAT}\n";
}
print FIN "\n";
system "mv Proteins_GENOME.txt /home/[illegible]/Virus-Eucariontes/[illegible]/LCS/LCS_func";

print FIN2 "\n";
sub byvspCAT {$f_sp_cat{$a} cmp $f_sp_cat{$b}}
foreach $prot_spCAT (sort byvspCAT keys %f_sp_cat) {
   print FIN2 "$prot_spCAT\t$f_sp_cat{$prot_spCAT}\n";
}
print FIN2 "\n";
system "mv Prot_SP_GENOME.txt /home/[illegible]/Virus-Eucariontes/[illegible]/LCS/LCS_func";
```

```
@dates= <ENTRA>;
chomp @dates;
$all= join ":", @dates;
@gi_ss= split ">", $all;

print SALE1 "\n";
foreach $junto (@gi_ss) {
   $junto=~ s/^>//;
   print SALE1 "$junto\n";
}
print SALE1 "$junto\n";

print SALE6 "\n";
foreach $gi (@gi_ss) {
   $gi=~ s/^>//;
   print "$gi\n";

   if ($gi=~ /^(.+\|.+\|)(\[.+\])(.+)(\(.+\))\s*complexity=(.+)\):(.+)/) {
            #print "$1\n";
            #print "*$2\n";
            #print "**$3\n";
            #print "***$4\n";
            #print "****$5\n";
            #print "*****$6\n";
            print SALE6 "$6\n";
            $juntas= $1.$2.$3;
            $form= $4.$6;
            $juntas_lcs{$juntas} .= $form;
            $firma= $2.$3." ".$6;
            $frec= $1.$2.$3;
            $prot= $2.$3;
            $complex= $2.$3." ".$5;
            $fre_firma_sp{$firma}++;
            $fre_prot{$3}++;
            $fre_firma{$6}++;
            $fre_prot_sp{$prot}++;
            $complexity{$complex}++;
            $frecuency{$frec}++;
            $LCS_gen{$2}++;
   }
}

print SALE6 "\n";
system "mv aa_LCS.txt /home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_aa";

print SALE7 "\n";
sub byk_10 {$a cmp $b}
foreach $clave (sort byk_10 keys %juntas_lcs) {
   print SALE7 ">$clave\n$juntas_lcs{$clave}\n\n";
}

print SALE7 "\n";
system "mv Too_seg-l.txt /home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_func";
@valores= values %frecuency;
foreach (@valores) {
   $veces{$_}++;
}

print SALE "\n";
sub byk {$a cmp $b}
foreach $ss_sp (sort byk keys %fre_firma_sp) {
   print SALE "$ss_sp \t| $fre_firma_sp{$ss_sp}\n";
}
print SALE "\n";
system "mv ClustPro_LCS.txt /home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_func";

print SAL "\n";
sub byv3 {$fre_firma{$a} cmp $fre_firma{$b}}
foreach $ss_firma2 (sort byv3 keys %fre_firma) {
   print SAL "$ss_firma2 \t| $fre_firma{$ss_firma2}\n";
}

print SAL "\n";
system "mv Clusters_LCS.txt /home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_func";

print SALE2 "\n";
sub byv2 {$fre_prot{$a} cmp $fre_prot{$b}}
foreach $prot2 (sort byv2 keys %fre_prot) {
   print SALE2 "$prot2 \t| $fre_prot{$prot2}\n";
}
print SALE2 "\n";
system "mv Proteins_LCS.txt /home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_func";

print SALE5 "\n";
sub byv4 {$fre_prot_sp{$a} cmp $fre_prot_sp{$b}}
foreach $prot_sp2 (sort byv4 keys %fre_prot_sp) {
   print SALE5 "$prot_sp2 \t| $fre_prot_sp{$prot_sp2}\n";
}
print SALE5 "\n";
system "mv Prot_SP_LCS.txt /home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_func";

print SALE3 "\n";
sub byv5 {$complexity{$a} cmp $complexity{$b}}
foreach $complej2 (sort byv5 keys %complexity) {
   print SALE3 "$complej2 \t| $complexity{$complej2}\n";
}
print SALE3 "\n";

print SALE4 "\n";
sub byv6 {$frecuency{$a} cmp $frecuency{$b}}
foreach $fre2 (sort byv6 keys %frecuency) {
   print SALE4 "$fre2 \t| $frecuency{$fre2}\n";
}

print SALE4A "\n____________________________________________\nNO. DE SEGMENTOS DENTRO DE LAS PROTEINAS\n\n";
sub byv7 {$veces{$a} cmp $veces{$b}}
foreach (sort byv7 keys %veces) {
   print SALE4A "S: $_ - P: $veces{$_}\n";
   $count_sec= $veces{$_};
   $Genes+= $count_sec;
   $Pre= $count_sec * $_;
   $Segmentos+= $Pre;
}

print SALE4A "____________________________________________\nTOTALES DE LCS\n\n";
print SALE4A "***Total Secuencias: $Genes\n";
print SALE4A "***Total Segmentos: $Segmentos\n";

@genomas= keys %LCS_gen;
$lcs_genomas= @genomas;

print SALE4A "\n***Genomas con al menos una proteina con LCS: $lcs_genomas\n\n";
print SALE10 "$lcs_genomas\n";


print SALE9 "\n";
sub byv9 {$LCS_gen{$a} cmp $LCS_gen{$b}}
foreach $gen (sort byv9 keys %LCS_gen) {
   print SALE9 "$gen \t| $LCS_gen{$gen}\n";
}
print SALE9 "\n";

system "grep -c '[>]' cat >cat.num";
open FH, "cat.num" || die "No pude abrir cat.num\n";
$por_LCS= <FH>;
chomp $por_LCS;
$LCS_HCS= int(($Genes/$por_LCS)*100);
print SALE8 "Proteinas: $por_LCS\n";
print SALE8 "Proteinas con LCS: $Genes\n";
print SALE8 "Porcentaje: $LCS_HCS%\n";
close FH;
```

3. Programa Locprot.pl

```
#! /usr/bin/perl

## PROGRAMA QUE HACE EL CALCULO DE LA LONGITUD DE LAS SS, LA POSICION RELATIVA DENTRO DEL ORF Y
COLOCA A LAS HCS EN UN ARCHIVO
## PARA EL ANALISIS DE LA COMPOSICION DE AMINOACIDOS

##  Elaborado por Irma Lozada (ilozada@fciencias.mx) y escrito en lenguaje de programación PERL.

## Necesita como archivos de entrada:
## 1) El archivo cat.seg.x que tiene al archivo cat analizado previamente con seg y la opción -x  (seg 12.0 1.9 2.1 cat -x).

