Tesis Doctoral

Motivos lineales en proteínas de virus humanos involucradas en el ciclo celular

Glavina, Juliana

2019

Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Química Biológica

Motivos lineales en proteínas de virus humanos involucradas en el ciclo celular

---

Tesis presentada para optar por el título de Doctor de la Universidad de Buenos Aires en el área Química Biológica

Lic. Juliana Glavina

Director de tesis: Dr. Ignacio E. Sánchez
Consejera de estudios: Dra. Valeria Levi

Lugar de trabajo: Departamento de Química Biológica, FCEN-UBA. Instituto de Química Biológica - Ciencias Exactas y Naturales. UBA-CONICET.

Ciudad Autónoma de Buenos Aires, 27 de marzo 2019.

Motivos lineales en proteínas de virus humanos involucradas en el ciclo celular

Los motivos lineales son elementos de secuencia que comúnmente se encuentran en dominios intrínsecamente desordenados. Consisten, en promedio, de cinco residuos que determinan la función y participan en interacciones proteína-proteína. Los virus se enfrentan a una presión de selección constante debido a ambientes cambiantes y a la respuesta inmune del hospedador. Es común que usen motivos para secuestrar la maquinaria celular mimetizando proteínas del hospedador. Se postula que estos motivos, al ser elementos de secuencia cortos, juegan un rol en la evolución adaptativa ya que adquieren o modifican su función con pocas mutaciones. Sin embargo, existen pocas evidencias que apoyen esta hipótesis.

Este trabajo se enfoca en el estudio de motivos lineales en dos proteínas virales, E7 de Papilomavirus y E1A de Adenovirus. El objetivo es investigar las relaciones genotipo-fenotipo en virus causantes de infecciones persistentes en humanos, utilizando para esto más de 100 secuencias para cada proteína. La amplia distribución de hospedadores, que involucra amniotas en el caso de E7 y mamíferos en el caso de E1A, y el número de motivos que presentan estas proteínas - ocho y doce respectivamente - permiten estudiar la relación entre motivos y fenotipos y el rol adaptativo de los motivos en la historia evolutiva viral.

Este estudio se realizó principalmente mediante tres tipos de análisis. Primero, se realizaron análisis de secuencia incluyendo co-evolución entre pares de residuos. Luego, se realizaron estudios de co-especiación globales y basados en eventos evolutivos - considerando co-especiación, duplicación, cambio de hospedador y extinción parcial. Por último, se realizaron pruebas estadísticas de asociación. Los análisis de secuencia permitieron determinar que la falta de una estructura globular no implica un menor grado de conservación de secuencia. El combinar este tipo de análisis con la filogenia viral permitió afirmar que los motivos en las proteínas estudiadas presentan numerosos eventos de aparición y desaparición. Los estudios de co-evolución en secuencia y las pruebas de asociación revelaron que los distintos motivos y regiones de las proteínas estudiadas no evolucionan de manera independiente. Por último, las pruebas de asociación aplicadas a estudios de filogenia revelaron que dos eventos evolutivos - cambio de hospedador y extinción parcial - y la aparición/desaparición de motivos tampoco son procesos independientes. En conjunto, los resultados obtenidos en este trabajo sugieren que los motivos lineales presentan una alta plasticidad evolutiva, independiente del contexto estructural, y establecen las bases de la contribución de los motivos lineales en la evolución adaptativa viral.

Palabras claves: MOTIVOS, PROTEÍNA, VIRUS, DESORDEN, SECUENCIAS, PAPILOMAVIRUS, ADENOVIRUS, ADAPTATIVO, CO-EVOLUCIÓN.

Linear motifs within human viral proteins involved in cell cycle

Linear Motifs are sequence elements commonly found within intrinsically disordered domains. They are on average five function-determining residues long and mediate protein-protein interactions. Viruses are under constant selection pressure because of their changing environment and host immune response. They usually use linear motifs to hijack the host cellular machinery by mimicking host proteins. Since motifs are short sequences, it is believed that they play a role in adaptive evolution by modifying or acquiring their function with a few mutations. However, there is scarce evidence to support this hypothesis.

This work focuses on the study of linear motifs within two viral proteins, E7 from Papillomavirus and E1A from Adenovirus. The main goal is to investigate genotype-phenotype relationships in viruses causing persistent infections in humans, using more than 100 sequences for each protein. The wide host distribution, which involves amniotes in the case of E7 and mammals in the case of E1A, and the number of motifs present in these proteins - eight and twelve respectively - enable us to study the relationship between motifs and phenotypes and the adaptive role of motifs in viral evolutionary history.

This work was performed mainly by applying three types of analyses. First, sequence analyses, including co-evolution between pairs of residues were performed. Then, co-speciation analyses were carried out, both global and event based - considering co-speciation, duplication, host-switch and partial extinction. Finally, statistical association tests were performed. Sequence analyses allowed to determine that the lack of globular structure does not correlate with a lower degree of sequence conservation. Combining these analyses with viral phylogenetics showed that the motifs from these proteins have a high number of appearance and disappearance events. Sequence co-evolution studies and association tests revealed that different motifs and regions from the proteins under study do not evolve independently. Last, association tests applied to phylogenetic studies revealed that two evolutionary events - host switch and partial extinction - and motif appearance/disappearance events are not independent either. Altogether, the results of this work suggest that linear motifs have a high evolutionary plasticity independent of the structural context and provide a molecular mechanism for adaptive viral evolution.

Keywords: MOTIFS, PROTEIN, VIRUS, DISORDER, SEQUENCES, PAPILLOMAVIRUS, ADENOVIRUS, ADAPTIVE, CO-EVOLUTION.

Agradecimientos

Las palabras que utilice y el tamaño de esta sección no alcanzan para reflejar la enorme gratitud ni el enorme cariño que siento por todas las personas aquí nombradas.

A Nacho, por embarcarse en este proyecto, darme un lugar donde pueda desarrollar mi doctorado, estar en las buenas y en las malas, por su confianza, su apoyo y sus enseñanzas.

A la Facultad de Ciencias Exactas y Naturales, por mi formación, por la excelente calidad académica y de investigación y por ser modelo de universidad pública. A todos los que día a día trabajan para que eso siga siendo posible.

A todos los integrantes del LFP. Por el tiempo que se tomaron para discutir mi trabajo y el suyo, por las risas y los mates. A Diego, por sus palabras duras pero ciertas, su entusiasmo constante y transmitir su pasión por la ciencia. A mis compañeros de laboratorio. A Gonza por el empuje continuo a superarme y valorarme. A Brenda por su cariño, alegría y apoyo. A Rocío, por todas nuestras charlas, discusiones y mates compartidos. A todos aquellos que pasaron por el laboratorio: Nico Palopoli, por todas nuestras charlas y proyectos, a Nina y Nico M. A Eze, María, Cesar y Lucio. A todos ellos por el apoyo continuo. Todos los integrantes del LFP además de hacer que el laboratorio funcione, nos empujamos a mejorar cada día. Juntos creamos un ambiente que es difícil de dejar y que se llega a extrañar apenas te vas cada día.

A todos los miembros de QB6, QB65, QB10. Al Dr. Pato Craig y al Dr. Ale Nadra por los aportes científicos. A los chicos que siempre estaban dispuestos a compartir alegrías, frustraciones, un poco de yerba cuando ya era tarde para comprar y un poco de tiempo para despejar dudas. A Ari A. y Sol por nuestras charlas. A Bucci, Naty y muchos más!

A todos los que colaboraron en esta tesis, por dedicarle el tiempo de hacer este proyecto juntos: Gonzalo de Prat-Gay, Leo A, Cris M, Vale R, Ernest, Ricardo, Lu C y de nuevo, Ro y Cesar!

A la Dra. Valeria Levi por todo su apoyo y dedicación, por sus palabras sinceras y su presencia constante, por su lucha y el empuje que le pone a la ciencia.

A Lucía Chemes, excelente científica y una de mis mentoras, por fomentar en mí el pensamiento crítico y el amor por la investigación, por todo su infinito apoyo, por conocer mis defectos y mis virtudes, por tomarse el tiempo de discutir mi trabajo, por enseñarme, empujarme a expresarme y por mostrarme que no es malo equivocarse. Por valorarme y motivarme siempre a hacer más. Por todos los sábados compartidos y por toda la confianza en mí.

A todo el grupo del EMBL, Bálint, Lena, Davide, Marc, Manjeet, Hugo, Malvika y Norman. En especial muchas gracias a Toby por toda su amabilidad, charlas, calidez, enseñanzas y por recibirme.

A Javi S, por todas nuestras juntadas, discusiones eternas y por escucharme siempre.

A mis compañeros de carrera, Vicky, Sebas, Yael, Naty y Raque. Por los días de estudio, mates y resúmenes compartidos, y en especial por seguir compartiendo muchísimos momentos lindos.

A Lucas P, por todos los años de amistad, porque siempre nos hacemos un tiempito aunque cueste y por todas las salidas.

A Lau L, por compartir el amor por la biología y hacerme sentir siempre acompañada. A Vane por su cariño, sus risas y los mates eternos.

A Raque y Teo, sin ellos yo no estaría acá. Raque, por toda su sabiduría, siempre será mi primera y mejor maestra. A Teo por su apoyo, empuje y discusiones. Por ser excelentes abuelos de Emma, cuidarla siempre que yo no podía y amarla. A mis 4 hermanos, mis compañeros de la vida y mis mejores amigos. A Pablo por

ser ejemplo de buscar lo que a uno le apasiona, sin importar dónde o cuánto cueste. A Andrés por iniciarme en el hermoso mundo de la programación y presentarme numerosos desafíos. A Mariángeles por hacerme sentir siempre querida, por mostrarme lo fuerte que se puede ser, por el aguante en las demostraciones a pedido y por todo el amor y cuidado que le brinda a Emma. A Ceci, por su bondad y lucha, por todos los mates que compartimos y por todas las noches en vela. A Walter por todo su cariño, las ricas comidas y las eternas charlas. A todos ellos, gracias por ser tan nerds, por divertirnos tanto juntos, por compartir locuras y estar siempre, en los momentos lindos y en los difíciles. A mis sobrinos, Alba, Raquel, Joaquín, Helena, Angué y Alan. Joaquín y Helena gracias por mantener latente durante tantos años la criatura que llevo adentro. A Angué por sus travesuras y que junto con Emma hicieron crecer la curiosidad inherente a mi y replantearme las cosas más simples.

A toda mi familia, tíos, primos y sobrinos segundos. Por nuestras gigantes y hermosas reuniones familiares, por nuestras locuras, risas y diversiones.

A mis amigos y hermanos del alma: Lala, Gerva y Edu. Por todas las noches compartidas, por todas las risas y lágrimas compartidas. Por ser los mejores tíos postizos de Emma, por amarla y cuidarla. A Gerva por ser mi IT incondicional. A Edu por entenderme y buscarme la solución cuando me encapricho con algo. A Lala mi confidente incondicional, por estar en desacuerdo siempre y quererme igual. Son mi gran apoyo. A los tres por desconectarme, por las salidas, por todas las nerdeadas que hicimos y por reírse conmigo de las cosas más simples.

A mi amiga de la infancia, Gisi. Por estar siempre, a pesar de todo, distancias y diferencias. Por las alegrías y tristezas compartidas. Por hacer mi vida más feliz.

A Carlos, Claudia y Vero por su acompañamiento y apoyo, y por todo su amor a Emma.

A Mariano, por acompañarme en esta montaña rusa que puede llegar a ser la vida, tanto en las subidas como en las bajadas, con tropezones y alegrías, por el café en las mañanas, por crecer a mi lado, por esta familia que hoy tenemos y por intentarlo.

Por último, a la personita más importante de mi mundo. A Emma ...

Gracias por llegar al mundo. Gracias por ser tan feliz. Gracias por tu sonrisa, tus cantos, tus bailes y tus juegos. Gracias por mostrarme la fuerza que llevo adentro. Gracias por tu paciencia. Gracias por ser complicada y compleja, y obligarme siempre a reinventar la vida para conectarme con vos. Gracias por recordarme lo importante que son las cosas simples. Gracias por que aún siendo tan chiquita me enseñas y sorprendes con algo nuevo todos los días. Pero por sobre todo gracias por todo el amor, dulzura, alegría y buena onda que irradiás a todos los que te rodean. Nuncas dejes de bailar, cantar y divertirte, pero por sobre todo nunca dejes de preguntar porqué. Te cuento que ya terminé de trabajar, gracias por esperar, ...

... esta tesis te la dedico a vos, Emma.

Trabajos presentados en el marco de esta tesis

- Glavina, J., Román, E. A., Espada, R., de Prat-Gay, G., Chemes, L. B., y Sánchez, I. E. (2018). *Interplay between sequence, structure and linear motifs in the adenovirus E1A hub protein*. Virology, 525(May):117–131

- Chemes, L. B., Glavina, J., Alonso, L. G., Marino-Buslje, C., de Prat-Gay, G., y Sánchez, I. E. (2012a). *Sequence evolution of the intrinsically disordered and globular domains of a model viral oncoprotein.* PLoS One, 7(10):e47661

- Chemes, L. B., Glavina, J., Faivovich, J., de Prat-Gay, G., y Sánchez, I. E. (2012b). *Evolution of linear motifs within the papillomavirus E7 oncoprotein.* Journal of molecular biology, 422(3):336–46

- Dinkel, H., Van Roey, K., Michael, S., Davey, N. E., Weatheritt, R. J., Born, D., Speck, T., Krüger, D., Grebnev, G., Kuban, M., Strumillo, M., Uyar, B., Budd, A., Altenberg, B., Seiler, M., Chemes, L. B., Glavina, J., Sánchez, I. E., Diella, F., y Gibson, T. J. (2014). *The eukaryotic linear motif resource ELM: 10 years and counting.* Nucleic acids research, 42(Database issue):D259–66

En preparación:

- Glavina, J., Rodríguez de la Vega, R., Risso V.A., Faivovich, J., de Prat-Gay, G., Chemes, L. B., y Sánchez, I. E. *E1A linear motifs contribute to adaptive Mastadenovirus evolution.*

Índice general

Capítulo 1

Introducción

En este capítulo introduzco los conceptos, aspectos y el contexto necesarios para comprender el trabajo realizado. Consta de seis secciones. En la primera sección describo parte del universo de las proteínas, haciendo enfoque principalmente en las proteínas o regiones desordenadas. Incluyo aspectos estructurales, funcionales y los métodos utilizados para el estudio de las mismas. La segunda sección está enfocada en los elementos de secuencia llamados motivos lineales. La tercera y la cuarta sección describen las proteínas centrales de este trabajo, la proteína E7 de papilomavirus y la proteína E1A de adenovirus. En la quinta sección describo cómo los virus utilizan los motivos lineales para asegurar su supervivencia, poniendo como ejemplo las interacciones con la proteína retinoblastoma reguladora del ciclo celular. En la última sección describo y fundamento la hipótesis y los objetivos de la tesis.

1.1. Proteínas intrínsecamente desordenadas

1.1.1. Definición de proteína

Una proteína es un polímero macromolecular construido a partir de 20 aminoácidos posibles unidos entre sí por un enlace peptídico. Los veinte aminoácidos constan de un núcleo central común y difieren en sus cadenas laterales (Figura 1.1), que les confieren propiedades fisicoquímicas particulares.

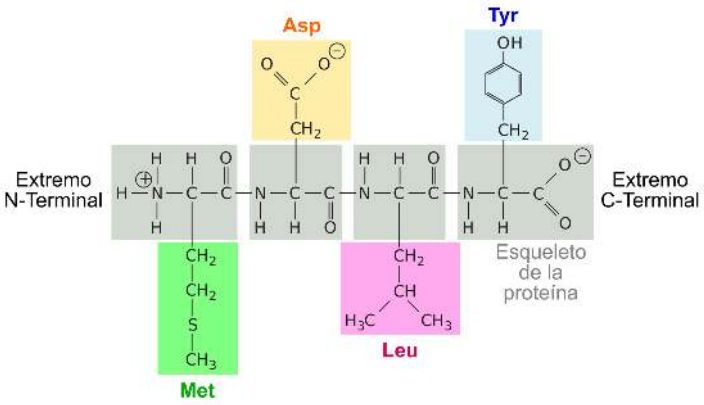


Figura 1.1: Estructura molecular de un péptido. Se muestra la estructura de cuatro aminoácidos unidos entre si, metionina (Met), ácido aspártico (Asp), leucina (Leu) y tirosina (Tyr). En gris está resaltado el esqueleto del péptido. En colores se resaltan las cadenas laterales de cada uno de los aminoácidos.

La estructura de una proteína se suele describir en cuatro niveles distintos. El orden de los aminoácidos de una proteína determina su estructura primaria. Las proteínas pueden tener plegamientos locales que definen su estructura secundaria en tres estados posibles: hélice α, lámina β o *coil*. La estructura tridimensional de una proteína, o plegamiento global, determina su estructura terciaria. Finalmente, las proteínas pueden ser monoméricas o adquirir una estructura cuaternaria combinándose con otras unidades del mismo tipo (formando homodímeros/trímeros/etc) o de distinto tipo, en cuyo caso se denominan heterodímeros/trímeros/etc. (Figura 1.2).

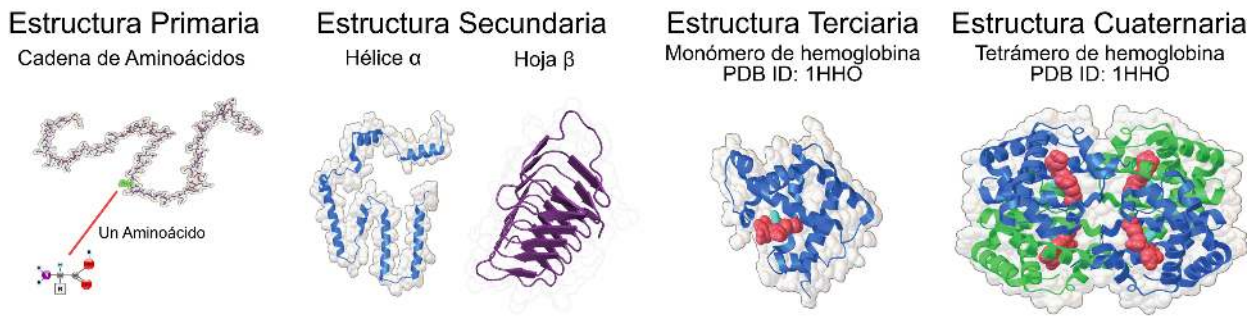


Figura 1.2: Esquema de la estructura primaria, secundaria, terciaria y cuaternaria de una proteína. La estructura primaria se un péptido está representada como una cadena de palillos. La estructura secundaria está representada como cintas. La estructura terciara y cuaternaria de la mioglobina están representadas en forma de cintas (PDB ID: 1HOO) (Shaanan, 1983). Figura adaptaba de la base de datos PDB.

La estructura de una proteína, sus interacciones intramoleculares y sus funciones están definidas tanto por la secuencia específica de aminoácidos como por el entorno (Anfinsen, 1973).

La longitud promedio de una proteína es aproximadamente 250 residuos. Si se consideran los 20 aminoácidos, una proteína de longitud promedio podría presentar 20^{250} secuencias posibles, aproximadamente 2 10^{325}. Este número es mayor que el número de átomos que se estiman que existen en el universo, entre 4 10^{78} y 6 10^{79}. Esto sugiere que las proteínas naturales no pueden explorar todo el espacio de secuencia disponible. Sin embargo, dicho número supone una exploración aleatoria de las combinaciones de aminoácidos. También, hay que considerar que algunos aminoácidos poseen similitudes en su carácter fisicoquímico y que las proteínas se encuentran en un entorno biológico con requerimientos funcionales y presiones evolutivas específicas, que varían en el espacio y tiempo y acotan el universo de secuencia explorable (Dryden *et al.*, 2008). Por lo tanto, cada proteína tendrá propiedades de secuencia, estructura y actividad biológica definidas por su evolución en el marco de un sistema biológico cambiante.

Las proteínas son moléculas muy versátiles que se encuentran en todos los sistemas biológicos. Participan en numerosos procesos biológicos: en los procesos metabólicos como catalizadores, en la localización de moléculas como transportadoras a través de membranas y en solución, en la transducción de señales y en el soporte mecánico de la célula. La amplia diversidad de secuencias, conformaciones y funciones hacen que las proteínas sean un atractivo objeto de estudio.

1.1.2. Desorden intrínseco

En el año 1958 se resolvió por primera vez la estructura de una proteína globular por difracción de rayos X, la mioglobina (Kendrew *et al.*, 1958). Desde entonces y hasta la fecha existen más de 147000 estructuras depositadas en la base de datos banco de datos de proteínas (PDB) (en inglés, *Protein Data Bank*). Este crecimiento continuo del conocimiento estructural llevó al desarrollo del paradigma secuencia-estructura-función: la secuencia de una proteína determina una única estructura, que a su vez determina la funcionalidad de la proteína (Figura 1.3, izquierda). En consecuencia, es muy común que se asocien directamente los términos proteína y dominio proteico a proteínas con función conocida y con una estructura tridimensional caracterizada por posiciones fijas de sus átomos que varían poco alrededor de un estado de equilibrio. Sin embargo, en las últimas décadas se descubrió que existe otra clase de proteínas y regiones proteicas que no poseen una única estructura tridimensional. Los resultados experimentales obtenidos por las técnicas más comunes no se podían explicar en el contexto de una estructura globular. Antes de 1950, había evidencia relacionada a la proteína de la leche, llamada caseína, que marcaba una diferencia con las proteínas globulares. La caseína mantenía sus propiedades en condiciones que favorecen el desplegado como ser el calentamiento prolongado y luego del tratamiento con agentes desnaturalizantes como urea o guanidinio (Oldfield y Dunker, 2014). Utilizando dispersión rotacional óptica, Jirgensons (1958) propuso clasificar las proteínas en base a sus conformaciones y su propuesta incluía una categoría denominada desorden basada en las propiedades anómalas de la proteína fosvitina de la yema del huevo, que poseía valores de dispersión bajos similares a los

de las proteínas desnaturalizadas. Arnone *et al.* (1971) determinan por cristalografía de rayos X la estructura de una nucleasa de *Staphylococcus aureus*, sin embargo, para algunos residuos no se podía asignar estructura. Los experimentos en la proteína del fibrinógeno realizados por Doolittle (1973) sugerían la presencia de regiones no plegadas en la proteína. En primer lugar, una región del extremo C-terminal de la cadena α mostraba un espectro de dicroísmo celular diferente al de las proteínas globulares y en segundo lugar, el fibrinógeno mostraba ser sensible a la degradación por proteasas. Bode *et al.* (1976) resolvieron la estructura del tripsinógeno mediante cristalografía de rayos X y observaron la falta de densidad electrónica en una región de 15 residuos en el extremo amino terminal sugiriendo alta movilidad. En conjunto, esta evidencia indica la existencia de regiones de proteínas y proteínas que existen como un conjunto heterogéneo de confórmeros que no pueden ser descriptos por un conjunto único de posiciones atómicas. Esta nueva clase de proteínas, a diferencia de las proteínas globulares que están ordenadas o estructuradas, se describen como desordenadas o desestructuradas y dan origen a la biología no estructural. Hoy se sabe que una gran fracción del proteoma de cualquier organismo consiste en secuencias polipeptídicas no estructuradas (Figura 1.3, derecha) llamadas regiones intrínsecamente desordenadas (IDRs) (en inglés, *Intrinsically Disordered Regions*). El conjunto del proteoma puede considerarse como un conjunto de proteínas con combinaciones de regiones estructuradas y desordenadas (Figura 1.3, abajo) (van der Lee *et al.*, 2014).

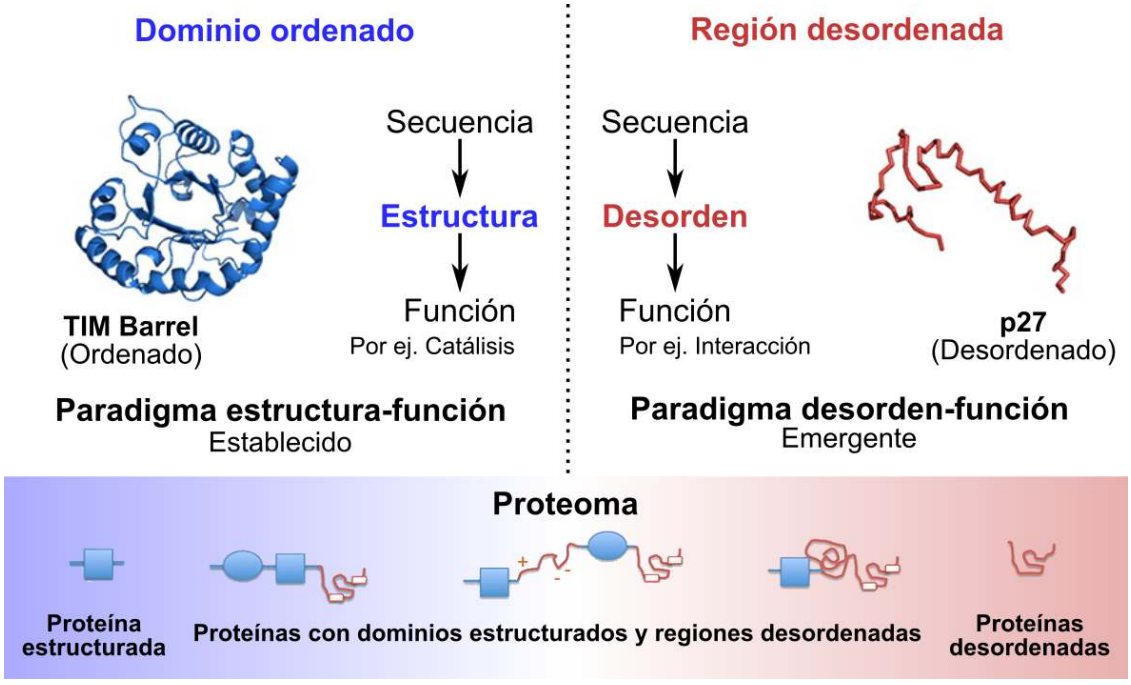


Figura 1.3: Evolución del paradigma estructura-función. Figura adaptada de van der Lee *et al.* (2014).

En Dunker *et al.* (2013) realizan una recolección de los numerosos nombres que fueron utilizados para describir estas proteínas (Figura 1.4). La mayoría de los nombres hacen referencia a la falta de plegado o definición estructural en el estado nativo y se refieren al desorden como carac-

terística intrínseca de estas proteínas, marcando la diferencia con la visión "clásica" de la proteína globular.

Figura 1.4: Nombres utilizados para describir las proteínas intrínsecamente desordenadas. Se muestran los distintos nombres que se utilizan en la literatura. A la derecha e izquierda están los nombres asociados a alguna característica observada en las proteínas intrínsecamente desordenadas. En el centro se indican los términos utilizados actualmente como combinaciones de adjetivos. Se señalan en rojo las palabras que mantienen la visión de las proteínas "clásicas". Figura adaptada de Dunker *et al.* (2013).

Los fragmentos proteicos desordenados presentan distintas longitudes. Pueden encontrarse en forma de conectores entre dos dominios globulares (*loops*, en inglés), ser regiones que abarcan más de 30 residuos de longitud, en cuyo caso se los llama IDRs, o bien toda la proteína puede carecer de una estructura ordenada. En este último caso, se la llama proteína intrínsecamente desordenada (IDP) (en inglés, *Intrinsically Disordered Protein*).

Según las predicciones computacionales realizadas sobre bases de datos inclusivas de secuencias proteicas, se sabe que el desorden estructural es abundante en todas las especies. Sus niveles son significativamente más altos en eucariotas que en procariotas. Se estima que entre el 10 y el 35 % de proteínas procariotas y el 15 al 45 % de proteínas eucariotas contienen IDRs de por lo menos 30 residuos de longitud. En humanos, el 31 % de las proteínas son 35 % IDPs y el 44 % contienen IDRs. Esto sugiere que existe una gran parte de las proteínas existentes que carecen de las características clásicas de las proteínas globulares y demanda un nuevo enfoque en el estudio de las proteínas. Por último, las IDPs/IDRs son más comunes en virus que en bacterias, en especial en virus con genomas muy compactos. Esto se debe a que la flexibilidad de estas regiones permite la interacción con múltiples proteínas del hospedador, maximizando la relación entre funcionalidad codificada y la información genética (Tokuriki *et al.*, 2009; Xue *et al.*, 2014).

Propiedades *estructurales* de las proteínas desordenadas

La información relacionada con las IDPs/IDRs comenzó a aumentar a partir de los años 90s luego de los primeros estudios de resonancia magnética nuclear (RMN) y biología computacional relacionados con IDPs/IDRs (Oldfield y Dunker, 2014). Las IDPs/IDRs estudiadas por RMN

muestran que existen como un conjunto de confórmeros que intercambian conformaciones en escalas de tiempo entre los pico y milisegundos, sin estructura terciaria y bajo contenido de estructuras secundarias (van der Lee *et al.*, 2014), como se muestra en la Figura 1.5.

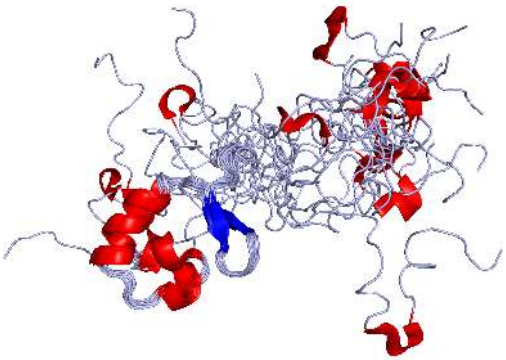


Figura 1.5: Proteína globular con una región intrínsecamente desordenada. Asamblea estructural de la proteína parcialmente desordenada At2g23090 de *Arabidopsis thaliana* (PDB: 1WVK). La estructura fue resuelta mediante resonancia magnética nuclear y se crearon 15 modelos. En rojo se indican los elementos de estructura secundaria hélice α y en azul las láminas β.

Existen numerosos ensayos experimentales que proveen evidencia directa de desorden (Eliezer, 2009; Marsh y Forman-Kay, 2012; Na *et al.*, 2018; Uversky, 2015). RMN provee información a nivel de residuo relacionada con la dinámica y la estructura de una proteína en solución. Por lo tanto, permite determinar desorden local, transiciones estructurales y desorden. Para una proteína globular los picos de resonancia individuales están esparcidos porque el contexto local magnético de cada núcleo difiere significativamente. Sin embargo, para una proteína desordenada los distintos picos tienen a estar más juntos porque el contexto es más similar debido a la alta movilidad. La dispersión de rayos X de ángulo pequeño (SAXS) permite evaluar las dimensiones y forma de una proteína midiendo las diferencias a nivel nanoescala en densidades electrónicas en una muestra. Se utiliza para determinar los parámetros hidrodinámicos, permitiendo determinar si una proteína está plegada o no. A nivel de molécula única la transferencia de energía de resonancia Forster (FRET) puede medir la dinámica y las conformaciones individuales no unidas del conjunto de confórmeros y los estados intermediarios cuando se induce el plegado. La espectroscopía por dicroísmo circular en el UV lejano (CD) (en inglés, *far UV CD*) permite determinar elementos de estructura secundaria. El espectro de CD de las IDPs presenta una banda fuerte negativa cerca de los 200 nm, similar a *random coil* (las hélices α presentan dos mínimos en 208 y 222 nm y un máximo en 190 nm, mientras que las lámina β presentan un mínimo en 217 nm y un máximo en 198 nm) y cambia radicalmente cuando se induce la formación de estructuras secundarias. La resonancia paramagnética electrónica (EPR) permite estudiar la estructura y dinámica de las proteínas utilizando marcación con nitróxidos. También se puede marcar el grupo tiol de cisteínas existentes o introducidas por mutagénesis dirigida. El espectro resultante está relacionado con la movilidad de la cadena en la región donde está la marca ya que está determinado por la distancia a otros centros paramagnéticos y accesibilidad al solvente. El espectro de una proteína IDP se

caracteriza por picos muy marcados debido a la alta movilidad de la marca. La proteólisis limitada permite determinar las regiones expuestas al solvente, es decir, cuan desplegada está la proteína. La dispersión dinámica de luz (DLS) estudia las fluctuaciones temporales de la intensidad de luz desviada debido a movimientos hidrodinámicos en solución. Por lo tanto, permite determinar el radio hidrodinámico de una proteína, R_H, también llamado radio de Stokes. El valor de R_H refleja el tamaño aparente adquirido por la proteína en solución. La comparación de los distintos radios en condiciones que favorezcan el desplegado permite determinar cuán plegada o desplegada está una proteína.

Parece razonable pensar que la alta flexibilidad de las IDPs/IDRs es una propiedad necesaria para su función. Algunas IDPs/IDRs pueden adoptar de manera transitoria estructuras secundarias. La población de confórmeros observada en determinadas condiciones puede variar, por ejemplo al modificarse el contexto o mediante modificaciones post-traduccionales (García-Alai *et al.*, 2006). Un ejemplo que caracteriza el cambio estructural frente a la modificación del contexto es que la mayoría de las IDPs/IDRs sufren una transición de desorden a orden al unirse a sus blancos proteicos. En la interacción participa sólo un fragmento de la IDPs/IDRs que, en general, es anfipático e interacciona de manera específica y con baja afinidad con uno o varios bolsillos de la superficie de su blanco proteico, permitiendo una rápida y espontánea disociación (Wright y Dyson, 2015). La formación del complejo implica al menos dos eventos: (1) La interacción con el blanco proteico y (2) el plegado de la IDPs/IDRs. Estos dos eventos pueden o no ocurrir de manera simultánea. El mecanismo de selección conformacional de la IDPs/IDRs por parte del blanco proteico postula que ocurre primero el plegado de la IDPs/IDRs y luego la unión. El mecanismo de ajuste inducido postula que la interacción con el blanco proteico induce el plegado de la IDPs/IDRs.

La teoría de paisajes energéticos (Bryngelson *et al.*, 1995) permite describir muchas características físicas y estructurales del plegado y estabilidad de las proteínas. En la Figura 1.6 se representa el paisaje energético para una proteína globular (izquierda) o desordenada (derecha). El eje *x* representa cada una de las conformaciones posibles y el eje *y* la energía libre. Las proteínas globulares poseen un paisaje energético en forma de embudo (Figura 1.6, izquierda), con un mínimo energético global que representa el estado nativo. En el caso de las IDPs/IDRs la diversidad conformacional se corresponde con múltiples mínimos isoenergéticos en un paisaje energético más chato (Figura 1.6, derecha).

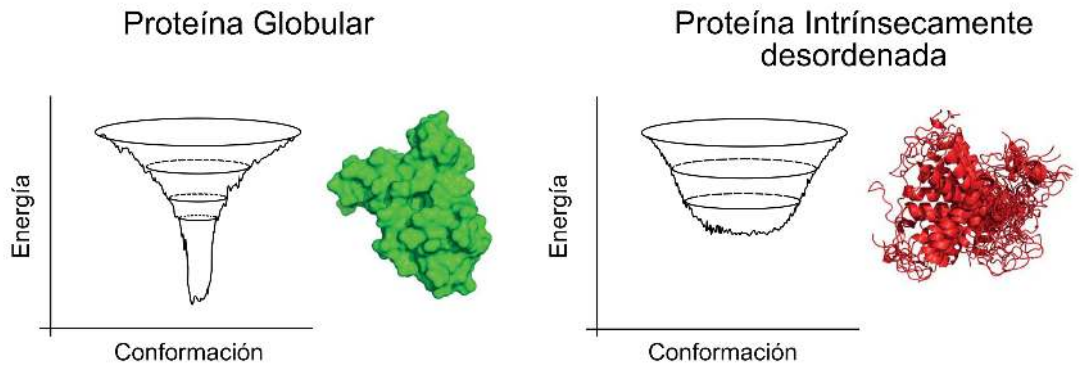


Figura 1.6: Paisajes energéticos conformacionales y representaciones de proteínas ordenadas y desordenadas. *Izquierda.* La visión clásica establece que una proteína puede describirse por una superficie similar a un embudo con un mínimo global que corresponde al estado nativo como el que se muestra en color verde en representación de superficie. Esta estructura es una asamblea estructural de 17 modelos obtenidos por RMN del dominio KIX de la proteína de unión a CREB (PDB: 1KDX). *Derecha.* Las IDPs poseen un paisaje energético caracterizado por mínimos isoenergéticos que representan los estados conformacionales de menor energía como el que se muestra en color rojo en representación de cintas. Esta estructura corresponde a la asamblea conformacional de 17 estados del dominio inhibidor de quinasas de la proteína p27 (PED: PED2AAA) obtenido a partir del análisis combinado por simulaciones de dinámica molecular y RMN. Figura adaptada de Pauwels *et al.* (2017).

En resumen, no es posible describir el estado nativo de las IDPs/IDRs como una estructura tridimensional promedio. Deben describirse a partir de una población de estados conformacionales diferentes separados por pequeñas barreras energéticas (Pauwels *et al.*, 2017). La descripción de estas asambleas requiere de la utilización conjunta de diversos métodos biofísicos. Estos métodos deben definir características como ser la estructura secundaria residual y la formación de contactos transitorios de largo alcance.

Propiedades de secuencia de las proteínas desordenadas

Como se dijo anteriormente, el experimento de Anfinsen (Anfinsen, 1973) sugiere que la estructura tridimensional de una proteína está codificada en la secuencia de la misma. Si bien la predicción de una estructura tridimensional a partir de la secuencia es aún un desafío sin resolver, es posible reconocer algunas características de secuencia de las proteínas globulares, como la existencia de un núcleo hidrofóbico. En el caso de las IDPs/IDRs es también esperable que su secuencia determine el grado de desorden y que presenten características de secuencia que las diferencie de las proteínas globulares. Así, el plegado o ausencia del mismo de una proteína en condiciones fisiológicas estaría determinado por su secuencia de aminoácidos.

Las IDPs/IDRs presentan una baja complejidad de secuencia y un desvío en la composición de aminoácidos en comparación a las proteínas globulares (Brown *et al.*, 2011). La primera observación realizada es que poseen un bajo contenido de aminoácidos hidrofóbicos y una alta proporción de aminoácidos cargados (Xie *et al.*, 1998) en comparación a las proteínas globulares (Figura 1.7). En particular, están enriquecidas en aminoácidos que favorecen el desorden como ser alanina, gli-

cina, serina, prolina, glutamina, ácido glutámico, lisina y arginina, y empobrecidas en aminoácidos que favorecen el orden como ser isoleucina, valina, leucina, fenilalanina, cisteína, triptofano, tirosina y asparagina (Uversky, 2017).

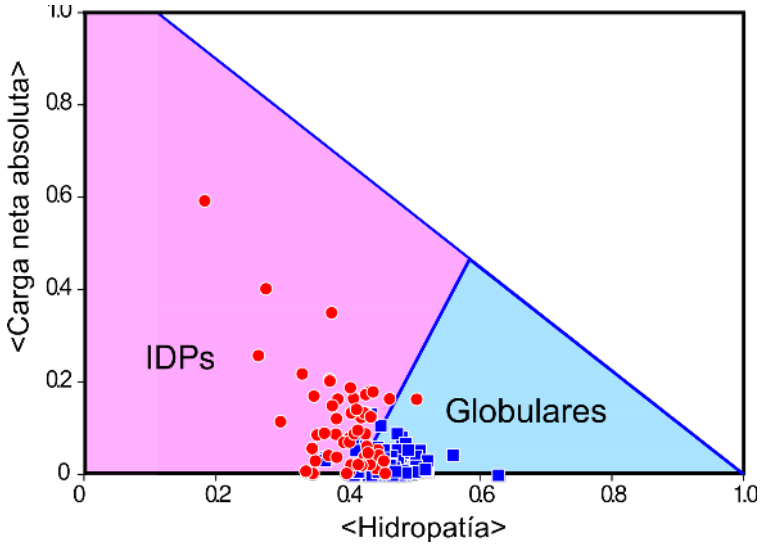


Figura 1.7: Relación entre la carga neta y la hidropatía en proteínas globulares y desordenadas. En el eje *x* se grafica la hidropatía media y en el eje *y* se indica la media de la carga neta absoluta. La hidropatía se define en base a las propiedades hidrofóbicas e hidrofílicas de cada aminoácido. Las áreas definidas para las proteínas globulares y desordenadas están indicadas en celeste y rosa respectivamente. Figura adaptada de Uversky (2013).

Los estudios que intentan relacionar la composición de secuencia con las asambleas conformacionales de las IDPs/IDRs parten de la definición de tres tipos de regiones: regiones polares, polielectrolitos y polianfolitos (Mao *et al.*, 2013). Las regiones polares están enriquecidas en aminoácidos polares como glutamina, asparagina, serina, glicina y prolina y son deficientes en residuos cargados o hidrofóbicos. Estas regiones polares tienden a colapsar prefiriendo la auto-solvatación y dan origen a una asamblea de conformaciones compactas con estabilidad similar llamadas glóbulos. La composición aminoacídica de los polielectrolitos está enriquecida en aminoácidos cargados y forman estructuras desordenadas expandidas. Los polianfolitos están enriquecidos también en residuos cargados, pero el número de cargas positivas y negativas es similar. La asamblea conformacional va a estar gobernada por la distribución lineal en la secuencia de residuos con carga opuesta. Si estos están segregados, la atracción electrostática entre cargas opuestas hará que colapsen resultando en una conformación de glóbulo. Pero si no están segregados las repulsiones y atracciones se balancean y dependiendo de la carga neta adoptará una conformación de glóbulo o una conformación conocida como *random-coil*. Para estudiar esta característica, Mao *et al.* (2010) utilizan un conjunto de polipéptidos ricos en arginina que corresponden a IDPs asociadas con la condensación de la cromatina durante la espermogénesis y el empaquetamiento de genomas virales. Utilizando simulaciones y experimentos de fluorescencia para estimar el radio de giro, muestran que la transición entre conformaciones de glóbulos, extendidas o *random-coil* es nítida. Estos resultados sugieren que pequeños cambios en la carga neta a través de modificaciones post-traduccionales

como fosforilación o acetilación puede producir la transición modulando la conformación (van der Lee *et al.*, 2014).

En resumen, la distribución de carga neta y aminoácidos hidrofóbicos en la secuencia de las IDPs/IDRs impide que puedan formar interacciones que estabilicen una única estructura y define un conjunto de asambleas conformacionales mayoritariamente desplegadas.

1.1.3. Bases de datos de proteínas desordenadas

Así como existe el PDB, múltiples bases de datos se crearon para almacenar la información obtenida a partir de experimentos con IDPs/IDRs. El primer lanzamiento de la base de datos de desorden proteico DisProt (en inglés, *Database of Protein Disorder*, `http://www.disprot.org`) fue en el año 2004. DisProt almacena los límites de las IDRs determinadas experimentalmente y la metodología utilizada, junto con la actividad y función biológica si se la conoce. DisProt es actualizada continuamente mediante una curación manual de la literatura. El último lanzamiento de la base de datos (DisProt 7.0) fue en el año 2017 (Piovesan *et al.*, 2017) con más de 800 entradas (Agosto, 2018). El primer lanzamiento de la base de datos MobiDB (`http://mobidb.bio.unipd.it`) fue en el año 2012. MobiDB combina tres tipos de datos relacionados con IDPs/IDRs: datos manualmente curados, evidencia indirecta de desorden y evidencia predicha a nivel computacional. Los datos manualmente curados se extraen de la base de datos DisProt, entre otras. La evidencia indirecta de desorden se infiere de las estructuras almacenadas en la PDB, considerando los residuos ausentes (en inglés, *missing residues*) y de alta temperatura (en inglés, *B-factors*) en estructuras determinadas por cristalización de rayos X. También se infiere de las regiones móviles en las asambleas de estructuras determinadas por RMN. Por último, MobiDB utiliza seis predictores computacionales de desorden con sus distintas configuraciones para realizar la anotación de regiones desordenadas cuando no existe otra información disponible. Además, informa un consenso pesado de desorden calculado a partir de los tres tipos de datos. El último lanzamiento de la base de datos (MobiDB 3.0) fue en el año 2017 (Piovesan *et al.*, 2018) y posee más de 12000 entradas curadas manualmente, más de 40000 entradas con evidencia indirecta y todas las entradas de la base de datos de secuencias de proteínas no redundantes UniParc (Agosto, 2018). La base de datos D2P2 (`http://d2p2.pro`) se creó en 2013 (Oates *et al.*, 2013) y almacena predicciones de desorden de genomas completos. Por último, FuzDB (`http://protdyn-database.org`) (Miskei *et al.*, 2017), se creó en 2015 y almacena dos tipos de evidencia experimental: (1) evidencia estructural, determinada por cristalografía de rayos X, RMN, dicroísmo circular, FRET, SAXS, entre otros y (2) evidencia bioquímica relacionada a la funcionalidad de la proteína, determinada por ensayos de afinidad, actividad transcripcional, proteolítica o enzimática, entre otros.

Otras bases de datos sobre IDPs se centran en información estructural. La base de datos IDEAL (`http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL`) fue creada en el año 2011. IDEAL recolecta la información de IDPs verificadas experimentalmente, enfocándose en las regiones que se pliegan luego de unirse a su blanco proteico, es decir, que poseen evidencia

experimental para el estado aislado desordenado y para el estado ordenado unido. Actualmente (Noviembre 2017) (Fukuchi *et al.*, 2014) incluye más de 900 entradas. El primer lanzamiento de la base de datos pE-DB (`http://pedb.vib.be`) fue en el año 2013 (Varadi *et al.*, 2014). pE-DB recolecta la información estructural de las asambleas de IDPs y proteínas desnaturalizadas determinadas por RMN o SAXS. La última versión (pE-DB 2.0) cuenta con 24 entradas con 60 asambleas con más de 25000 estructuras proteicas (Agosto 2018).

En resumen, las bases de datos de IDPs son numerosas y poseen distintas características. Su desarrollo se incrementó en la última década, facilitando el estudio de las IDPs y permitiendo profundizar el conocimiento de esta nueva clase de proteínas cuyas características moleculares y funcionalidades recién se están empezando a comprender.

1.1.4. Algoritmos de predicción de proteínas desordenadas

Uno de los mayores desafíos en el campo de las proteínas es la predicción de la estructura tridimensional de la proteína a partir de su estructura primaria, incluyendo aquellas proteínas que son total o parcialmente desordenadas. Mientras que las proteínas globulares adquieren una única estructura nativa, las proteínas desordenadas son un conjunto de estructuras tridimensionales. La predicción de IDRs a partir de la secuencia de aminoácidos permite un análisis rápido y abarcativo de distintas proteínas permitiendo establecer hipótesis sobre la presencia de desorden en las proteínas (Dunker *et al.*, 2008; van der Lee *et al.*, 2014). La importancia que adquirieron las IDRs/IDPs en los últimos años llevó a su incorporación en el experimento mundial comunitario para la predicción estructural proteica CASP (en inglés, *Critical Assessment of Techniques for Protein Structure Prediction*, `http://predictioncenter.org`). La primera incorporación fue en el CASP5 en el año 2002 y su última incorporación fue en el CASP10 en el año 2014. Las CASP se realizan cada dos años y consisten en evaluar en forma comparativa y objetiva los métodos de predicción estructural desarrollados por distintos grupos de investigación sobre un conjunto de proteínas entregado. El objetivo principal es contribuir al desarrollo de métodos que permitan predecir la estructura a partir de la secuencia, los cuales mejoraron sustancialmente en los últimos años.

Existen numerosos métodos de predicción, pero en general se basan en tres estrategias: (1) predicción de desorden a partir de composición de secuencia, (2) a partir del aprendizaje automatizado (en inglés, *machine learning*) sobre estructuras determinadas por cristalografía de rayos X y (3) los meta-predictores que integran los resultados predichos por diferentes métodos.

Entre los algoritmos que se basan en composición de secuencia podemos nombrar IUPred (Dosztányi *et al.*, 2005a,b; Mészáros *et al.*, 2018), que aplica un campo de energía desarrollado a partir de un gran número de proteínas con estructura determinada obtenidas de PDB y SLIDER (Peng *et al.*, 2014), que analiza las características fisicoquímicas de los aminoácidos y la complejidad y composición de la secuencia. El primer algoritmo desarrollado en base al aprendizaje automatizado fue PONDR (Obradovic *et al.*, 2003; Romero *et al.*, 1997), entrenado a partir de un grupo estructuras de proteínas globulares y atributos de secuencia asociados a residuos no resuel-

tos. En los años siguientes se desarrollaron mediante aprendizaje automatizado numerosos algoritmos que "aprenden" características de secuencia a partir de un grupo de estructuras. Por ejemplo, GlobPlot (Linding *et al.*, 2003b) fue entrenado estudiando la tendencia de un residuo a adquirir determinada estructura secundaria, hélices α o láminas β. DisEMBL (Linding *et al.*, 2003a) y DISOPRED2 (Jones y Ward, 2003; Ward *et al.*, 2004) además estudian el grado de movilidad determinado según el factor de temperatura del C_α y la tendencia a ser un residuo no resuelto en la estructura. El servidor Spritz (Vullo *et al.*, 2006) y sus algoritmos derivados CSpritz (Walsh *et al.*, 2011) y ESpritz (Walsh *et al.*, 2012) incluyen en su conjunto de entrenamiento proteínas desordenadas determinadas experimentalmente y la flexibilidad estructural determinada a partir de datos de RMN. SPOT-Disorder (Hanson *et al.*, 2017) fue entrenado para poder diferenciar, además de las características estructurales, los patrones evolutivos de secuencia. Los algoritmos meta-predictores son más recientes. Por ejemplo, metaPrDos (Ishida y Kinoshita, 2008) y MFPd (Mizianty *et al.*, 2011) utilizan siete predictores diferentes y aprenden de las predicciones de manera automatizada mediante redes vectoriales de soporte (SVMs) (en inglés, *Support Vector Networks*).

Los algoritmos de predicción de desorden existentes son variados y se basan en distintas características ya que no existe un estándar definido para la asignación de regiones desordenadas. Daughdrill *et al.* (2011) estudian la correlación entre los valores de RMN y los predictores de desorden. Los valores de NHNOE obtenidos por RMN permiten medir la dinámica del esqueleto de la proteína con resolución a nivel de residuo. Utilizando como modelo el dominio de transactivación desordenado de p53 proveniente de seis organismos distintos y los valors de NHNOE Daughdrill *et al.* (2011) observan una buena correlación en la identificación a nivel de residuo en las predicciones de tres predictores de desorden entrenados diferentes, incluyendo IUPred. Dado un conjunto de prueba de proteínas con estructuras conocidas, los mejores predictores identifican con una precisión del 80 % las regiones globulares y/o desordenadas. Sin embargo, esta precisión depende del conjunto de datos de prueba analizado ya que muchos de ellos están siendo entrenados a partir de un conjunto de estructuras donde los residuos terminales son no resueltos o poseen alta movilidad (Dosztányi, 2018).

En esta tesis se utiliza el algoritmo de predicción IUPred, ya que a diferencia de los otros algoritmos se basa en una estimación de energía y captura las diferencias fundamentales entre las propiedades físicas de regiones ordenadas y desordenadas (Dosztányi, 2018).

1.1.5. Funciones características de las proteínas intrínsecamente desordenadas

Las IDRs/IDPs presentan una gran diversidad de funciones y son responsables de muchas patologías. Determinar la relación secuencia-estructura-función es uno de los grandes desafíos en proteínas globulares y la ausencia de estructura en las IDRs/IDPs impone el desarrollo de un nuevo paradigma.

A partir de una curación manual de la literatura relacionada con más de 150 proteínas que contenían regiones desordenadas de más de 30 residuos demostradas experimentalmente, en Dun-

ker *et al.* (2002) determinaron 28 funciones diferentes asociadas. Estas funciones incluyen interacciones proteína con proteína, con ácidos nucleicos - ácido desoxirribonucleico (ADN), ácido ribonucleico (ARN), ARN de transferencia (ARNt), ARN ribosomal (ARNr), ARN mensajero (ARNm) y ARN genómico - y con lípidos, unión a sustratos, cofactores y metales, por último, exposición de sitios de modificación post-traduccional como acetilación, glicosilación, metilación y fosforilación, entre otros. Un análisis bioinformático de la base de datos SwissProt reveló una correlación entre IDRs/IDPs predichas y 238 de las 710 palabras claves que identifican las distintas funciones proteicas (Xie *et al.*, 2007), demostrando el amplio repertorio funcional. La mayoría de las funciones se relacionaban con señalización y procesos celulares claves para la supervivencia, como ser la diferenciación celular y la regulación de la transcripción, traducción y el ciclo celular. Las funciones de las IDRs/IDPs pueden agruparse en tres grandes grupos (Figura 1.8) que a su vez pueden clasificarse en seis clases funcionales según si hay o no interacción con un blanco y cuanto se mantiene en el tiempo la interacción (Tompa, 2005; van der Lee *et al.*, 2014).

Figura 1.8: Clasificación funcional de las regiones intrínsecamente desordenadas. La función de las IDRs puede clasificarse en tres grandes grupos. Las IDRs pueden no unirse a un blanco y cumplir funciones únicamente como cadenas entrópicas (rojo), pueden unirse de manera transiente a un blanco (verde) o pueden unirse de manera más permanente como proteínas efectoras (azul). Figura adaptada de van der Lee *et al.* (2014).

En el primer grupo se incluyen aquellas IDRs que cumplen su función por ser cadenas entrópicas y no establecen una interacción con un blanco proteico (Figura 1.8, rojo). Es el caso de los conectores que permiten la movilidad de los dominios que conectan. En el segundo grupo se incluyen aquellas IDRs que cumplen su función mediante la interacción transiente con un blanco proteico (Figura 1.8, verde). Este grupo abarca dos categorías principales, las chaperonas y las IDRs que llevan a cabo su función mediante la exposición de sitios de modificaciones post-traduccionales. La exposición de un sitio de modificación post-traduccional se ve facilitada por la alta flexibilidad de las IDRs, ya sea como sustrato enzimático que recibe la modificación o como sitio de reconocimiento por una proteína efectora una vez modificado (Dyson y Wright, 2005; Wright y Dyson, 1999). El tercer y último grupo abarca a aquellas IDRs que ejercen su función mediante

una interacción más estable con un blanco proteico e incluye tres categorías, las proteínas efectoras, ensambladoras y colectoras (Figura 1.8, azul). Las proteínas efectoras interactúan con otras proteínas y modifican su actividad. Por ejemplo, la proteína E1A de adenovirus forma complejos con las proteínas celulares proteína retinoblastoma (pRb) o proteína de unión a la proteína CREB (CBP) (en inglés, *CREB Binding Protein*), con cooperatividad positiva o negativa dependiendo en la disponibilidad de los sitios de interacción. La interacción con pRb disminuye la probabilidad de unión a CBP y la unión de CBP disminuye la probabilidad de unión a pRb (Ferreon *et al.*, 2013). En la célula infectada, la proteína E1A participa de procesos que involucran a CBP de manera independiente de pRb modulando la activación o represión de la transcripción. E1A también participa en procesos que involucran a pRb de manera independiente de CBP modulando la progresión del ciclo celular. Por último, E1A también posee funciones específicas de diferenciación que involucra procesos que dependen de CBP y pRb. La cooperatividad positiva en la formación del complejo ternario podría aumentar la acetilación de pRb mediada por CBP promoviendo la salida permanente del ciclo celular y la diferenciación de la célula hospedadora. La cooperatividad negativa podría facilitar la formación de complejos binarios y la interacción con otros blancos proteicos. Ambos tipos de cooperatividad permitirían entonces la formación de distintos complejos moleculares que facilitarían la creación de un contexto que favorezca la replicación viral (Ferreon *et al.*, 2013). Las proteínas ensambladoras actúan atrayendo las distintas proteínas que forman grandes complejos proteicos. Por último, las proteínas colectoras almacenan y neutralizan pequeños ligandos, como ser el almacenamiento de calcio en la caseína de la leche. Estas tres categorías y la exposición de sitios de modificación son en su mayoría ejercidas gracias a la presencia de motivos lineales, objeto de estudio de esta tesis y descritos en detalle en una sección aparte (véase Sección 1.2).

Proteínas desordenadas y patologías. Las IDRs/IDPs están involucradas en numerosas enfermedades debido a la amplia variedad de funciones en las que están involucradas y el rol que ejercen en los procesos celulares vitales para la supervivencia. Las IDRs/IDPs están asociadas a cáncer y enfermedades cardiovasculares, así como a enfermedades generadas por patógenos como los virus (Uversky, 2011). Otro grupo de enfermedades está asociado a la formación de agregados proteicos tóxicos formados por las IDPs, incluyendo las enfermedades de Parkinson y Alzheimer (Huang y Stultz, 2009). Las perturbaciones que sufren las IDPs que llevan al desarrollo de una patología pueden ser mutaciones puntuales que produzcan una transición de desorden a orden en la IDP o modifiquen sitios de modificación post-traduccional o de interacción con proteínas blanco produciendo un cambio (Forman-Kay y Mittag, 2013).

Ventajas de las proteínas desordenadas. Desde el reconocimiento de su existencia han surgido múltiples propuestas para explicar las proteínas desordenadas en términos de ventajas evolutivas. Las supuestas ventajas evolutivas de las proteínas desordenadas derivan principalmente de su alta flexibilidad, que les permitiría cumplir sus funciones de manera óptima. Liu y Huang (2014) describe diferentes ventajas, algunas de las cuales están relacionadas con las funciones descriptas al

comienzo de la sección. Por ejemplo, las IDRs/IDPs sobrellevan restricciones estéricas en la unión a una proteína blanco debido a la alta flexibilidad, facilitan las modificaciones post-traduccionales por exposición de los sitios y permiten la existencia de conectores flexibles. Además, las IDPs logran interacciones de alta especificidad, baja afinidad y alta tasa de disociación con sus proteínas blanco gracias al acoplamiento entre el plegado y la interacción. Se razona que estas propiedades son esenciales para la transducción de señales y la regulación de procesos celulares. Las IDPs previenen la agregación debido a la baja hidrofobicidad y elevada carga neta de sus secuencias. En contraste a las proteínas ordenadas, vulnerables a perder su estructura en condiciones no nativas, las IDPs tienen mayor tendencia a mantener una asamblea funcional de conformaciones en condiciones extremas. Las proteínas ordenadas poseen en principio un espacio de secuencia más reducido que las desordenadas debido a la necesidad de formar una estructura ordenada soluble. Por último, las IDPs economizan el genoma y los recursos proteicos. La interfaz de interacción en los complejos proteína-proteína es similar en tamaño para las IDPs y las proteínas ordenadas, pero la secuencia utilizada para crear la misma interfaz de interacción es más corta para las IDPs, ya que interaccionan a través de estructuras extendidas sin núcleo hidrofóbico. Esta ventaja se ve reflejada en los virus, cuyo genoma es pequeño y cuyo contenido de IDRs/IDPs es alto. Otro mecanismo que favorece un menor tamaño de genoma gracias al desorden es la superposición de regiones que cumplen distintas funciones adoptando diferentes conformaciones. Hasta el momento, la mayor parte de las ventajas evolutivas expuestas parecen razonables pero no se han podido poner a prueba de manera directa.

1.1.6. Evolución de proteínas desordenadas

Una de las grandes preguntas que existen en la relación a las IDPs es la forma en que evolucionan. En el caso de las proteínas ordenadas la evolución se ve restringida con el objetivo de mantener una estructura globular y una función. Las proteínas desordenadas no necesitan mantener una estructura única. En este caso, la restricción evolutiva está dada por mantener una estructura desordenada determinada y, al igual que en el caso de las globulares, la función.

La evolución de las proteínas puede estudiarse comparando secuencias proteicas homólogas que comparten un ancestro común (Brown *et al.*, 2011). Los modelos evolutivos de proteínas se pueden crear comparando posición a posición la frecuencia de cada aminoácido y analizando la probabilidad de cambio de un aminoácido por otro, reflejando las mutaciones puntuales que son aceptadas por la evolución por tener un efecto neutro o positivo en la función de la proteína. Las IDRs/IDPs tienen, en general, una tasa de evolución más rápida que las proteínas ordenadas, incluyendo cambios en sustituciones de aminoácidos, expansiones de repeticiones e inserciones y deleciones (van der Lee *et al.*, 2014). Otro estudio, que compara modelos evolutivos construidos a partir de proteínas ordenadas y desordenadas, muestra que las IDPs tienen mayor probabilidad de aceptar un cambio no conservativo en comparación a las proteínas ordenadas (Brown *et al.*, 2010). En particular, los aminoácidos leucina, tirosina, triptofano y prolina se encuentran más conservados en las IDRs que otros. Otro estudio, comparando la tasa evolutiva de proteínas homólogas

predichas desordenadas dentro del proteoma humano reveló que los residuos predichos desordenados poseen una mayor tasa evolutiva que los que se presentan en regiones ordenadas (van der Lee *et al.*, 2014). Sin embargo, el rango de tasas evolutivas de los residuos desordenados es mucho más amplio que el de las regiones ordenadas y se superponen, sugiriendo cierto grado de conservación de algunos residuos de las proteínas desordenadas (van der Lee *et al.*, 2014). Esta conservación probablemente esté relacionada con relevancia funcional y pertenencia a regiones que median interacciones proteína-proteína como los motivos lineales (véase Sección 1.2).

En relación a la evolución de las propiedades estructurales y funcionales de las proteínas desordenadas, se han realizado principalmente estudios basados en simulaciones y en predicciones computacionales. Un estudio compara la conservación de estructura secundaria y desorden intrínseco predichos sobre un conjunto de secuencias desordenadas y globulares luego de someterlas a procesos de mutación puntual *in silico* bajo tres modelos evolutivos diferentes basados en proteínas globulares (Schaefer *et al.*, 2010). Las estructuras secundarias parecen ser robustas, mientras que las regiones desordenadas parecen ser más susceptibles a perder la característica de desorden y las proteínas ordenadas tienden a mantener la característica de orden. Otro trabajo evaluó la conservación de IDRs desde tres aspectos distintos: perfil de desorden predicho, conservación de aminoácidos según la tendencia a favorecer orden o desorden y la superposición entre regiones ordenadas y desordenadas. El resultado reveló que a pesar de las diferencias en secuencia las IDRs pueden llevar a cabo la misma función en distintos organismos, ya que muchas de sus características dependen de la composición de aminoácidos y no de la secuencia en sí misma (Toth-Petroczy *et al.*, 2008). Este resultado sugiere una alta permisividad en la secuencia.

Aún no existe una metodología de consenso para abordar el estudio evolutivo de proteínas desordenadas, debido a sus características tan diferentes de las proteínas globulares. Los estudios realizados hasta el momento concluyen de manera similar que una IDP con una función determinada puede presentar más mutaciones neutras y positivas que una proteína ordenada debido a la redundancia de secuencia permitida para alcanzar un fenotipo ventajoso (Liu y Huang, 2014). Sin embargo, no resulta del todo claro cuáles son aquellas características de las IDRs/IDPs sometidas a la presión de selección o si ésta es mayor o menor sobre unos residuos que otros.

1.2. Motivos Lineales

La diversidad de las funciones celulares llevadas a cabo por las proteínas se ve representada por el alto repertorio de módulos proteicos de interacción con propiedades de interacción distintivas (Van Roey *et al.*, 2014). Los dominios globulares son los módulos de interacción más conocidos pero no son los únicos. En las décadas pasadas se reveló que las IDRs/IDPs tienen un rol central en los procesos celulares. Muchas de las funciones de las IDRs/IDPs son llevadas a cabo por sitios de interacción que carecen de una conformación predefinida y median la mayoría de los procesos de regulación dentro de la célula. Los elementos funcionales más comunes son pequeñas regiones de aminoácidos continuos llamados motivos lineales. El término motivo hace referencia a una figura o diseño que se repite y fue apropiado por la biología para referirse a un patrón de nucleótidos o aminoácidos que corresponde a un módulo funcional autónomo (Van Roey y Davey, 2015) que se repite en una clase de moléculas relacionadas o no. Por lo tanto, un motivo proteico puede definirse como un patrón repetido en un conjunto de cadenas polipeptídicas que desempeña una función.

Los motivos lineales son fragmentos de secuencia cortos de entre tres y doce residuos funcionales en promedio que median interacciones proteína-proteína. Un ejemplo fácilmente reconocible son las secuencias proteicas que corresponden a señales de localización, transportando a la proteína a su organela blanco mediante la interacción con otra proteína. Secuencias de este tipo como la señal de localización nuclear (NLS) (en inglés, *Nuclear Localization Signal*) (Kalderon *et al.*, 1984) fueron adquiriendo importancia por su rol funcional y siendo reconocidas de manera temprana como un patrón de secuencia (Dingwall y Laskey, 1991). La Tabla 1.1 muestra la secuencia de algunos ejemplos de NLS.

Nro. Acceso	Proteína	Secuencia de NLS	Organismo
Q969H0	FBXW7	SKRRRT	*Homo sapiens* (Humano)
P38398	BRCA1	KRKRRP	*Homo sapiens* (Humano)
P25054	APC	PKKKKP	*Homo sapiens* (Humano)
Q16644	MAPKAPK3	NKRRKK	*Homo sapiens* (Humano)
Q62315	Jarid2	RKRPRL	*Mus musculus* (Ratón doméstico)
P18870	JUN	CRKRKL	*Gallus gallus* (Gallina)
P05411	JUN	SRKRKL	Avian sarcoma virus 17
P03255	E1A	CKRPRP	Human adenovirus 5
P03096	VP2	QKKKRR	Murine polyomavirus
P03073	Antígeno Large T	RKRPRP	Murine polyomavirus
P04608	tat	GRKKRR	Human immunodeficiency virus 1
P20919	rev	SKRRRK	Equine infectious anemia virus

Tabla 1.1: Secuencias de localización nuclear. Se muestran las distinas secuencias proteicas correspondientes al péptido de localización nuclear en proteínas de distintos organismos. En la primera columna y segunda columna se indica el número de acceso de Uniprot y la proteína correspondiente. En la tercera columna la secuencia reportada y en la cuarta el organismo correspondiente. Los ejemplos se obtuvieron de la base de datos de motivos lineales eucariotas (ELMdb) (en inglés, *Eukaryotic Linear Motif Database*).

A simple vista, las secuencias parecen compartir como característica una alta densidad de carga positiva y sugieren la existencia de un patrón de secuencia.

Recién en el año 1990 (Fred Dice, 1990) se reconoce a los motivos lineales en general como elementos funcionales particulares y se los agrupa para su estudio, introduciendo la característica de lineales en contraposición a la estructura tridimensional de un dominio globular. A partir de entonces y hasta la fecha se utilizaron diferentes nombres siendo los más relevantes mini-motivos (Kadaveru *et al.*, 2008), motivos lineales cortos o motivos lineales eucariotas (Davey *et al.*, 2012). En la Figura 1.9 se listan algunas de las palabras comúnmente utilizadas para describirlos como ser secuencia corta de aminoácidos (Kalderon *et al.*, 1984), secuencia consenso (Dingwall y Laskey, 1991), motivos de secuencia proteicos (Fred Dice, 1990), sitios cortos funcionales (Puntervoll *et al.*, 2003), patrones de secuencia cortos (Neduva *et al.*, 2005), elementos modulares o interruptores de interacción (Neduva y Russell, 2005), patrones de secuencia (Dinkel y Sticht, 2007), segmentos peptídicos cortos (Petsalaki *et al.*, 2009), motivos de unión cortos (Tan *et al.*, 2006), motivo de reconocimiento peptídico (Liu *et al.*, 2007), motivos peptídicos lineales funcionales (Dinkel y Sticht, 2007) y motivos cortos lineales funcionales (Walsh *et al.*, 2011).

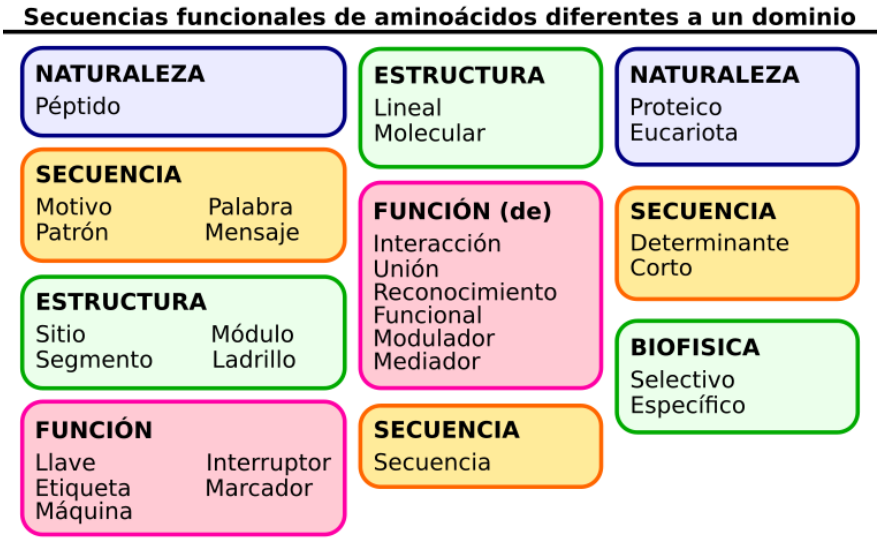


Figura 1.9: Nombres utilizados para describir los motivos lineales. Se muestran los distintos nombres que se utilizan en la literatura agrupados según a que característica hacen referencia. En general, los nombres surgen de la combinación de una palabra de la primera columna con una o dos palabras de la segunda y tercera columna.

Algunos de los términos utilizados para su descripción no son del todo exactos. Es muy común el uso indistinto de motivos lineales eucariotas y motivos lineales cortos (SLiMs) (en inglés, *Short Linear Motifs*). Ambos términos prestan a confusión. El término corto aparece en contraposición a la longitud de secuencia que posee un dominio, pero sigue siendo subjetivo. En segundo lugar, los motivos lineales no son exclusivos de organismos eucariotas. Está bien establecida la presencia de motivos lineales en virus (Davey *et al.*, 2011b), donde son una mecanismo principal para secuestrar la maquinaria celular en las células infectadas (véase Sección 1.5). También se han iden-

tificado motivos lineales de bacterias que median interacciones patógeno-hospedador. Por ejemplo, *Salmonella spp* utiliza un motivo que inactiva las MAP quinasas del hospedador (Zhu *et al.*, 2007), inhibiendo la activación de la respuesta inmune. Llamarlas secuencia consenso o palabras no las describe con exactitud ya que se pierde la variabilidad de secuencia que presentan. La utilización de la palabra módulo tampoco es útil, ya que no los diferencia de los dominios globulares. Incluso el término lineal que se utiliza para diferenciarlos de los dominios globulares no es del todo correcto, ya que muchos de los motivos lineales se pliegan luego de unirse a su blanco proteico.

La mayoría del conocimiento de los motivos lineales surge del estudio de unas pocas proteínas con IDRs caracterizadas experimentalmente, sugiriendo que el número de motivos lineales es mucho mayor del que se conoce (Van Roey *et al.*, 2014). La estimación de motivos lineales en el proteoma humano en base a las características de los distintos predictores (véase Sección 1.2.2) estima que existen más de 100000 posibles motivos en el proteoma humano, y más de un millón de posibles sitios de modificación post-traduccional (Tompa *et al.*, 2014). Sin embargo, dada la alta plasticidad evolutiva que pueden presentar los motivos lineales, no todos ellos necesariamente son funcionales. La mejor forma de poder lograr identificar mejor los motivos lineales es caracterizando motivos especie específicos que no estén evolutivamente conservados pero que podrían ser relevantes en el modelo de estudio. De esta manera se ampliaría el conocimiento de la diversidad y funcionalidad de los mismos (Tompa *et al.*, 2014). El estudio de los motivos lineales es aproximadamente contemporáneo con el estudio de las proteínas desordenadas. Ambos campos están ligados íntimamente. Aproximadamente un tercio del proteoma humano consiste en IDRs, donde se encuentran el 80 % de los motivos lineales (Davey *et al.*, 2012). Los roles de muchas de estas regiones y los motivos lineales correspondientes aún no se caracterizaron. Podríamos suponer que, por lo tanto, las características moleculares y la funcionalidad de los motivos lineales recién se empiezan a comprender.

Propiedades de los motivos lineales

La característica principal de los motivos lineales es la capacidad de codificar una interacción funcional y específica con unos pocos residuos. Los motivos lineales tienen un promedio de seis residuos funcionales (mínimo uno, máximo 23) en la base de datos ELMdb (Davey *et al.*, 2012). La mayoría de los motivos lineales son una secuencia monopartita, es decir, un único segmento de aminoácidos contiguos determina la función.

El bajo número de residuos implicados en un motivo lineal tiene varias consecuencias. En primer lugar, es común la existencia de falsos positivos que concuerdan en secuencia con un motivo lineal pero no colocalizan con su proteína blanco debido a la compartimentalización celular o especificidad tisular. En segundo lugar, la superficie de contacto entre un motivo lineal y el dominio globular tiene menores áreas enterradas ($\sim$ 500 Å^2) que las interacciones entre dominios globulares ($\sim$ 1150 Å^2). Los motivos lineales se unen a sus proteínas blanco con una afinidad relativamente baja, con constantes de disociación entre 1 y 10 μM, mientras que los dominios globulares pueden alcanzar constantes de disociación en el orden picomolar (Van Roey *et al.*, 2014). En el estado

libre los motivos lineales son la mayoría altamente flexibles y carecen de una estructura terciaria estable, lo que les permite adaptarse a la estructura de diferentes proteínas blanco mediante el acoplamiento de plegado e interacción. La especificidad de unión depende de la función biológica del motivo lineal y varía desde una única proteína blanco a ser altamente promiscuo. La especificidad y afinidad de un motivo lineal va a estar dada por factores que son intrínsecos, como los residuos que forman el motivo lineal, y extrínsecos, como la localización subcelular.

Representación de los motivos lineales

Los motivos lineales pueden representarse de diferentes maneras. En un principio la descripción de los motivos lineales se realizaba mediante secuencias consenso, indicando los aminoácidos más frecuentes y con una `x` o un punto "`.`" las posiciones donde los aminoácidos presentan frecuencias de ocurrencia bajas o similares entre sí. Por ejemplo, el motivo de unión a pRb se indicaba `LxCxE`. Sin embargo, este método no refleja la variabilidad de secuencia que puede presentar un motivo lineal. Por ello fue rápidamente dejado de usar y en general se utiliza únicamente para darle un nombre al motivo lineal como es el caso del motivo LxCxE. Una mejor descripción de la variabilidad de secuencia en los motivos lineales se logra utilizando expresiones regulares. En una expresión regular existen dos tipos de posiciones. Las posiciones fijas, determinantes de la funcionalidad del fragmento de la secuencia y las posiciones comodín, muy variables, que parecieran no determinar o participar en la funcionalidad de dicho fragmento (véase Sección 2.1.5). Las posiciones fijas son poco variables y permiten distintos aminoácidos con igual probabilidad. Por ejemplo, la expresión regular para el sitio de fosforilación de quinasa de caseína II (CKII) (en inglés, *Casein Kinase II*) es `[ST]..[DE]` indicando que en la primera posición puede existir tanto una serina como una treonina y en la última posición tanto un ácido aspártico como un glutámico. Los residuos continuos en secuencia al núcleo del motivo se denominan posiciones adyacentes. En algunos casos, estas posiciones participan en la interacción modulando la afinidad pero no son estrictamente necesarias para la interacción. Por último, otra forma para lograr una mayor descripción es utilizar los logos de secuencia (Schneider y Stephens, 1990). Un logo de secuencia es una representación gráfica de un alineamiento de secuencias (véase Sección 2.1.9), en el eje x se indica la posición del alineamiento y en el eje y el contenido de información (véase Sección 2.1.8) que indica el grado de conservación de cada posición. La altura de cada una de las letras es directamente proporcional a la frecuencia de un aminoácido en esa posición. A diferencia de una expresión regular el logo de secuencias permite identificar cuán más frecuente es cada uno de los aminoácidos presente en una posición fija.