## Genera como archivos de salida:

## 1) aa_HCS.txt (redirigido a: /home/irmine/Virus-[Eucariontes/Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_aa)
##   Archivo que contiene los aminoácidos que componen a las HCS para el análisis composicional de los mismos.

## 2) Length_SEC.txt (redirigido a: /home/irmine/Virus-[Eucariontes/Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
##   Tamaño total en residuos de aminoácidos de todos los ORFs que presentan al menos una SS.

## 3) Localization_LCS.txt (redirigido a: /home/irmine/Virus-[Eucariontes/Procariontes]/[Tipo de genoma]/[Grupo
taxonómico]/LCS/LCS_seg)
##   Despliega la localización de los segmentos de las SS dentro del ORF que las contiene por separado y mostrando en cada uno
##   la longitud del ORF, el inicio y final de la cobertura (en porcentaje) de la SS dentro del mismo.

## 4) Length_LCS.txt (redirigido a: /home/irmine/Virus-[Eucariontes/Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
##   Archivo que presenta un formato más accesible para la extracción de resultados sobre la posición relativa (porcentaje)
##   de inicio a fin de las SS dentro de cada ORF.

## 5) LCS_bps.txt (redirigido a: /home/irmine/Virus-[Eucariontes/Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_seg)
##    Incluye varios datos:
## a) Tamaño total del proteoma en residuos de aminoácidos.
## b) Tamaño total de las HCS y las LCS de forma conjunta y separada en residuos de aminoácidos.
## c) Porcentaje de la contribución neta de las LCS al proteoma.
## d) Porcentaje de la contribución neta de las LCS EN LOS ORFs que las contienen.
## e) Número de segmentos de SS en todos los proteomas de ese grupo viral y el intervalo en el tamaño de estos.
## f) Número de segmentos de HCS en todos los proteomas de ese grupo viral y el intervalo en el tamaño de estas.



open SALE, ">aa_HCS.txt" || die "no pude";
open SALE2, ">Length_SEC.txt" || die "no pude";
open SALE3, ">Localization_LCS.txt" || die "no pude";
open SALE4, ">Length_LCS.txt" || die "no pude";
open SALE5, ">LCS_bps.txt" || die "no pude";
open OUT,"cat.seg.x" || die "no pude";

@cat= <OUT>;
$fasta= join "", @cat;
$fasta=~ s/>/\>/g;
@sec_fasta= split /\>/, $fasta;

@ini_hx= ();
@fin_hx= ();
$y= 0;
$z= 0;
foreach (@sec_fasta) {
   chomp;
   $_=~ /(\w+)/;
   $gi= $1;
   s/\w+//;

   @hcs= ();
   @chx= ();
   $w= 0;
   $q= 0;
   @hcs_x= ();
   @lcs_x= ();
   @sec_join= ();
```

```
$cat= "";
if (s+) {
      @sec_join= split /\-/, $_;

      foreach (@sec_join) {
        chomp $_;
        $cat.= $_;
      }

      $complete= $gi."\n".$cat;
      $length_sec= length($cat);
      $secuen_total+= $length_sec;
      $cat=~ s/\s/9/g;
      $cat=~ s/\D/&/g;

      if ($cat=~ /^&+/) {
        @hcs_x= split /9+/, $cat;
        foreach $h (@hcs_x) {
              $length_hcs= length($h);
              $hcs[$w++]= $length_hcs;
              $size_hcs[$xx++]= $length_hcs;
              $hcs_long_total+= $length_hcs;
        }
        @lcs_x= split /&+/, $cat;
        shift @lcs_x;
        foreach $s (@lcs_x) {
              $length_lcs= length($s);
              $lcs[$q++]= $length_lcs;
              $size_lcs[$yy++]= $length_lcs;
              $lcs_long_total+= $length_lcs;
        }
        chomp @hcs;
        chomp @lcs;
        $ini= 0;
        $inicio= 0;
        $fin= 0;
        $z= 0;
        for ($p= 0; $p<= $#lcs; $p++) {
              $inicio= $fin + $hcs[$z++] + 1;
              $fin= ($inicio -1) + $lcs[$p];
              $N_pos= int(($inicio / $length_sec) * 100);
              $C_pos= int(($fin / $length_sec) * 100);
              print SALE2 "$length_sec\n";
              $ini_lcs[$y++]= $N_pos;
              $fin_lcs[$o++]= $C_pos;
              #print "$gi | $length_sec |Inicio: $inicio |Fin: $fin |$N_pos% - $C_pos%\n";
              print SALE3 "$gi | $length_sec |Inicio: $inicio |Fin: $fin |$N_pos% - $C_pos%\n";
        }
      }
      if ($cat=~ /^9+/) {
        @lcs_x= split /&+/, $cat;
        foreach $s (@lcs_x) {
              $length_lcs= length($s);
              $lcs[$q++]= $length_lcs;
              $size_lcs[$yy++]= $length_lcs;
              $lcs_long_total+= $length_lcs;
        }
        @hcs_x= split /9+/, $cat;
        shift @hcs_x;
  foreach $h (@hcs_x) {
    $length_hcs= length($h);
    $hcs[$w++]= $length_hcs;
    $size_hcs[$xx++]= $length_hcs;
              $hcs_long_total+= $length_hcs;
  }
  chomp @hcs;
  chomp @lcs;
  $ini= 0;
  $inicio= 0;
  $fini= 0;
  $fin= 0;
```

```
        $z= 0;

        for ($p= 0; $p<= $#hcs; $p++) {
                $inicio= $fin + 1;
                $fin= $fin2 + $lcs[$z++] + 1;
          $N_pro= int(($inicio / $length_sec) * 100);
          $C_pro= int(($fin / $length_sec) * 100);
                print SALE2 "$length_sec\n";
          $ini_lcs[$y++]= $N_pro;
          $fin_lcs[$o++]= $C_pro;
                #print "Sgi | $length_sec |Inicio: $inicio |Fin: $fin |$N_pro% - $C_pro%\n";
          print SALE3 "Sgi | $length_sec |Inicio: $inicio |Fin: $fin |$N_pro% - $C_pro%\n";
                $fin= $fin + $hcs[$p];
                $fin2= $fin;
            }
          }
    }
    s/\n//g;
    @sec_join= split /\*/, $_;
    $aa_hcs= "";
    foreach (@sec_join) {
          chomp $_;
          $aa_hcs.= $_;
          $doble_hcs= $aa_hcs;
    }
    print SALE "\n$aa_hcs\n";
    $length_prot= length($doble_hcs);
    $total_prot+= $length_prot;
}

print SALE4 "___INICIO___\n";
foreach $pot (@ini_lcs) {
    print SALE4 "$pot\n";
}
print SALE4 "\n___FIN___\n";
foreach $cen (@fin_lcs) {
    print SALE4 "$cen\n";
}

$all_total= $total_prot + $lcs_long_total;
print SALE5 "Tamanio total en aas de Genoma: $all_total\n";
print SALE5 "Tamanio total en aas de HCS con LCS: $hcs_long_total\n";
print SALE5 "Tamanio total en aas de LCS: $lcs_long_total\n";
print SALE5 "$lcs_long_total / $all_total\n";
$procen= ($lcs_long_total / $all_total)*100;
print SALE5 "El porcentaje de LCS neto al proteoma: $procen%\n";
print SALE5 "$lcs_long_total / $hcs_long_total\n";
$procen_razon= ($lcs_long_total / $hcs_long_total)*100;
print SALE5 "El porcentaje del LCS neto en las proteinas que las contienen: $procen_razon%\n";

$long_size= @size_lcs;
@size_lcs2= sort {$a <=> $b} @size_lcs;
$first_lcs= shift (@size_lcs2);
$last_lcs= pop (@size_lcs2);
print SALE5 "\n_____\nTAMANIO DE LAS SECUENCIAS SIMPLES\n\n";
print SALE5 "No. de segmentos de LCS: $long_size\n";
print SALE5 "$first_lcs - $last_lcs\n";
print SALE5 "\nLongitud en aas de las LCS: $lcs_long_total\n\n";

$long_size_h= @size_hcs;
@size_hcs2= sort {$a <=> $b} @size_hcs;
$first_hcs= shift (@size_hcs2);
$last_hcs= pop (@size_hcs2);
print SALE5 "\n_____\nTAMANIO DE LAS SECUENCIAS DE ALTA COMPLEJIDAD\n\n";
print SALE5 "No. de segmentos de HCS: $long_size_h\n";
print SALE5 "$first_hcs - $last_hcs\n";
print SALE5 "\nLongitud en aas de las HCS: $hcs_long_total\n\n";

system "mv aa_HCS.txt /home/irmine/Virus-Eucariontes/dsDNA/Polyomaviridae/LCS/LCS_aa";
```

---

4. Programa seg-l.pl

```
#!/usr/bin/perl

## EJECUTA EL ALGORITMO SEG CON PARAMETROS 12.0 1.9 2.1 Y LA OPCION -L DE FORMA AUTOMATIZADA PARA
UN DETERMINADO NUMERO DE ARCHIVOS QUE CONTENGAN
## UNA O MAS SECUENCIAS EN FORMATO FASTA

## Elaborado por Irma Lozada (ilozada@ibt.unam.mx) y escrito en leguaje de programacion PERL.

## Necesita como archivo de entrada:
## 1) El archivo cat (se encuentra en: /home/irmine/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo
taxonómico]/LCS/LCS_seg/)
##    Este archivo contiene todos los proteomas concatenados en formato FASTA para ser evaluados por la opción -l del algoritmo SEG
##    (seg 12.0 1.9 2.1 cat -l >cat.seg-l)

## Genera como archivo de salida:
## 1) cat.seg-l  (Esta opcion permite obtener un archivo en el que se encuentren solo la secuencia de las SS,
##    la localización en residuos de aminoácidos dentro del ORF y el valor de complejidad de la SS detectada).
##    Este archivo es ampliamente utilizado para analizar diversas características de las SS (ver programa Fre_LCS.pl).


$a[0]="cat";

foreach $T (0..0) {
   chomp ($a[$T]);
   print "analizo $a[$T]\n";
   $nombre= $a[$T];
   print "$nombre\n";
   $salida= $a[$T].".seg-l";
   system "seg $nombre 12.0 1.9 2.1 -l > $salida";
}
```

---

5. Programa seg-x.pl

```
#!/usr/bin/perl

## EJECUTA EL ALGORITMO SEG CON PARAMETROS 12.0 1.9 2.1 Y LA OPCION -L DE FORMA AUTOMATIZADA PARA
UN DETERMINADO NUMERO DE ARCHIVOS QUE CONTENGAN
## UNA O MAS SECUENCIAS EN FORMATO FASTA

## Elaborado por Irma Lozada (ilozada@ibt.unam.mx) y escrito en leguaje de programacion PERL.

## Necesita como archivo de entrada:
## 1) El archivo cat (se encuentra en: /home/irmine/Virus-[Eucariontes|Procariontes]/[Tipo de genoma]/[Grupo
taxonómico]/LCS/LCS_seg/)
##    Este archivo contiene todos los proteomas concatenados en formato FASTA para ser evaluados por la opción -x del algoritmo SEG
##    (seg 12.0 1.9 2.1 cat -x >cat.seg-x)

## Genera como archivo de salida:
## 1) cat.seg-x  (Esta opción permite obtener un archivo en el que se encuentren las secuencias evaluadas con SEG en formato FASTA.
##    Donde la secuencia de las SS es sustituida por caracteres del tipo "x").
##    Este formato es utilizado para cuantificar el numero de ORFs que presentan al menos una SS y la localización relativa de la
##    misma dentro de este (ver programas Locprot.pl y azar_ss.pl


$a[0]="cat";
foreach $T (0..0) {
   chomp ($a[$T]);
   print "analizo $a[$T]\n";
   $nombre= $a[$T];
   print "$nombre\n";
   $salida= $a[$T].".seg-x";
   system "seg $nombre 12.0 1.9 2.1 -x > $salida";
}
```

APÉNDICE VII

Sobre el análisis de la composición de aminoácidos de los proteomas, las LCS y las secuencias de alta complejidad.

Programa aacomp.pl

```
#!usr/bin/perl

## PROGRAMA QUE HACE TODO EL ANALISIS COMPOSICIONAL EN BASE A AACOMP DESDE LA RAIZ AL
SUBDIRECTORIO CORRESPONDIENTE POR GRUPO TAXONÓMICO VIRAL.