Clasificación y funciones de los motivos lineales

Si bien las funciones de los motivos lineales son muy diversas, todas giran alrededor de la mediación de interacciones proteína-proteína. Pueden funcionar como señales de localización celular direccionando a la proteína a un compartimento particular, reclutar enzimas que alteran el estado químico del motivo por modificaciones post-traduccionales, regular la estabilidad de una proteína

y promover el reclutamiento de factores que facilitan la formación de complejos proteicos (Davey *et al.*, 2012; Diella *et al.*, 2008). La Figura 1.10 ejemplifica una forma de interacción entre un motivo lineal (azul) y un dominio globular (gris).

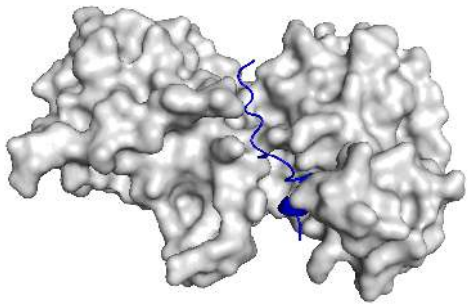


Figura 1.10: Estructura tridimensional de un motivo lineal unido a su blanco proteico. En representación de superficie y color gris se muestra el dominio bolsillo de la proteína retinoblastoma. En representación de cintas y color azul se muestra el péptido correspondiente a la proteína E2F que contiene el motivo lineal pRb_ABGroove que media la interacción (PDB ID: 1N4M) (Lee *et al.*, 2002).

La base de datos ELMdb considera seis categorías funcionales de motivos lineales: (1) sitios de clivaje (2) motivos de degradación (3) sitios de acoplamiento molecular (en inglés, *docking*) (4) ligandos (5) sitios de modificación post-traduccional y (6) motivos de señalización. Cada categoría contiene distintas clases de motivos con una expresión regular y cada clase de motivo agrupa instancias del motivo reportadas en la literatura y curadas manualmente. Por ejemplo, dentro de los sitios de clivaje se encuentran clasificados los sitios de reconocimiento de las proteasas caspasa 3 y 7 que juegan un rol principal en la apoptosis celular. Los sitios de degradación son aquellos cuya deleción o mutación aumenta la vida media de la proteína donde se encontraban naturalmente. Los mejores caracterizados son la caja de destrucción D-box y la caja KEN. Ambos son reconocidos por el complejo ubiquitin ligasa que promueve la anafase APC/C y selecciona proteínas que regulan el ciclo celular para su degradación en el proteosoma. Los sitios de acoplamiento molecular son sitios que median mecanismos adicionales de interacción y facilitan la especificidad de unión a una proteína blanco. Dentro de esta categoría encontramos, por ejemplo, los sitios de acoplamiento para las proteínas quinasas activadas por mitógenos, MAPK. El consenso de fosforilación de MAPK consiste en una serina o treonina seguida por una prolina (`[ST]P`). Estos sitios se encuentran ampliamente distribuidos en las proteínas y, por lo tanto, la existencia de un sitio adicional de interacción permite aumentar la especificidad de la fosforilación. Los motivos lineales que funcionan como sitios ligando específicamente son aquellos que participan en el ensamblado de complejos macromoleculares. En esta categoría se incluye al motivo LxCxE de unión a pRb cuya función más estudiada es la regulación del ciclo celular (véase Sección 1.5.1). El motivo LxCxE se encuentra en numerosas quinasas, histona deacetilasas y metil transferasas y permite la interacción con pRb. Los motivos lineales que son sitios de modificación post-traduccional consisten en patrones de secuencia que son reconocidos para su modificación covalente. Dentro de

esta categoría se encuentran los sitios de glicosilación, fosforilación y acetilación, entre otros. Por último, los motivos de señalización son aquellos que permiten la localización de la proteína a un compartimento especializado de la célula o su retención en el mismo, como la NLS nombrada al comienzo de la sección.

Estas seis categorías funcionales pueden clasificarse en dos grandes grupos, los motivos que funcionan como ligandos (categorías 2, 3, 4 y 6) y los motivos que funcionan como sitios de modificación post-traduccional (categorías 1 y 5). Los motivos que funcionan como ligandos son motivos regulatorios que controlan la localización, estabilidad de proteínas y el ensamblado de complejos macromoleculares. Los motivos que funcionan como sitios de modificación post-traduccional pueden ser sitios donde se modifique covalentemente la proteína, ocurra un clivaje u otra modificación estructural.

1.2.1. Bases de datos de motivos lineales

En los últimos años la importancia que adquirieron los motivos lineales llevó a la creación de distintas bases de datos enfocadas en recolectar la variabilidad de secuencia e información experimental relacionada a motivos lineales.

La base de datos ELMdb fue creada en el año 2003 (Puntervoll *et al.*, 2003) y es actualizada continuamente (Gouw *et al.*, 2018). Actualmente cuenta con 268 clases de motivos funcionales y 3,078 ejemplos de motivos lineales recolectados de la literatura y curados manualmente. La base de datos utilizada por el algoritmo MiniMotif Miner (Balla *et al.*, 2006; Lyon *et al.*, 2018) incluye motivos involucrados en interacciones proteína-proteína, modificaciones post-traduccionales y localización de proteínas. Su última actualización fue en el año 2018 (Lyon *et al.*, 2018) y no puede ser accedida por el usuario. La base de datos utilizada por Scansite (Obenauer *et al.*, 2003; Yaffe *et al.*, 2001) construida a partir de motivos determinados experimentalmente es actualizada continuamente, pero solamente puede ser accedida por motivo y no puede ser descargada por el usuario. La base de datos LMPID (en inglés, *Linear Motif mediated Protein Interaction Database*) es una base de datos curada manualmente de motivos lineales que median interacciones proteína-proteína (Sarkar *et al.*, 2015). Actualmente cuenta con más de 1500 ejemplos de motivos que participan en más de 2200 interacciones proteína-proteína. Otras bases de datos, como PDZ-Base (Beuming *et al.*, 2005), recolectan información de interacciones relacionadas a un motivo o dominio en particular.

En resumen, las bases de datos de motivos lineales no son muy numerosas y no todas están disponibles. En los últimos años, debido a a las mejoras en las técnicas experimentales, la información biológica disponible ha aumentado en gran escala. La capacidad de capitalizar esos datos biológicos y disponerlos a la comunidad científica facilita y acelera el entendimiento de procesos y elementos novedosos como los motivos lineales.

1.2.2. Algoritmos de predicción de motivos lineales

Dada la complejidad y el alto número de experimentos necesarios para poder identificar un motivo lineal (véase Sección 2.1.5) y la importancia funcional, en las últimas décadas se desarrollaron diversos métodos bioinformáticos que facilitan la identificación de posibles motivos lineales. Existen dos tipos de identificadores de motivos lineales, los algoritmos que llevan a cabo la identificación *de novo* y los que se basan en las bases de datos de motivos lineales pre-existentes. Estos últimos a su vez pueden dividirse entre los que buscan para una proteína dada un conjunto de motivos lineales pre-definidos en la base de datos y los que buscan en un gran grupo de proteínas un motivo consenso (Edwards y Palopoli, 2015).

MiniMotifMiner (Balla *et al.*, 2006; Mi *et al.*, 2012), QuasiMotifFinder (Gutman *et al.*, 2005) y la base de datos ELMdb (Gouw *et al.*, 2018) toman como entrada una secuencia proteica ingresada por el usuario y realizan la búsqueda de motivos a partir de una base de datos propia de carácter generalista. Scansite (Obenauer *et al.*, 2003; Yaffe *et al.*, 2001) y PhosphoELM (Dinkel *et al.*, 2011) realizan búsquedas del mismo tipo usando base de datos propias relacionadas un sub-grupo particular de motivos lineales. Por otro lado, SLiMSearch (Davey *et al.*, 2011a; Krystkowiak y Davey, 2017) y ScanProsite (de Castro *et al.*, 2006) toman como entrada un motivo ingresado por el usuario y buscan ejemplos del motivo en una base de datos de proteínas. Dentro de los algoritmos que predicen motivos lineales *de novo* podemos nombrar DILIMOT (Neduva y Russell, 2006) y SLiMFinder (Edwards *et al.*, 2007) y su derivado QSLimFinder (Palopoli *et al.*, 2015). Estos algoritmos toman como entrada un conjunto de secuencias de proteínas no relacionadas que comparten una característica funcional y buscan patrones sobre-representados en las secuencias que puedan ser responsables de dicha función. Si bien los algoritmos desarrollados para identificar motivos lineales son, en general, propensos a los falsos positivos, su desarrollo permitió profundizar y conocer la variabilidad de los mismos.

1.2.3. Evolución de motivos lineales

La mayoría de los motivos lineales se encuentran en regiones desordenadas. La evolución de los dominios globulares ocurre de manera lenta y está restringida por la necesidad de mantener una estructura rígida (Neduva y Russell, 2005). Por el contrario, las IDRs pueden evolucionar rápidamente presentando como restricción la preservación de asambleas conformacionales y sitios de unión (Brown *et al.*, 2011; Chemes *et al.*, 2015, 2012a) ya que carecen de la fuerte restricción estructural que poseen los dominios globulares para explorar el espacio de secuencia. Este contexto estructural en el cual se encuentran los motivos lineales sugiere que son elementos de evolución rápida. La mayoría de los dominios que unen motivos lineales tienen baja especificidad, ya que contactan unos pocos residuos que pueden tener propiedades fisicoquímicas similares (Davey *et al.*, 2015, 2012). Si bien los aminoácidos adyacentes al motivo pueden modular de manera indirecta la afinidad por el blanco proteico, existen pocas restricciones sobre ellos. Dado que el número de residuos funcionales de los motivos lineales presentes en las proteínas desordenadas es pequeño

y que la pérdida o ganancia de función de un motivo lineal puede ocurrir fácilmente por una mutación puntual, está establecida la creencia de que pueden tener una alta tendencia a evolucionar de manera convergente en proteínas no relacionadas (Davey *et al.*, 2012; Diella *et al.*, 2008). Dado el bajo número de restricciones estructurales y de secuencia, es esperable que los motivos lineales ocurran al azar y que un proteoma dado posea numerosos motivos que sean complementarios a un mismo sitio de unión (Davey *et al.*, 2015).

La plasticidad evolutiva de los motivos lineales está limitada por la selección natural (Neduva y Russell, 2005). Si el motivo lineal presenta una ventaja adaptativa, la presión de selección es positiva y ocurre la generación de una nueva instancia de un motivo lineal funcional. Esta selección positiva puede ejemplificarse con la proteína mitocondrial de señalización antiviral (MAVS) (en inglés, *Mitochondrial AntiViral-Signaling protein*). MAVS forma parte del mecanismo antiviral de la inmunidad innata que en células de mamífero sensa ARN viral y desencadena la respuesta vía interferón. Muchos virus son capaces de inhibir esa respuesta. Por ejemplo, el virus de la hepatitis C (HCV) utiliza la proteasa NS3/4A para clivar la proteína MAVS en humanos. Un estudio evolutivo en distintas proteínas MAVS de primates reveló que existe una selección positiva para determinadas mutaciones en el sitio de clivado que inhiben la acción de la proteasa de HCV por disminución de la afinidad (Patel *et al.*, 2012). Por otro lado, si el motivo lineal produce interacciones proteína-proteína no beneficiosas o deletéreas, la presión de selección es negativa y no ocurre la fijación del motivo lineal en la población. La selección negativa de un motivo es más difícil de mostrar porque no hay evidencia directa del proceso. Sin embargo, un estudio comparando la abundancia de motivos reportados para todos los proteomas contra secuencias generadas al azar a partir de los mismos proteomas reveló que un 3 % de los motivos analizados son menos frecuentes en la naturaleza que lo esperado, sugiriendo la existencia de un proceso de selección negativa (Via *et al.*, 2007).

Hasta la fecha se realizaron pocos estudios que pongan directamente a prueba las creencias establecidas de evolución rápida y convergente y el rol adaptativo de los motivos lineales. Un estudio realizado en la base de datos ELMdb reveló que 50 % de los motivos ocurren en proteínas no relacionadas (Davey *et al.*, 2012). Un estudio de los sitios de N-glicosilación a nivel de proteoma en diferentes especies de eucariotas reveló que si bien todos los sitios de N-glicosilación poseían características comunes, como ser el motivo de secuencia, las restricciones estructurales y la localización celular, un gran porcentaje de dichos sitios apareció luego de la divergencia filogenética entre plantas, hongos, nematodos, insectos y vertebrados (Zielinska *et al.*, 2012). Estos resultados apoyan la hipótesis de que los motivos lineales son elementos que evolucionan de manera rápida y convergente. Sin embargo, 5 % de los motivos lineales reportados en ELMdb están conservados en una posición dada para una secuencia proteica tanto en humanos como en levaduras (Davey *et al.*, 2012). Algunos motivos funcionalmente relevantes se conservan en linajes recientes (Davey *et al.*, 2012) y otros están conservados a lo largo de grandes distancias evolutivas, sugiriendo que existen restricciones a los cambios en los motivos lineales una vez que la utilización del motivo está distribuida ampliamente en el proteoma (Van Roey *et al.*, 2014).

1.3. Proteína E7 de papilomavirus

1.3.1. La familia *Papillomaviridae*

Los papilomavirus (PVs) son virus desnudos icosaédricos y poseen un genoma ADN doble cadena circular de entre 5 y 8 Kb. El comité internacional de taxonomía de virus (ICTV) (en inglés, *International Committe on Taxonomy of Viruses*) los agrupó dentro de la familia *Papillomaviridae*. Sus hospedadores incluyen una amplia variedad de vertebrados: peces, reptiles, aves y mamíferos. Los PV infectan el epitelio mucoso y queratinizado y producen lesiones denominadas condilomas o papilomas y verrugas, respectivamente. Algunos PVs están asociados al cáncer cervical uterino y a la formación de tumores en el tracto urogenital y las vías aéreas superiores.

Taxonomía de los papilomavirus

La clasificación de la familia *Papillomaviridae* está basada en la identidad de secuencia nucleotídica del marco de lectura abierto (ORF) L1 (de Villiers *et al.*, 2004). El análisis filogenético de PVs basado en el alineamiento de las secuencias de los ORF E1, E2, L2 y L1 concatenadas, incluyendo un representante de cada especie de los 53 géneros, apoya la existencia de por lo menos dos subfamilias que comparten más del 45 % de identidad en el ORF L1 (Figura 1.11). La subfamilia *Firstpapillomavirinae* incluye más de 50 géneros y 130 especies que infectan vertebrados amniotas y la subfamilia *Secondpapillomavirinae* incluye un único género y una única especie, que infecta peces (Van Doorslaer *et al.*, 2018).

La subfamilia *Firstpapillomavirinae* agrupa PVs que comparten más del 60 % de identidad en L1. Los géneros son nombrados según el alfabeto griego, por ejemplo, *Alpha-*, *Beta* , *Gammapapillomavirus*). Cuando se completa el primero y segundo ciclo del alfabeto se utilizan los prefijos *Dyo-* y *Treis-*. Por ejemplo, *Dyoalphapapillomavirus* una vez que se utilizó *Omegapapillomavirus* y *Treisalphapapillomavirus* luego de que se utilizó *Dyoomegapapillomavirus*. Dentro de cada género, los PVs que comparten entre el 60 y 70 % de identidad en L1 se agrupan en especies que están nombradas igual que el género y un número. Por ejemplo, *Alphapapillomavirus 1*, *Treisiotapapillomavirus 1*. Cada especie está compuesta por distintos tipos virales que comparten entre el 71 y 89 % de identidad en el ORF L1. A su vez, cada tipo agrupa subtipos que comparten una identidad entre el 90 y 98 % en el L1 y/o variantes que comparten una identidad mayor al 98 % en el L1.

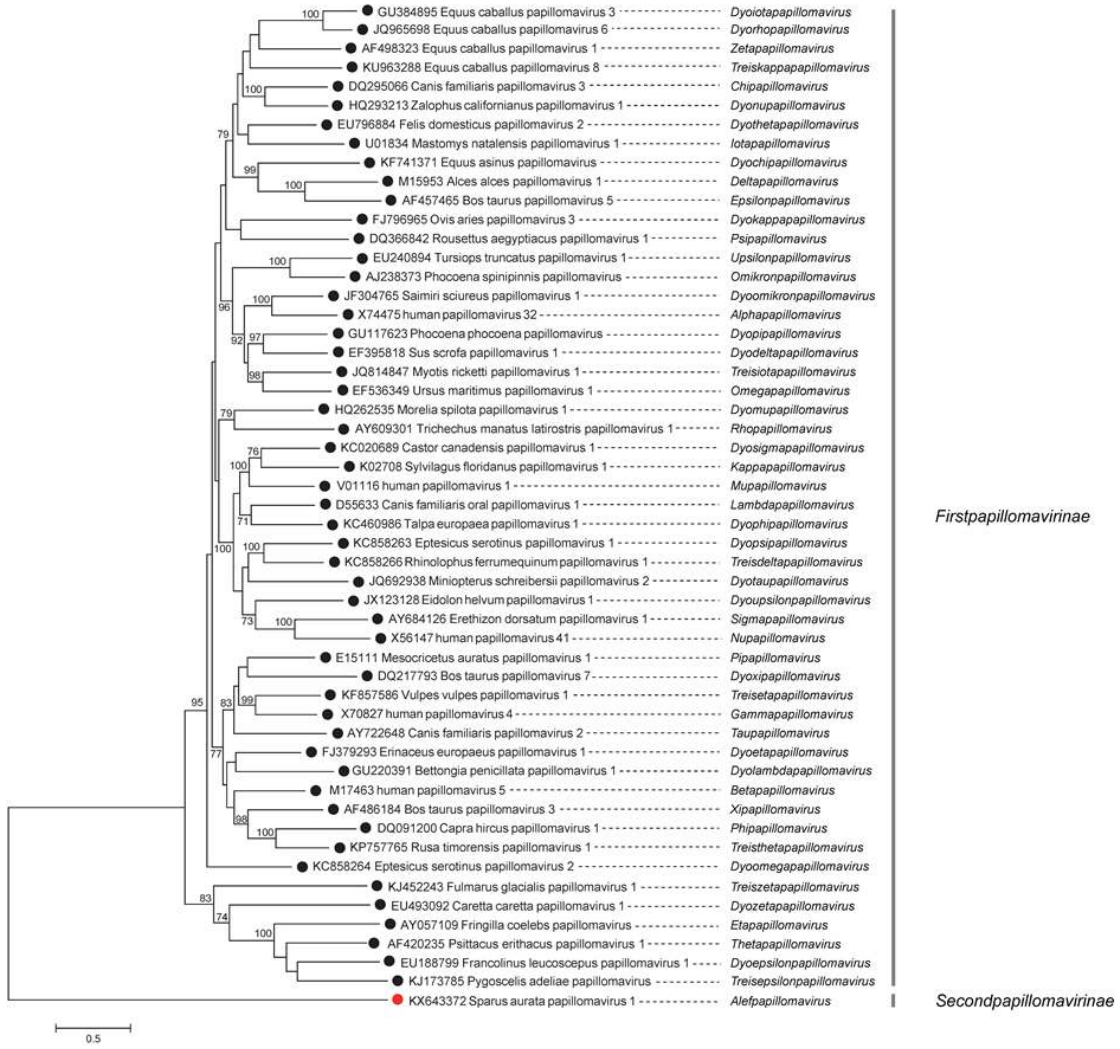


Figura 1.11: Árbol filogenético de papillomavirus construido a partir de los genes concatenados E1-E2-L2 y L1. A partir de las secuencias nucleotídicas de los marcos de lectura abierta E1, E2, L2 y L1 alineadas y concatenadas se construyó un árbol por máxima verosimilitud (en inglés, *maximum likelihood*). Se muestran los nodos con un soporte mayor al 70 % generado por remuestreo. Figura adaptada de `http://www.ictv.global/report/papillomaviridae`.

Dentro de la subfamilia *Secondpapillomavirinae* los géneros son nombrados según los abyades semíticos. Actualmente, la subfamilia *Secondpapillomavirinae* posee un único género denominado con la primera letra de este alfabeto, *Alefpapillomavirus*, que incluye una única especie, *Alefpapillomavirus 1*, y un único serotipo, Sparus aurata papillomavirus 1.

1.3.2. Estructura genómica y proteínas de los papilomavirus

Los PVs son virus desnudos con una cápside icosaédrica de ~ 60nm de diámetro. La cápside está formada por dos proteínas, la proteína L1 y la proteína L2. 360 copias de la proteína L1 forman los 72 pentámeros y doce moléculas de la proteína L2 se encuentran en cada uno de los vértices del icosaedro (Figura 1.12). La cápside posee una simetría T= 7 1\4 y los pentámeros ocupan

posiciones hexavalentes. En el interior de la cápside se encuentra el genoma ADN, asociado a histonas celulares.

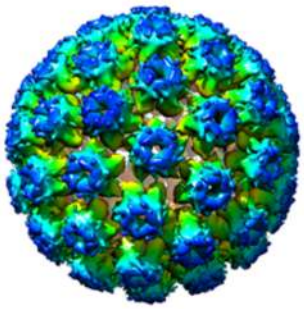


Figura 1.12: Estructura externa de la cápside del virión de HPV16. Superficie de la cápside de HPV16 derivada de la reconstrucción obtenida por cryo-EM a una resolución de 4.5 Å coloreada por distancia radial al centro de la cápside siendo azul lo más externo (~ 29nm) y rojo lo más interno (~ 20nm) (PDB ID: 5KEP) (Guan *et al.*, 2017). Figura adaptada de `http://www.ictv.global/report/papillomaviridae`.

El genoma de PV es un ADN circular doble cadena de ~ 8 kb (Figura 1.13) y codifica entre seis y nueve proteínas. La organización del genoma está conservada en los diferentes PVs. Los genes virales están codificados en una sola cadena del genoma viral, que funciona como templado para la transcripción. Los genes codificantes están clasificados como tempranos, E (en inglés, *early*), o tardíos, L (en inglés, *late*), según su ubicación en el genoma. La región temprana codifica para proteínas regulatorias, incluyendo aquellas necesarias para la iniciación de la duplicación del ADN. La región tardía codifica para los genes L1 y L2, que se expresan únicamente en las células infectadas que producen partículas virales. El origen de replicación del ADN se encuentra en una región no codificante de aproximadamente 1 kb denominada región larga de control (LCR) (en inglés, *Long Control Region*), que incluye además sitios de unión para proteínas regulatorias (Zheng y Baker, 2006).

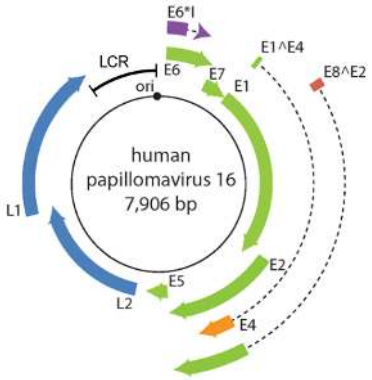


Figura 1.13: Esquema de la organización genómica de HPV16. Los distintos marcos de lectura abiertos están indicados con flechas de color según pertenecen a los genes tempranos (verde, violeta, rojo y naranja) o tardíos (azul). Las líneas punteadas indican las secuencias intrónicas. El círculo negro indica el origen de replicación del genoma viral. En negro se indica la LCR y con un punto la ubicación del origen de replicación. Figura adaptada de `http://www.ictv.global/report/papillomaviridae`.

La expresión de las proteínas virales está asociada a la diferenciación del tejido del epitelio escamoso estratificado (Figura 1.14). En las etapas tempranas de la infección se expresan las proteínas E1 a E7. La proteína E1 es una ADN helicasa esencial para la replicación y amplificación del genoma viral (Bergvall *et al.*, 2013). E2 es la proteína reguladora principal del ciclo viral y participa en la regulación e iniciación de la duplicación de genoma viral (McBride, 2013). La proteína E1∧E4 es el producto de la traducción del ARNm formado por *splicing* que incluye los cuatro primeros codones de E1 y a E4 completa y participa a los procesos de síntesis y amplificación del genoma viral (Doorbar, 2013). La proteína hidrofóbica transmembrana E5 actúa modulando la actividad de proteínas celulares, contribuyendo a la transformación y fase productiva del ciclo viral. E5 interacciona por ejemplo con la ATPasa vacuolar y afecta la expresión del complejo mayor de histocompatibilidad de clase I (MHCI) (DiMaio y Petti, 2013). Las proteínas E6 y E7 participan en la usurpación de la maquinaria celular, generando un ambiente propicio para la replicación viral. Sin embargo, no todos los PVs poseen la región codificante para ellas (Van Doorslaer, 2013). Por último, el exón E8 utiliza el mismo sitio aceptor del *splicing* generando el ARNm para la proteína inhibidora de la replicación viral y expresión génica E8∧E2.

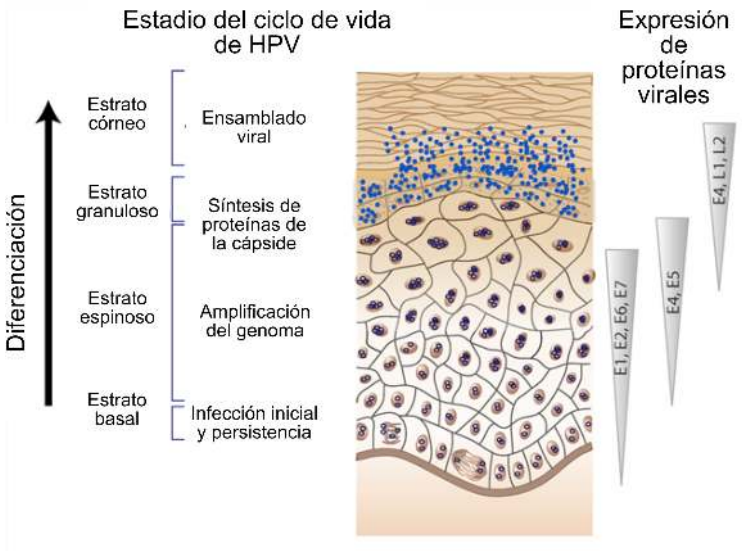


Figura 1.14: Esquema de la organización del ciclo de vida de papilomavirus. A la izquierda se muestran las distintas capas del epitelio escamoso estratificado. A la derecha se muestra la expresión temporal de las proteínas virales en cada estrato y el nivel de expresión está esquematizado como triángulos. En la infección inicial del estrato basal se expresan las proteínas de la etapa temprana relacionadas con la regulación. El aumento de los niveles de expresión de estas proteínas facilitan la amplificación y mantenimiento del genoma. La expresión de las proteínas L1 y L2 es mayor en los estratos granuloso y córneo donde ocurre el ensamblado de las partículas virales. Figura adaptada de `http://www.ictv.global/report/papillomaviridae`.

En las etapas tardías, se expresan las proteínas de la cápside L1 y L2. Mientras que L1 es el componente estructural principal de la cápside, L2 es el componente minoritario y participa en el ensamblado de las partículas virales (Buck *et al.*, 2013; Wang y Roden, 2013).

1.3.3. Ciclo de vida de los papilomavirus

Los PVs poseen un amplio rango de hospedadores. Sin embargo, cada tipo viral es en general altamente específico de una especie e infecta únicamente células del epitelio escamoso estratificado. La infección viral puede dividirse en distintas fases que están separadas por el estado de diferenciación de la célula epitelial (Figura 1.14), restringiendo espacial y temporalmente la síntesis de proteínas regulatorias, del genoma viral, de proteínas de la cápside y el proceso del ensamblado viral. Dado que el estrato basal es el único donde ocurre la división celular, el virus debe infectar estas células para poder establecer una infección persistente. En este estrato se expresan las proteínas tempranas del virus, mientras que la síntesis de las proteínas de la cápside, la replicación del ADN y el ensamblado viral ocurren únicamente en las células ya diferenciadas de los estratos granuloso y córneo.

Infección inicial. La interacción inicial entre el virión y la célula del epitelio basal ocurre mediante la unión a proteoglicanos de la membrana basal expuestos en sitios de lesiones o permeabilización. Esta unión induce un cambio conformacional en la cápside que expone el extremo N-terminal de L2. Este extremo de L2 es clivado por furina, llevando a la exposición de un sitio de unión en la cápside para un receptor aún no identificado de la superficie celular de los queratinocitos. Si bien se sabe que las cápsides son endocitadas, se desconoce la vía involucrada. En el endosoma ocurre un desnudamiento parcial del virus. Los genomas, unidos a L2, escapan del endosoma y son transportados mediante motores de dineína hacia el núcleo, donde entran por disrupción de la membrana nuclear durante la mitosis (Aksoy *et al.*, 2017). En el núcleo ocurre la transcripción del genoma viral, asociada al programa de diferenciación celular del epitelio escamoso. La expresión génica de los PVs está altamente regulada al nivel de la transcripción por promotores que se encuentran en la región LCR. El procesamiento del ARN por poliadenilación y *splicing* alternativo del ARNm lleva a la producción diferencial de ARNm en diferentes células (Schwartz, 2013). La región LCR posee elementos de regulación que responden a factores celulares y factores virales como E2, que actúan en conjunto modulando la transcripción viral, la duplicación y el mantenimiento a largo plazo del genoma viral (McBride, 2017). El genoma viral es mantenido como un plásmido multi-copia, replicando durante la fase S en sincronía con la célula y asegurando una infección persistente de las células basales de la epidermis.

Fase replicativa. La fase replicativa del virus está estrechamente conectada al programa de diferenciación del epitelio escamoso. La síntesis de las proteínas de la cápside ocurre únicamente en queratinocitos diferenciados, ya que los genes tardíos están regulados por un promotor que sólo se activa en estas células y que se encuentra dentro de la región codificante de la proteína E7 (Graham, 2017). Una vez que ocurre la diferenciación de los queratinocitos con el genoma viral mantenido de manera episomal, se activa la expresión de los genes tardíos L1 y L2 y aumenta el nivel de la transcripción de los genes tempranos E1, E4, E6 y E7 (Moody, 2017). La transcripción de los genes L1 y L2 también está regulada a nivel post-transcripcional por elementos regulatorios nega-

tivos en la región 3' no codificante del ARNm que se unen a factores celulares específicos ausentes luego de la diferenciación del queratinocito (Moody, 2017). Una vez sintetizadas, las proteínas L1 y L2 son transportadas al núcleo, donde ocurre el ensamblado viral. La cápside viral se ensambla de manera estable frente a la presencia del genoma viral circular y del tamaño correcto (~ 8 kb).

Los queratinocitos son células diferenciadas que ya no se dividen ni duplican el ADN. Las proteínas E6 y E7 reactivan la maquinaria de duplicación en beneficio de la replicación viral. La duplicación del genoma viral requiere de la unión de la proteína E1 al origen de duplicación del ADN, la cual es estabilizada por interacción con la proteína auxiliar E2. E1 recluta la maquinaria de iniciación de la duplicación del ADN del hospedador al origen de replicación viral (Moody, 2017). La duplicación del genoma viral es bidireccional, aunque no se comprende bien el mecanismo. El mantenimiento de esta fase puede prolongarse por meses o años. Finalmente, los viriones son liberados al ambiente a medida que las células epiteliales se desprenden.

Inmortalización y transformación celular. La transformación celular consiste en el cambio de una célula normal a una célula de replicación celular masiva perdiendo el control del ciclo celular y evadiendo la muerte celular programada. Este proceso puede ocurrir por mutaciones en el ADN o por una infección viral. Para el inicio o mantenimiento del proceso de transformación celular no es necesario que el genoma viral se integre en el genoma del hospedador, pudiendo ser mantenido como un plásmido. Sin embargo, la transformación es dependiente de una expresión continua de los genes virales. Las proteínas E6 y E7 son las principales involucradas en este proceso, a través mecanismos que involucran a las proteínas celulares p53 y pRb y crean un ambiente favorable para la replicación celular en células diferenciadas. Mientras que los PVs de alto riesgo, como HPV16 y HPV18, pueden llevar a cabo la inmortalización o transformación de la célula, los PVs de bajo riesgo, como HPV6 y HPV11, no pueden (Schutze *et al.*, 2014). Por esta razón se considera a las proteínas E6 y E7 de los tipos de alto riesgo oncoproteínas. Una diferencia entre los dos grupos de riesgo es que la proteína E6 interacciona con la proteína celular p53 en los tipos de alto riesgo, mientras que no lo hace en los tipos de bajo riesgo. Esta interacción bloquea la función transcripcional de p53 y promueve la degradación dependiente de ubiquitina de p53 (Scheffner *et al.*, 1990). Por lo tanto, al igual que el antígeno large T del poliomavirus SV40 y la proteína E1B de adenovirus, E6 inhibe la apoptosis e interfiere con las funciones regulatorias de p53 en el ciclo celular. El rol de la proteína E7, objeto de estudio de esta tesis, en este proceso será discutido en una sección aparte. Por último, es importante aclarar que la transformación celular y la replicación viral son mutuamente excluyentes, sugiriendo que el progreso hacia la inmortalización celular no es el camino típico de la infección por PV.

1.3.4. Patologías de papilomavirus

Los PVs están en general bien adaptados a sus hospedadores. La transmisión de HPV es por contacto y la forma de infección más común ocurre por transmisión sexual. La infección en el tracto genital ocurre por transmisión sexual y afecta la piel y mucosa, abarcando vagina, cervix y ano.

La incidencia de la infección es mayor en el grupo femenino de la población activo sexualmente (Crosbie *et al.*, 2013; Serrano *et al.*, 2018). La mayoría de las infecciones por PV son subclínicas, aunque algunos tipos virales pueden causar cáncer cervical, anal, oral o de cuello. Según su tropismo los PVs pueden dividirse en cutáneos o mucocutáneos. A su vez, los tipos virales mucocutáneos pueden estar asociados con lesiones benignas o malignas (Cubie, 2013). Se estima que el 90 % de las infecciones por PV en el cuello cervical son eliminadas luego de dos años post-infección. Debido a su relevancia clínica, los datos epidemiológicos disponibles corresponden principalmente al género *Alphapapillomavirus*. Este género abarca la mayoría de los PVs que infectan a humanos y a aquellos PVs responsables de cáncer.

Papillomavirus y cáncer. La distribución global de los PVs contribuye a que más del 90 % de los casos de cáncer cervical a nivel global estén asociados a HPV (Figura 1.15), con una alta prevalencia de los tipos virales de alto riesgo de la especie *Alphapapillomavirus 9*, HPV16 y HPV31, y *Alphapapillomavirus 7*, HPV18 y HPV45. Los tipos virales de bajo riesgo, como HPV11 y HPV6, se encuentran asociados con menor frecuencia a lesiones malignas. En los casos de cáncer atribuidos a HPV los tumores se desarrollan muchos años después de la infección inicial. Por lo tanto, la infección persistente es necesaria para el desarrollo de cáncer invasivo. El mantenimiento del fenotipo transformado depende de la expresión continua de algunos de los genes virales, principalmente las proteínas E6 y E7.

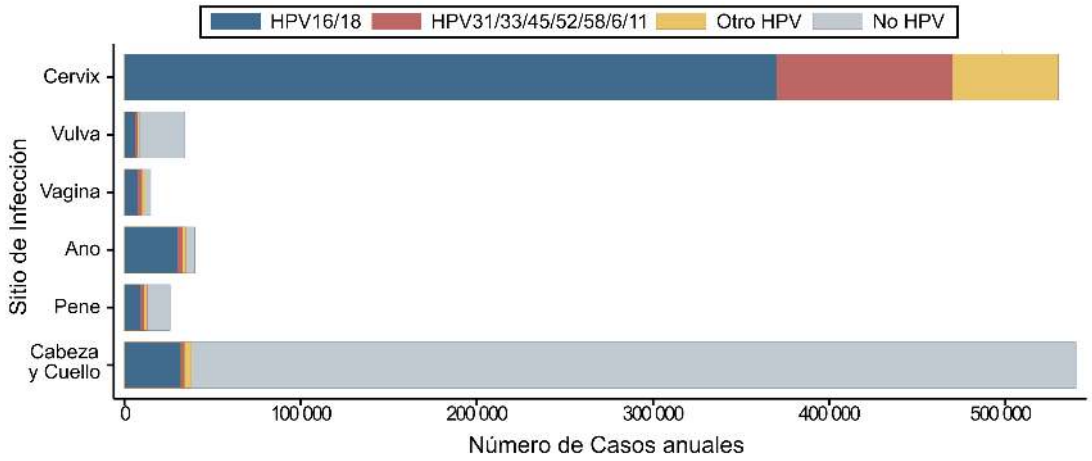


Figura 1.15: Número de casos anuales de cáncer atribuidos a la infección por HPV según el sitio de infección y tipo viral. Figura adaptada de Serrano *et al.* (2018).