##  Elaborado por Irma Lozada (ilozada@ibt.unam.mx) y escrito en leguaje de programación PERL.

##  Ubicación de trabajo de este programa, donde se encuentran los archivos de entrada y salida:
##  /home/irmine/Virus-[Eucariontes[Procariontes]/[Tipo de genoma]/[Grupo taxonómico]/LCS/LCS_aa

##  Necesita como archivos de entrada:
##  1) aa_HCS.txt
##  2) aa_LCS.txt
##  3) cat

##  Genera como archivos de salida:
##  1) aacomp_HCS (Archivo que contiene el análisis de la composición de aminoácidos para las HCS)
##  2) aacomp_LCS  (Archivo que contiene el análisis de la composición de aminoácidos para las LCS)
##  3) aa_GENOME.txt  (Archivo que contiene sólo los aminoácidos de cat)
##  4) aacomp_GENOME  (Archivo que contiene el análisis de la composición de aminoácidos de los proteomas completos contenidos en
cat)
##  5) Date_$_  (Archivos que tienen un formato especial para el análisis estadístico de los datos de 1,2 y 4)


my @directorios= qw(/home/irmine/Virus-Eucariontes/ssRNA+/Pomovirus/LCS/LCS_aa);

foreach my $i (@directorios){
   chdir $i;
   opendir DIR, ".";
   my @archivos= readdir DIR;
   &baja(@archivos);
   closedir DIR;
}

sub baja{
   my @archivos= @_;
   if($#archivos > 1){
      @archivos= @archivos[2..$#archivos];
      foreach(@archivos){
         if($_ =~ /\~/){
            system "rm $_";
         }
      }
            system "aacomp aa_HCS.txt >aacomp_HCS";
            system "aacomp aa_LCS.txt >aacomp_LCS";
            system "grep -v '[>]' cat >aa_GENOME.txt";
            system "aacomp aa_GENOME.txt >aacomp_GENOME";

      my @cadenas;
      @archivos= sort @archivos;
      foreach(@archivos){
               @aa= ();
               @mol= ();
               @wt= ();
               $s= 0;
               $e= 0;
               $r= 0;
         if($_ =~ /^aacomp_.+/){
            open ARCHIVO1, $_;
                  print "$_\n";
                  @cadenas= <ARCHIVO1>;
                  $aa= shift(@cadenas);
                  shift(@cadenas);


            foreach $i (@cadenas){
                     chomp $i;
                     @aacomp= split /\s+/, $i;
                     $aa[$s++]= $aacomp[1];
                     $mol[$e++]= $aacomp[4];
                     $wt[$r++]= $aacomp[5];
            }
            close ARCHIVO1;
                     open OUT1, ">Date_$_" || die "No pude abrir";
                     print OUT1 "___aa___\n";
                     foreach $aa (@aa) {
                        print OUT1 "$aa\n";
                     }
                     print OUT1 "\n___mol___\n";
                     foreach $mol (@mol) {
                        print OUT1 "$mol\n";
                     }
                     print OUT1 "\n___wt___\n";
                     foreach $wt (@wt) {
                        print OUT1 "$wt\n";
                     }
            }
            close OUT1;
         }
      }
   }
```

APÉNDICE VIII

Sobre las categorías funcionales de las secuencias virales y, de aquellas que presentaron, al menos, una LCS.

---

1. Programa Saca_ORFs_SS.pl

funcion_final.pl

```
#!/usr/bin/perl

##PROGRAMA QUE HACE LAS COMPARACIONES DE LAS REGIONES DE LAS SECUENCIAS SIMPLES EN LOS PFAMS
ENCONTRADOS EN LOS PROTEOMAS VIRALES

open PFAMS, "all_virus_hmm_dominios" || die "No pude abrir all_virus_hmm_dominios\n";
open LCS, "lcs_all_virus_dic03" || die "No pude abrir lcs_all_virus_dic03\n";
open VCOGS, "hash_all_VCOGs" || die "No pude abrir hash_all_VCOGs\n";

@pfams= <PFAMS>;     ##Aqui todo los datos en bruto de los pfams
chomp @pfams;

@dates= <LCS>;       ##Aqui las secuencias simples, nombre, ubicacion, formato fasta
chomp @dates;
$all= join ":", @dates;
@gi_ss= split ":>", $all;

@vcogs= <VCOGS>;     ##Aqui pfams desglosados por funcion en el NCBI
chomp @vcogs;

##AQUI HAGO UN HASH DE LOS PFAMS CON SUS RESPECTIVAS FUNCIONES
foreach $vo_fu (@vcogs) {
   if ($vo_fu=~ /^(PF.+)\t(.+)/) {
         $funciones{$1}= $2;
   }
}

##AQUI HAGO UN HASH DE LOS NOMBRES CON SUS RESPECTIVOS PFAMS O NO_HITS
foreach $gi (@pfams) {
   #print "$gi\n";

   if ($gi=~ /^(.+)\t(PF.+|no.?hits)$/) {
      #print "$1\n";
      #print "*$2\n";
            $edit_2= ":".$2.":";
            $edit_2=~ s/::/:/;
      $pfams{$1}.= $edit_2;
   }
}

##AQUI HAGO UN HASH DE LOS NOMBRES DE LAS SECUENCIS Y SUS RESPECTIVAS SECUENCIAS SIMPLES Y
LOCALIZACION
foreach $ss (@gi_ss) {
   #print "$ss\n";
   $ss=~ s/:\{/::\{/;

   if ($ss=~ /^(.+){::+$)/) {
         $llave_ss= $1;
         #print "**$llave_ss\n";
         $reg_ss= $2;
         $reg_ss=~ s/::/:/;
         #print "+++$reg_ss\n";
         $local_lcs{$llave_ss}= $reg_ss;
   }
}

@keys_pfams= keys %pfams;
@keys_local_lcs= keys %local_lcs;
@funcs_pfams= keys %funciones;
   chomp @keys_pfams;
   chomp @keys_local_lcs;
   chomp @funcs_pfams;

##RECORRO LAS LLAVES DE SECUENCIAS SIMPLES Y DESPUES LAS DE LA LOS PFAMS
foreach $key_1 (@keys_local_lcs) {

   foreach $key_2 (@keys_pfams) {
         #print "*******$key_2\n";

         if ($key_1 eq $key_2) {
            #print "$key_1\n";
            #print "$key_2\n";
            $key_count= $key_2;
            $key_count=~ s/-.+//;
            $count_family{$key_count}++;  ###LAS LLAVES QUE ENTRAN (SECUENCIAS SIMPLES CON PFAMS) SON EL
CONTEO POR FAMILIA

            open RESULT, ">>key_count_no_hits";
            open RESULT1, ">>key_count_lcs_pfams";
            open RESULT2, ">>SS_DENTRO_PFAM";
            open RESULT3, ">>SS_IZQUIERDA_PFAM";
            open RESULT4, ">>SS_DERECHA_PFAM";
            open RESULT5, ">>SS_FUERAi_PFAM";
            open RESULT6, ">>SS_FUERAd_PFAM";
            open RESULT7, ">>PFAM_DENTRO_SS";

            $valor_lcs= $local_lcs{$key_1};
            $valor_pfam= $pfams{$key_2};

            if ($valor_pfam eq ":no hits:") {   ##AQUI FILTRO TODOS LO SQUE QUE NO SON NO HITS
                  $count_no{$key_count}++;
                  #print "$valor_pfam\n";
                  print RESULT ">$key_2\n";
                  print RESULT "$valor_lcs\n";
                  print RESULT "$valor_pfam\n";
            }
            else {                        ##AQUI EMPIEZA LO BUENO
                  @split_lcs= ();
                  @split_pfam= ();
                  @num_lcs= ();

#                 print "$key_2\n";
#           print "$valor_lcs\n";
#           print "$valor_pfam\n";

                  @split_lcs= split(":", $valor_lcs);
                  @split_pfam= split(":", $valor_pfam);

                  foreach $lcs (@split_lcs) {
                     @data_pfam= ();
                     @sss= ();
                     @pfs= ();

                     if ($lcs=~ /^\{(.+)\}([a-z]+)/) {
                           #print "***$1\n";
                           #print "$2\n";
                           $par_lcs= $1;
                           $eg_lcs= $2;
                           $reg_lcs{$1}= $2;
                           foreach $all_pfams (@split_pfam) {
                              $all_pfams=~ s/:+//gis;
                              #print "$all_pfams\n";
                              @data_pfam= split("\t", $all_pfams);
                              chomp @data_pfam;
                              $no_pfams= $data_pfam[0];
```

```
            #print "$data_pfam[1]\n";
            $fam= $data_pfam[1];
#print "$data_pfam[2]\n";
            @ass= split "-", $par_lcs;
            @pfs= split "-", $fam;
            chomp @ass;
            chomp @pfs;
            #print "$ass[0]\n";
            #print "$pfs[1]\n";
            #<STDIN>;
            foreach $fn (@funcs_pfam){
                    if ($ss_pfam eq $fn){
                      $recoge_func= $funciones{$fn};
                      $recoge_func.= "/";
                    }
            }

            if ($data_pfam[2] =~ /(0.01|0.00 +|e-)/ && $data_pfam[2] !~ /e\+/) {
#                   print "xxx$data_pfam[2]xxx\n";
                    if ($ass[0]>=$pfs[0] && $ass[1]<=$pfs[1]){
        print "ss_dentro_pfam\n";
        print RESULT2 ">$key_2\n";
                                print RESULT2 "LCS:$par_lcs\n";
        print RESULT2 "PFAM:$fam\n";
        print RESULT2 "$eg_lcs\n";
                                print RESULT2 "$recoge_func\n";
                                $family= $key_2;
                                $family=~ s/-.+//;
                                $family_func2{$family} .= $recoge_func;
                                $count_family2{$family}++;
                    }
                    else {
                      if ($ass[0]<=$pfs[0] && $ass[0]<$pfs[1] && $ass[1]>$pfs[1]){
                                print "ss_izquierda_pfam\n";
                                print RESULT4 ">$key_2\n";
                                print RESULT4 "LCS:$par_lcs\n";
                                print RESULT4 "PFAM:$fam\n";
                                print RESULT4 "$eg_lcs\n";
                                print RESULT4 "$recoge_func\n";
                                $family= $key_2;
        $family=~ s/-.+//;
        $family_func4{$family} .= $recoge_func;
                                $count_family4{$family}++;
                      }
                      else {
                                if ($ass[0]>=$pfs[1] && $ass[1]>$pfs[1]){
                                  print "ss_fuera_derecha_pfam\n";
                                  print RESULT6 ">$key_2\n";
                                  print RESULT6 "LCS:$par_lcs\n";
                                  print RESULT6 "PFAM:$fam\n";
                                  print RESULT6 "$eg_lcs\n";
                                  print RESULT6 "$recoge_func\n";
                                  $family= $key_2;
                                  $family=~ s/-.+//;
                                  $family_func6{$family} .= $recoge_func;
                                  $count_family6{$family}++;
                                }
                                else {
                                  if ($ass[0]<$pfs[0] && $ass[1]>$pfs[0] && $ass[1]<=$pfs[1]){
                                      print "ss_derecha_pfam\n";
                                      print RESULT3 ">$key_2\n";
                                      print RESULT3 "LCS:$par_lcs\n";
                                      print RESULT3 "PFAM:$fam\n";
                                      print RESULT3 "$eg_lcs\n";
                                      print RESULT3 "$recoge_func\n";
                                      $family= $key_2;
                                      $family=~ s/-.+//;
                                      $family_func3{$family} .= $recoge_func;
                                      $count_family3{$family}++;
                                  }
                                  else {
```

```
                                  if ($ass[0]<$pfs[0] && $ass[1]>$pfs[1]){
                                    print "pfam_dentro_ss\n";
                                    print RESULT7 ">$key_2\n";
                                    print RESULT7 "LCS:$par_lcs\n";
                                    print RESULT7 "PFAM:$fam\n";
                                    print RESULT7 "$eg_lcs\n";
                                    print RESULT7 "$recoge_func\n";
                                    $family= $key_2;
                                    $family=~ s/-.+//;
                                    $family_func7{$family} .= $recoge_func;
                                    $count_family7{$family}++;
                                  }
                                  else {
                                    if ($ass[0]<$pfs[0] && $ass[1]<$pfs[0]){
                                            print "ss_fuera_izquierda_pfam\n";
                                            print RESULT5 ">$key_2\n";
                                            print RESULT5 "LCS:$par_lcs\n";
                                            print RESULT5 "PFAM:$fam\n";
                                            print RESULT5 "$eg_lcs\n";
                                            print RESULT5 "$recoge_func\n";
                                            $family= $key_2;
                                            $family=~ s/-.+//;
                                            $family_func5{$family} .= $recoge_func;
                                            $count_family5{$family}++;
                                    }
                                  }
                                }
                              }
                            }
                          }
                        }
                    }
                    else {
#                                   print "no - $_ -\n";
#                                   <STDIN>;
                    }
                  }
              }
            }
            @num_lcs= keys %eg_lcs;
            foreach $x (@num_lcs) {
              #print "****$x\n";
            }
        }
        #<STDIN>;
      }
  }
}