Por otro lado, la mayoría de las infecciones causadas por HPV, incluyendo los tipos virales de alto riesgo, son auto-limitadas o no progresan en cáncer aún cuando la infección es persistente. Esto sugiere que otros factores ambientales contribuyen a la progresión maligna, como ser la inmunosupresión del hospedador. A nivel celular, las proteínas E6 y E7 juegan un papel principal en la progresión de la infección al desarrollo de cáncer. Ambos genes se expresan a partir de un único promotor y sus niveles relativos de expresión están regulados por *splicing* alternativo. La

integración del ADN viral en el genoma de la célula infectada está asociada a una deleción de un fragmento amplio del mismo que incluye una región del gen E2. Dado que E2 es un regulador negativo de la expresión de las proteínas E6 y E7, la pérdida de E2 en el genoma integrado se refleja en un aumento de la transcripción de los genes E6 y E7 y una expresión elevada de las proteínas E6 y E7 (Ho *et al.*, 2011).

1.3.5. E7 funciones, estructura y motivos preexistentes

La proteína E7 de papilomavirus comparte similitudes funcionales con la proteína E1A de adenovirus y el antígeno T del poliomavirus SV40. Las tres proteínas poseen actividades transformantes e interaccionan con pRb y las proteínas bolsillo relacionadas (Chellappan *et al.*, 1992). La interacción de la proteína E7 con pRb es responsable de la inducción de la síntesis de ADN viral y la proliferación celular. La inmortalización y transformación de la célula infectada inducida por E7 es consecuencia tanto de la interacción de E7 con las proteínas bolsillo como su interacción con numerosos blancos proteicos (véase Sección 3.2), involucrados en crecimiento celular y transformación (McIntyre *et al.*, 1996), transcripción, apoptosis y síntesis de ADN (Massimi *et al.*, 2001; Münger *et al.*, 2001). Por ejemplo, E7-HPV16 interactúa con la proteína p27^{kip1} que inhibe el crecimiento celular inducido por el factor de crecimiento TGFβ en queratinocitos, desactivando el arresto celular asociado al mismo (Münger *et al.*, 2001).

Estructura de dominios de E7

E7 es una proteína pequeña con una longitud promedio de 100 residuos. La arquitectura de dominios de E7 consiste en dos dominios: un dominio N-terminal intrínsecamente desordenado (E7N, residuos 1 a 40 de E7-HPV16) (García-Alai *et al.*, 2007) y un dominio globular C-terminal (E7C, residuos 51-98 de HPV16) (Figura 1.16), también conocido como región conservada 3 (CR3) (Ohlenschläger *et al.*, 2006). El dominio E7N consiste en dos regiones conservadas denominadas CR1 y CR2. El dominio E7C media la homodimerización y posee un sitio CxxC de coordinación de zinc por monómero (Alonso *et al.*, 2002; Liu *et al.*, 2006; Ohlenschläger *et al.*, 2006).

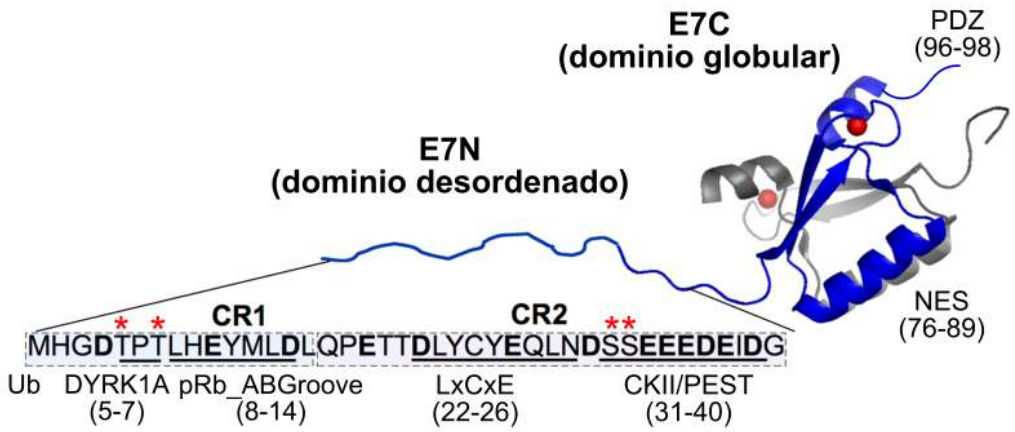


Figura 1.16: Representación esquemática de la proteína E7. Se muestra a la derecha la estructura diméri-ca del dominio E7C de HPV45 (PDB ID: 2F8B) (Ohlenschläger *et al.*, 2006) obtenida por RMN. Cada cada monómero esta representado como *cartoon* en azul y gris. Los átomos de cinc está representados como esferas rojas. Se indica además la ubicación relativa de los motivos lineales del dominio globular (NES y PDZ), con las posiciones correspondientes a la proteína HPV16 indicadas entre paréntesis. El dominio E7N desordenado de uno de los monómeros está representado a la izquierda como un cordón con la secuencia de aminoácidos de HPV16 detallada en negro. Las regiones conservadas CR1 y CR2 están remarcadas. En la parte inferior se indican el sitio N-terminal de ubiquitinación (Ub) y los motivos lineales del dominio E7N (DYRK1A, pRb_ABGroove, LxCxE, CKII y la región acídica que contiene la PEST), con las posiciones correspondientes a HPV16 indicadas entre paréntesis. En forma de asteriscos se indican las treoninas 5 y 7 fosforiladas por DYRK1A, y las S31 y S32 fosforiladas por CKII. Figura adaptada de Noval *et al.* (2013).

Motivos lineales de E7

E7 es una proteína rica en motivos lineales. Hasta la fecha se describieron un total de seis motivos lineales en el dominio desordenado E7N y tres motivos en el dominio globular E7C. En la región CR1 existe un motivo de ubiquitinación (Reinstein *et al.*, 2000). E7 puede ser ubiquitinada en el extremo N-terminal y degradada, independientemente de la presencia de lisinas. Los mecanismos aún no son del todo entendidos (Ben-Saadon *et al.*, 2004; Oh *et al.*, 2004). El motivo pRb_ABGroove interacciona con el bolsillo AB de pRb (véase Sección 1.5.1). Por último, la región CR1 presenta un sitio de fosforilación para la quinasa de tirosina de especificidad dual regulada por fosforilación 1A (DYRK1A) (Liang *et al.*, 2008). La proteína E7 de HPV16 es fosforilada por DYRK1A en las treoninas 5 y 7. Esta interacción estabiliza a E7 e inhibe su degradación, contribuyendo a la actividad oncogénica de E7.

En la región CR2 de E7 existe un motivo LxCxE que interacciona con una región del subdominio B de la proteína pRb (Lee *et al.*, 1998; Wang *et al.*, 2010). Este sitio presenta a continuación dos sitios de fosforilación para CKII (Firzlaff *et al.*, 1989). La proteína E7 de HPV16 puede ser fosforilada *in vivo* e *in vitro* en las serinas 31 y 32 (Barbosa *et al.*, 1990; Smotkin y Wettstein, 1987). Estos sitios de fosforilación están inmersos en una región acídica que además contiene una secuencia de degradación PEST (Rechsteiner y Rogers, 1996). En la región CR3 de E7 se describieron tres motivos lineales: el motivo CxxC que media la interacción con el zinc, una señal de

exportación del núcleo (NES)(en inglés, *Nuclear Export Signal*) (Knapp *et al.*, 2009) que media la exportación del núcleo y un motivo de unión a dominios PDZ (Tomaić *et al.*, 2009).

Características biofísicas de E7

Los estudios conformacionales de la proteína E7 de HPV16 revelaron que el dímero de la proteína E7 posee una conformación plegada pero extendida y que puede sufrir múltiples transiciones conformacionales, incluyendo la disociación del dímero (Alonso *et al.*, 2002). La composición de secuencia de E7 y del dominio E7N refleja las características de una proteína intrínsecamente desordenada (véase Sección 1.1.2). El análisis biofísico del dominio E7N reveló una alta flexibilidad conformacional, ya que posee distintos estados de equilibrio que varían según las condiciones químicas del medio y los estados de fosforilación (García-Alai *et al.*, 2007). El fragmento de la región CR1 del E7N, que incluye el motivo pRb_ABGroove, y la región acídica del CR2 adquieren de manera transiente una estructura secundaria de hélice-α. En el caso de la región acídica, la transición es dependiente del pH. A pH neutro la hélice-α abarca al motivo LxCxE de unión a pRb (residuos 21-29) mientras que la región acídica (residuos 33-38) posee una estructura de poliprolina tipo II (PII). En medio ácido, ocurre una transición de PII a hélice-α debido a un cambio en la carga de los residuos ácidos (Noval *et al.*, 2013) y favorecida por la fosforilación en los residuos serina 31 y serina 32 (Chemes *et al.*, 2010). Por otro lado, la estructura de E7C fue determinada para las proteínas E7 de los serotipos HPV1 y HPV45 (Liu *et al.*, 2006; Ohlenschläger *et al.*, 2006) revelando que E7C posee una estructura globular altamente conservada.

1.4. Proteína E1A de *Mastadenovirus*

1.4.1. La familia *Adenoviridae*

Los adenovirus (AdVs) son un grupo de virus desnudos icosaédricos cuyo genoma está formado por un ADN doble cadena lineal entre 25 y 50Kb (Davison *et al.*, 2003). El primer AdV fue aislado por Rowe *et al.* (1953) a partir de cultivos celulares de las glándulas adenoides de un niño. Desde entonces se sabe que estos virus están esparcidos a nivel global y son responsables de hepatitis canina, queratoconjuntivitis, neumonía intersticial y gastroenteritis. El análisis de los ARNm de los AdVs llevó al descubrimiento del *splicing* del ARNm. La ICTV los clasificó dentro de la familia *Adenoviridae* y sus hospedadores incluyen una amplia variedad de vertebrados, como ser mamíferos, reptiles, aves, anfibios y peces (Harrach *et al.*, 2012).

Clasificación filogenética

La familia *Adenoviridae* está compuesta por cinco géneros, que inicialmente se definieron acorde al hospedador que infectaban. Los avances en las técnicas de secuenciación en las últimas décadas llevaron a redefinir este criterio. Actualmente, diversas regiones son utilizadas para determinar las relaciones filogenéticas de los AdV como ser el hexón (Kohl *et al.*, 2012), la proteasa (Kovács *et al.*, 2003), y la ADN polimerasa (Davison *et al.*, 2003). Los cinco géneros aceptados por la ICTV se ven confirmados por todos los análisis filogenéticos (Figura 1.17) (King *et al.*, 2018).

Dentro de cada género, los virus se agrupan en especies nombradas según el hospedador al que infectan y con letras del alfabeto. El criterio de demarcación de especies incluye, en primer lugar, la distancia filogenética en la secuencia de aminoácidos de la ADN polimerasa. Dadas dos secuencias, se consideran que pertenecen a especies diferentes si dicha distancia es > 5–$15\,\%$. Otros criterios de demarcación de especies son la especificidad de hospedador, el análisis de restricción enzimática, la hibridización del ADN, el contenido GC en el genoma, la neutralización cruzada, la posibilidad de recombinación, la hemoaglutinación, la oncogenicidad en ratones y la transformación de cultivos primarios celulares.

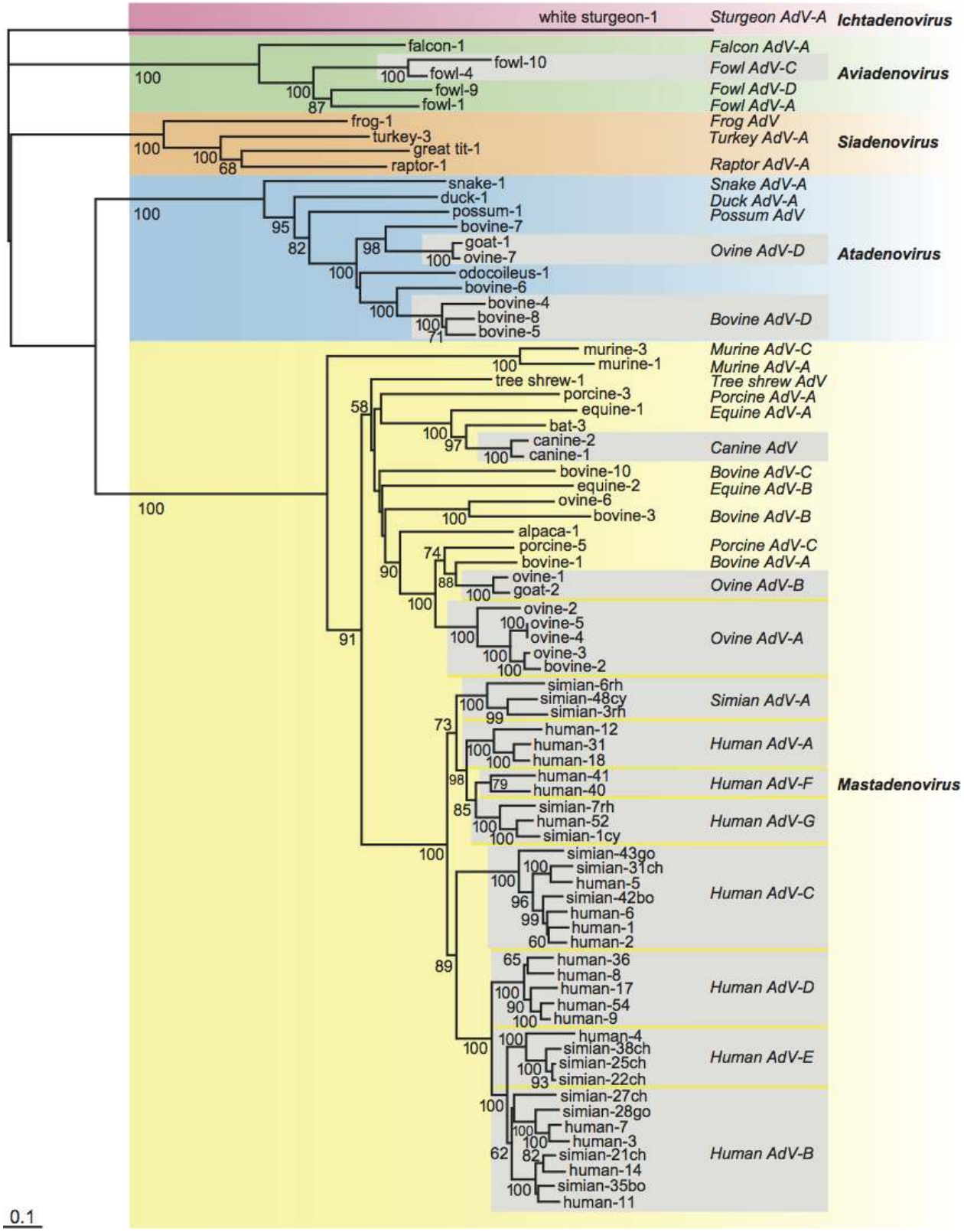


Figura 1.17: Árbol filogenético de adenovirus construido a partir de una matriz de distancias de secuencia de aminoácidos de la proteína del hexón. Para aquellas especies de adenovirus con más de tres serotipos sólo se muestran serotipos seleccionados (resaltado en gris). No se incluyen aquellos serotipos de adenovirus que infectan primates con eventos de recombinación homóloga en la proteína del hexón. Las abreviaciones son bo, bonobo; ch, chimpancé; go, gorila; cy, macaco cangrejero; rh, macaco Rhesus. Los nombres de las especies están abreviados reemplazando la palabra adenovirus por AdV. Se muestran los valores de remuestreo mayores a 50. Figura adaptada de Harrach *et al.* (2012).

Dentro de cada especie se agrupan los distintos serotipos o tipos virales. Según los criterios de la ICTV, los serotipos se diferencian en base a ensayos de neutralización complementados con evidencia sustancial biofísica, bioquímica o filogenética. Sin embargo, el criterio de utilizar los ensayos de neutralización para la clasificación de serotipos es aún polémico y se encuentra en revisión. Las reacciones de inhibición de la hemoaglutinación y neutralización están basadas en la interacción con las proteínas de la cápside viral, la proteína de la fibra y del hexón respectivamente. Dado que existen un número elevado de eventos de recombinación homóloga entre serotipos, estos métodos pueden llevar a la clasificación incorrecta de los serotipos virales (Singh *et al.*, 2012).

El género de adenovirus con mayor número de representantes es *Mastadenovirus*, que incluye sólo AdVs que infectan a mamíferos. Los miembros de *Mastadenovirus* están clasificados en más de 20 especies. Entre ellas, siete especies contienen a todos los AdVs que infectan a humanos y a algunos primates no humanos y nueve especies contienen a los restantes AdVs que infectan primates no humanos. El género *Aviadenovirus* incluye AdVs que infectan únicamente aves. El género *Atadenovirus* fue nombrado así por que presenta alto contenido A+T en el genoma y es el género con mayor rango de hospedadores, incluyendo aves, mamíferos, marsupiales y reptiles (Benkö y Harrach, 2003). Los AdVs que pertenecen al género *Siadenovirus* infectan aves, tortugas y anfibios y poseen un gen que codifica para una posible sialidasa en el extremo izquierdo del genoma (Benkö *et al.*, 2002; Davison *et al.*, 2003). Por último, el género *Ichtadenovirus* es el que se definió más recientemente, con un único miembro que infecta peces y posee el genoma más largo de todos los AdVs conocidos (Kovács *et al.*, 2003). En esta tesis se hace enfoque en las especies y serotipos agrupadas en el género *Mastadenovirus*.

Los serotipos de AdV tienen una especificidad de hospedador restringida, acotada a una única especie o especies altamente relacionadas. Por un lado, las similitudes entre los árboles filogenéticos de AdV y su hospedador, creados a partir de la proteasa viral y la subunidad menor del ARNr mitocondrial respectivamente, sugieren que los cinco géneros de AdV coespecian con sus hospedadores (Benkö y Harrach, 2003). Por otro lado, el alto número de recombinaciones observadas entre los distintos serotipos sugiere que un alto número de eventos de cambio de hospedador contribuye también a la diversidad observada en AdV (Benkö y Harrach, 2003). Estudios recientes realizados en AdV que infectan primates también contribuyen a esta hipótesis (Hoppe *et al.*, 2015).

1.4.2. Estructura genómica y proteínas de los adenovirus

Los AdVs son virus desnudos con una cápside icosaédrica de $\sim$ 90nm de diámetro. La cápside está formada por tres proteínas principales, el hexón, el pentón y la fibra, y siete proteínas minoritarias, IIIa, VI, VIII, IX, pVIn, pVIIn2 y VI. En cada vértice de la cápside (Figura 1.18) se ubica una estructura pentamérica de la proteína pentón (en rojo) acompañada por una fibra trimérica. El resto de la superficie de la cápside está recubiera por 240 proteínas triméricas del hexón (en verde). En el interior de la cápside se encuentra el genoma ADN doble cadena, con seis proteínas adicionales: IVa2, V, VII, μ, la proteína terminal y la proteinasa viral (Dai *et al.*, 2017).

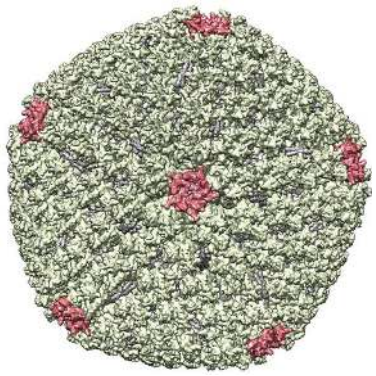


Figura 1.18: Estructura externa de la cápside del virión de HAdV5. Se muestran en color verde las proteínas del hexón y en color rojo las proteínas del pentón en un mapa de densidad electrónica del virión de HAdV5 obtenido por cryo-EM a una resolución de 3.2Å. (PDB ID: 6B1T). Figura adaptada de Harrach *et al.* (2012)

El genoma de AdV es un ADN lineal doble cadena de entre 26 y 45 kb (Figura 1.19) (Davison *et al.*, 2003).

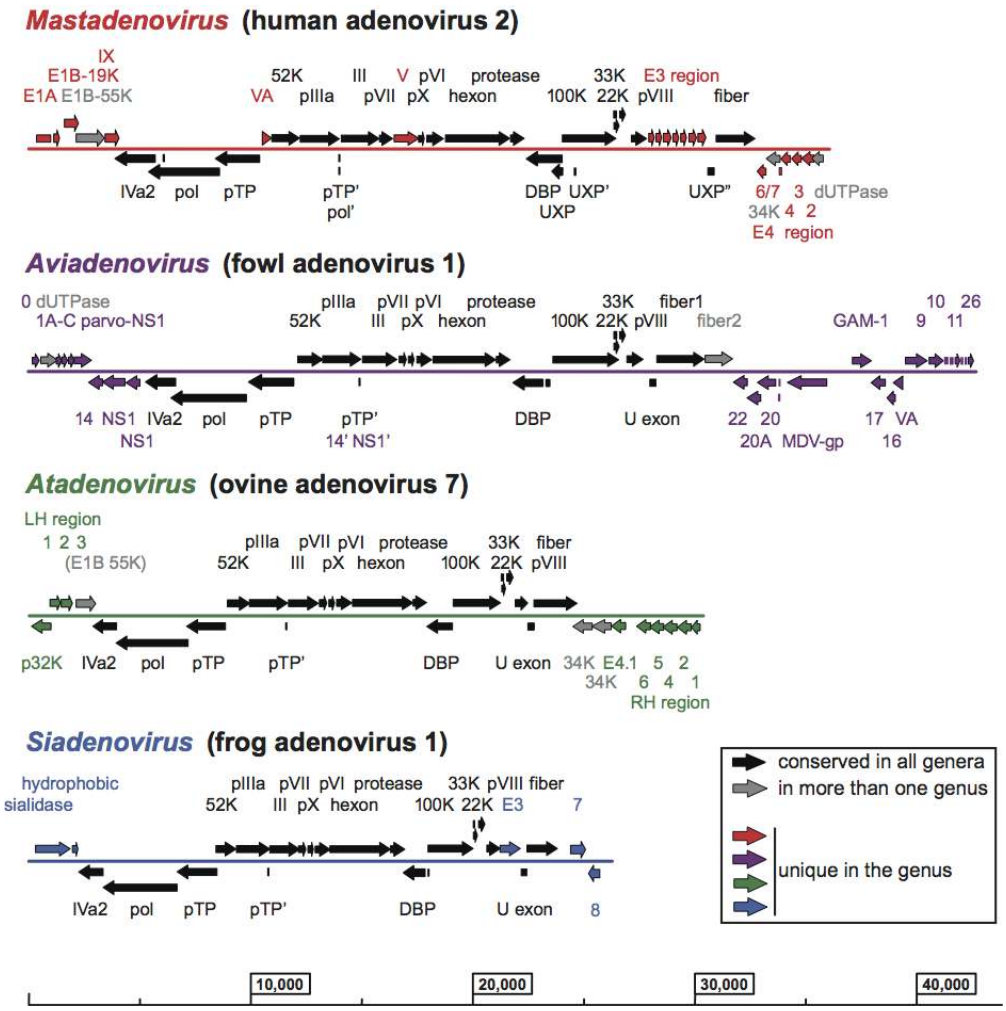


Figura 1.19: Esquema de la organización genómica de miembros de cuatro géneros de adenovirus. Las flechas negras indican los genes conservados en todos los géneros, las flechas grises indican presencia en más de un género y las coloreadas indican genes específicos de cada género. Figura adaptada de Harrach *et al.* (2012).

En el extremo de cada cadena 5' se encuentra unida covalentemente la proteína terminal que sirve como cebador en la replicación del ADN. Por convención, la región E1 del genoma de *Mastadenovirus* se dibuja a la izquierda. Los genes virales son codificados en ambas cadenas del genoma. Aquellos comunes a todos los géneros se encuentran en la parte central del genoma e incluyen tres genes homólogos que son requeridos para la replicación del ADN (la proteína terminal, la ADN polimerasa y la proteína de unión a ADN simple cadena) y los componentes principales de la estructura del virión, con excepción de la proteína V que se encuentra únicamente en el género *Mastadenovirus*. Otra característica común a todos los genomas de los géneros de AdV es la presencia en cada extremo del genoma viral de secuencias terminales invertidas (ITR) (en inglés, *Inverted Terminal Repeat*). Las ITRs funcionan como origen de duplicación del ADN y permiten el apareamiento de bases entre las cadenas individuales del ADN que se circularizan durante la síntesis asimétrica.

El genoma de los *Mastadenovirus* posee cinco unidades de transcripción temprana (E1A, E1B, E2, E3, y E4), cuatro unidades de transcripción intermedias que se transcriben al comienzo de la duplicación del ADN (IX, IVa2, L4 intermedia, y E2 tardía) y una unidad de transcripción tardía que es procesada para generar cinco familias de ARNm tardíos (L1 a L5). La ARN polimerasa III transcribe el gen VA (ARN viral asociado) y la ARN polimerasa II los restantes genes. Estos últimos además dan origen a múltiples ARNm por *splicing* alternativo y sitios de poliadenilación alternativos en el caso de los genes de E2 tardíos y E3.

Los genes tempranos se transcriben antes del comienzo de la duplicación del ADN y generan en la célula del hospedador un ambiente propicio para la replicación viral. Por ejemplo, el gen E1A codifica para dos proteínas principales que activan la transcripción e inducen que la célula del hospedador entre a la fase S del ciclo celular. La proteína E1B codifica dos proteínas que bloquean la apoptosis, el gen E2 codifica para tres proteínas que participan directamente en la replicación del ADN y el producto del gen E3 modula la respuesta del hospedador a la infección. Los genes tardíos se expresan luego de la duplicación del ADN y están involucrados en la producción y ensamblado de los componentes de la cápside.

1.4.3. Ciclo de vida de los adenovirus

El ciclo de vida de AdV puede dividirse para su estudio en dos fases separadas por el comienzo de la duplicación del ADN. La fase temprana ocurre antes de la duplicación del ADN e incluye la adsorción, la entrada y transporte de la partícula viral al núcleo y la expresión de los genes tempranos. La fase tardía comienza con la duplicación del ADN e incluye la expresión de los genes tardíos y el ensamblado de las partículas virales.

Fase temprana. La interacción inicial entre la partícula viral y la célula del hospedador está dada por una amplia diversidad de receptores. Algunas especies de AdV comienzan su ciclo mediante la interacción de la proteína de la fibra de la cápside viral con la proteína transmembrana de la célula del hospedador receptor de adenovirus y virus Coksackie B (CAR) (en inglés, *Cocksackie*

B and Adenovirus Receptor) (Bergelson *et al.*, 1997; Roelvink *et al.*, 1998; Tomko *et al.*, 1997). Los receptores CAR se expresan de manera abundante casi todos los órganos y son un componente de las uniones estrechas de las células epiteliales. El serotipo HAdV37 de la especie *Human adenovirus D* se une a moléculas de ácido siálico (Lenman *et al.*, 2015). Otras especies, como *Human adenovirus B* y *Human adenovirus D*, pueden unirse al receptor CD46 (Gaggar *et al.*, 2003; Liu *et al.*, 2018; Marttila *et al.*, 2005), que también se encuentra presente en la mayoría de los tipos celulares. La entrada de la partícula viral ocurre por endocitosis mediada por receptor, dependiente o independiente de clatrina según la especie. En ambos casos, el desensamblado de la partícula viral por pérdida de las fibras comienza en la superficie celular y continúa en el endosoma por cambios estructurales de la partícula viral debido al bajo pH. Por un mecanismo aún desconocido, las partículas virales semi-desnudas son liberadas al citoplasma y transportadas al núcleo, donde se produce el desnudamiento completo y ocurre la transcripción del ADN viral. La expresión de genes en la fase temprana induce a la célula a entrar en la fase S del ciclo celular, proveyendo un ambiente propicio para la replicación viral. El primer gen en expresarse es el gen de región temprana 1A (E1A). La proteína E1A juega uno de los roles más importantes en este proceso y es objeto de estudio de esta tesis, por lo que sus funciones serán discutidas en una sección aparte (véase Sección 1.4.5). Después se expresan el resto de los genes tempranos, E1B, E2, E3 y E4. El amplio repertorio proteico de los AdV se genera porque la mayoría de los transcriptos de AdV sufren *splicing* alternativo. El gen E1B codifica para dos proteínas anti-apoptóticas, E1B-55k y E1B-19k. E1B-55k promueve la ubiquitinación y degradación de la proteína p53 supresora de tumores, mientras que E1B-19k mimetiza la proteína celular Bcl-2 y bloquea la apoptosis uniéndose a BAK y BAX. Además, se desatan mecanismos de protección de la célula infectada frente a las defensas del sistema inmune del hospedador. Por último, en la fase temprana también se da la síntesis de los genes virales necesarios para la duplicación del ADN.

Fase tardía. A medida que se acumulan los elementos moleculares necesarios para la duplicación del ADN, comienza la fase tardía. El ADN de AdV se replica mediante un mecanismo que requiere un conjunto mínimo de proteínas: la proteína cebadora de la síntesis de ADN pTP terminal, la ADN polimerasa viral y la proteína de unión a ADN simple cadena, DBP. Al mismo tiempo, se sintetizan las proteínas estructurales que permitirán el ensamblado del virus. En el citoplasma se ensamblan los pentones pentaméricos y la fibra trimérica, que serán transportados al núcleo para el empaquetamiento del ADN viral en la cápside. Finalmente, la progenie viral es liberada de las células infectadas por lisis celular, permitiendo la expansión del virus en el tejido infectado.

Persistencia y latencia de adenovirus. La latencia implica que el genoma de AdV se mantiene en la célula, sin integrarse, de manera episomal, y que algunos genes se expresan de manera basal. Se cree que la infección persistente o latente de AdV explicaría las enfermedades observadas en hospedadores inmunocomprometidos. Sin embargo, el mecanismo de persistencia o latencia de AdV en los tejidos linfáticos o amígdalas aún no se conoce. Estudios recientes sugieren que el

mecanismo de persistencia está relacionado con la inhibición de la expresión de la proteína E1A en presencia de interferón (Zheng *et al.*, 2016).

1.4.4. Patologías de adenovirus

Los AdV poseen una capacidad de infección y replicación en su hospedador moderadamente específica. Los AdV que infectan a humanos replican en cultivos de células de rata (Ginsberg *et al.*, 1989), ratón, perro, cerdos y hamsters (Jogler *et al.*, 2006), pero en menor medida y con una alta expresión de las proteínas tempranas y baja o nula expresión de los genes tardíos. Las líneas celulares derivadas de humanos - HEK293, HeLa y A549 - son el mejor sistema de cultivo para la producción abundante de HAdV (Graham *et al.*, 1977). En cultivos celulares se observa que los AdV producen rápidamente un efecto citopático y un rápido desprendimiento de la monocapa celular. A nivel celular, la infección por AdV causa inhibición de la síntesis del ADN celular, ARNm y proteínas, produciendo cambios visibles en la morfología celular como ser agrandamiento del núcleo con inclusiones formadas por las partículas virales (Tollefson *et al.*, 1996).

La transmisión del virus se produce a través de la ruta fecal-oral, aerosoles o en contacto con superficies contaminadas (Fox *et al.*, 1969). Así, la vía de entrada de AdV al hospedador es por boca, nasofaringe o conjuntiva ocular. Por ejemplo, en niños la infección por los serotipos HAdV1, HAdV2, HAdV5 y HAdV6 puede estar presente meses, especialmente en heces, y en consecuencia producir la transmisión endémica por la ruta fecal-oral. Las infecciones endémicas por serotipos que causan queratoconjuntivitis se esparcen comúnmente en natatorios por agua contaminada y en consultorios médicos por instrumentos oftalmológicos contaminados. Si bien se identificaron diferentes receptores compartidos entre varios serotipos de AdV, no se conoce aún por qué algunos serotipos causan enfermedades en algunos órganos y no en otros. Los conocimientos adquiridos hasta el momento mediante experimentos *in vitro*, *in vivo* o *in silico* no explican los mecanismos del tropismo tisular ni las patologías órgano-específicas.

El sitio inicial de replicación de AdV son las amígdalas y adenoides de la orofaringe, aunque también se incluye el epitelio respiratorio no ciliado en serotipos que causan enfermedades respiratorias. La expansión de la infección se debe al aumento de la permeabilidad entre las células como consecuencia de la interacción entre la proteína de la fibra de AdV y los homodímeros CAR, que facilita el acceso de las partículas virales al torrente sanguíneo luego de la lisis celular. Los AdV pueden infectar y replicar en varios lugares del tracto respiratorio, ojo o tracto gastrointestinal y en algunos casos infectan la vejiga, el hígado y otros órganos como ser páncreas, miocardio o sistema nervioso central.

En general, los AdV están bien adaptados a sus hospedadores y causan infecciones asintomáticas o leves. En humanos, la mayoría de las infecciones son subclínicas con formación de anticuerpos de protección específicos contra el serotipo. Sin embargo, el virus puede persistir durante meses, mediante mecanismos aún no del todo claros (King *et al.*, 2016), en el tracto gastrointestinal y respiratorio (Lion, 2014). En humanos, la infección de AdV varía entre esporádica y epidémica correlacionando con el serotipo viral y la edad (niños o adultos) de la población susceptible. Se

estima que los AdV causan el 8 % de las enfermedades virales relevantes y 5 a 10 % de las enfermedades febriles. A continuación se describen brevemente las patologías comúnmente asociadas a AdV.

Adenovirus y enfermedades respiratorias. La primera caracterización de AdV ocurrió durante una epidemia de infección aguda respiratoria por los serotipos HAdV4 o HAdV7 entre soldados durante la segunda guerra mundial, favorecida por la superpoblación en cuarteles y fatiga de los soldados (Dingle y Langmuir, 1968). Los AdV son responsables del 7 % de las infecciones del tracto respiratorio superior en niños, los síntomas incluyen congestión nasal, rinitis y tos. Estos síntomas pueden estar acompañados por conjuntivitis (fiebre faringoconjuntiva). Los AdV también causan enfermedades del tracto respiratorio inferior y son responsables del 10 % de las neumonías en niños. La mayoría de los pacientes se recuperan, pero algunas epidemias tuvieron alta mortalidad. En Argentina, en los años 80, adenovirus tuvo una mortalidad del 35 % de los casos por el serotipo HAdV7 (Murtagh *et al.*, 1993).

Adenovirus y enfermedades oculares. La conjuntivitis aguda puede darse como parte de la fiebre faringoconjuntiva o aislada. Es una enfermedad leve y se espera una recuperación sin secuelas. Epidemiológicamente, las infecciones suelen ser esporádicas y en grupos cerrados como dentro de una familia o grupos de personas que comparten natatorios. Por otro lado, la queratoconjuntivitis (EKC) es una enfermedad muy contagiosa y puede derivar en conjuntivitis hemorrágica. El contagio ocurre en instalaciones médicas o natatorios y se caracteriza por conjuntivitis acompañada de edema en pestañas, dolor, lagrimeo excesivo y fotofobia, con infiltración en la córnea produciendo la opacidad característica de esta enfermedad (Hamada *et al.*, 2008; Lynch y Kajon, 2016).

Adenovirus y enfermedades urinarias. La cistitis hemorrágica aguda por infección de AdV es una enfermedad autolimitada que ocurre casi exclusivamente en varones jóvenes o pacientes trasplantados inmunosuprimidos. El síntoma que la caracteriza es la hematuria, por lo que se la confunde frecuentemente con enfermedades más severas de los riñones (Lynch y Kajon, 2016; Yokose *et al.*, 2009).

Adenovirus y enfermedades del sistema nervioso central. Ocasionalmente se aísla AdV del fluido cerebroespinal o del cerebro en autopsias en pacientes inmunosuprimidos o inmunocompetentes con meningoencefalitis o meningitis. Algunos casos de meningoencefalitis asociada a AdV ocurrieron luego de una neumonía en individuos inmunocompetentes (Schwartz *et al.*, 2018).

Adenovirus y enfermedades del tracto gastrointestinal. La gastroenteritis o inflamación del estómago e intestino se caracteriza por fiebre, vómitos y diarrea. Los principales serotipos causantes de diarrea infantil son HAdV40 y HAdV41. La asociación de AdV con esta enfermedad fue tardía, ya que la presencia de AdV en las heces se da tanto en individuos enfermos como individuos sanos con una infección subclínica. La infección puede causar una gastroenteritis aguda con o sin

síntomas respiratorios. La incidencia es variable y ocurre con mayor frecuencia en niños menores de 4 años. Su prevalencia es menor a la de rotavirus, de la cual es indistinguible clínicamente. Algunas complicaciones son colitis hemorrágica, hepatitis y pancreatitis (Lynch y Kajon, 2016).

Adenovirus y enfermedades del sistema circulatorio. La miocarditis es una enfermedad inflamatoria del miocardio que se caracteriza por necrosis de los miocitos. Los AdV pueden infectar miocitos en cultivo y entrar a la célula utilizando el receptor CAR. Si bien la infección sintomática del corazón no es lo más común, se reportaron numerosos casos donde se observa una asociación entre miocarditis y la infección por AdV (Bowles *et al.*, 2003; Valdés *et al.*, 2008).