###RECORDAR IMPRIMIR EN ARCHIVO DE SALIDA %COUNT_NO###

@ss_dentro_pfam= keys %family_func2;
@ss_derecha_pfam= keys %family_func3;
@ss_izquierda_pfam= keys %family_func4;
@ss_fueraI_pfam= keys %family_func5;
@ss_fueraD_pfam= keys %family_func6;
@pfam_dentro_ss= keys %family_func7;

open SALE2, ">FUNCION_ss_dentro_pfam";
open SALE3, ">FUNCION_ss_derecha_pfam";
open SALE4, ">FUNCION_ss_izquierda_pfam";
open SALE5, ">FUNCION_ss_fueraI_pfam";
open SALE6, ">FUNCION_ss_fueraD_pfam";
open SALE7, ">FUNCION_pfam_dentro_ss";

$f_recoge= "";
@recoge= ();
@key_recoge= ();
sub byk2 {$a cmp $b}
foreach $clave (sort byk2 keys %family_func7) {
```

```
    print SALE2 "\n::\n";
    print SALE2 ">$slave\n";
    print SALE2 "Count_Family:$count_family2{$slave}\n";
    $f_vowgs= $family_func2{$slave};
    @vowgs= split ":", $f_vowgs;
    chomp @vowgs;
    $vo= "";
    %count_vowgs= "";
    foreach $vo (@vowgs) {
            $vo=~ s/.//gis;
            $count_vowgs{$vo}++;
    }
    @key_vowgs= keys %count_vowgs;
    $valor= ();
    @each= ();
    $letra= "";
    %FIN= "";
    sub by12 {$count_vowgs{$a} cmp $count_vowgs{$b}}
    foreach $vo_no2 (sort by12 keys %count_vowgs) {
            print SALE2 "$vo_no2: \t $count_vowgs{$vo_no2}\n";
            $valor= $count_vowgs{$vo_no2};
      @each= split "", $vo_no2;
      foreach $letra (@each) {
        $FIN{$letra}+= $valor;
            }
    }
    sub bykey2 {$a cmp $b}
    foreach $let2 (sort bykey2 keys %FIN) {
            print SALE2 "$let2:\$FIN{$let2}\n";
    }
}

$f_vowgs= "";
@vowgs= ();
@key_vowgs= ();
sub by13 {$a cmp $b}
foreach $slave3 (sort by13 keys %family_func3) {
    print SALE3 "\n::\n";
    print SALE3 ">$slave3\n";
    print SALE3 "Count_Family:$count_family3{$slave3}\n";
    $f_vowgs= $family_func3{$slave3};
    @vowgs= split ":", $f_vowgs;
    chomp @vowgs;
    $vo= "";
    %count_vowgs= "";
    foreach $vo (@vowgs) {
      $vo=~ s/.//gis;
      $count_vowgs{$vo}++;
    }
    @key_vowgs= keys %count_vowgs;
    $valor= ();
    @each= ();
    $letra= "";
    %FIN= "";
    sub by13 {$count_vowgs{$a} cmp $count_vowgs{$b}}
    foreach $vo_no3 (sort by13 keys %count_vowgs) {
            print SALE3 "$vo_no3: \t $count_vowgs{$vo_no3}\n";
            $valor= $count_vowgs{$vo_no3};
      @each= split "", $vo_no3;
      foreach $letra (@each) {
        $FIN{$letra}+= $valor;
            }
    }
    sub bykey13 {$a cmp $b}
    foreach $let3 (sort bykey3 keys %FIN) {
            print SALE3 "$let3:\$FIN{$let3}\n";
    }
}

$f_vowgs= "";
@vowgs= ();
@key_vowgs= ();
sub by14 {$a cmp $b}
foreach $slave4 (sort by14 keys %family_func4) {
    print SALE4 "\n::\n";
    print SALE4 ">$slave4\n";
    print SALE4 "Count_Family:$count_family4{$slave4}\n";
    $f_vowgs= $family_func4{$slave4};
    @vowgs= split ":", $f_vowgs;
    $vo= "";
    %count_vowgs= "";
    chomp @vowgs;
    foreach $vo (@vowgs) {
      $vo=~ s/.//gis;
      $count_vowgs{$vo}++;
    }
    @key_vowgs= keys %count_vowgs;
    $valor= ();
    @each= ();
    $letra= "";
    %FIN= "";
    sub by14 {$count_vowgs{$a} cmp $count_vowgs{$b}}
    foreach $vo_no4 (sort by14 keys %count_vowgs) {
            print SALE4 "$vo_no4: \t $count_vowgs{$vo_no4}\n";
      $valor= $count_vowgs{$vo_no4};
      @each= split "", $vo_no4;
      foreach $letra (@each) {
        $FIN{$letra}+= $valor;
      }
    }
    sub bykey14 {$a cmp $b}
    foreach $let4 (sort bykey4 keys %FIN) {
            print SALE4 "$let4:\$FIN{$let4}\n";
    }
}

$f_vowgs= "";
@vowgs= ();
@key_vowgs= ();
sub by15 {$a cmp $b}
foreach $slave5 (sort by15 keys %family_func5) {
    print SALE5 "\n::\n";
    print SALE5 ">$slave5\n";
    print SALE5 "Count_Family:$count_family5{$slave5}\n";
    $f_vowgs= $family_func5{$slave5};
    @vowgs= split ":", $f_vowgs;
    chomp @vowgs;
    $vo= "";
    %count_vowgs= "";
    foreach $vo (@vowgs) {
      $vo=~ s/.//gis;
      $count_vowgs{$vo}++;
    }
    @key_vowgs= keys %count_vowgs;
    $valor= ();
    @each= ();
    $letra= "";
    %FIN= "";
    sub by15 {$count_vowgs{$a} cmp $count_vowgs{$b}}
    foreach $vo_no5 (sort by15 keys %count_vowgs) {
            print SALE5 "$vo_no5: \t $count_vowgs{$vo_no5}\n";
      $valor= $count_vowgs{$vo_no5};
      @each= split "", $vo_no5;
      foreach $letra (@each) {
        $FIN{$letra}+= $valor;
      }
    }
    sub bykey15 {$a cmp $b}
    foreach $let5 (sort bykey5 keys %FIN) {
            print SALE5 "$let5:\$FIN{$let5}\n";
    }
}
```