Adenovirus y cáncer. En 1962, Trentin y colegas observaron que HAdV12 podía causar tumores en hámsters recién nacidos, siendo la primera demostración de que un virus humano podría ser el agente causal de cáncer (Trentin *et al.*, 1962). Además de la especie *Human adenovirus A*, los serotipos virales pertenecientes a *Human adenovirus B* pueden inducir tumores en hámsteres recién nacidos, mientras que la especie *Human adenovirus D* puede inducir tumores mamarios en ratas. Para que ocurra la transformación son suficientes las regiones E1A y E1B de los HAdV. En células transformadas y tumores inducidos en roedores se observan estas regiones integradas al genoma celular. En humanos no hay evidencia fuerte que defina a los AdV como agentes causales de cáncer. No se observó ADN, ARN o proteínas de AdV en tumores de pacientes humanos. Se cree que puede existir una asociación con leucemia linfoblástica aguda y la infección uterina de AdV, pero la evidencia hasta el momento no es contundente. Con la excepción de las líneas celulares HEK293, retinoblastos embrionarios humanos y células del fluido amniótico, no se observó transformación de células humanas que sigan el modelo observado en roedores. Sin embargo, aún no se descarta que exista un mecanismo de transformación "golpeo y corro" (en inglés, *hit and run*) mediante el cual AdV cause un cambio en las células infectadas que resulten en cáncer sin que el genoma viral sea retenido (Berk, 2013). En humanos no se observa proliferación debida a transformación en infecciones agudas de AdV. Sin embargo, el tejido linfático presenta hipertrofia y se observen centros germinales activos. Hasta la fecha, la evidencia existente no es suficiente para apoyar la asociación entre adenovirus y cáncer.

1.4.5. E1A funciones, estructura y motivos preexistentes

La proteína E1A se encuentra únicamente en el género *Mastadenovirus*. Luego de la infección, es la primera proteína en expresarse y es esencial para la producción de partículas virales (Jones y Shenk, 1979). Luego de asociar a HAdV12 como agente causal de tumores (Trentin *et al.*, 1962), se determinó que la región E1A y E1B del genoma eran responsables de las propiedades oncogénicas y HAdV fue utilizado para estudiar muchos de los procesos biológicos fundamentales como ser regulación del ciclo celular y splicing del ARNm.

El transcripto primario de E1A da origen a cinco isoformas en el serotipo viral HAdV5 y HAdV2 (Figura 1.20), nombradas según su coeficiente de sedimentación: 13S, 12S, 11S, 10S y

9S. Las isoformas codifican para 5 proteínas de 289, 243, 217, 171 y 55 residuos, respectivamente (Miller *et al.*, 2012; Stephens y Harlow, 1987). Todos las isoformas con excepción de la 9S mantienen el marco de lectura (Virtanen y Pettersson, 1983).

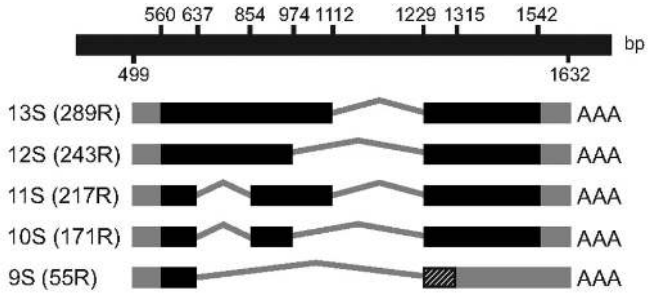


Figura 1.20: Esquema de los transcriptos de E1A de HAdV5. Se muestran los distintos ARNm generados por splicing alternativo a partir del transcripto primario de E1A. A la izquierda se indican los coeficientes de sedimentación y entre paréntesis el número de residuos de la proteína correspondiente. En el centro se muestra un esquema indicando las regiones no codificantes (en gris) y codificantes (en negro) de los exones (rectángulos). A la derecha se indica la cola de poli-adenina. Figura adaptada de Miller *et al.* (2012).

Durante las etapas tempranas de la infección, las isoformas de 289 y 243 residuos son las que presentan mayores de niveles de expresión (Perricaudet *et al.*, 1979; Stephens y Harlow, 1987). La única diferencia entre ambas proteínas es la ausencia de un segmento de 46 residuos en la proteína más pequeña. Ambas proteínas se encuentran en el núcleo y en el citoplasma y son responsables de la mayoría de las actividades de E1A (Rowe *et al.*, 1983; Turnell *et al.*, 2000). Entre ellas se incluye la inducción de la entrada a la fase S del ciclo celular y la activación de los genes virales (Berk *et al.*, 1979; Jones y Shenk, 1979; Montell *et al.*, 1984, 1982; Winberg y Shenk, 1984). Estas funciones llevan a generar un ambiente celular adecuado esencial para la replicación viral (Bayley y Mymryk, 1994; Berk, 2005; Flint y Shenk, 1997; Gallimore y Turnell, 2001; Pelka *et al.*, 2008). Esta tesis se enfoca en el estudio de la isoforma de 289 residuos de E1A para poder comparar la evolución en el dominio globular y en los dominios desordenados.

Estructura de dominios de E1A

Estudios de biología molecular (Kimelman *et al.*, 1985; van Ormondt y Hesper, 1983) y bioinformática (Avvakumov *et al.*, 2004, 2002) de las secuencias de E1A de 34 serotipos virales distintos que infectan a humanos revelaron cuatro regiones de alta similitud de secuencia denominadas región conservada 1 (CR1), 2 (CR2), 3 (CR3) y 4 (CR4) y una región N-terminal menos conservada (Figura 1.21). Los dominios CR1 y CR2 presentan regiones de secuencia con similitud a las proteínas E7 de papilomavirus y el antígeno Large T de poliomavirus. Además de los dominios clásicos se identificó una región que separa los dominios CR2 y CR3 y es un determinante oncogénico de la proteína E1A de HAdV12 (Telling y Williams, 1994). Las regiones auxiliares 1 y 2 fueron caracterizadas en HAdV5 como co-reguladores de la transcripción de los genes tempranos virales (Bondesson *et al.*, 1992).

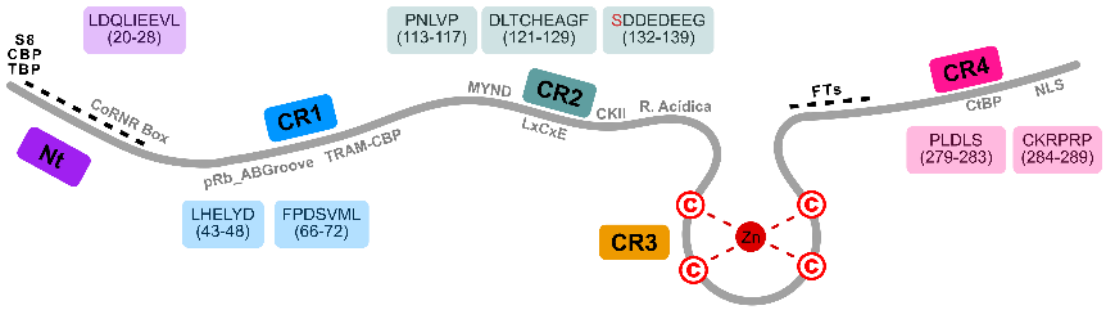


Figura 1.21: Representación esquemática de la proteína E1A. Se indica la posición relativa y el esquema de dominios de la proteína E1A de HAdV5. El átomo de zinc coordinado por las cuatro cisteínas (C roja) en el dominio CR3 está representado como un círculo rojo. Se indican además la ubicación relativa de los motivos lineales de los dominios desordenados, con las posiciones correspondientes a la proteína HAdV5 indicadas entre paréntesis. En forma de recuadro se indica la secuencia correspondiente a cada motivo lineal de cada región: N-terminal (CoRNR Box), CR1 (pRb_ABGroove), CR2 (MYND, LxCxE, CKII y Región Acídica) y CR4 (CtBP y NLS). En rojo se indica la serina 132 fosforilable por CKII. Las líneas punteadas indican el sitio de interacción de las proteínas CBP, TBP y S8 en el dominio N-terminal y el sitio de interacción de los factores de transcripción (FTs) en el dominio CR3.

Motivos lineales de E1A

Estudios de biología molecular mostraron que la proteína E1A posee diez motivos lineales, identificados en los serotipos HAdV12 o HAdV5 de las especies *Human adenovirus A* y *Human adenovirus C* respectivamente. Estos motivos son responsables de la interacción con distintas proteínas del hospedador.

Análisis de mutagénesis en el dominio N-terminal de HAdV5 determinaron que numerosas posiciones altamente conservadas en la región N-terminal de E1A HAdV5 eran relevantes para la interacción con diferentes blancos celulares (Boyd *et al.*, 2002; Rasti *et al.*, 2005) sugiriendo la existencia de un motivo lineal que media la interacción con CBP, la proteína de unión a la región TATA (TBP) (en inglés, *TATA Binding Protein*) y la subunidad S8 del proteosoma. Esta región se superpone con un motivo caja del receptor del corepresor nuclear (CoRNR Box) (en inglés, *CoRepressor Nuclear Receptor box*) (Meng *et al.*, 2005; Phelan *et al.*, 2010) que modula la interacción con receptores nucleares. El dominio CR1 posee un motivo pRb_ABGroove que media la interacción con el dominio AB de pRb (Dyson *et al.*, 1992a; Liu y Marmorstein, 2007) (véase Sección 1.5.1) y una región que interactúa con el motivo de unión a la región TRAM de CBP (Ferreon *et al.*, 2009). El dominio CR2 contiene un motivo de unión a los dominios MYND (Ansieau y Leutz, 2002; Hateboer *et al.*, 1995; Isobe *et al.*, 2006) y un motivo LxCxE que se une a un surco en el subdominio B de pRb (Dyson *et al.*, 1992a; Whyte *et al.*, 1988b), un sitio de fosforilación por la quinasa CKII (Whalen *et al.*, 1996) y una región acídica que coopera en la unión al motivo LxCxE en la unión a pRb (Palopoli *et al.*, 2018) (véase Sección 1.5.1). Mediante el dominio CR3, la proteína E1A interactúa con reguladores de la transcripción incluyendo sitios de unión para factores de transcripción del hospedador (Chatton *et al.*, 1993; Liu y Green, 1994) y factores asociados a TBP (Geisberg *et al.*, 1995; Mazzarelli *et al.*, 1997). En el CR3 se encuentra también

el motivo CxxC que coordina el zinc. Se describieron dos motivos en el dominio CR4. Estudios *in vivo* e *in vitro* demostraron la existencia de un motivo de unión a la proteína de unión al extremo C-terminal (CtBP) (en inglés, C-terminal Binding Protein) (Boyd *et al.*, 1993; Cohen *et al.*, 2013; Molloy *et al.*, 2007, 2006, 1998; Schaeper *et al.*, 1995). CtBP es un co-represor de la transcripción que participa en la regulación del ciclo celular y fue identificado por primera vez como una proteína de interacción con E1A (Boyd *et al.*, 1993). Continuo en secuencia se describió una NLS (Köhler *et al.*, 2001; Lyons *et al.*, 1987; Madison *et al.*, 2002) que interactúa con la proteína importina $\alpha - 1$.

Características biofísicas de E1A

Estudios de RMN revelaron que los dominios CR1, CR2 y CR4 de E1A de HAdV5 son intrínsecamente desordenados, mientras que el dominio N-terminal y el CR3 se pliegan en una estructura de hélice α poco entendida (Hošek *et al.*, 2016; Pelka *et al.*, 2008). Sin embargo, no se reportó aún ninguna estructura cristalográfica para estos dos dominios. El dominio CR3 contiene dos motivos CxxC que coordinan la unión a un zinc (Culp *et al.*, 1988), similares a los presentes en el dominio globular de la proteína E7. Sin embargo, ambos dominios no presentan similitud de secuencia más allá de estos motivos. Estudios estructurales utilizando fragmentos de las proteínas E1A de HAdV5 o HAdV12 muestran que el dominio CR1 (Ferreon *et al.*, 2009), CR4 (Molloy *et al.*, 2000) y los dominios N-CR1 (Haberz *et al.*, 2016) son intrínsecamente desordenados cuando están aislados del resto de la proteína, apoyando la evidencia obtenida por RMN.

La proteína E1A puede ser descrita globalmente como intrínsecamente desordenada y, al igual que la proteína E7 de papilomavirus, posee transiciones de desorden a orden. Por ejemplo, en presencia de trifluoroetanol los sitios de unión para las proteínas CtBP y TBP adoptan conformaciones β y α respectivamente (Molloy *et al.*, 2000, 1998, 1999). De igual manera, tanto el dominio CR1 como la fusión de los dominios N-CR1 adquieren una estructura ordenada luego de la unión al blanco proteico CBP (Ferreon *et al.*, 2009; Haberz *et al.*, 2016). Estas transiciones conformacionales probablemente regulen la interacción simultánea con múltiples blancos proteicos (Ferreon *et al.*, 2013).

1.5. Motivos Lineales y Virus

Uno de los campos más interesantes en la biología es el estudio del mecanismo de infección de los distintos patógenos. La relación entre parásito y hospedador está principalmente basada en las interacciones proteína-proteína que permiten la comunicación entre ambos y juegan un rol principal en el establecimiento de la infección. El estudio de estos mecanismos es interesante desde dos puntos de vista. En primer lugar permite identificar potenciales blancos terapéuticos y en segundo lugar permite profundizar en el conocimiento de los mecanismos de interacción proteína-proteína.

La interacción entre un patógeno y su hospedador conlleva al establecimiento de una red de interacciones entre proteínas del patógeno y proteínas del hospedador, cuya principal consecuencia es la generación de cambios de los procesos celulares del hospedador hacia procesos que favorezcan la supervivencia del patógeno. Para lograr estos cambios, muchas de las proteínas del patógeno poseen la capacidad de interactuar con proteínas altamente conectadas del hospedador o que conectan módulos funcionales en el hospedador (cuellos de botella) (Dyer *et al.*, 2008). En particular, los patógenos virales logran su supervivencia manipulando distintos mecanismos celulares. Pueden tomar el control del ciclo celular para asegurar la transcripción del genoma viral, interactuando por ejemplo con factores de transcripción, o interactuar con blancos proteicos claves para regular procesos como la apoptosis, el transporte celular o la evasión del sistema inmune. La interacción entre las proteínas virales y las proteínas del hospedador se establece en muchos casos gracias a la presencia de motivos lineales en la proteína viral que mimetizan los motivos lineales del hospedador (Davey *et al.*, 2011b; van der Lee *et al.*, 2014). La facilidad de evolución por convergencia de los motivos lineales y su aparición en proteínas no relacionadas permite que los patógenos virales posean proteínas que mimetizan los motivos lineales de las proteínas del hospedador y compiten en afinidad. En consecuencia pueden manipular interacciones proteína-proteína y los procesos celulares del hospedador (Davey *et al.*, 2011b; van der Lee *et al.*, 2014), generando un contexto que favorece su supervivencia.

Los patrones evolutivos de los motivos lineales en general están poco caracterizados y se han estudiado poco en patógenos. Sin embargo, la mayoría de los estudios existentes sugieren que los motivos lineales en virus son responsables de una reorganización de las redes de interacción del hospedador y poseen un rol adaptativo en la evolución viral.

Alrededor del 30 % de los motivos en ELMdb pueden encontrarse en proteínas de virus no relacionados y muchos motivos lineales se identifican en diferentes grupos de patógenos (Chemes *et al.*, 2015). Estas dos evidencias sugieren que los motivos lineales evolucionaron de manera convergente a través de grupos de patógenos.

Utilizando la proteína E7 de papilomavirus como modelo, nuestro grupo de trabajo realizó la reconstrucción de la historia de motivos lineales en 217 secuencias de la proteína E7. Los papilomavirus coevolucionaron con sus hospedadores amniotas, proveyendo una filogenia bien establecida de ~350 millones de años. El patrón evolutivo de los motivos lineales fue variable. Mientras que algunos motivos altamente conservados presentaron un único evento de aparición, cuatro motivos

lineales dentro de E7 mostraron múltiples eventos de aparición independientes en ramas profundas y recientes, dando evidencia directa de evolución convergente dentro de una filogenia viral (Chemes *et al.*, 2012b). La proteína E7 no es el único caso, el estudio de motivos presentes en la proteína P de los *Paramyxovirus* mostró que los motivos STAT-1 de Nipah y Measles, considerados inicialmente homólogos, aparecieron independientemente dos veces a lo largo de la historia evolutiva. El estudio de instancias de motivos que evolucionan por convergencia dentro de una familia viral requiere una construcción cuidadosa de los alineamientos a analizar y las filogenias virales (Chemes *et al.*, 2015). Una conexión interesante para hacer con el estudio evolutivo de los motivos lineales, es observar en forma paralela la asociación entre el repertorio de motivos lineales y los cambios fenotípicos virales, como virulencia, persistencia, tropismo celular y hospedador (Chemes *et al.*, 2015). Esta asociación deriva de estudios estadísticos entre la presencia o ausencia de un determinado motivo y un conjunto de rasgos fenotípicos. En la proteína E7 de papilomavirus se observó una asociación entre la ausencia del motivo LxCxE y la infección a hospedadores del orden Perissodactyla (Chemes *et al.*, 2012b). Esta evidencia sugiere un rol adaptativo de los motivos lineales en la evolución viral. Los motivos lineales de las proteínas del patógeno pueden funcionar de manera coordinada y su detección puede determinarse estudiando patrones de asociaciones de manera estadística. Un ejemplo de funcionamiento coordinado es el caso de los motivos CKII y LxCxE y el motivo CKII con la región acídica los cuales co-ocurren de manera significativa (Chemes *et al.*, 2012b). Sin embargo, estos estudios presentan ciertos desafíos. Es necesario contar con una buen conocimiento filogenético de la familia viral y con un buen alineamiento de las proteínas a analizar. Además, con el objetivo de no obtener falsos negativos, es necesario considerar que los motivos muchas veces no mantienen su localización dentro de la misma proteína e incluso entre distintas proteínas.

En resumen, el estudio de los motivos lineales en combinación con herramientas filogenéticas ayudaría a comprender el rol de los motivos lineales en la evolución viral.

1.5.1. Ejemplo de interacción proteína del hospedador-proteína viral: La proteína retinoblastoma

El mecanismo principal que permite la supervivencia de los patógenos en la célula infectada es el secuestro de la maquinaria celular mediante motivos lineales que mimetizan los del hospedador (Davey *et al.*, 2011b). En algunos casos puede ocurrir que estas interacciones sean mediadas entre múltiples superficies del dominio globular de la proteína del hospedador y motivos lineales en la proteína viral (Ferreon *et al.*, 2013; Jansma *et al.*, 2014). Los detalles de una interacción entre un motivo viral y un dominio globular del hospedador pueden variar entre patógenos debido al contexto.

La proteína retinoblastoma (pRb) supresora de tumores, también conocida como p105, forma parte de la familia de las proteínas bolsillo (en inglés, *pocket protein*) junto con las proteínas p107 y p130. Dada la alta similitud de secuencia, los tres miembros de la familia comparten numerosas funciones celulares. En esta sección nos enfocaremos principalmente en la función de pRb.

pRb fue idenficada inicialmente como uno de los genes cuya mutación era responsable de la predisposición al tumor pediátrico retinoblastoma. Se sabe que la mayoría de los distintos tipos de cáncer ocurren con la modificación de la función de pRb, ya sea por mutación directa del gen codificante o por una alteración en la expresión de los reguladores del funcionamiento de pRb (Dick y Rubin, 2013). Más de 500 mutaciones diferentes fueron identificadas en el gen codificante para pRb en tumores de retinoblastoma (Lohmann, 1999; Valverde *et al.*, 2005).

El gen codificante para pRb fue clonado hace más de 30 años (Friend *et al.*, 1986). Desde entonces, pRb fue identificada como un regulador universal del ciclo celular, con un rol central en el control del pasaje de la fase G1 a la fase S de replicación del ADN. pRb participa en la regulación del ciclo celular regulando de manera negativa la proliferación, diferenciación, senescencia y apoptosis celular (Burkhart y Sage, 2008; Classon y Dyson, 2001; Classon y Harlow, 2002; Cobrinik, 2005; Dick *et al.*, 2018; Dick y Rubin, 2013). Estas funciones son mediadas por la interacción con más de 100 proteínas, de manera dependiente del tipo celular y el estadío del ciclo celular (Dick, 2007; Morris y Dyson, 2001). Su función principal la lleva a cabo mediante la interacción con la familia de factores de transcripción E2F, reprimiendo la expresión génica regulada por ellos y mediante la interacción con co-reguladores transcripcionales que modifican la estructura de la cromatina. La actividad de pRb es inhibida por la actividad de quinasas dependiente de ciclinas y por proteínas de virus oncogénicos como E7 de papilomavirus, E1A de adenovirus y el antígeno Large T de poliomavirus.

Estructura de la proteína retinoblastoma. La arquitectura de dominios de pRb en humanos (Figura 1.22) consiste de un dominio N-terminal estructurado (RbN, residuos 53-355), formado por dos subdominios A y B (RbNA, 53-245 y RbNB, 267-355), unido mediante el conector RbIDL (residuos 356-379) al dominio bolsillo (RbAB, residuos 380-787) y un dominio C-terminal intrínsecamente desordenado (RbC, residuos 788-928). El dominio RbAB contiene dos subdominios A (residuos 380-581) y B (residuos 643-787) que están conectados por un conector de aproximadamente 60 residuos de longitud (RbPL) (Burke *et al.*, 2012).

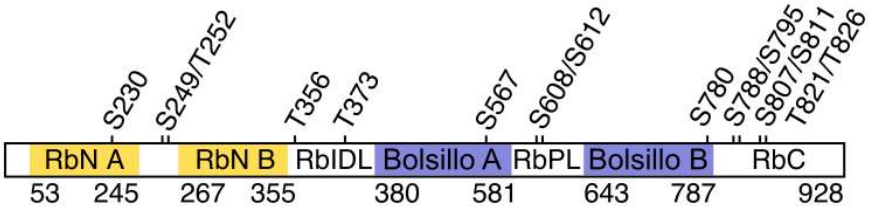


Figura 1.22: Representación esquemática de la proteína retinoblastoma. Se muestra la estructura de dominios de la proteína retinoblastoma. Los dominios estructurados, RbN y el dominio RbAB, se muestran en amarillo y azul respectivamente. Las regiones desordenadas, incluyendo los conectores RbIDL y RbPL y el dominio RbC, no están coloreados. Los sitios de fosforilación de CDK conservados se indican en la parte superior. Figura adaptada de Burke *et al.* (2012).

En humanos, el dominio RbAB de pRb fue identificado como la mínima región necesaria para unir las proteínas virales como E1A de adenovirus, E7 de papilomavirus y el antígeno Large T de poliomavirus (Dyson *et al.*, 1992a,b; Münger *et al.*, 1989). Cada subdominio A y B posee una estructura similar a ciclina e interactúan entre ellos por una superficie no covalente, conformando una unidad estructural (Figura 1.23). Entre ambos dominios se forma un bolsillo denominado bolsillo AB.

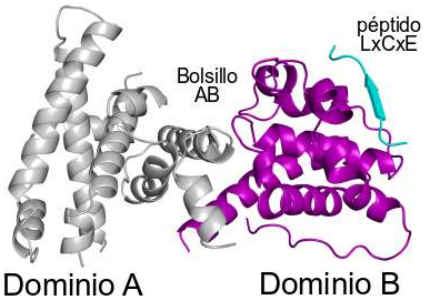


Figura 1.23: Estructura del dominio RbAB. Se muestra la estructura cristalográfica del dominio RbAB (PDB ID: 1GUX) (Lee *et al.*, 1998). Cada dominio está representado en forma de cintas con colores distintos. El péptido LxCxE de la proteína E7 de papilomavirus está representado en forma de cintas en color celeste. Se indica además la ubicación del bolsillo AB.

Funciones de la proteína retinoblastoma. La función más estudiada y mejor caracterizada de pRb es el control de la transición en el ciclo celular de la fase G1 a la fase S mediante la interacción con factores de transcripción de la familia E2F (Figura 1.24). Las proteínas E2F son factores de transcripción que regulan la expresión de genes involucrados en la progresión del ciclo celular, como ciclina A (Schulze *et al.*, 1995) o la ADN polimerasa (DeGregori *et al.*, 1995) y en la apoptosis, como Arf (Guo *et al.*, 2001).

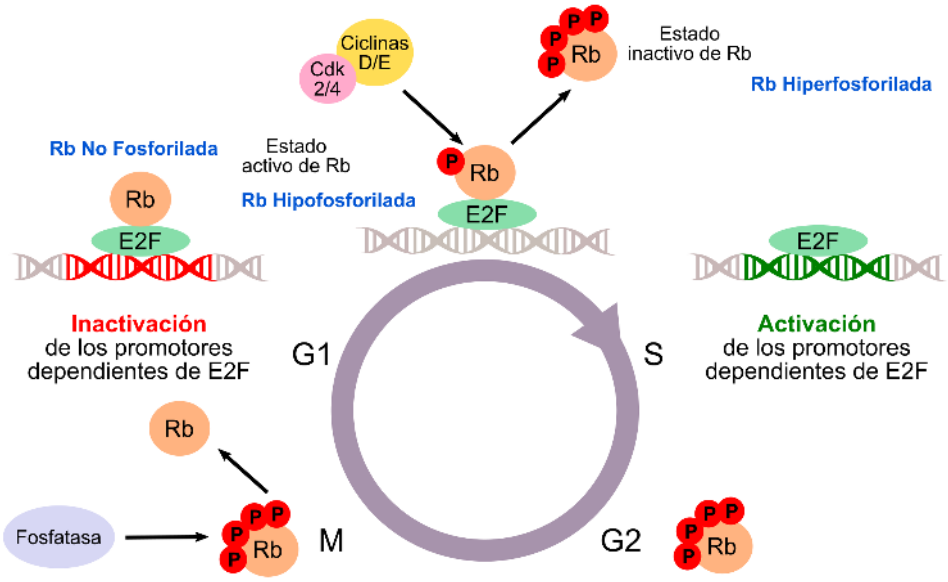


Figura 1.24: Regulación de la progresión de la fase G1 a S por el complejo entre la proteína retinoblastoma y el factor de transcripción E2F. Figura adaptada de `https://viralzone.expasy.org/`.

En mamíferos, la familia E2F tiene un total de 8 miembros (Figura 1.25), que pueden agruparse en dos subfamilias: las proteínas E2F1-3 funcionan como activadores de la transcripción, mientras que E2F4-8 actúan como represores. Las proteínas E2F1-5 interaccionan mediante el dominio TA-PB con miembros específicos de la familia de las proteínas bolsillo. El dominio de unión a las proteínas bolsillo está ausente en E2F6-8. Todas las proteínas de la familia E2F poseen al menos un dominio de unión al ADN (DBD). Las proteínas E2F1-6 se unen al ADN como heterodímeros con las proteínas DP1 y DP2 con las cuales interactúan a través del dominio de unión a las proteínas DP (DP-BD). Las proteínas E2F7-8 son consideradas atípicas, ya que contienen dos subdominios de unión al ADN diferentes (DBD1 y DBD2) (Morgunova *et al.*, 2015). Las proteínas E2F1-3 presentan un dominio de unión a ciclina A, ausente en E2F4-8. E2F1-3 interaccionan con pRb, E2F-5 interacciona con p130 mientras que E2F-4 se asocia con pRb, p130 y p107 (Lee *et al.*, 2002; Liban *et al.*, 2016).

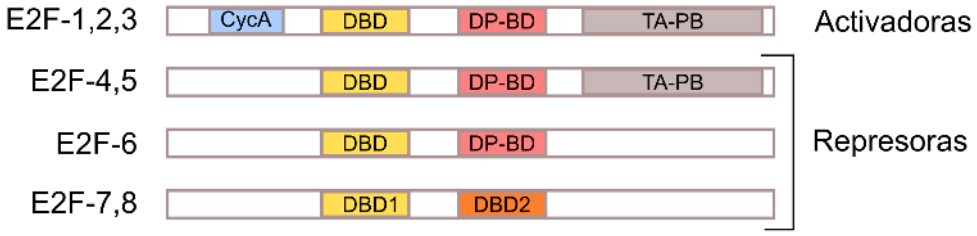


Figura 1.25: Representación esquemática de la organización de dominios de la familia E2F. Se muestra la estructura de dominios de la familia de factores de transcripción E2F: el dominio de unión a ciclinas (CycA, azul), el dominio de unión al ADN (DBD, amarillo y naranja), el dominio de unión a las proteínas DP (DP-BD rojo) ausente en E2F-7 y 8 y el dominio de transactivación y unión a las proteínas bolsillo (TA-PB, gris). Figura adaptada de Morgunova *et al.* (2015).

La asociación de pRb, p107 y p130 con E2F está regulada por la fosforilación de quinasas dependientes de ciclinas (CDKs). Existen al menos 15 sitios de fosforilación por ciclinas en pRb (Figura 1.22), que modifican la afinidad de pRb por E2F y otros blancos proteicos (Burke *et al.*, 2012). A lo largo del ciclo celular pRb existe en tres estados. (1) No fosforilada, (2) el estado activo hipofosforilada y (3) el estado no activo hiperfosforilada (Figura 1.24). En el estado no fosforilado pRb se encuentra unida a los factores de transcripción E2F. La fosforilación comienza en la fase G1 por los complejos Cdk4/Ciclina D y Cdk2/Ciclina E. El estado hipofosforilado se considera el estado activo de pRb respecto a la capacidad de inhibir la progresión del ciclo celular. En este estado, pRb mantiene su capacidad de interacción con E2F. La fosforilación progresiva por los complejos Cdk4/Ciclina D y Cdk2/Ciclina E, deriva en la hiperfosforilación de pRb. La hiperfosforilación de pRb induce un cambio conformacional que inhibe la interacción con E2F.

La hiperfosforilación de pRb se considera un punto de no-retorno en el ciclo celular, comprometiendo a la célula a completar la división celular (Burke *et al.*, 2012). La liberación de E2F permite la activación de los genes involucrados en la progresión del ciclo celular permitiendo la entrada a la fase S del ciclo celular. Al final de la fase M, pRb se desfosforila, se une nuevamente a E2F y recupera su actividad represora.

Motivos lineales de unión a la proteína retinoblastoma

Las numerosas funciones de pRb son llevadas a cabo a través de la interacción con más de 100 blancos proteicos (Dick, 2007; Morris y Dyson, 2001). Si bien se conocen numerosas mutaciones en pRb que llevan al desarrollo de cáncer, como la mutación no sinónima H549Y (Figura 1.26, estrella amarilla) (Liu *et al.*, 1995; Valverde *et al.*, 2005), el conocimiento de los mecanismos moleculares y especificidad de las interacciones de pRb es limitado, así como los sitios de contacto específicos. En muchos casos la interacción está mediada por más de un dominio de pRb (Chemes *et al.*, 2010; Dyson *et al.*, 1992a; Magnaghi-Jaulin *et al.*, 1998; Stokes *et al.*, 2007). Por ejemplo, la proteína E2F interactúa a través del dominio de transactivación (E2F-TA) con el dominio RbAB y mediante una región cercana al dominio de unión al ADN con el dominio RbC.

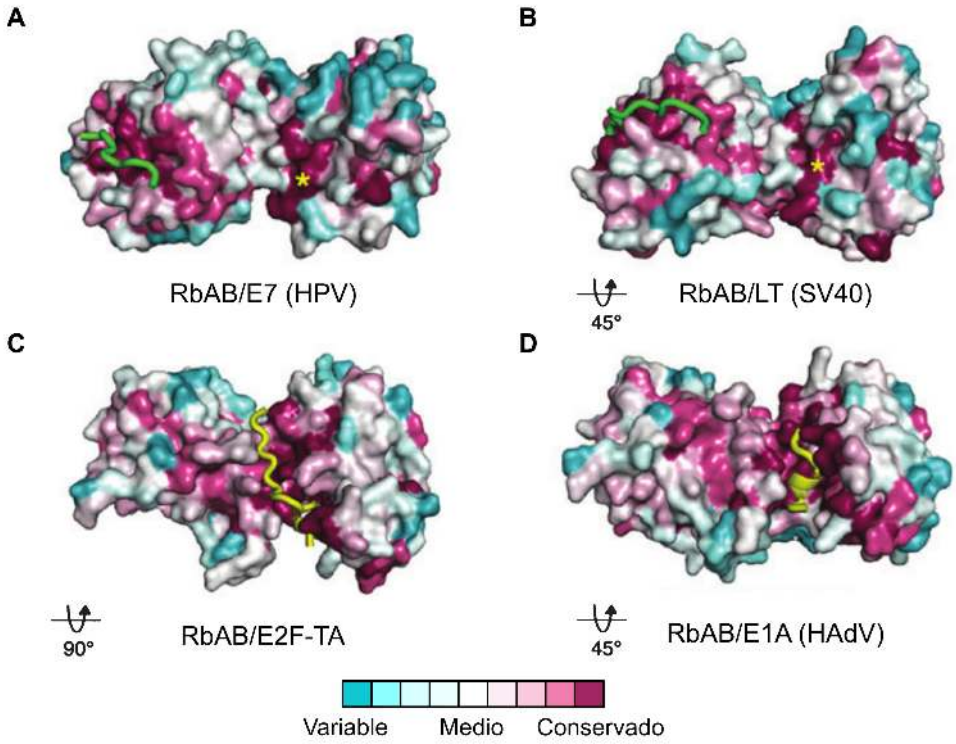


Figura 1.26: Conservación de las superficies de interacción en la Proteína retinoblastoma. A partir de un alineamiento del dominio RbAB de 46 especies de vertebrados diferentes se calcularon los índices de conservación utilizando Consurf (Ashkenazy *et al.*, 2016). El grado de conservación se muestra en una escala 9 colores (inferior), siendo celeste el más variable y rosa fuerte el más conservado. Las estructuras corresponden al dominio RbAB en complejo con las siguientes proteínas: (A) RbAB-E7 (PDB ID: 1GUX) (Lee *et al.*, 1998), (B) RbAB-LT (PDB ID: 1GH6) (Kim *et al.*, 2001), (C) RbAB-E2F-TA (PDB ID: 1N4M) (Lee *et al.*, 2002) y (D) RbAB-E1A (PDB ID: 2R7G) (Liu y Marmorstein, 2007). El asterisco amarillo marca la posición de la mutación H549Y. Las flechas indican la rotación de la molécula en el eje x entre dos imágenes consecutivas. Figura adaptada de (Chemes *et al.*, 2010).

Los motivos lineales poseen un rol principal en los procesos de señalización celular como mediadores de interacciones proteína-proteína, señalización intracelular, modificaciones post- traduccionales y degradación (Tompa *et al.*, 2014). pRb no es una excepción. Muchos motivos lineales presentes en pRb regulan su función, incluyendo sitios de fosforilación por quinasas dependientes

de ciclinas (Lees *et al.*, 1991) y otras quinasas (Delston *et al.*, 2011; Inoue *et al.*, 2007), sitios de desfosforilación por la fosfatasa 1 (Hirschi *et al.*, 2010), sitios de unión a proteínas reguladoras de la respuesta al daño en el ADN (Carr *et al.*, 2014), la señal de localización nuclear (Fontes *et al.*, 2003) y la señal de degradación (Tedesco *et al.*, 2002), entre otros.

Dos motivos lineales, el motivo LxCxE y el motivo pRb_ABGroove, median la interacción de pRb y numerosos blancos proteicos, uniéndose a dos sitios distintos de interacción altamente conservados (Figura 1.26) dentro del dominio RbAB (Chemes *et al.*, 2010; Lee *et al.*, 1998). Las proteínas que contienen el motivo lineal LxCxE interactúan con un sitio ubicado en el subdominio B del dominio RbAB (Figura 1.26 A y B) (Kim *et al.*, 2001; Lee *et al.*, 1998), mientras que las proteínas que contienen el motivo pRb_ABGroove interactúan con un sitio ubicado en el surco entre los subdominios A y B (Figura 1.26 C y D) (Lee *et al.*, 2002; Liu y Marmorstein, 2007).

Hasta la fecha, existen 18 proteínas celulares de mamíferos y plantas reportadas en la base de datos ELMdb con instancias del motivo LxCxE (Tabla 1.2) y evidencia bioquímica suficiente para demostrar su interacción con pRb. Estas proteínas están involucradas en procesos como regulación de la cromatina y modificación de histonas.

La proteína pRb hipofosforilada se une a E2F y recluta histonas desacetilasas y metiltransferasas que reprimen la expresión de los genes controlados por E2F.

Proteína	**Función**	**Nombre en Uniprot**	**Organismo**
KDM5A	Histonas desacetilasa	KDM5A_HUMAN	*Homo sapiens*
ARI4A	Histonas desacetilasa	ARI4A_HUMAN	*Homo sapiens*
HDAC1	Histonas desacetilasa	HDAC1_HUMAN	*Homo sapiens*
HDAC2	Histonas desacetilasa	HDAC2_HUMAN	*Homo sapiens*
NDC80	Proteína del cinetocoro	NDC80_HUMAN	*Homo sapiens*
HBP1	Regulador transcripcional	HBP1_RAT	*Rattus novergicus*
EID1	Regulador transcripcional	EID1_HUMAN	*Homo sapiens*
PRDM2	Histona metiltransferasa	PRDM2_HUMAN	*Homo sapiens*
PPR26	Fosfastasa	PPR26_HUMAN	*Homo sapiens*
BRM/SMCA2	Activador de la transcripción	SMCA2_HUMAN	*Homo sapiens*
BRG1/SMCA4	Activador de la transcripción	SMCA4_HUMAN	*Homo sapiens*
Brg1	Activador de la transcripción	Q63928_9MURI	*Mus musculus*
G1/S Ciclina D1	Ciclina	CCND1_HUMAN	*Homo sapiens*
G1/S Ciclina D2	Ciclina	CCND2_HUMAN	*Homo sapiens*
G1/S Ciclina D3	Ciclina	CCND3_HUMAN	*Homo sapiens*
Ciclina D1-1	Ciclina	CCD11_ARATH	*Arabidopsis thaliana*
Ciclina D2-1	Ciclina	CCD21_ARATH	*Arabidopsis thaliana*
Ciclina D3-1	Ciclina	CCD31_ARATH	*Arabidopsis thaliana*

Tabla 1.2: Blancos celulares con el motivo LxCxE de unión a retinoblastoma. Tabla adaptada de Palopoli *et al.* (2018).

El motivo pRb_ABGroove fue redefinido más recientemente (Chemes *et al.*, 2012b) como producto de esta tesis (véase Sección 3.2). Este motivo consiste en un hélice anfipática corta que se une en la interfaz entre los subdominios A y B del dominio RbAB llamado bolsillo AB. Existen

únicamente seis instancias celulares reportadas (Tabla 1.3). Dentro de esas instancias se encuentra el conector RbPL de pRb.

Proteína	Función	Nombre en Uniprot	Organismo
EF1	Factor de transcripción	E2F1_HUMAN	*Homo sapiens*
EF2	Factor de transcripción	E2F2_HUMAN	*Homo sapiens*
EF3	Factor de transcripción	E2F3_HUMAN	*Homo sapiens*
EF4	Factor de transcripción	E2F4_HUMAN	*Homo sapiens*
EF5	Factor de transcripción	E2F5_HUMAN	*Homo sapiens*
pRb	Factor de transcripción	RB_HUMAN	*Homo sapiens*

Tabla 1.3: Blancos celulares con el motivo pRb_AbGroove de unión a retinoblastoma. Tabla adaptada de Palopoli *et al.* (2018).

El bolsillo AB donde se une E2F-TA participa del mecanismo de inactivación de pRb mediado por los complejos Ciclina/CDK. El conector flexible RbPL que une ambos subdominios posee un sitio de fosforilación (S608) que es reconocido por la quinasa dependiente de ciclinas 4 (CDK4) (Inoue *et al.*, 2007). Luego de la fosforilación el conector adquiere la estructura de hélice anfipática que mimetiza el motivo pRb_ABGroove y desplaza la interacción con E2F (Burke *et al.*, 2012). Este desplazamiento permite la liberación de E2F y la activación de la transcripción de los genes dependientes de E2F.

1.5.2. Modulación viral de la actividad de la proteína retinoblastoma

Muchas de las funciones de pRb fueron inicialmente dilucidadas por la interacción con tres proteínas virales: el antígeno Large T de simian poliomavirus (SV40) (DeCaprio *et al.*, 1988), y las proteínas E1A de adenovirus (Whyte *et al.*, 1989) y E7 de papilomavirus (Dyson *et al.*, 1989), ambas objeto de estudio de esta tesis. Actualmente se sabe que numerosos virus poseen proteínas capaces de intervenir en la red de interacciones de pRb, interactuando directamente con la proteína (Palopoli *et al.*, 2018) o en algún paso anterior en la regulación del funcionamiento de pRb. Algunas de estas proteínas y su forma de interacción se resumen en la Figura 1.27.

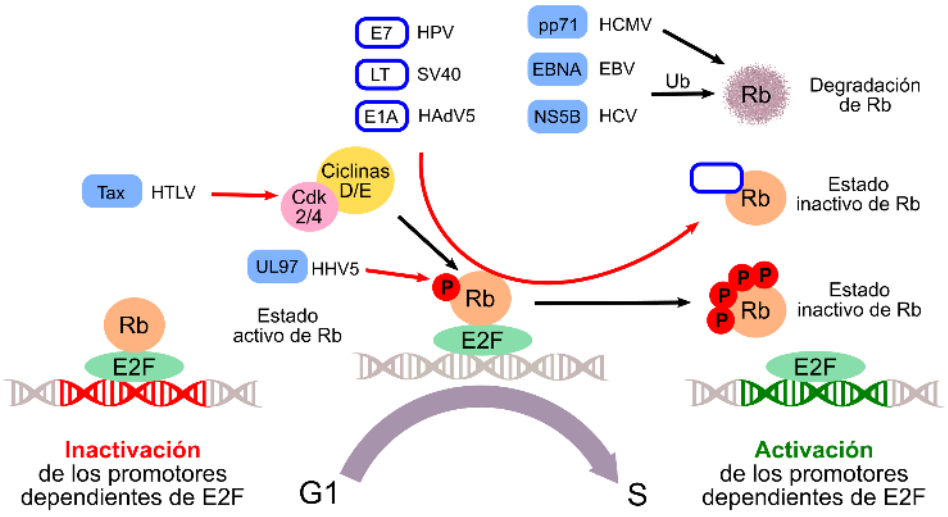


Figura 1.27: Intervención viral en la regulación de la progresión de la fase G1 a S por el complejo entre la proteína retinoblastoma y el factor de transcripción E2F. Figura adaptada de `https://viralzone.expasy.org/`.

Las proteínas virales E7 de papilomavirus, E1A de adenovirus y el antígeno Large T del poliomavirus SV40 además poseen el motivo pRb_ABGroove e interactúan directamente con pRb hipofosforilada disociando el complejo con E2F. E7 además induce la degradación de pRb. La proteína UL97 del citomegalovirus humano tiene actividades similares a los complejos formados por ciclinas y las quinasas dependientes de ciclinas. Interactúa con pRb mediante un motivo LxCxE (Prichard *et al.*, 2008) y la fosforila e inactiva pRb estimulando la progresión del ciclo celular (Hume *et al.*, 2008). Por otro lado, la proteína Tax del virus humano de la leucemia de células T tipo I (HTLV-I) induce la expresión de un grupo de genes celulares que codifican para proteínas involucradas en el control del ciclo celular, como ser las ciclinas D y E y las quinasas dependientes de ciclinas CDK2 y CDK4, que llevan a la fosforilación y consecuente inactivación de pRb (Iwanaga *et al.*, 2001). Además de la hiperfosforilación o desplazamiento de la interacción con E2F, otras proteínas virales inducen la degradación proteosomal de pRb. Por ejemplo, la proteína EBNA3C del virus Epstein Barr (Virus del herpes humano 4) (Knight *et al.*, 2005), la proteína NS5B del virus de la hepatitis C (Munakata *et al.*, 2007, 2005) y la proteína E7 de papilomavirus (Boyer *et al.*, 1996; Huh *et al.*, 2007) inducen la degradación de pRb dependiente de la ubiquitinación mediante el complejo SCF-ubiquitin ligasa, la ubiquitin ligasa E6AP y el complejo ubiquitin-ligasa culin 2, respectivamente. Otra proteína que induce la degradación de pRb es la proteína pp71 de citomegalovirus (Kalejta y Shenk, 2003) pero pareciera ser independiente de ubiquitinación. Actualmente hay reportadas 13 proteínas virales (Blancos virales que poseen el motivo LxCxE de unión a retinoblastoma) que poseen el motivo LxCxE de unión a retinoblastoma (Palopoli *et al.*, 2018) y que actúan modulando la actividad de pRb.

Proteína	Nombre en Uniprot	Virus	Familia	Genoma
Antígeno Large T	LT_SV40	Simian Virus 40	*Polyomaviridae*	ADNdc
Antígeno Large T	B8ZX42_9POLY	Merkel cell polyomavirus	*Polyomaviridae*	ADNdc
E7	VE7_HPV16	Human papillomavirus 16	*Papillomaviridae*	ADNdc
E1A	E1A_ADE05	Human adenovirus 5	*Adenoviridae*	ADNdc
UL97	GCVK_HCMVA	Human cytomegalovirus (HHV-5)	*Herpesviridae*	ADNdc
Wsv069	Q77J89_WSSVS	Shrimp white spot syndrome Virus	*Nimaviridae*	ADNdc
Wsv056	Q77J94_WSSVS	Shrimp white spot syndrome Virus	*Nimaviridae*	ADNdc
MC007	Q98178_MCV1	Molluscum contagiosum virus	*Poxviridae*	ADNdc
RepA	REPA_BEYDV	Bean yellow dwarf virus	*Geminiviridae*	ADNsc
RepA	REPA_WDVS	Wheat dwarf virus	*Geminiviridae*	ADNsc
RepA	REPA_MSVS	Maize streak virus	*Geminiviridae*	ADNsc
Clink	CLINK_FBNY1	Faba bean necrotic virus	*Nanoviridae*	ADNsc
UNK protein	Q9WKM8_BBTV	Banana bunchy top virus	*Nanoviridae*	ADNsc

Tabla 1.4: Blancos virales con el motivo LxCxE de unión a retinoblastoma. Tabla adaptada de Palopoli *et al.* (2018). En la última columna se indica si el genoma viral es ADN doble cadena (ADNdc) o simple cadena (ADNsc).

En resumen, algunas proteínas virales inducen la hiperfosforilación de pRb induciendo la expresión de por ejemplo, las ciclinas o interactuando con pRb, mientras que otras inducen su degradación dependiente o independiente de ubiquitina y otras compiten por E2F, desplazando la interacción.

Interacción entre E7 y la proteína retinoblastoma

La interacción de la proteína E7 de HPV16 con pRb está mediada por el motivo pRb_ABGroove de la región CR1 y el motivo LxCxE de la región CR2, que interactúan con el dominio RbAB, y por la región CR3, que interactúa con el dominio RbC y con el dominio RbAB. La interacción del motivo LxCxE y la región CR3 son suficientes para el desplazamiento de E2F (Huang *et al.*, 1993). Según las determinaciones realizadas por titulación de anisotropía de fluorescencia la afinidad entre el motivo LxCxE y pRb, residuos 16-31 de E7-HPV16, es alta (5.1 ± 1.3 nm) (Chemes *et al.*, 2010), mientras que la constante de interacción del dominio E7C aislado con el dominio RbAB está en el orden micromolar (2700 ± 600nm) (Chemes *et al.*, 2010; Liu *et al.*, 2006). Por otro lado, la fosforilación de Ser31 y Ser32 aumenta la afinidad por el dominio AB de pRb (1.8 ± 0.4nm) (Chemes *et al.*, 2010). La región CR2 de E7 es intrínsecamente desordenada y presenta elementos de estructura secundaria de poliprolina tipo II (PII). El aumento en la afinidad probablemente esté relacionado con el cambio conformacional inducido por la fosforilación, que aumenta el contenido de PII estabilizando una conformación extendida que optimiza la interacción (Chemes *et al.*, 2010). Por último, la región CR1 de E7 que contiene el motivo pRb_ABGroove interactúa de manera independiente con pRb *in vitro*, probablemente porque en el monómero la longitud del conector entre ambos motivos es demasiado corta para permitir una interacción simultánea (Chemes *et al.*, 2010). Sin embargo, en un dímero de E7 probablemente el motivo LxCxE de la región CR2 de una

molécula de E7 interactúe con el surco LxCxE de pRb, mientras que el motivo pRb_ABGroove de la región CR1 de la segunda molécula de E7 se una al mismo sitio de E2F en el bolsillo AB de pRb.

Interacción entre E1A y la proteína retinoblastoma

De manera similar a la proteína E7 de papilomavirus, la interacción de E1A con pRb está mediada por el motivo pRb_ABGroove de la región CR1 y el motivo LxCxE de la región CR2, que interactúan con distintas superficies del dominio RbAB (Figura 1.21). Un estudio de titulación por calorimetría isotérmica abarcando el dominio CR1, que incluye el motivo pRb_ABGroove, reveló que la constante de afinidad por el dominio pRb era aproximadamente 1μM. Ensayos de competencia por la interacción con pRb revelaron que el desplazamiento de E2F era más efectivo en presencia del dominio CR2, que incluye el motivo lineal LxCxE (Liu y Marmorstein, 2007). Así, ambos motivos actuarían de manera cooperativa logrando una interacción de alta afinidad con pRb. Finalmente, estudios semi-cuantitativos combinados con mutagénesis revelaron que la fosforilación de la Ser132 por la quinasa de caseína II aumenta la afinidad por el dominio AB de pRb, al igual que en el caso de la proteína E7 de papilomavirus (Whalen *et al.*, 1996). Los aminoácidos L43, L46 y L47 de E1A forman parte de una hélice anfipática que se une a una superficie hidrofóbica en pRb (Lee *et al.*, 2002) y son críticos para la interacción con pRb y el desplazamiento de E2F (Liu y Marmorstein, 2007). Adicionalmente, el residuo ácido adyacente al motivo D46 y el residuo H44, en una posición variable de la expresión regular pero altamente conservado en E1A, forman puentes de hidrógeno con pRb (Liu y Marmorstein, 2007; Palopoli *et al.*, 2018).

Tanto la proteína E1A como E7 compiten por la interacción de pRb con E2F induciendo la entrada en la fase S del ciclo celular, sin embargo, existen efectos específicos. Por ejemplo, la proteína E1A induce la acetilación y fosforilación de pRb mientras que E7 induce la degradación de pRb (Chemes *et al.*, 2015; Felsani *et al.*, 2006; Wang *et al.*, 1991).

1.6. Fundamentación, hipótesis y objetivos

En la introducción se estableció que el estudio de las proteínas desordenadas es un campo novedoso. A diferencia de las proteínas globulares, existen muchas características de las proteínas desordenadas que aún no son del todo claras. Al mismo tiempo, se introdujo los elementos de interacción denominados motivos lineales, que también se comenzaron a estudiar en los últimos años. Por lo tanto, la definición de un motivo lineal en particular en el tiempo no es estática, y cambia rápidamente a medida que nuevos experimentos son llevados a cabo. Por último, se introdujeron las características de la proteína pRb como ejemplo de funcionamiento de los motivos lineales y su participación en el ciclo celular en relación a nuestros dos objetos de estudio, las proteínas virales E7 y E1A.

En la presente tesis se propone estudiar las propiedades de conservación de secuencia de regiones desordenadas y las características evolutivas de los motivos lineales que se encuentran inmersos en dichas secuencias.

En el estudio experimental de los motivos lineales es común considerarlos como unidades independientes aunque, como ya se mostró en la introducción, pueden actuar de manera cooperativa. En la literatura disponible se propone, aunque con poca evidencia que lo respalde, que los motivos lineales pueden evolucionar por convergencia, que la ganancia o pérdida funcional ocurre fácilmente por mutaciones puntuales y que esto los hace excelentes candidatos para poseer un rol adaptativo en la evolución. Esto sugiere fuertemente que una forma posible de estudiar la evolución de los motivos lineales es en el marco de la filogenia de los organismos involucrados. Parece aconsejable elegir un modelo sometido a presiones de selección fuertes y para el que se puedan observar cambios en rasgos fenotípicos. Los virus son un excelente modelo en este caso porque el cambio de hospedador o tropismo a lo largo de la filogenia constituye un cambio definido en el fenotipo viral. Además, tanto el sistema inmune del hospedador como los cambios de hospedador o tropismo implican una presión de selección. Por lo tanto, la hipótesis principal de este trabajo de tesis es que es posible explicar la variabilidad fenotípica de los virus a partir del estudio del repertorio de los motivos lineales en las proteínas virales, lo cual permitirá mejorar el conocimiento sobre los mecanismos evolutivos subyacentes. Como objeto de estudio se eligieron las proteínas E7 de papilomavirus y E1A de adenovirus. La elección se basó en que estas proteínas poseen dominios intrínsecamente desordenados y numerosos motivos lineales que interrumpen el ciclo celular. Además, estas dos proteínas pertenecen a los virus papilomavirus y adenovirus de alta importancia clínica. En consecuencia existe una gran disponibilidad de información de secuencia y clínica, definiendo a ambas proteínas como excelentes objetos de estudio.

Los estudios aquí presentados contribuyen al avance del conocimiento acerca de las propiedades de conservación de secuencia de proteínas desordenadas y de conservación de secuencia, coocurrencia y evolución de motivos lineales, asentando las bases de una nueva forma de estudio.

1.6.1. Objetivo de la tesis

Objetivo general

El objetivo general de esta tesis fue el estudio de la evolución de las secuencias y los motivos lineales de las proteínas E7 y E1A.

Objetivos específicos

- **Objetivo 1. Creación de una base de datos representativa de la variabilidad de secuencia y clínica para las proteínas E7 y E1A.**

  Este objetivo incluyó, en primer lugar la recolección de secuencias para cada proteína, de manera que sean representativas de la familia o el género correspondiente. En segundo lugar, se realizó una búsqueda sistemática en la literatura para recolectar los datos fenotípicos necesarios.

- **Objetivo 2. Estudio comparativo de conservación de secuencia y desorden entre los dominios desordenados y ordenados de ambas proteínas.**

  Para llevar a cabo este objetivo fue necesaria la creación y curación manual de alineamientos de secuencia, realizar una redefinición de dominios para cada proteína y establecer numeraciones de secuencia para usar como coordenadas. Por último, fue necesario establecer el método para medir conservación y elegir un algoritmo de predicción de desorden.

- **Objetivo 3. Estudio de la variabilidad de motivos a nivel de secuencia.**

  Se propuso estudiar los motivos presentes en las proteínas E7 y E1A prototípicas, por lo cual fue necesario revisar las definiciones de las expresiones regulares utilizadas en la literatura. Esto implicó descartar algunos de los motivos previamente reportados, redefinir las expresiones regulares o definir expresiones *de novo* en base a la evidencia experimental existente. Por último, en base a dichas definiciones se realizó una búsqueda en texto de emparejamiento de patrones para ver la distribución del motivo en el resto de las secuencias.

- **Objetivo 4. Estudio de la co-ocurrencia de motivos a nivel de secuencia.**

  Dado que algunos motivos presentan un acoplamiento funcional, es esperable que ambos motivos estén sujetos a la misma presión de selección y sean seleccionados positiva o negativamente de manera conjunta. Esto se vería reflejado en una asociación a nivel de secuencia. Por lo tanto, se estudió estadísticamente la asociación entre los motivos.

- **Objetivo 5. Estudio de la correlación motivo fenotipo.**

  Utilizando los datos fenotípicos recolectados en la base de datos, se realizaron estudios estadísticos para determinar la asociación de la presencia de los motivos lineales con determinados rasgos fenotípicos, como tropismo y hospedador.

- **Objetivo 6. Estudio de la evolución de motivos a lo largo de la filogenia.**

  El desarrollo de este objetivo implicó en primer lugar la creación de un árbol filogenético para el género *Mastadenovirus* que abarcara todos los serotipos virales que serían utilizados en el análisis. En segundo lugar, se analizó la reconstrucción de la historia evolutiva de los motivos lineales para evaluar la presencia o ausencia de motivos lineales de la proteína E1A en los ancestros de los serotipos actuales y poder definir la existencia o ausencia del motivo en cada una de las ramas del árbol.

- **Objetivo 7. Estudio de la correlación de eventos de aparición/desaparición entre motivos y entre eventos de aparición/desaparición de motivos y eventos evolutivos.**

  Para llevar acabo este objetivo se realizó un test estadístico para evaluar la asociación entre los eventos de aparición y desaparición de los distintos motivos en las ramas del árbol. En segundo lugar, fue necesario realizar un estudio de co-evolución virus-hospedador para evaluar cuáles eran los eventos evolutivos involucrados en la evolución viral y por último evaluar si existía una asociación significativa entre el evento evolutivo y la aparición o desaparición de los motivos.

Capítulo 2

Métodos

En este capítulo describo y discuto las herramientas y formatos utilizados en este trabajo de tesis para el análisis bioinformático de secuencias.

Las secciones se organizan según el tipo de dato de partida: secuencias y alineamientos, estructuras y filogenias. En cada sección, se describen las distintas fuentes de donde se obtuvieron los datos, cómo se almacenan y las herramientas bioinformáticas utilizadas para su procesamiento y representación, el funcionamiento de las herramientas y los análisis estadísticos realizados.

Los experimentos de simulación para la predicción de una estructura globular del dominio CR3 de la proteína E1A de adenovirus fueron realizados por el Dr. Ernesto A. Román, la filogenia de *Mastadenovirus* fue realizada por el Dr. Ricardo Rodríguez de la Vega, la reconstrucción de secuencias ancestrales de la proteína E1A de *Mastadenovirus* fue realizada por la Dra. Valeria A. Risso. En estos casos mi labor como autora de esta tesis fue proporcionar los datos de entrada, acompañar la ejecución del experimento, analizar los resultados y/o tomarlos como punto de partida para los experimentos descriptos en la tesis. Se incluyen aquí los métodos utilizados para facilitar la interpretación de los resultados.

2.1. Secuencias y alineamientos

2.1.1. Taxonomía

Los datos taxonómicos de los virus estudiados son acorde con los lineamientos del ICTV disponible en `https://talk.ictvonline.org/`. Cuando el serotipo en cuestión no había sido clasificado aún por el ICTV, los datos se obtuvieron y curaron de la bibliografía disponible. La taxonomía de la familia *Papillomaviridae* se consultó en junio 2011, una nueva consulta y actualización de la misma se realizó en abril 2014. La taxonomía de la familia *Adenoviridae* se consultó en diciembre 2012.

2.1.2. Recopilación de secuencias

Las secuencias de las proteínas analizadas y genomas utilizados en este trabajo de tesis se obtuvieron de la bases de datos públicas, como GenBank o Swiss-Prot, accediendo a través de el navegador de Centro Nacional de Información Biotecnológica (NCBI, *National Center for Biotechnology Information*), disponible en `http://www.ncbi.nlm.nih.gov`.

Las secuencias de la proteína E7 de la familia *Papillomaviridae* se recolectaron en junio 2011, una nueva consulta y actualización de la base de datos se realizó en abril 2017. La recolección de las secuencias de la proteína E1A de la familia *Adenoviridae* se realizó en diciembre 2012. En ambos casos, si para un serotipo viral no se encontró la secuencia de la proteína correspondiente en NCBI, se verificó si existía un genoma reportado para el serotipo y se buscó un marco de lectura abierto que codifique para la proteína. De igual manera, en agosto 2014 se recolectaron las secuencias de los genomas reportados para los serotipos virales de *Mastadenovirus* para la construcción de la filogenia. Las listas de secuencias incluidas en las cuatro bases de datos se pueden ver en el Apéndice A.

En todos los casos, la recolección se realizó con el objetivo de obtener un conjunto de secuencias que abarquen los géneros, especies y serotipos descriptos en la literatura disponible. De esta manera se obtuvo un conjunto de secuencias representativas de la familia viral. Debido al desarrollo de los métodos de secuenciación en los últimos años y la importancia clínica de los virus analizados existe una sobrerrepresentación en las bases de datos de los serotipos virales que infectan humanos y una mayor cantidad de secuencias de variantes de los serotipos de mayor relevancia clínica. Por lo tanto, para mantener una representación balanceada, se recolectó únicamente una secuencia por serotipo viral.

2.1.3. Generación y curación de alineamientos múltiples de secuencias

Generación de alineamientos múltiples de secuencias

Para poder comparar secuencias de proteínas es muy útil poder identificar las posiciones que son genéticamente equivalentes. Esto permite, por ejemplo, identificar posiciones ocupadas por aminoácidos iguales o que tienen propiedades en común entre las distintas proteínas. Una de las

herramientas más utilizadas para este fin son los alineamientos múltiples de secuencias. En un alineamiento, cada secuencia proteica es escrita en una línea y se hacen coincidir aquellas posiciones genéticamente equivalentes insertando sitios vacíos (o *gaps*) donde sea necesario denotados por un guión, " − ". La inserción de sitios vacíos se realiza con el objetivo de maximizar la similitud global entre pares de secuencias del alineamiento. De esta manera, los alineamientos permiten relacionar regiones similares entre distintas proteínas de una misma familia o regiones conservadas entre proteínas distantes.

La generación y validación de alineamientos de secuencias proteicas es aún un problema abierto. En la actualidad existen múltiples métodos computacionales relevantes para la generación de alineamientos. Algunos de los algoritmos más utilizados son Clustal (Thompson *et al.*, 1994), MUSCLE (Edgar, 2004) y T-Coffee (Notredame *et al.*, 2000). En la Figura 2.1 se muestran seis fragmentos de secuencias, no alineadas y alineadas, de la proteína E7 de seis serotipos del género *Alphapapillomavirus*. Los fragmentos corresponden a la región donde se encuentra el motivo LxCxE de unión a la pRb estudiado en este trabajo.

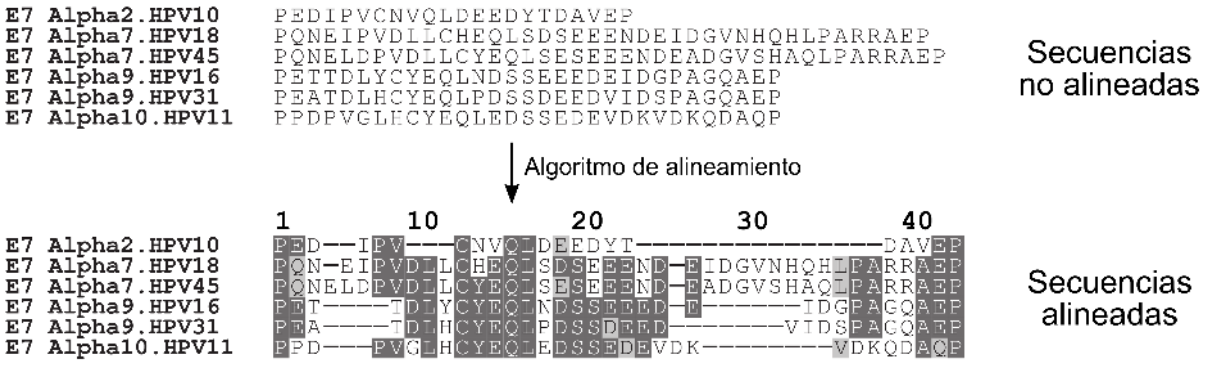


Figura 2.1: Alineamiento múltiple de secuencias de proteínas. *Arriba*, seis fragmentos de secuencias de la proteína E7 de papilomavirus antes de utilizar el algoritmo de alineamiento. *Abajo*, las mismas secuencias alineadas. Los guiones representan los sitios vacíos que se agregan para aumentar las coincidencias entre secuencias. Las posiciones de las secuencias alineadas están coloreadas resaltando los residuos más conservados (negro), similares (gris) y no similares (blanco).

Alineamiento sin sitios vacíos. Dadas dos secuencias, cuando se inserta un sitio vacío en una posición de una de las dos secuencias se puede decir que no es posible identificar que esa posición sea genéticamente equivalente para ambas secuencias. En consecuencia, en regiones con posiciones con alto contenido de sitios vacíos no se puede asegurar que sean genéticamente equivalentes. En estos casos, en lugar de utilizar el alineamiento completo, se utiliza lo que se conoce como un alineamiento sin sitios vacíos. Para construir este alineamiento, una vez obtenido el alineamiento curado, se calcula el porcentaje de sitios vacíos por posición y se eliminan aquellas columnas que poseen un porcentaje de sitios vacíos mayor a determinado umbral. En este trabajo, para los alineamientos de las secuencias recolectadas, eliminamos aquellas posiciones con más del 30 % de sitios vacíos.

Curación de alineamientos múltiples de secuencias

Los algoritmos desarrollados hasta el momento para la construcción de alineamientos se diseñaron, en gran parte, en base al conocimiento adquirido de características estructurales de proteínas globulares y se validaron en base a alineamientos estructurales (Thompson *et al.*, 2005).

La construcción de un alineamiento de proteínas desordenadas (véase Sección 1.1) es más compleja que en el caso de proteínas globulares. Los alineamientos de proteínas globulares suelen consistir en bloques conservados que se corresponden con elementos de estructura secundaria, unidos por secuencias conectoras poco conservadas pero relativamente cortas. Por otro lado, los alineamientos de muchas proteínas desordenadas consisten en bloques conservados de motivos lineales, más cortos que un típico elemento de estructura secundaria, unidos por secuencias conectoras poco conservadas y de longitud mayor que en proteínas globulares. Como consecuencia de estas características, los algoritmos de alineamiento más usados no son los más adecuados para proteínas desordenadas. Por lo tanto, es necesario realizar una curación manual de los alineamientos obtenidos en base a conocimientos experimentales previos. Esto implica una recopilación exhaustiva de literatura sobre las características de las proteínas, recolectando información relacionada con: motivos lineales definidos, posiciones iniciales y finales de dominios, relaciones filogenéticas, estructuras secundarias de determinadas regiones y observación de la presencia de sitios vacíos.

Una vez obtenida la información necesaria se procede a la curación del alineamiento. Esto implica reacomodar aquellas posiciones para las cuales se posee información relevante. En la Figura 2.2 se muestra la curación del alineamiento mostrado en la Figura 2.1.

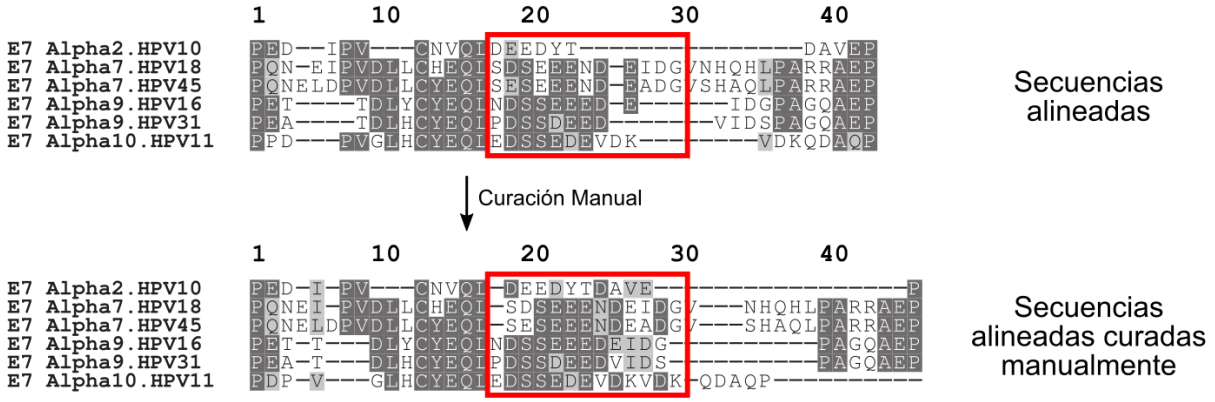


Figura 2.2: Curación de alineamiento múltiple de secuencias de proteínas. *Arriba.* Seis fragmentos de secuencias de la proteína E7 de papilomavirus alineadas con el algoritmo de alineamiento como se mostró en la Figura 2.1. *Abajo*. Las mismas secuencias que arriba alineadas luego de la curación manual. En ambos casos, la región que contiene al motivo de fosforilación de la quinasa de caseína II (CKII) y a la región acídica está resaltada en rojo. Los guiones representan los sitios vacíos agregados para aumentar las coincidencias entre secuencias. Las posiciones de las secuencias alineadas están coloreadas resaltando los residuos más conservados (negro), similares (gris) y no similares (blanco).

En la proteína E7 de HPV16 de papilomavirus se identificó la existencia de un motivo de fosforilación de la CKII, `[ST]..[DE]` seguido de una región rica en aminoácidos ácidos (`D` y `E`). Por lo tanto, las serinas correspondientes al motivo y los ácidos cercanos se alineron manualmente.

En este trabajo se utiliza el algoritmo MUSCLE (Edgar, 2004) para la generación de los alineamientos de las proteínas E7 y E1A. La descripción del formato y los alineamientos curados manualmente con y sin sitios vacíos utilizados en este trabajo de tesis se encuentran disponibles en el Apéndice B.

2.1.4. Identidad de pares de secuencias alineadas

En las siguientes secciones se describen diversos análisis, cualitativos y cuantitativos, que pueden realizarse sobre un alineamiento de secuencias. Por ejemplo, se puede usar el alineamiento para cuantificar cuán similares son entre sí las secuencias de dos proteínas. El porcentaje de identidad de secuencias (% Id) es el porcentaje de aminoácidos iguales en cada una de las posiciones del alineamiento para dos proteínas. Dado un alineamiento de largo L y n secuencias el porcentaje de identidad entre la secuencia *seq1*, de longitud L_1, y la secuencia *seq2*, de longitud L_2, se calcula como:

$$\%\ \text{Id}\ (seq1, seq2) = \sum_{j=1}^{L} \frac{i_j}{K} \times 100 \tag{2.1}$$

donde i es 1 si ambos aminoácidos en la posición j son iguales, 0 si ambos aminoácidos en la posición j son distintos o hay sitios vacíos en alguna de las dos secuencias proteicas, y K es el $max(L_1, L_2)$, es decir la longitud de la secuencia más larga. Las posiciones que son sitios vacíos en ambas secuencias no se incluyen.

A continuación se calcula el porcentaje de identidad entre los fragmentos de secuencia de E7 de Human papillomavirus 10 (HPV10) y Human papillomavirus 18 (HPV18) de longitud 22 y 40 respectivamente utilizando el alineamiento de la Figura 2.3.

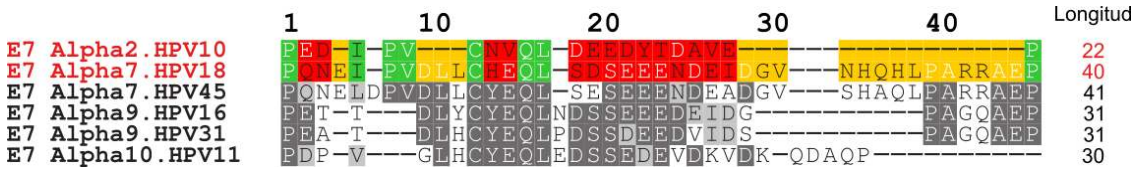


Figura 2.3: Cálculo de Identidad. Los nombres de las secuencias utilizadas como ejemplo se señalan en rojo. En verde se destacan los residuos idénticos, en rojo los residuos distintos y en naranja cuando alguna de las dos secuencias comparadas posee un sitio vacío. Los guiones representan los sitios vacíos que se agregan para aumentar las coincidencias entre secuencias. Las posiciones de las secuencias alineadas están coloreadas resaltando los residuos más conservados (negro), similares (gris) y no similares (blanco).

Como se indicó previamente, se suma un 1 por cada posición con aminoácidos iguales (posiciones verdes), 0 si los aminoácidos son distintos (posiciones rojas) o si hay un sitio vacío en alguna de las dos secuencias proteicas (posiciones naranjas). Las posiciones en donde ambas secuencias poseen sitios vacíos no son tenidas en cuenta. El total de aminoácidos idénticos es ocho. La longitud a considerar es la mayor, es decir, 40 y por lo tanto el porcentaje de identidad entre HPV10 y HPV18 es 20 %. De esta manera, el porcentaje de identidad permite establecer un umbral

a partir del cual definir si fragmentos o dominios de un par de proteínas son homólogos o no entre sí.

2.1.5. Expresiones regulares

Puede ser útil representar la variabilidad observada de aminoácidos en las distintas posiciones de un alineamiento de secuencia mediante expresiones regulares. En una expresión regular existen dos tipos de posiciones: Las posiciones fijas, determinantes de la funcionalidad del fragmento de la secuencia y las posiciones comodín, muy variables, que parecieran no determinar o participar en la funcionalidad de dicho fragmento. Existen numerosas reglas para el uso de expresiones regulares. En la Tabla 2.1 se listan las principales y las más utilizadas en este trabajo.

Símbolo	Definición
.	Cualquier aminoácido es permitido
[XY]	Solo los aminoácidos X e Y son permitidos
[^XY]	Los aminoácidos X e Y están prohibidos
min,max	Número mínimo y máximo de veces que se puede repetir una posición
^X	El aminoácido X se encuentra en el extremo N-terminal
X$	El aminoácido X se encuentra en el extremo C-terminal
(AB)|(CD)	Se encuentran, o bien, los aminoácidos AB, o bien, los aminoácidos CD

Tabla 2.1: Reglas para el uso de expresiones regulares.

Por ejemplo, para describir que en una posición fija hay un único aminoácido permitido, se coloca la letra del aminoácido correspondiente. Si en una posición hay dos o más aminoácidos permitidos se colocan las letras de los aminoácidos permitidos entre corchetes. Para describir las posiciones comodín donde cualquier aminoácido es permitido sin ningún tipo de restricción se utiliza un punto, “ . ”. Si en la posición comodín existe algún tipo de restricción, se indican las letras de los aminoácidos no permitidos entre corchetes colocando un acento circunflejo, “ ^ ”, antes de los aminoácidos. A continuación y a modo de ejercicio, se realiza una interpretación de dos expresiones regulares de motivos lineales.

Interpretación de la expresión regular del motivo pRb_ABGroove. La expresión regular del motivo pRb_ABGroove de unión a pRb es:

```
[IVLA].[NQDE][IVLFMYA][IVLFMYA][IVLA]{0,1}[AHKTNQDES]
```

y se interpreta de la siguiente manera, en la posición:

1. [IVLA]: Los aminoácidos I, V, L o A son permitidos, pero ningún otro.
2. . : Cualquiera de los 20 aminoácidos es permitido.
3. [NQDE]: Los aminoácidos N, Q, D o E son permitidos, pero ningún otro.

4. `[IVLFMYA]`: Los aminoácidos `I`, `V`, `L`, `F`, `M`, `Y` o `A` son permitidos, pero ningún otro.

5. `[IVLA]{0,1}`: Los aminoácidos `I`, `V`, `L` o `A` son permitidos, pero pueden estar ausentes (la longitud mínima es cero) o presentes una vez (la longitud máxima es uno).

6. `[AHKTNQDES]`: Los aminoácidos entre corchetes están permitidos.