```
$f_vocgs= "";
@vocgs= ();
@key_vocgs= ();
sub byk6 {$a cmp $b}
foreach $clave6 (sort byk6 keys %family_func6) {
   print SALE6 "\n /n";
   print SALE6 ">$clave6\n";
   print SALE6 "Count_Family $count_family6{$clave6}\n";
   $f_vocgs= $family_func6{$clave6};
   @vocgs= split ",", $f_vocgs;
   chomp @vocgs;
   $vo= "";
   %count_vocgs= "";
   foreach $vo (@vocgs) {
      $vo=~ s/ //gis;
      $count_vocgs{$vo}++;
   }
   @key_vocgs= keys %count_vocgs;
   $valor= ();
   @each= ();
   $letra= "";
   %FIN= "";
   sub byv6 {$count_vocgs{$a} cmp $count_vocgs{$b}}
   foreach $vo_no6 (sort byv6 keys %count_vocgs) {
            print SALE6 "$vo_no6: \t $count_vocgs{$vo_no6}\n";
      $valor= $count_vocgs{$vo_no6};
      @each= split "", $vo_no6;
      foreach $letra (@each) {
         $FIN{$letra} += $valor;
      }
   }
   sub bykey6 {$a cmp $b}
   foreach $let6 (sort bykey6 keys %FIN) {
            print SALE6 "$let6:\t$FIN{$let6}\n";
   }
}

$f_vocgs= "";
@vocgs= ();
@key_vocgs= ();
sub byk7 {$a cmp $b}
foreach $clave7 (sort byk7 keys %family_func7) {
   print SALE7 "\n /n";
   print SALE7 ">$clave7\n";
   print SALE7 "Count_Family $count_family7{$clave7}\n";
   $f_vocgs= $family_func7{$clave7};
   @vocgs= split ",", $f_vocgs;
   chomp @vocgs;
   $vo= "";
   %count_vocgs= "";
   foreach $vo (@vocgs) {
      $vo=~ s/ //gis;
      $count_vocgs{$vo}++;
   }
   @key_vocgs= keys %count_vocgs;
   $valor= ();
   @each= ();
   $letra= "";
   %FIN= "";
   sub byv7 {$count_vocgs{$a} cmp $count_vocgs{$b}}
   foreach $vo_no7 (sort byv7 keys %count_vocgs) {
            print SALE7 "$vo_no7: \t $count_vocgs{$vo_no7}\n";
      $valor= $count_vocgs{$vo_no7};
      @each= split "", $vo_no7;
      foreach $letra (@each) {
         $FIN{$letra} += $valor;
      }
   }
   sub bykey7 {$a cmp $b}
   foreach $let7 (sort bykey7 keys %FIN) {
            print SALE7 "$let7:\t$FIN{$let7}\n";
   }
}
```

```
[invitad@LUCA by_familias]$ more matar
#!/bin/tcsh
foreach madre (*_mas_depurated)
./tableando.pl $madre
end
```

```
[invitad@LUCA by_familias]$ more matar
#!/bin/tcsh
foreach madre (*_mas_depurated)
./tableando.pl $madre
end

[invitad@LUCA by_familias]$ more medif.pl
#! /usr/bin/perl

open FH, "<$ARGV[0]" or die "cannot open $ARGV[0]";
open SALE, ">$ARGV[0]_mas" or die "no se abrio salida";

@linear= <FH>;

foreach $chido (@linear) {
      if ($chido=~ /PF.+ +\w+\( +\n/) {
      $chido =~ s/$2//;
      print SALE "$chido\n"

      }else{
      print SALE "$chido";
      }
}

close FH;
close SALE;
```

```
[jrosine@LUCA by_familias]$ more tablando.pl
#!/usr/bin/perl

open DOMAINS, "</home/jdiaz/DEMA/HMMING/dominios_en_familias_mas_de_dos" or die "cannot open $ARGV[0]";
open TEST, "<$ARGV[0]" or die "cannot open $ARGV[0]";
open SALELETR, ">>matrix_by_familias_letter" or die "no se abrio salida";
open SALENUMB, ">>matrix_by_familias_numbr" or die "no se abrio salida";

if ($ARGV[0] =~ /(\w+_\w+)(_mas_de.+)/){
print SALELETR ">$1\n";
print SALENUMB ">$1\n";
}

@test = <TEST>;
foreach $line (@test) {
        if ($line =~ /(PF.+)(.+)/){
        $hash{$1} = $1;
        }
}


while (<DOMAINS>){
        chomp;
        $cada=$_;
        if ($hash{$cada}){
        print SALELETR "A";
        print SALENUMB "1";
        }else{
        print SALELETR "-";
        print SALENUMB "0";
        }
}
 print SALELETR "\n";
 print SALENUMB "\n";


close TEST;
close DOMAINS;
close SALELETR;
close SALENUMB;


[jrosine@LUCA HMMING]$ more recoge_machcados_de_hmm.pl
#!/usr/bin/perl

$lista = $ARGV[0];

open IN, "<$lista";
open OUTSI, ">$ARGV[0]_si_dominios";
open OUTNO, ">$ARGV[0]_no_dominios";

@partes = (<IN>);
$all = join ("",@partes);
@cada = split ("//", $all);

foreach $each (@cada) {
#print "$each"; <STDIN>;
        if ($each =~ /(no hits above thresholds)/){
        @no = split ("\n", $each);
                foreach $cada (@no) {
                        if ($cada =~ /Query sequence: (.+)/){
                        print OUTNO "$1\tno hits above thresholds\n";
                        }
                }

        }elsif ($each =~ /(.+\d+\/\d+.+)/){
        @si = split ("\n", $each);
                foreach $cadasi (@si) {
                #print "$cadasi"; <STDIN>;
                        if ($cadasi =~ /(Query sequence:\s+)(\w+_\w+.+)/){
                        $seq = $2;
                        print OUTSI "\n$seq";
                        }elsif ($cadasi =~ /(.+)(.+\d+\/\d+)(.+)/){
                        $dom = $1;
                        #$hash{$seq} .= $dom;
                        print OUTSI "\t$dom";
                        }

                }

        }
}

print "\n";


close IN;
close OUTSI;
close OUTNO;


[jrosine@LUCA HMMING]$ more saca_para_familias.pl
#!/usr/bin/perl

$lista = $ARGV[0];

open IN, "<$lista";
open OUT, ">$ARGV[0]_dominios";

@partes = (<IN>);
$all = join ("",@partes);
@cada = split ("//", $all);

foreach $each (@cada) {
#print "$each"; <STDIN>;
        if ($each =~ /(no hits above thresholds)/){
        @no = split ("\n", $each);
                foreach $cada (@no) {
                        if ($cada =~ /Query sequence: (.+)(.\d+.+)/){
                        #print "\n$2\tno hits above thresholds";
                        }
                }

        }elsif ($each =~ /(.+\d+\/\d+.+)/){
        @si = split ("\n", $each);
                foreach $cadasi (@si) {
                #print "$cadasi"; <STDIN>;
                        if ($cadasi =~ /(Query sequence:\s+)(\w+_\w+)(.+)/){
                        $seq = $2;
                        }elsif ($cadasi =~ /(.+)(.+\d+\/\d+)(.+)/){
                        $dom = $1;
                        $hash{$seq} .= $dom;