Es importante resaltar de este ejemplo que el motivo puede tener una longitud variable de cinco o seis residuos, dependiendo de si la posición 5 está ausente o presente.

Interpretación de la expresión regular del motivo señal de localización nuclear. La expresión regular del motivo señal de localización nuclear es:

`[^DE]K[RK][KRP][KR][^DE]`

En este caso, lo importante a resaltar son la primera y la última posición del motivo, `[^DE]`, donde los aminoácidos `D` y `E` están prohibidos.

2.1.6. Identificación de motivos lineales y definición de expresiones regulares

La identificación de un motivo lineal y la definición de la expresión regular correspondiente son dos procesos que se realizan en conjunto. La primera parte consiste en evaluar la información relacionada a la región de una proteína que media una interacción proteína-proteína donde es posible que se encuentre el motivo lineal y proponerlo como tal. La segunda parte consiste en definir las posiciones responsables de la interacción y definir los residuos que pueden estar presentes en cada una de las posiciones para poder establecer una expresión regular que describa al motivo.

Identificación del motivo lineal

La identificación de un motivo lineal puede describirse como un proceso de cuatro etapas. La primer etapa consiste en la recopilación exhaustiva de literatura y la identificación de todas las proteínas que pueden poseer el motivo lineal de interés, partiendo de una posible instancia del motivo, es decir, la identificación de un posible motivo lineal en una proteína determinada. La información de partida es, por un lado, la proteína que posee el motivo y, por otro lado, el blanco proteico correspondiente. Para poder identificar nuevas instancias del motivo y recolectar la mayor evidencia experimental posible que describa la interacción, se realiza una búsqueda que capture la información relacionada con: (1) la familia de proteínas que pueden poseer el motivo y (2) la familia de posibles blancos proteicos. Además, la búsqueda debe orientarse a literatura relacionada con interacciones reportadas para dichas proteínas y debe incluir a los autores cuyo trabajo de investigación está relacionado con las proteínas de interés. Por último, la búsqueda puede ampliarse de dos maneras: (1) utilizando bases de datos de interacciones proteína-proteína, como por ejemplo IntAct (Orchard *et al.*, 2014), y (2) bases de datos de motivos lineales, como por ejemplo la ELMdb (Gouw *et al.*, 2018) (véase Sección 1.2.1). Esta etapa finaliza realizando una primera evaluación de la literatura obtenida. Se descarta la literatura no relacionada con interacciones proteína-proteína y

la relacionada a interacciones entre dominios globulares y se retiene la literatura que describe las interacciones entre dominios globulares y un posible motivo lineal. Como resultado se obtiene una lista de posibles instancias del motivo lineal y la literatura asociada.

La segunda etapa consiste en evaluar la evidencia experimental recolectada para cada una de las posibles instancias del motivo. El primer paso consiste en determinar si las interacciones reportadas poseen evidencia experimental que valide la interacción. Es decir, que se descartan aquellas proteínas como posible instancia para las cuales la evidencia experimental de interacción es únicamente indirecta por ejemplo, colocalización, experimentos de coprecipitación utilizando los extractos celulares, doble híbrido o genes reporteros, y se mantiene la literatura relacionada con evidencia experimental de interacción directa. Estos últimos incluyen experimentos de coprecipitación *in vitro*, medición de constantes de afinidad, ensayos *in vitro* de competencia, determinación de estructuras tridimensionales y ensayos de mutagénesis, entre otros. Dentro de los experimentos de mutagénesis, si la única evidencia experimental reportada es la pérdida de interacción debido a la deleción de la región que involucra el motivo, se descarta la proteína como posible instancia. Se consideran los experimentos de mutagénesis a nivel posición, como la mutación secuencial por alanina, ya que permiten determinar los residuos relevantes para la interacción. Sin embargo, el uso de mutaciones no conservativas o, por ejemplo, la mutación de un residuo con carga positiva por uno con carga negativa o viceversa, no proveen evidencia sólida de la relevancia del residuo y se consideran información secundaria. Un cambio radical en la región del motivo puede modificar otros factores que afecten la interacción. Por ejemplo, se puede inducir la modificación de la estructura secundaria, crear impedimentos estéricos en el contexto estructural o un contexto de cargas que impida la interacción. Por último, se evalúa si la evidencia experimental incluye ensayos funcionales. Por ejemplo, se evalúa si la mutación de un motivo de señalización celular cambia la localización de la proteína o si la mutación de un sitio de modificación post-traduccional impide la modificación correspondiente, teniendo en cuenta, cuando es posible si la funcionalidad es evaluada tanto *in vitro* como *in vivo*. Al final de esta etapa, se cuenta con la evidencia experimental correspondiente para cada posible instancia que define una interacción mediada por un posible motivo lineal y las posibles posiciones involucradas en la interacción.

En la tercera etapa se evalúa para cada posible instancia del motivo el contexto estructural y el grado de conservación. En general, se espera que un motivo lineal se encuentre en una región desordenada (véase Sección 1.2). Hay dos formas principales de evaluar el contexto estructural, se puede utilizar un predictor de desorden (véase Sección 1.1.4) o bien se puede realizar una búsqueda de estructuras relacionadas con la instancia para visualizar si la región de interés está en un dominio globular o no. Por otro lado, dado que un motivo lineal determina alguna de las funciones de la proteína que lo posee se espera, en general, que esté conservado en el conjunto de proteínas homólogas. Para esto es necesario para cada instancia del motivo crear un alineamiento de proteínas homólogas, por ejemplo, creado a partir de una búsqueda con BLAST o partiendo de un alineamiento propio. Luego, se evalúa la conservación de las posiciones que median la interacción, por ejemplo, cuantificando el contenido de información por posición (véase Sección 2.1.10)

(Schneider *et al.*, 1986).

Por último, se evalúa toda la evidencia experimental recolectada para cada posible instancia. Se define si la evidencia experimental, el contexto estructural predicho u observado y el grado de conservación son suficientes para considerar que existe una interacción mediada por un motivo lineal. De ser así, al final de esta etapa se cuenta con un número de instancias y la evidencia experimental relacionada que permite determinar las posiciones involucradas en la interacción, por ejemplo, los experimentos de mutagénesis y estructuras reportadas para cada una de las posibles instancias. Por lo tanto, se puede proceder a definir una expresión regular para el motivo lineal.

Definición de expresiones regulares

La definición de una expresión regular puede dividirse en tres etapas. La primera etapa consiste en la creación de un alineamiento de todas las instancias del motivo seleccionadas al final de la etapa anterior. En este alineamiento se incluye la región de la instancia que conforma el núcleo del motivo, es decir, la región que abarca las posiciones que participan en la interacción y además se incluyen las regiones flanqueantes para evaluar si existe alguna otra posición conservada entre las distintas instancias que podría estar involucrada en la interacción.

La segunda etapa consiste en evaluar la evidencia experimental para cada una de las posiciones de la región considerada. A nivel de secuencia, para cada posición es necesario considerar el grado de conservación y las posibles mutaciones conservativas que se pueden aceptar en base a la evidencia experimental y a lo observado en el alineamiento. A nivel estructura, si se cuenta con una estructura cristalográfica es necesario observar para cada posición los elementos de estructura secundaria si existen, si las posiciones que median la interacción se encuentran a una distancia compatible con un contacto, estudiar la movilidad del residuo según los factores de temperatura o la desviación de la raíz media cuadrática (RMSD) (en inglés, *Root Mean Square Deviation*), y el contexto estructural del residuo, incluyendo la accesibilidad o exposición al solvente y la relación entre las características químicas del residuo y el ambiente hidrofóbico o polar en el que se encuentra.

La tercera etapa consiste en definir en base a todo lo observado para las posiciones posiciones fijas y las posiciones comodín de la expresión regular las variaciones y restricciones aceptadas. En las posiciones fijas se incluyen las posiciones para las cuales existe evidencia experimental suficiente que indique su participación en la interacción. Para definir los aminoácidos que se permiten en una posición fija se utilizan los residuos presentes en las instancias reportadas. Luego, cada una de las posiciones puede degenerarse incluyendo otros residuos, teniendo en cuenta las características químicas del mismo, tamaño y carga, y la presencia en el conjunto de instancias conocidas y sus proteínas homólogas. Las posiciones comodín son aquellas que pertenecen a la región del motivo, pero que no tienen evidencia experimental que indique su participación en la interacción. A veces en esta posición puede aceptarse cualquier aminoácido o bien pueden existir restricciones en base al contexto estructural. Por ejemplo, si el motivo adquiere una estructura de hélice α al unirse a su blanco proteico, se indica que en las posiciones comodín no puede estar presente una

prolina.

Finalmente, se obtiene una expresión regular lo suficientemente amplia para describir un motivo lineal presente en distintas proteínas y que facilite la futura búsqueda en otras proteínas.

2.1.7. Búsqueda de motivos

Una vez construido el alineamiento de secuencias y definidas las expresiones regulares para cada uno de los motivos se realizó una búsqueda basada en texto del motivo lineal para cada una de las secuencias y motivos.

2.1.8. Teoría de la información molecular

En esta sección describimos los conceptos principales de la teoría de la información molecular, como introducción necesaria a las herramientas derivadas de ella que se describen en secciones posteriores.

Teoría de la información

La teoría de la información molecular es una adaptación a la biología de la teoría de la información desarrollada por Claude Shannon en el campo de las comunicaciones (Shannon, 1948). La teoría de Shannon abarca el estudio de la transmisión, procesamiento y almacenamiento de la información. La información implica la transmisión de un mensaje, por un canal, a partir de ciertos datos iniciales. El mensaje que se recibe es uno seleccionado a partir de un conjunto finito de mensajes posibles. El concepto principal de la teoría de la información es que la cantidad de información es cuantificable, independientemente del contenido del mensaje.

La cuantificación de la cantidad de información transmitida se hace símbolo a símbolo. Antes de la transmisión, se desconoce cuál es el símbolo que se va a transmitir. Por lo tanto, podría ser cualquiera de los símbolos incluidos en el alfabeto utilizado. Después de la transmisión, se sabe que el símbolo transmitido es uno solo de los incluidos en el alfabeto. Si se trata de un canal ruidoso, después de la transmisión se podría decir algo como *la probabilidad de que el símbolo sea una A es 85 %, una B 10 % y una C 5 %*. En cualquier caso, ha disminuido la incertidumbre del símbolo respecto del estado anterior a la transmisión.

Shannon pone esta idea en ecuaciones a través de su definición de la entropía. En el marco de esta teoría, la cantidad de información transmitida, R, es la diferencia entre la entropía antes, H_{antes}, y después, $H_{posterior}$, de la transmisión del símbolo.

$$R = H_{antes} - H_{posterior} \tag{2.2}$$

dado un número finito de estados posibles (un alfabeto finito), la entropía total H es el promedio

de la entropía de cada estado pesado por la probabilidad de ocurrencia del estado, es decir:

$$H = -\sum_{i=1}^{n} p_i \; log \; p_i \tag{2.3}$$

donde: p_i es la probabilidad del estado i y n es el número total de estados. La base del logaritmo define las unidades de la medición. Si la base es 2, la unidad se llama bits, si la base es 10, la unidad es digits y si por el contrario se utiliza el logaritmo natural la unidad sería nats o nits. La unidad comúnmente utilizada es bits.

Teoría de la información molecular

El objetivo de la teoría es describir cuantitativamente la interacción entre proteínas y ácidos nucleicos. Se explota la observación experimental de que muchas proteínas son específicas, es decir, unen de manera específica un subconjunto de todas las secuencias posibles de ADN de una cierta longitud. Estas secuencias de ADN de una longitud dada se analizan como mensajes. Dentro de estos mensajes, se distinguen nucleótidos, que se analizan de la misma manera que los símbolos individuales de la teoría original. En otras palabras, la formación del complejo proteína-ADN sería equivalente a una "transmisión" de un mensaje. Esta analogía con la teoría de la información de Shannon llevó al Dr. Thomas Schneider al desarrollo de la teoría de la información molecular (Schneider *et al.*, 1986).

Cuando una proteína globular interactúa con el ADN, se puede definir dos estados. Primero, un estado previo a la interacción en el que la proteína se encuentra libre y se desconoce la secuencia de nucleótidos con la que va a interactuar. Para cada posición de la secuencia reconocida, en el estado inicial (libre) el símbolo puede ser uno entre cuatro posibles: adenina (`A`), citosina (`C`), guanina (`G`) o timina (`T`). Segundo, un estado posterior en el que la proteína forma un complejo específico con el ADN. Esto ocurre solamente con un cierto subconjunto de secuencias de una cierta longitud. Por lo tanto, para cada posición de la secuencia reconocida, en el estado final (complejo proteína-ADN) el subconjunto de secuencias reconocidas nos permite calcular la probabilidad de observar adenina (`A`), citosina (`C`), guanina (`G`) o timina (`T`).

El análisis anterior permite cuantificar la información transmitida en la formación del complejo proteína-ADN de la siguiente manera. Dado un alineamiento de *N* secuencias de ADN, la frecuencia de cada base *b* en cada posición *i* es *f(b,i)*. El contenido de información (IC) para la posición *i*, aplicando la Ecuación 2.2 estará dado por:

$$IC(i) = H_{antes} - \Big(H_{posterior} + e(n_i)\Big) \tag{2.4}$$

$$IC(i) = log_2 \, 4 - \left(-\sum_{b=A}^{T} f(b,i) \; log_2 \; f(b,i)\right) - e(n_i) \tag{2.5}$$

donde $log_2 4$ es la entropía antes de la unión de la proteína globular al ADN, H_{antes}, y el término $\sum_{b=A}^{T} f(b,i)\ log_2\ f(b,i)$ es el promedio de las entropías de cada base b pesado por la probabilidad de cada base b en la posición i. El tercer término es una corrección por tamaño muestral chico donde n_i es el número de secuencias de la posición i. Este valor puede variar entre uno y el número de secuencias totales en el alineamiento, N. La corrección por tamaño muestral se calcula como:

$$e(n_i) = \frac{s-1}{2\ n_i\ ln\ 2} \tag{2.6}$$

donde s es el tamaño del alfabeto. Para el ADN el tamaño del alfabeto es 4 ya que existen cuatro estados posibles: adenina (A), citosina (C), guanina (G) y timina (T). En base a esta corrección, la incerteza de una posición aumenta si el número de secuencias es chico y, por lo tanto, el contenido de información disminuye.

En la Figura 2.4 (izquierda), se muestra un alineamiento ficticio de secuencias de ADN. Las secuencias utilizadas en este ejemplo se eligieron con el objetivo de resaltar dos casos que ayudan a comprender con mayor facilidad el concepto general. El primer caso consiste en entender como varía el contenido de información para un número constante de secuencias según como varía la variabilidad de secuencia. El segundo caso consiste en entender cómo afecta el número de secuencias al valor de contenido de información.

Posición	*n*	Fr. absoluta				Fr. relativa			
		A	C	G	T	A	C	G	T
1	8	8	0	0	0	1	0	0	0
2	8	2	0	6	0	0.25	0	0.75	0
3	8	4	4	0	0	0.5	0.5	0	0
4	8	2	2	2	2	0.25	0.25	0.25	0.25
5	1	1	0	0	0	1	0	0	0
6	2	1	1	0	0	0.5	0.5	0	0
7	4	2	2	0	0	0.5	0.5	0	0
8	4	1	1	1	1	0.25	0.25	0.25	0.25

Figura 2.4: Datos iniciales para el cálculo del contenido de información. *Izquierda.* Alineamiento ficticio de ADN en base al cuál se calcula el contenido de información. Los guiones representan los sitios vacíos. Las posiciones de las secuencias están coloreadas resaltando los residuos más conservados (gris oscuro), similares (gris claro) y no similares (blanco). *Derecha* Se muestran las frecuencias absolutas y relativas obtenidas a partir del alineamiento construido a la izquierda.

En la Figura 2.4 (derecha) se muestran los valores de las frecuencias absolutas y relativas para cada uno de los cuatro estados. Utilizando estas frecuencias y la Ecuación 2.5 se calcula el contenido de información. Los resultados se muestran en la Tabla 2.2.

Posición (i)	$-\ f(b,i)\ log_2\ f(b,i)$				$H_{posterior}$	$H_{antes} - H_{posterior}$	$e(n_i)$	$IC(i)$
	A	C	G	T				
1	0	0	0	0	0	2	0.27	1.73
2	0.50	0	0.31	0	0.81	1.19	0.27	0.92
3	0.50	0.50	0	0	1	1	0.27	0.73
4	0.50	0.50	0.50	0.50	2	0	0.27	0
5	0	0	0	0	0	2	2.16	0
6	0.50	0.50	0	0	1	1	1.08	0
7	0.50	0.50	0	0	1	1	0.54	0.46
8	0.50	0.50	0.50	0.50	2	0	0.54	0

Tabla 2.2: Cálculo del contenido de información. Con los datos provenientes de la Figura 2.4, considerando cuatro estados, ($s = 4$), y utilizando la Ecuación 2.5

En la Tabla 2.2 se indica la posición del alineamiento (Figura 2.4), la entropía de cada base ($-\ f(b,i)\ log_2\ f(b,i)$), la incerteza de la posición ($H_{posterior}$), el contenido de información sin la correción por el tamaño de la muestra ($H_{antes} - H_{posterior}$), la corrección por el tamaño de la muestra ($e(n_i)$) y el contenido de información con la corrección por el tamaño de la muestra ($IC(i)$).

Para entender como varía el contenido de información para un número constante de secuencias y su relación con la variabilidad de secuencia hay que a centrarse en las posiciones 1 a 4 del alineamiento (Figura 2.4). En la Figura 2.4 se puede observar para estas posiciones que el número de secuencias es constante y que únicamente varía la composición de bases. En la Tabla 2.2 se observa que a medida que aumenta la variabilidad de secuencia de una posición para un número constante de secuencias disminuye el contenido de información, ya que aumenta la incerteza de la posición ($H_{posterior}$), como ocurre para las posiciones 1 a 3, hasta hacerse cero cuando los cuatro estados tienen igual frecuencia (posición 4, Tabla 2.2). En este caso, el contenido de información corregido es menor a cero y se considera que su valor es cero.

Para entender la relación entre el número de secuencias y el valor de contenido de información es necesario observar todas las posiciones del alineamiento (Figura 2.4). Las posiciones 1 y 5 tienen igual composición de bases y difieren únicamente en el número de secuencias que disminuye de ocho a uno. El contenido de información sin corregir es igual en ambos casos (Tabla 2.2). La corrección por el tamaño de la muestra (Ecuación 2.6) aumenta al disminuir el número de secuencias y en consecuencia el contenido de información corregido disminuye (Tabla 2.2). De la misma manera se pueden comparar las posiciones 3, 6 y 7. Si bien las frecuencias relativas de las bases nucleotídicas son iguales, el contenido de información corregido por el tamaño de la muestra disminuye a medida que disminuye el número de secuencias. El contenido de información con y sin corrección no se ve afectado por el número de secuencias cuando los cuatro estados tienen igual frecuencia relativa (posiciones 4 y 8, Tabla 2.2).

En resumen, el contenido de información varía de manera intuitiva al cambiar tanto el número de observaciones como las frecuencias relativas de A, C, G y T, y adquiere un valor nulo cuando el

número de secuencias es muy bajo o cuando las frecuencias relativas de los distintos estados son iguales.

El razonamiento utilizado para la aplicación de la teoría de la información a secuencias de ADN puede aplicarse de igual forma a proteínas, considerando que el número de estados posibles, s, para una posición del alineamiento está dado por el tamaño del alfabeto de las proteínas definido por 20 aminoácidos. El contenido de información queda definido entonces como:

$$IC(i) = log_2\ 20 - \left(- \sum_a f(a,i)\ log_2\ f(a,i) \right) - e(n_i) \tag{2.7}$$

donde $f(a,i)$ es la frecuencia relativa del aminoácido a en la posición i. El valor máximo de $IC(i)$ es $\approx 4.32\ bits$ ($\log_2\ 20$) cuanto más chico es $e(n_i)$ (cuanto mayor es el número de secuencias), y el valor mínimo es cero.

2.1.9. Logos de secuencia

Las expresiones regulares brindan una información incompleta acerca de los aminoácidos observados en cada posición del alineamiento y sus secuencias. Por ejemplo, en una expresión regular es imposible definir para una posición fija con dos aminoácidos posibles cuál es la diferencia entre sus frecuencias observadas, y por lo tanto se asume que ambos aminoácidos son igualmente probables. Una forma alternativa para describir la variabilidad de secuencia observada en un alineamiento es utilizar los logos de secuencia (Schneider y Stephens, 1990), una herramienta basada en la teoría de la información molecular que supera algunas de las limitaciones de las expresiones regulares.

Un logo de secuencia es una representación gráfica de un alineamiento de secuencias. En la Figura 2.5 se muestra el logo de secuencia construido a modo de ejemplo a partir de la región correspondiente al motivo LxCxE de unión a la pRb del alineamiento de la Figura 2.1. En el eje x se indica la posición del alineamiento y en el eje y se indica el contenido de información medido en bits (véase Sección 2.1.8).

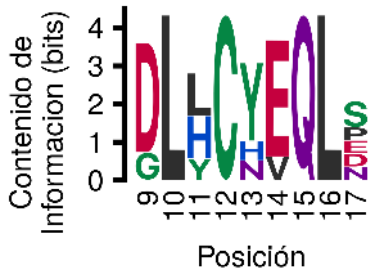


Figura 2.5: Logos de secuencia. En la parte superior, se muestra la secuencia consenso correspondiente al alineamiento de la Figura 2.1. La coloración de las letras destaca las características químicas de cada aminoácido.

Un logo de secuencia se construye colocando una columna de letras para cada posición. En cada columna se incluyen los aminoácidos presentes en el alineamiento en esa posición. La altura total

de la columna corresponde a la conservación de la posición medida como contenido de información por posición en bits calculado según la Ecuación 2.7 (véase Sección 2.1.8). La altura de cada una de las letras pertenecientes a la columna es proporcional a la frecuencia de cada letra en el alineamiento y está dada por:

$$h(a,i) = f(a,i)\ IC(i) \tag{2.8}$$

donde: $h(a,i)$ es la altura del aminoácido a en la posición i, $f(a,i)$ es la frecuencia del aminoácido a en la posición i e $IC(i)$ es el contenido de información de la posición i calculado según la Ecuación 2.7. En cada columna, las letras están ordenadas según la frecuencia, la cual disminuye desde la parte superior, donde se encuentra el aminoácido más frecuente, hacia la parte inferior donde se encuentra el aminoácido menos frecuente. La coloración de las letras puede ser elegida según la conveniencia. Por ejemplo en este caso se utilizó la coloración por características químicas de los aminoácidos.

Los logos de secuencia presentan una gran ventaja para visualizar un alineamiento de secuencias en contraposición al uso de expresiones regulares ya que permiten visualizar una mayor cantidad de propiedades de las secuencias, incluyendo el grado de conservación, frecuencias y características de los aminoácidos presentes. Por esta razón, son el método elegido para visualizar los alineamientos de las distintas regiones y dominios a lo largo de esta tesis. Los logos se generaron utilizando WebLogo (Crooks *et al.*, 2004) y los alineamientos sin sitios vacíos correspondientes están disponibles en el Apéndice B.

2.1.10. Conservación de secuencia

A lo largo de la historia evolutiva, una secuencia proteica puede sufrir mutaciones que den lugar a un cambio en la estabilidad de plegado, en la capacidad de interacción con otras moléculas o en la función de la proteína. Estas mutaciones pueden ser conservativas, cuando un aminoácido es reemplazado por uno de características fisicoquímicas similares de manera que no se produce un cambio drástico en fenotipo, o no conservativas, que son menos frecuentemente observadas en la naturaleza debido a los efectos negativos en la supervivencia del individuo. Esto da a lugar a que secuencias de proteínas relacionadas presenten cierto grado de variación a nivel de residuo. Los residuos que se reemplazan sin generar un cambio detectable en las características fisicoquímicas o funcionales de la proteína suelen estar poco conservados, y viceversa. Una vez construido un alineamiento de secuencias relacionadas, se puede medir la conservación de secuencia como contenido de información en bits utilizando la Ecuación 2.7:

$$IC(i) = log_2\ 20 - \left(- \sum_a f(a,i)\ log_2\ f(a,i) \right) - e(n_i) \tag{2.7 revisitada}$$

Continuando con la analogía utilizada antes (véase Sección 2.1.8), una posición conservada representaría un mensaje que está siendo transmitido. Este mensaje será recibido, por ejemplo,

por otra proteína en una interacción proteína-proteína o bien, por la misma proteína, por ejemplo, mediando un contacto intra-secuencia. Los sitios totalmente conservados tendrán el máximo contenido de información para un alfabeto de 20 símbolos, $IC(i) \approx 4.32$ bits, y disminuirá a medida que disminuya la conservación de secuencia.

Se midió la conservación de secuencia utilizando los alineamientos sin sitios vacíos (Apéndice B) de las distintas regiones y dominios de las proteínas E7 y E1A utilizando el contenido de información por posición.

2.1.11. Análisis estadístico de la conservación de aminoácidos

Los histogramas del contenido de información por posición se construyeron de la siguiente manera. Los valores posibles para el contenido de información con el alfabeto de aminoácidos varían entre cero y $\log_2 20$ bits (aproximadamente 4.32 bits). Se definieron intervalos de 0.5 bits entre 0 y 4.5 bits. Estos histogramas se utilizaron para clasificar grupos de posiciones provenientes de tres grupos de datos: (1) Dos grupos de posiciones de E7 de los alineamientos sin sitios vacíos correspondientes al dominio E7N y al dominio E7C, (2) ocho grupos de posiciones de la proteína E1A de los alineamientos sin sitios vacíos pertenecientes a cada uno de los dominios y regiones; y (3) cuatro grupos posiciones de la proteína E1A de los alineamientos sin sitios vacíos agrupadas en fijas, comodín, adyacentes y "otras" en relación la definición de la expresión regular del motivo lineal. Para cada subgrupo de posiciones se cuantificaron cuantos valores del contenido de información correspondían a cada intervalo. Estas frecuencias absolutas se convirtieron en frecuencias relativas dividiendo por el número total de posiciones considerado en cada subgrupo y se visualizó en forma de gráfico de barras.

Dentro de cada grupo de posiciones se evaluaron las diferencias en el promedio de la conservación de aminoácidos medido como el contenido de información promedio ($\overline{IC_i}$) entre los subgrupos. En primer lugar, se utilizó la prueba de Shapiro-Wilk (Shapiro y Wilk, 1965) para evaluar normalidad en los distintos conjuntos de datos (véase Sección E.2 y Sección F.3). Dado que no todos los conjuntos de datos seguían una distribución normal (valor $p \leq 0.05$), se calcularon los valores p para las diferencias en ($\overline{IC_i}$) utilizando la prueba de permutación (Good, 2006) que se detalla a continuación.

Prueba de permutación

La prueba de permutación es una prueba no paramétrica mediante la cual se construye la distribución de la muestra remuestrando los datos observados. A partir de esto, se puede calcular un estadístico que permita evaluar la hipótesis nula. Dadas dos muestras A y B, con m y n observaciones respectivamente, el estadístico se construye según:

$$\Theta = |\overline{X}_A - \overline{X}_B| \tag{2.9}$$

donde $\overline{X}_A$ y $\overline{X}_B$ son los valores promedio para cada muestra. La construcción de la distribución

por permutaciones se realiza de la siguiente manera:

1. Se combinan los valores observados para cada muestra y se construye un conjunto, C, de datos con $m + n$ valores.

2. Del conjunto C, se toman dos muestras A' y B', de m y n observaciones cada una, sin reemplazo y se calcula la diferencia entre los valores promedio de las dos muestras $\overline{X}_{A'}$ y $\overline{X}_{B'}$.

3. El punto anterior se repite 10000 veces.

Los valores p se calculan como la fracción de veces que el valor absoluto de la diferencia entre los valores promedios las muestras A' y B' es mayor o igual que el valor absoluto de la diferencia observada en el número total de permutaciones. A los valores p obtenidos se aplicó la correción de Benjamini-Hochberg para comparaciones múltiples (véase Sección 2.5.2).

Las hipótesis planteadas son:

$$H_o : \overline{IC_i} \textit{ del subgrupo } A = \overline{IC_i} \textit{ del subgrupo } B.$$
$$H_a : \overline{IC_i} \textit{ del subgrupo } A \neq \overline{IC_i} \textit{ del subgrupo } B.$$

Dos pares de valores de $\overline{IC_i}$ son diferentes, es decir se rechaza la hipótesis nula, si el valor p corregido es menor que el umbral de 0.05 elegido (véase Sección E.2 y Sección F.3).

2.2. Estructuras

2.2.1. Estructuras tridimensionales

Las estructuras tridimensionales de proteínas individuales y complejos proteicos reportadas en la literatura se recolectaron en formato `pdb` de la base de datos pública PDB disponible en `http://www.rcsb.org`. A lo largo de esta tesis, se indica el uso de las estructuras con su identificador (PDB ID) y la referencia bibliográfica correspondiente si posee. La descripción del formato y la lista de estructuras se encuentra disponible en el Apéndice C.

Visualización de estructuras tridimensionales. Para visualizar, trabajar y crear las figuras de las estructuras tridimensionales se utilizó el programa Pymol disponible en `pymol.org`. Pymol es un sistema de visualización de moléculas que interpreta las coordenadas informadas en el archivo `pdb` y construye un modelo tridimensional de la molécula. En la Figura 2.6 se muestra un fragmento de la estructura determinada para el dominio CR3 de la proteína E7 (PDB ID: 2F8B) (Ohlenschläger *et al.*, 2006).

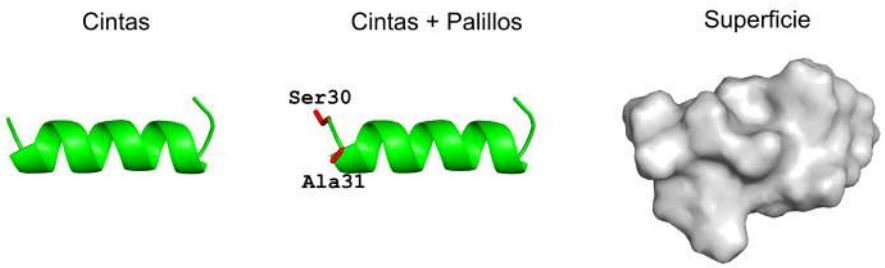


Figura 2.6: Visualización de Moléculas. Se muestra un fragmento de la estructura del dominio CR3 de la proteína E7 (PDB ID: 2F8B) (Ohlenschläger *et al.*, 2006). *Izquierda*. Se utilizó el modelo de visualización de cintas (en inglés, *cartoon*). *Centro*. Se combinó el modelo de cintas y palillos (en inglés, *sticks*). En palillos se muestran las cadenas laterales de los residuos serina y alanina en las posiciones 30 y 31 respectivamente mostrados en el archivo `pdb` de la Figura C.1. *Derecha*. Se muestra la representación de superficie (en inglés, *surface*).

Pymol permite elegir al usuario representaciones gráficas y colores. Por ejemplo, la representación de cintas (izquierda, Figura 2.6) permite visualizar los elementos de estructura secundaria, la representación de palillos permite ver la ubicación y orientación de los residuos (centro) y la representación de superficie permite visualizar el volumen ocupado por la molécula (derecha). De esta manera, uno puede elegir la representación gráfica de mayor conveniencia para resaltar la característica de interés.

2.2.2. Mapa de contactos

Los contactos entre residuos de la proteína relevantes para su estructura y función ocurren entre residuos cercanos en el espacio. Dadas las coordenadas x, y y z de los átomos de cada aminoácido, se puede calcular la distancia euclídea entre pares de átomos y definir un umbral de distancia a partir del cual se considera que los residuos están en contacto. En algunos casos, la distancia de contacto se toma entre entre los C_α o C_β del residuo o el centro de masa del mismo. Los valores de umbral utilizados en la literatura varían entre 5 Å y 10 Å. En este trabajo se indica el umbral utilizado de manera individual para cada caso. En el ejemplo de la Figura 2.7, se utilizó un umbral de 7 Å entre dos C_α para determinar si los residuos están en contacto.

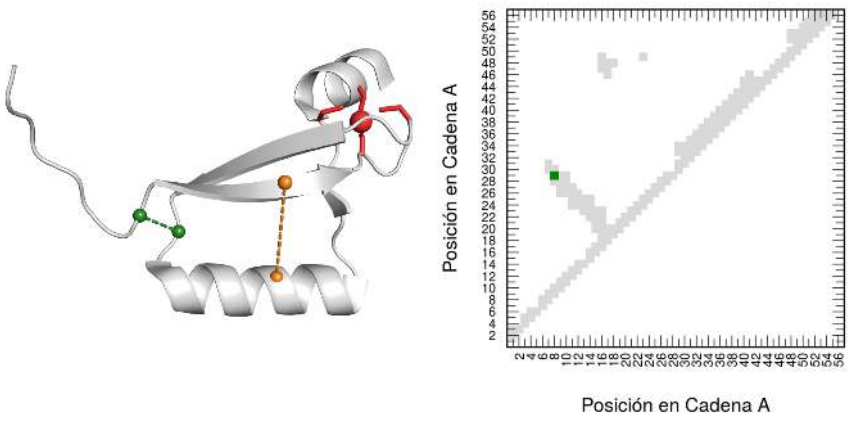


Figura 2.7: Mapa de contactos en proteínas. *Izquierda*. Estructura globular del dominio CR3 de la proteína E7 de papilomavirus (PDB: 2F8B) (Ohlenschläger *et al.*, 2006)). La cadena A está representada en forma de cinta en color gris. El átomo de zinc y las cisteínas que lo coordinan están representados en color rojo como esferas y palillos respectivamente. Sobre la cadena A, se muestran los C_α de dos residuos en contacto (verde, distancia: 4.7 Å), y los C_α de dos residuos no en contacto (naranja, distancia: 11.1 Å). Los C_α está representados en forma de esferas. *Derecha*. Mapa de contactos para la cadena A. El umbral utilizado es de 7 Å. En gris se indican los contactos dentro de la cadena A. En verde se muestra el contacto identificado en la estructura de la derecha. El par naranja no está presente ya que no está formando un contacto.

Un mapa de contactos (derecha, Figura 2.7) es una representación en dos dimensiones de la estructura tridimensional de una proteína. Dada la estructura tridimensional de una proteína de largo, L, se crea una matriz de L x L. Cada fila i, y cada columna j representan uno de los residuos de la proteína ordenados desde el extremo N-terminal al C-terminal. Cada posición (i, j) de la matriz vale 1 o 0 dependiendo si el residuo i y el residuo j, están o no en contacto, respectivamente. Resulta suficiente analizar únicamente la región superior de la diagonal, ya que es una matriz simétrica.

2.2.3. Predicción de desorden

En ausencia de una estructura tridimensional obtenida mediante experimentos, como suele ser el caso de las proteínas intrínsecamente desordenadas, podemos inferir diversas características estructurales de una secuencia mediante métodos computacionales. Por ejemplo, el algoritmo IUPred (Mészáros *et al.*, 2018) para la predicción de desorden intrínseco en proteínas da una predicción del grado de desorden residuo a residuo. Dicha predicción correlaciona de manera apreciable con parámetros de flexibilidad determinados mediante resonancia magnética nuclear (Daughdrill *et al.*, 2011).

IUPred se basa en dos conceptos centrales. Primero, que la estructura primaria de una proteína determina su estructura tridimensional nativa, que se corresponde con un mínimo energético global en el espacio conformacional. Segundo, que las proteínas globulares poseen un mayor número de residuos capaces de formar contactos energéticamente favorables que las proteínas desordenadas.

El algoritmo estima para cada aminoácido su capacidad de formar una estructura globular estabilizada por interacciones favorables con los aminoácidos próximos a él en una ventana de secuencia de 21 residuos. En ausencia de una estructura globular conocida, las energías de interacción se aproximan mediante una expresión cuadrática que depende de la composición aminoacídica de la ventana. Así, la energía calculada para un aminoácido depende de su entorno de secuencia de la siguiente manera:

$$\frac{E_{estimado}}{L} = \sum_{ij=1} n_i P_{ij} n_j \tag{2.10}$$

Donde n_i es la frecuencia del aminoácido del tipo i en la secuencia y se calcula como $\frac{N_i}{L}$ donde N_i es el número de aminoácidos del tipo i en la secuencia y L la longitud de secuencia. P es la matriz predictora de energía, formada por 20 filas, i, y 20 columnas, j, una por cada tipo de aminoácido. Cada una de las posiciones de la matriz, P_{ij}, indica la dependencia energética entre los aminoácidos del tipo i y j. Estos valores son calculados de forma que aproxima de manera óptima la energía de proteínas globulares de estructura conocida cuando dicha energía se calcula a partir de los contactos observados en la estructura.

En resumen, el algoritmo IUPred utiliza una secuencia de aminoácidos en formato `FASTA` como entrada y calcula el perfil de energía de contacto a lo largo de la secuencia utilizando la matriz de predicción de energía posición específica. Los valores obtenidos son transformados en una puntuación probabilística donde el mínimo, 0, indica completamente ordenada y el máximo, 1, indica completamente desordenada. Los residuos que tienen puntuaciones mayores a 0.5 pueden considerarse desordenados.

En la Figura 2.8 se muestra el resultado de la predicción de IUPred para la proteína E7 de HPV16 de papilomavirus. El dominio globular en la región C-terminal se predice ordenado (posición 51 en adelante), por el contrario la región N-terminal posee regiones desordenadas (en color rojo).