                        }
                }
        }
}
print "\n";

close IN;
close OUT;

#@keys= keys %hash;
#foreach $uno(@keys) {
#print "$uno $hash{$uno}\n";
#}

sub byk {$a cmp $b}
foreach $ec (sort byk keys %hash) {
   print "\n>$ec\n$hash{$ec}\n"; #<STDIN>;
}
```

```
[irenisa@LUCA Blasting]$ more busca.pl
#! /usr/bin/perl

open FH, "$ARGV[0]" or die "cannot open $ARGV[0]";

open SALENO, ">>/home/irma/Perfiles_Virus/Blasting/no" ) die "cannot open no_jalaron_en_blast";
open SALESI, ">>/home/irma/Perfiles_Virus/Blasting/si" ) die "cannot open no_jalaron_en_blast";

@linear= <FH>;
$junto= join ("", @linear);

if ($junto=~ /QUERY/ && $junto=~ /  Database /) {
      #print ".";
      print SALESI "$ARGV[0]\n";
      print ".";

}else{
      print SALENO "$ARGV[0]\n";
      print "$ARGV[0]\n";
}

close FH;
close SALE;
```

```
[irenisa@LUCA Blasting]$ more sustar
#!/bin/tcsh
foreach readie (ec*)
./LEE.pl $readie
end
```

```
[irenisa@LUCA Blasting]$ more metedor.pl
#!/usr/bin/perl

#Comentarios del programa a: Irma Lozada (irma_lozada@correo.unam.mx)
#y a: J. Javier Díaz Mejía (jdiaz@ifc.unam.mx)

#en este arreglo pueden poner las dos rutas de los directorios de comparado y completo

my @directorios = qw(/home/irma/Perfiles_Virus/Blasting/comparaciones);
#open SALECHEC, ">/home/irma/Perfiles_Virus/Blasting/todo";

foreach my $i (@directorios){
   chdir $i;
   opendir DIR, ".";
   my @archivos= readdir DIR;
   &baja (@archivos);
   closedir DIR;
}

sub baja{
            #print "me metí @_";
   my @archivos= @_;
   if ($#archivos> 1){
      @archivos= @archivos[2..$#archivos];
      foreach (@archivos){
         if ($_ =~ /\~/){
            system "rm $_";
         }
      }


      foreach (@archivos){
         chomp;

### Aquí se ponen el archivo (s) a los que les quiera hacer  cierta madre
            if ($_ =~ /ARCHIVO/){
            system "mv $_ /home/irma/Perfiles_Virus/Blasting/mah";
            }
      }
   }
}

#close SALECHEC;
```

```
[irenisa@LUCA Blasting]$ more modif.pl
#! /usr/bin/perl

open FH, "$ARGV[0]" or die "cannot open $ARGV[0]";
open SALE, ">depurado_$ARGV[0]" ) die "cannot open";

@linear= <FH>;

foreach $chido (@linear) {


      if ($chido=~ /(^Query= )(.+)(\n)/) {
            $name=$2;

      }elsif ($chido =~ /^QUERY/) {
      $chido =~ s/^QUERY.+\n//;
      print SALE "\n";
      }elsif ($chido =~ /^[A-Z]{1}[0-9]/ && $chido !~ /^BLASTP/ && $chido !~ /^Reference/ && $chido !~ /^Stephen/ && $chido !~
/^Database/ && $chido !~ /^Searching/ && $chido !~ /^Sequences/ && $chido !~ /^Lambda/ && $chido !~ /^Matrix/ && $chido !~
/^Gap/ && $chido !~ /^Number/ && $chido !~ /^T / && $chido !~ /^A / && $chido !~ /^X1 / && $chido !~ /^X2 / && $chido !~ /^X3/
&& $chido !~ /^S1 / && $chido !~ /^S2 /) {
      print SALE "$chido";


      }
}

close FH;
close SALE;
```

```
[illegible] PROGRAMAS_FUNCIONES]$ more saca_dominios_from_pfam.pl
#!/usr/bin/perl

################################################################################
### Este programa depura el OUTFILE de una busqueda usando HMMER contra PFAM
################################################################################

################################################################################
### Necesita:
###
### 1) Archivo de salida de HMMER contra PFAM (OUTFILE_de_HMMER), aqui llamado (FH)
###
### 2) Lista de equivalencias de nombre y numero del modelo de PFAM, aqui llamado (IDS):
###     2-Hacid_dh       PF00389
###     2-Hacid_dh_C     PF02826
###     2-oxoacid_dh     PF00198
###     2-ph_phosp       PF04029
###     2_5_RNA_ligase   PF02834
###     2OG-FeII_Oxy     PF03171
################################################################################

################################################################################
### Uso:
###
### saca_dominios_from_pfam.pl OUTFILE_de_HMMER
################################################################################

################################################################################
### Proporciona:
###
### nombre_de_la_secuencia numero_o_nombre_del_dominio_PFAM posicion e-value
################################################################################


$entra = $ARGV[0];


open FH, "<$entra" or die "no pude abrir el archivo de entrada\n";
open IDS, "</home/[illegible]/DATABASES/PFAM/IDs_PFv11" or die "no abrio ids\n"; #Modificar el pwd de IDS si es necesario


################################################################################
### Se forma el hash con los nombres de los modelos PFAM y sus numeros equivalentes
###

while (<IDS>) {
	if ($_ =~ /[illegible]/) {
	$ids{$1} = $3;
	}
}

###
################################################################################

################################################################################
while (<FH>) {
### Se recorre el archivo OUTFILE de HMMER buscando el nombre de las secuencias ($seq)
	if ($_ =~ /^Query sequence: +(.+)(\n)/) {
	$seq = "$2";
		if ($seq =~ [illegible]) {
		}else{
		$seq = $seq;
		}
### Se busca el o los dominios PFAM encontrados ($pf)
	}elsif ($_ =~ /^[illegible]/) {
	$pf = $1;
		if ($ids{$pf}) {
### Se imprime el nombre_de_la_secuencia numero_de_dominio posicion e-value
		print "$seq\t$ids{$pf} $3-$7 $19\n";
		}else{
### Si no se tiene su equivalente del nombre y numero de modelo PFAM
### Se imprime el nombre_de_la_secuencia nombre_de_dominio posicion e-value
		print "$seq\t$1 $3-$7 $19\n";
		}
### Se busca si no se encontraron dominios en esa secuencia y se omiten los resultados de esa secuencia
	}elsif ($_ =~ /no hits above thresholds/) {
	$seq = "";
	}
}
```