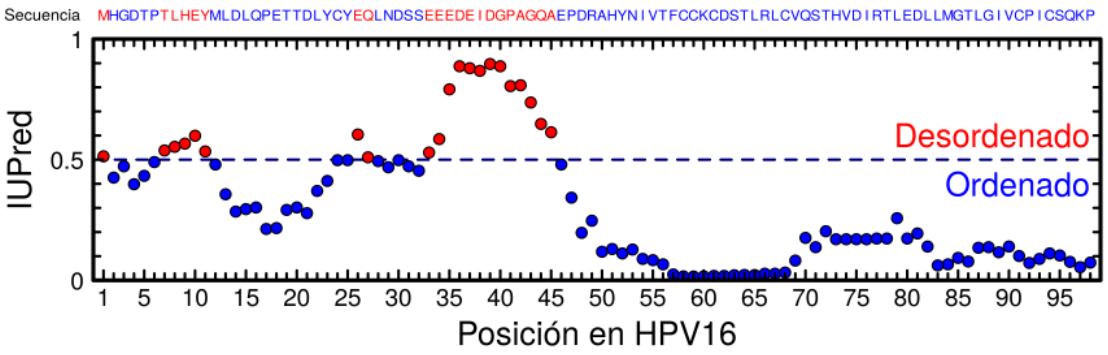


Figura 2.8: Predicción de desorden para la proteína E7 de HPV16 de papilomavirus utilizando IUPred. Se utilizó IUPred con la opción *long*. La secuencia de la proteína E7 se muestra en la parte superior. Las posiciones predichas como desordenadas (IUPred ≥ 0.5) se muestran en rojo y las ordenadas (IUPred < 0.5) en azul. La línea azul punteada indica el umbral de 0.5.

IUPred muestra una mejor capacidad predictiva que otros métodos, se basa en principios físicos transparentes y condice con la determinación de desorden por los métodos experimentales habituales. Es destacable que la predicción realizada por IUPred depende únicamente del tipo de aminoácido y el entorno de secuencia en el cual se encuentra, lo cual permite predecir proteínas intrínsecamente desordenadas basándose únicamente en su secuencia.

Las predicciones de desorden se realizaron utilizando las secuencias de las proteínas E7 y E1A de los alineamientos con sitios vacíos (véase Apéndice B) y el algoritmo IUPred.

2.2.4. Comparación del grado de desorden

Los valores de IUPred por posición fueron promediados para las posiciones correspondientes a los alineamientos sin sitios vacíos pertenecientes a dos grupos: (1) Dos subgrupos de posiciones de E7 de los alineamientos sin sitios vacíos correspondientes al dominio E7N y al dominio E7C y (2) ocho subgrupos de posiciones de la proteína E1A de los alineamientos sin sitios vacíos pertenecientes a cada uno de los dominios y regiones.

Dentro de cada grupo de posiciones se evaluó si los valores promedio de IUPred obtenidos para los dos grupos eran mayores o menores al umbral de 0.5 del predictor. En primer lugar, se utilizó la prueba de Shapiro-Wilk (Shapiro y Wilk, 1965) para evaluar normalidad en los distintos conjuntos de datos de IUPred. Para ambos dominios de la proteína E7, se rechazó la hipótesis nula que establecía que los datos provenían de una distribución normal (valor $p < 0.05$) (véase Sección E.3.1). Para los dominios y regiones de la proteína E1A, no hubo evidencia suficiente para rechazar la hipótesis nula que establecía que los datos provenían de una población que seguía una distribución normal (véase Sección F.4.1). Sin embargo, un gráfico Q-Q comparando los cuantiles de los datos en el eje vertical a una población normal en el eje horizontal sugirió que el conjunto de datos no seguían una distribución normal (véase Sección F.4.2). Por lo tanto, se utilizó el método de remuestreo (Good, 2006) para determinar si los valores promedio de IUPred diferían significativamente con el umbral de 0.5 del predictor.

Método de remuestreo

Se utilizó la técnica de remuestreo para generar intervalos de confianza del 99 % para la diferencia entre valores promedio y un valor fijo. Brevemente, se remuestrean los valores promedios por posición 10000 veces con reemplazo y se calcula el promedio para cada remuestreo. Luego, se ordenan de mayor a menor los 10000 estimadores como:

$$\overline{x}_1 \leq \overline{x}_2 \ldots \overline{x}_{9999} \leq \overline{x}_{10000} \tag{2.11}$$

por lo tanto, el intervalo de confianza del 99 % deseado es $(\overline{x}_{50}, \overline{x}_{9950})$.

Cuando el intervalo de confianza no incluye el umbral de 0.5 se puede decir que el valor promedio de IUPred del conjunto de posiciones difiere de manera significativa de 0.5 con un valor

$p < 0.01$. Si el límite inferior del intervalo de confianza ($\overline{x}_{50}$) es mayor a 0.5 el grupo de posiciones pertenece a una región desordenada. Si el límite superior del intervalo de confianza ($\overline{x}_{9950}$) es menor a 0.5 el grupo de posiciones pertenece a una región ordenada.

2.2.5. Coevolución en secuencias

IUPred predice el grado de orden de una cadena polipeptídica a partir de su secuencia, sin especificar una estructura. Si se cuenta con un alineamiento de un número suficiente de secuencias homólogas, los análisis de coevolución permiten predecir qué pares aminoácidos están cercanos en el espacio. Durante el periodo de trabajo de esta tesis se produjeron cambios en el estado del arte de los análisis de coevolución. Por ese motivo presento aquí los dos métodos principales de predicción: información mutua e información directa utilizados para el análisis de coevolución de secuencias.

Los dos métodos usados en esta tesis para analizar coevolución se basan en el mismo concepto central. Cuando una proteína manifiesta una mutación desfavorable que implica una disminución en la estabilidad del plegado o en la interacción con otra proteína, esta mutación puede ser compensada por mutaciones en otros aminoácidos que interactúen con el mismo en la estructura plegada, o mutaciones que participen en la misma interacción proteína-proteína o en la superficie de unión en la proteína blanco, de manera que se preserva o restaura la estabilidad o actividad. Por lo tanto, cuando las mutaciones son dentro de la misma proteína, al visualizar el conjunto de secuencias, se espera que estos pares de aminoácidos presenten patrones de sustitución restringidos y correlacionados. En la Figura 2.9 se puede observar un alineamiento "ficticio" de proteínas. El par de posiciones 5 y 10 (celeste) muestran un patrón de coevolución, cuando en la posición 5 hay una valina en la posición 10 hay una leucina. La sustitución de la valina por una leucina en la posición 5 correlaciona con una sustitución de una leucina por un glutámico. Otro patrón de sustitución es evidente para el par de posiciones 2 y 7 (verde), cuando en la posición 2 hay un glutámico, glutamina o aspártico, en la posición 7 hay un aspártico, una asparagina o una alanina respectivamente. Ningún otro patrón de sustitución correlacionado se observa para el resto de las posiciones.

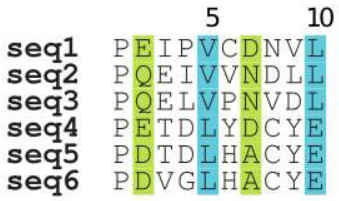


Figura 2.9: Coevolución de secuencias en un alineamiento de proteínas "ficticio". Se muestran dos pares de posiciones que coevolucionan, el par de posiciones 2 y 7 (en verde) y el par de posiciones 5 y 10 (en celeste).

El estado del arte muestra que la mayor parte de los patrones observados pueden interpretarse en términos de contactos entre aminoácidos de la misma cadena polipeptídica.

Información mutua

La información mutua aplica la teoría de la información para medir la correlación de las mutaciones entre dos posiciones y así inferir la coevolución entre pares de residuos (Gloor *et al.*, 2005). Dado un alineamiento múltiple de secuencias, se puede calcular: (1) la frecuencia marginal, o relativa, de cada aminoácido A en la posición i del alineamiento, $f_i(A_i)$,(2) la frecuencia marginal de cada aminoácido B en la posición j del alineamiento, $f_j(B_j)$ y (3) la frecuencia conjunta de observar el aminoácido A en la posición i y el aminoácido B en la posición j al mismo tiempo, $f_{ij}(A_i, B_j)$. El cálculo abarca $q = 21$ valores posibles de residuos, es decir, los 20 aminoácidos y los sitios vacíos. Estas frecuencias son usadas para calcular la información mutua entre las posiciones i y j, IM_{ij}, según:

$$IM_{ij} = \sum_{A_i}^{q} \sum_{B_j}^{q} f_{ij}(A_i, B_j) \ln \left(\frac{f_{ij}(A_i, B_j)}{f_i(A_i) f_j(B_j)} \right) \tag{2.12}$$

Cuando las posiciones no están correlacionadas, es decir, cuando las frecuencias son independientes, se cumple que $f_{ij}(A_i, B_j) = f_i(A_i) \cdot f_j(B_j)$. El argumento del logaritmo da 1, y por lo tanto, la información mutua toma su valor mínimo $IM_{ij} = 0$. Cuando las posiciones están completamente correlacionadas, se cumple que $f_i(A_i) = f_j(B_j) = f_{i,j}(A_i, B_j)$. Simplificando y reemplazando, se obtiene que la información mutua máxima es igual al valor de entropía, H_i o H_j, de las posiciones i o j, en nits (véase Ecuación 2.3).

Luego se aplica el método de corrección del producto promedio (APC, *Average Product Correction*) de Dunn *et al.* (2008) para reducir la señal de información mutua de fondo para cada par de posiciones. Dunn *et al.* (2008) definen el APC como:

$$APC_{ij} = M_{ij} \frac{MI_{i.} MI_{.j}}{MI_{..}} \tag{2.13}$$

donde MI_{ij} es el valor de información mutua entre el par de residuos i y j, $MI_{i.}$ es el valor promedio de información mutua del residuo i a todos los otros residuos del alineamiento, $MI_{.j}$ es el valor promedio de información mutua del residuo j a todos los otros residuos del alineamiento y $MI_{..}$ es el valor promedio de información mutua sobre todos los pares de residuos del alineamiento.

Los valores de información mutua son traducidos a un valor Z comparando los valores de información mutua de cada par de residuos a un valor promedio de valores de información mutua calculado a partir de alineamientos de secuencia mezclando los residuos dentro de cada posición, manteniendo las posiciones de los sitios vacíos fijas. Finalmente, se seleccionan aquellos pares de residuos ij que están a una distancia mayor o igual a 4 y tienen un valor $Z > 6$.

Información directa

La correlación entre posiciones puede estar dada por efectos directos o indirectos. En la Figura 2.10 se muestran los residuos A y B que interaccionan de manera directa y los residuos A y

C que interaccionan de manera indirecta mediados por B. Es decir, dos residuos pueden no estar en contacto y aún así tener un alto valor de coevolución debido a un efecto indirecto dado, por ejemplo, por interacciones con un tercer residuo en común (Weigt *et al.*, 2009).

Figura 2.10: Representación esquemática de interacción directa e indirecta. Se esquematizan tres residuos A, B y C. Los pares de residuos A y B, B y C interaccionan de manera directa. El par A y C interaccionan de manera indirecta mediados por B.

La información directa, a diferencia de la información mutua, permite distinguir las correlaciones directas de las indirectas (Morcos *et al.*, 2011). El primer paso en el cálculo de información directa es construir un modelo de Potts (Wu, 1982) en el que se asigna una energía a cada secuencia de largo L de acuerdo con el siguiente Hamiltoniano H (función de energía):

$$H(s) = \sum_{i}^{L} h_i(A_i) + \sum_{i}^{L} \sum_{j>i}^{L} j_{ij}(A_i, B_j) \tag{2.14}$$

se llama campos locales $h_i(A_i)$ a las energías de cada aminoácido A en cada posición de secuencia i y valores de acoplamiento $j_{ij}(A_i, B_j)$ a las energías de interacción entre el par de aminoácidos A y B situados en las posiciones de secuencia i y j. El modelo de Potts es una generalización del modelo de Ising para N spins interactuando en una estructura ordenada. Cuando $N = 2$, es equivalente al modelo de Ising.

En el marco de este modelo, la probabilidad de que ocurra una secuencia está dada por:

$$P(\vec{s}) = \frac{1}{Z} exp \left[- \sum_{i}^{L} h_i A_i - \sum_{i}^{L} \sum_{j>i}^{L} j_{ij}(A_i, B_j) \right] \tag{2.15}$$

donde Z es la función de partición asociada:

$$Z = \sum_{\vec{s}} exp \left[- \sum_{i}^{L} h_i A_i - \sum_{i}^{L} \sum_{j>i}^{L} j_{ij}(A_i, B_j) \right] \tag{2.16}$$

Bajo este modelo general la probabilidad conjunta de ocurrencia, $P_{ij}^{dir}(A_i, B_j)$, de dos aminoácidos, A y B, en dos posiciones, i y j, puede expresarse como:

$$P_{ij}^{dir}(A_i, B_j) = \frac{1}{Z_{ij}} exp \left(-j_{ij}(A_i, B_j) + h_i(A_i) + h_j(B_j) \right) \tag{2.17}$$

donde $j_{ij}(A_i, B_j)$ son los valores de acoplamiento, Z_{ij} es la función de partición para dos posiciones, y los parámetros $h_i(A_i)$ y $h_j(B_j)$ son los campos locales estimados.

El valor de información directa entre los pares de posiciones i y j, ID_{ij}, se calcula utilizando las probabilidades conjuntas de ocurrencia del aminoácido A en la posición i y del aminoácido B en la posición j de la siguiente manera:

$$ID_{ij} = \sum_{A_i}^{q} \sum_{B_i}^{q} P_{ij}^{dir}(A_i, B_j) \ \ln\left(\frac{P_{ij}^{dir}(A_i, B_j)}{f_i(A_i) f_j(B_j)}\right) \tag{2.18}$$

La información directa aumenta a medida que la correlación entre las posiciones aumenta. La principal diferencia con el cálculo de información mutua es que se usan los valores de probabilidades conjuntas de ocurrencia $P_{ij}^{dir}(A, B)$, generados por la coevolución directa entre estos pares de residuos, en lugar de las frecuencias conjuntas de los aminoácidos A y B en la posición i y j, que incluyen efectos tanto directos como indirectos.

Estimación de los campos locales y valores de acoplamiento. Se describe a continuación la determinación de los campos locales y los valores de acoplamiento a partir de un alineamiento múltiple de secuencias, Estos parámetros son necesarios para el cálculo de las probabilidades conjuntas de ocurrencia $P_{ij}^{dir}(A, B)$ y los valores de información directa ID_{ij}.

El número de términos del Hamiltoniano está dado por q^L, donde q es el número de residuos a considerar y L es la longitud de la secuencia. Por ejemplo, para un alfabeto de 20 aminoácidos y una secuencia de largo 5, el número de términos es mayor a 3 millones. Puesto que no resulta práctico determinar un número tan elevado de términos, se realiza una aproximación de campo medio que reduce el sistema a analizar (Morcos *et al.*, 2011). En este marco, se define la probabilidad de interacción directa como:

$$P_{ij}^{dir}(A_i, B_j) = \frac{1}{Z_{ij}} exp\left(-j_{ij}(A_i, B_j) + \tilde{h}_i(A_i) + \tilde{h}_j(B_j)\right) \tag{2.19}$$

donde los parámetros $\tilde{h}_i(A_i)$ y $\tilde{h}_j(B_j)$ son los campos locales estimados mediante la aproximación de campo medio y sustituyen a $h_i(A_i)$ y $h_j(B_j)$.

Para obtener los valores $j_{ij}(A_i, B_j)$ se calcula la matriz de correlación $C_{ij}(A_i, Bj)$:

$$C_{ij}(A_i, Bj) = f_{ij}^{p}(A_i, B_j) - f_i^{p}(A_i) f_j^{p}(B_j) \tag{2.20}$$

donde los parámetros $f_{ij}^{p}(A_i, B_j)$, $f_i^{p}(A_i)$ y $f_j^{p}(B_j)$ son la frecuencia conjunta y las frecuencias marginales observadas en el alineamiento de secuencias usado en cada caso.

Para que la matriz de correlación sea invertible, se aplican pseudocuentas a dichas frecuencias

de la siguiente manera:

$$f_{ij}^{p}(A_i, B_j) = \begin{cases} \left(\frac{\sigma}{q^2} + f_{ij}(A_i, B_j)\right) \frac{1}{\sigma + N_{ef}} & \text{si } i \neq j \\ f_{ij}(A_i, B_j) \frac{1}{\sigma + N_{ef}} & \text{si } i = j \text{ y } A_i \neq B_j \\ \left(f_{ij}(A_i, B_j) + \frac{\sigma}{q}\right) \frac{1}{\sigma + N_{ef}} & \text{si } i = j \text{ y } A_i = B_j \end{cases} \tag{2.21a}$$

$$f_{i}^{p}(A_i) = \left(\frac{\sigma}{q} + f_i(a_i)\right) \frac{1}{\sigma + N_{ef}} \tag{2.21b}$$

$$f_{j}^{p}(B_j) = \left(\frac{\sigma}{q} + f_j(B_j)\right) \frac{1}{\sigma + N_{ef}} \tag{2.21c}$$

donde q es el número de aminoácidos considerados incluyendo el sitio vacío (21 en nuestro caso), N_{ef} es el número efectivo de secuencias, es decir, la suma de la puntuación de secuencias (véase después), y σ es un parámetro de pseudocuentas elegido como $\sigma = N_{ef}$.

Una vez obtenida la matriz con la Ecuación 2.20, se calcula la matriz inversa y se obtienen los valores de acoplamiento bajo la aproximación de campo medio:

$$j_{ij} = \begin{cases} -C_{ij}^{-1}(A_i, B_j) & \text{si } i \neq j \\ 0 & \text{si } i = j \end{cases} \tag{2.22}$$

donde $C_{ij}^{-1}(A_i, B_j)$ es el elemento (i, j, A_i, B_j) de la matriz inversa obtenida con la Ecuación 2.20. Para una proteína de largo L y un alfabeto de q símbolos se obtiene una matriz de $Lq \times Lq$.

Para obtener los campos locales $\tilde{h}_i(B_i)$ y $\tilde{h}_j(B_j)$, se varían sus valores de manera secuencial hasta obtener unas frecuencias marginales estimadas similares a las frecuencias marginales calculadas a partir del alineamiento. Los pasos a seguir son:

1. Se fijan los parámetros: $\tilde{h}_i(a_i) = \tilde{h}_j(b_j) = 0$

2. Se estiman las frecuencias marginales con esos parámetros:

$$f_{i}^{est}(A_i) = \frac{1}{Z_{ij}} \sum_{B_j}^{q} exp\left(-j_{ij}(A_i, B_j) + \tilde{h}_i(A_i) + \tilde{h}_j(B_j)\right) \tag{2.23}$$

$$f_{j}^{est}(B_j) = \frac{1}{Z_{ij}} \sum_{A_i}^{q} exp\left(-j_{ij}(A_i, B_j) + \tilde{h}_i(A_i) + \tilde{h}_j(B_j)\right) \tag{2.24}$$

en ambos casos se normaliza dividiendo por la función de partición de dos posiciones:

$$Z_{ij} = \sum_{i} \sum_{j} \sum_{A_i} \sum_{B_j} exp\left(-j_{ij}(A_i, B_j) + \tilde{h}_i(A_i) + \tilde{h}_j(B_j)\right) \tag{2.25}$$

donde $j_{ij}(A_i, B_j)$ son los valores de acoplamiento bajo la aproximación de campo medio.

3. Renovación de los parámetros según el error en estimación de frecuencia:

$$\tilde{h}_i(A_i) \longleftarrow \tilde{h}_i(A_i) + \epsilon(f_i(A_i) - f_i^{est}(A_i)) \tag{2.26}$$

$$\tilde{h}_j(B_j) \longleftarrow \tilde{h}_j(B_j) + \epsilon(f_j(B_j) - f_j^{est}(B_j)) \tag{2.27}$$

4. Se repiten los pasos 2 y 3 hasta que el error de estimación de la frecuencia sea menor o igual a 10^{-6}.

Una vez obtenidos de esta manera los campos locales y los valores de acoplamiento se puede calcular la probabilidad conjunta directa según la Ecuación 2.19 y finalmente la información directa.

Puntuación de secuencias. Para evitar la sobrerrepresentación de secuencias muy similares entre sí y enfatizar la diversidad en el alineamiento múltiple de secuencias al realizar el cálculo de información directa, se le asignó un peso a cada secuencia según Henikoff y Henikoff (1994). A diferencia de otros métodos que se basan en el cálculo de distancias entre secuencias, este método se basa en la variabilidad de cada una de las posiciones. Para representar la diversidad de una posición, se le brinda a cada uno de los residuos presentes en una posición la misma fracción del peso. Es decir, dada una posición i, del alineamiento, con A residuos diferentes, un residuo presente en una única secuencia contribuye con $\frac{1}{r}$ a la puntuación de la secuencia, mientras que un residuo que está representado en n secuencias contribuye con $\frac{1}{nr}$ a la puntuación de esas secuencias. Por lo tanto, la puntuación de cada secuencia está dada por la suma de las contribuciones de cada posición. Así, los residuos más comunes contribuyen menos a la puntuación total de las secuencias.

Modelo nulo. Dado el número finito de secuencias que se utilizan en los alineamientos, es posible que ocurran correlaciones poco confiables o espúreas. Para disminuir esto, se construye un alineamiento múltiple con posiciones independientes. En concreto, se construye un nuevo alineamiento mezclando los residuos dentro de cada posición, manteniendo las frecuencias marginales de los aminoácidos dentro de cada posición constantes y eliminando las correlaciones. Se calcula

la información directa para este nuevo alineamiento y se restan los valores de información directa para el modelo nulo a los calculados en el alineamiento original (Espada *et al.*, 2015).

Selección de pares de residuos. Finalmente, los valores de la diferencia entre la información directa y el modelo nulo son transformados en valores Z como:

$$Z = \frac{x_i - \overline{x}}{\sigma} \tag{2.28}$$

donde: x_i es el valor de la diferencia entre el valor de información directa y el modelo nulo para el par de residuos i, $\overline{x}$ y σ son el promedio y el desvío respectivamente del conjunto de diferencias de información directa y el modelo nulo para todos los pares de residuos. Finalmente, se seleccionan para analizar aquellos pares de residuos con $Z \geq 3$.

2.2.6. Teoría de polímeros

La estabilidad de una proteína globular está determinada por los contactos establecidos entre los residuos que la conforman. A partir del estado desplegado de una proteína, el proceso de plegado se iniciaría por la formación de un contacto entre dos residuos que están a una distancia (L) en la cadena desplegada. En las proteínas desordenadas, o incluso en proteínas globulares con conectores entre los distintos dominios, las cadenas polipeptídicas poseen cierta flexibilidad, no hay una conformación discreta y por lo tanto una descripción aproximada puede lograrse mediante funciones de distribución.

El modelo de cadena entrópica (Zhou, 2003) considera una proteína como una cadena continua que cambia de dirección de forma aleatoria (cadena de gusano, *worm like chain*, en inglés). El cambio de dirección tiene lugar dentro de un radio de curvatura constante. Se utilizó este modelo para una proteína intrínsecamente desordenada ya que permite calcular una constante de equilibrio aparente para la formación de un contacto físico entre dos aminoácidos que están L residuos separados en la proteína.

La Figura 2.11 explica brevemente el principio del modelo. Dadas dos moléculas libres en solución A y B se asocian de manera intermolecular, con una constante de asociación $K_{intermolecular}$. Si esas dos moléculas están unidas covalentemente por un conector de longitud L la unión intramolecular depende de una constante de equilibrio aparente, $K_{intramolecular}$.

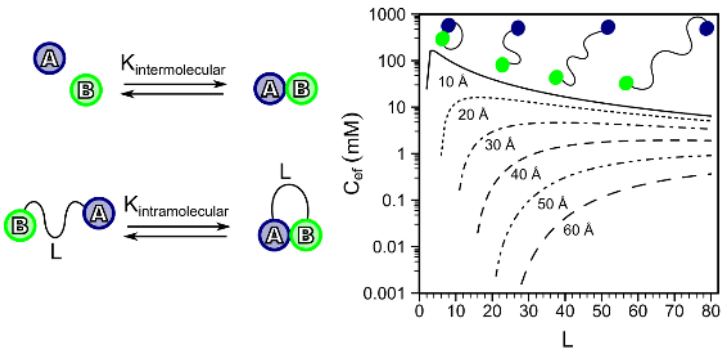


Figura 2.11: Modelo de cadena entrópica. *Izquierda.* Equilibrio de interacción del residuo A y B en solución (arriba) o unidos por un conector (abajo). *Derecha.* Dependencia de C_{ef} con la longitud de la cadena L para distintos valores de r indicados en la bajo las curvas. Figura adaptada de (Zhou, 2003).

La relación entre ambas constantes se llama concentración efectiva C_{ef} y está dada por:

$$C_{ef} = \frac{K_{intramolecular}}{K_{intermolecular}} \tag{2.29}$$

La C_{ef} depende de tres parámetros: (1) la longitud necesaria para que los cambios en dirección de cada residuo no estén correlacionados l_p (también llamada longitud de persistencia del conector), para proteínas este valor es de 3 Å y por debajo de este valor la cadena se comporta como una varilla rígida, (2) la longitud de la cadena, l_c, también llamada longitud de contorno, en el caso de proteínas es 3.8 Å por residuo y (3) la distancia de contacto r entre los C_α de dos aminoácidos.

Según la distancia de contacto utilizada existe un umbral de L debajo del cual la cadena no posee flexibilidad para que los residuos entren en contacto. A partir de ese umbral, C_{ef} aumenta hasta alcanzar un máximo en la longitud de cadena óptima, y luego disminuye (Figura 2.11, *derecha*).

La ecuación empírica C_{ef} se validó en diversos modelos (Borcherds *et al.*, 2017; Zhou, 2004) y puede escribirse como:

$$\begin{aligned} C_{ef} = \left(\frac{10^7}{6.022}\right)\left(\frac{3}{4\pi l_p l_c}\right)^{\frac{3}{2}} exp\left(\frac{-3r^2}{4l_p l_c}\right)\left(1 - \frac{5l_p}{4l_c} + \frac{2r^2}{{l_c}^2} - \frac{33r^4}{80l_p {l_c}^3} - \frac{79{l_p}^2}{160{l_c}^2}\right. \\ \left. - \frac{329r^2 l_p}{120{l_c}^3} + \frac{6799r^4}{1600{l_c}^4} - \frac{3441r^6}{2800l_p {l_c}^5} + \frac{1089r^8}{12800{l_p}^2 {l_c}^6}\right) \end{aligned} \tag{2.30}$$

reorganizando la Ecuación 2.29,

$$K_{intramolecular} = C_{ef} K_{intermolecular} \tag{2.31}$$

Por lo tanto, $K_{intramolecular}$ es el producto de la constante de asociación en equilibrio de dos aminoácidos libres en solución, $K_{intermolecular}$ y la penalidad entrópica por restringir las distancias posibles entre ambos aminoácidos, C_{ef}.

Utilizando los pares de contactos predichos por información directa para la proteína E1A se

calculó $K_{intramolecular}$ comparando el número de contactos predichos con el número de todos los contactos posibles:

$$K_{intramolecular}(L) = \frac{Número\ de\ contactos\ predichos\ a\ distancia\ L}{Número\ de\ contactos\ posibles\ a\ distancia\ L} \tag{2.32}$$

donde L es la distancia a la cual se encuentran los residuos considerados en contacto. Dado que el número de contactos es bajo, se realizó el promedio de $K_{intramolecular}$ sobre distintos tamaños de ventana de residuos.

2.2.7. Predicción estructural del dominio CR3 de E1A

El procedimiento para la predicción de la estructura del dominio CR3 se resume en la Figura 2.12. Partiendo de un alineamiento de E1A, se calcularon los valores de información directa, y se seleccionaron los pares de residuos con $Z \geq 3$ que pertenecían al CR3 (véase Sección 2.2.5). La estructura inicial a partir de la cual se realizó el modelado del dominio CR3 de E1A se construyó a partir de la secuencia del CR3 de HAdV8 que era la única que no presentaba sitios vacíos en el alineamiento (véase Sección B.2, CR3 sin sitios vacíos). Se utilizó la secuencia de residuos como entrada en el programa Flexible Meccano (Ozenne *et al.*, 2012) que puede generar utilizando la secuencia de aminoácidos un conjunto de confórmeros basados en los potenciales conformacionales y en el volumen de exclusión específicos de cada aminoácido.

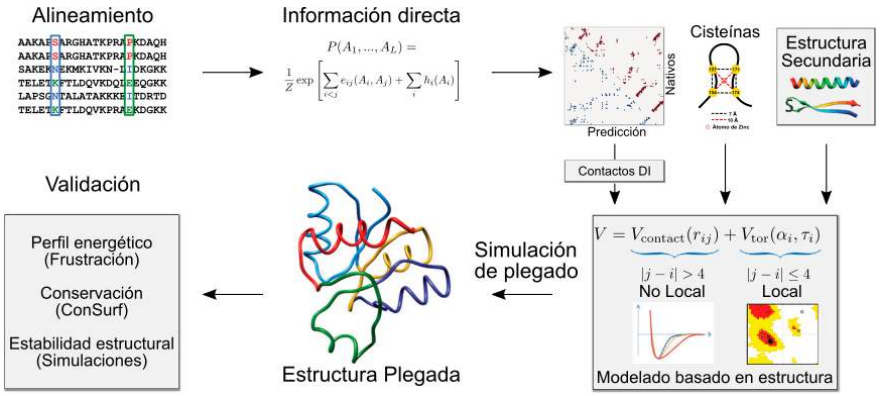


Figura 2.12: Metodología para la predicción estructural del dominio CR3 de E1A. Figura adaptada de (Sułkowska *et al.*, 2012).

Del ensamble conformacional se eligió una estructura al azar a partir de la cual se realizó la predicción de un modelo. En el modelado cada residuo se representa como una única esfera centrada en el C_α. La cadena polipeptídica se construye uniendo los residuos adyacentes utilizando un potencial que define la longitud del enlace y las restricciones de ángulos a través de potenciales armónicos. La estructura secundaria está codificada en el potencial de ángulos dihedros y en el de contactos nativos. Las predicciones geométricas se realizaron según Sułkowska *et al.* (2012). Se consideraron seis predicciones de contacto adicionales para definir el motivo de unión a zinc, restringiendo las distancias de los pares de cisteínas enfrentados, seguidos en secuencia y los pares

cruzados. Para las simulaciones de dinámica molecular se utilizó GROMACS 4.5.1 (Pronk *et al.*, 2013) utilizando un protocolo de simulación del recocido como en Sułkowska *et al.* (2012) y Clementi *et al.* (2000). Las estructuras colapsadas durante el enfriado se recuperaron para su análisis.

Validación del modelo estructural

Se validó el modelo en tres aspectos. En primer lugar, se utilizó la frustración configuracional para analizar el perfil energético del modelo estructural. Brevemente, la frustración configuracional compara la contribución de un par de residuos dado a la estabilización energética con las estadísticas de energías generadas a partir de haber utilizado pares de residuos con otra identidad en la misma ubicación. Si la contribución a la estabilización del par original no se diferencia de la mayoría de las alternativas, entonces la frustración es neutra. Si el par aporta a la estabilización más que la mayoría, se considera que está mínimamente frustrado y, si por el contrario, es lo suficientemente desestabilizante en comparación a las otras posibilidades el par de residuos está frustrado, es decir, presenta conflictos energéticos (Ferreiro *et al.*, 2007).

En segundo lugar se evaluó la conservación de secuencia en superficie utilizando el servidor ConSurf (Ashkenazy *et al.*, 2016) (véase Sección 2.2.8) utilizando como datos de entrada el modelo estructural y el alineamiento sin sitios vacíos del dominio CR3 de la proteína E1A.

Por último, se evaluó la estabilidad estructural. Para esto se reconstruyeron las cadenas laterales de las estructuras colapsadas utilizando el programa Pulchra (Rotkiewicz y Skolnick, 2008) y se realizó la minimización y simulaciones cortas con un paso de 2 fs utilizando TIP3P con agua explícita y el campo de fuerza electrostática AMBER99SB y malla de partículas de Ewald. Luego se calculó el RMSD a lo largo de la trayectoria utilizando GROMACS 4.5 (Pronk *et al.*, 2013) y VMD (Humphrey *et al.*, 1996).

2.2.8. Conservación de secuencia en la superficie de estructuras

El grado de conservación de cada residuo depende de la importancia funcional y del contexto estructural. Por lo tanto, los análisis de conservación son una herramienta muy útil ya que pueden revelar la contribución de una posición a la funcionalidad y estructura de la proteína. El servidor ConSurf (Ashkenazy *et al.*, 2016), disponible en `http://consurf.tau.ac.il` permite estimar la conservación evolutiva de la secuencia de aminoácidos en una estructura dada basada en las relaciones filogenéticas entre las secuencias homólogas partiendo de una secuencia de aminoácidos. El primer paso es la identificación de homólogos a partir de un alineamiento de secuencias o realizando una búsqueda en BLAST con la secuencia original. A partir del alineamiento se construye un árbol filogenético utilizando el algoritmo de unión de vecinos (NJ) (en inglés, *Neighbor-Joining*). ConSurf estima la tasa evolutiva de cada posición basándose en las relaciones evolutivas entre la proteína y sus homólogos y considerando la similitud de residuos. El índice de conservación es la tasa evolutiva y los valores obtenidos son divididos en una escala discreta de nueve intervalos de 1 a 9 para su visualización. Una tasa de evolución baja implica que los residuos están conservados (intervalos 5 a 9, siendo 9 el más conservado), una tasa de evolución rápida implica

que la posición es variable y los residuos están poco conservados (intervalos entre 1 y 4, siendo 1 el más variable).

Para las proteínas E7 y E1A se utilizó en ambos casos una estructura y un alineamiento como datos de entrada de ConSurf. Para la proteína E7, se utilizó el alineamiento sin sitios vacíos (Apéndice B) del dominio E7C y la estructura del dominio E7C (PDB ID: 2F8B) (Ohlenschläger *et al.*, 2006). Para E1A, se utilizó el alineamiento sin sitios vacíos (Apéndice B) del dominio CR3 y la estructura predicha para este dominio.

2.2.9. Homólogos estructurales

El servidor TopSearch (Sippl y Wiederstein, 2012) permite la búsqueda de homólogos estructurales. Dada una estructura, TopSearch realiza una búsqueda en la PDB y devuelve una lista de estructuras de proteínas ordenada por la similitud estructural. Para la proteína E7, se utilizó la estructura de la cadena A del dominio E7C (PDB ID: 2F8B,A) (Ohlenschläger *et al.*, 2006). Para E1A, se utilizó la estructura predicha para el dominio CR3.

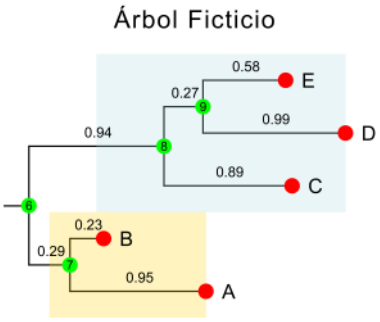
2.3. Análisis filogenéticos

2.3.1. Árboles filogenéticos

Un árbol filogenético representa una hipótesis acerca de las relaciones evolutivas entre un grupo de organismos. Se puede construir utilizando las características morfológicas (forma del cuerpo), bioquímicas o moleculares de los organismos. En las siguientes secciones describo la notación utilizada para referirse a las distintas regiones de un árbol filogenético y el funcionamiento de algoritmos utilizados como introducción necesaria a los experimentos que se describen en secciones posteriores.

Notación de árboles filogenéticos

Al construir un árbol se organizan los organismos de interés en grupos basados en los caracteres derivados compartidos entre ellos que los diferencian del ancestro. Las secuencias de genes o proteínas pueden compararse entre organismos y usarse para construir árboles filogenéticos. Las especies cercanas por lo general tienen pocas diferencias en sus secuencias, mientras que las menos emparentadas tienden a tener más. Una forma de medir las diferencias entre las distintas especies es utilizando matrices de distancia. La Figura 2.13 muestra un árbol filogenético ficticio que representa las relaciones evolutivas entre cinco secuencias A, B, C, D y E de cinco organismos distintos.



Matriz de Distancias

	A	B	C	D	E
A	0.00	1.18	3.07	3.44	3.03
B	1.18	0.00	2.35	2.72	2.31
C	3.07	2.35	0.00	2.15	1.74
D	3.44	2.72	2.15	0.00	1.57
E	3.03	2.31	1.74	1.57	0.00

Figura 2.13: Árbol filogenético ficticio. *Izquierda.* Relaciones evolutivas entre las secuencias A, B, C, D y E. Las hojas del árbol con un círculo rojo. Los nodos internos numerados de 6 a 9 se representan con círculos verdes. Dos clados están resaltados con rectángulos celeste y naranja. En cada rama se indica el largo de la misma. *Derecha.* Matriz de distancias entre las secuencias A, B, C, D y E correspondiente al árbol mostrado a la izquierda.

En un árbol filogenético, las secuencias actuales o extantes (círculos rojos) se encuentran en los extremos de las líneas a las que consideramos las ramas del árbol, es decir, en las hojas del árbol. Las ramas más cercanas a las hojas se consideran las ramas menos profundas o más recientes, y las más lejanas las ramas más profundas del árbol o menos recientes. Cada punto de ramificación se llama nodo interno (círculos verdes) y representa un evento de divergencia o separación de un grupo hipotético ancestral en dos grupos descendientes. Cada nodo interno es el ancestro común, o secuencia ancestral, más reciente de los grupos que descienden de ese nodo. Por ejemplo, el nodo 7 es el ancestro común más reciente de las secuencias A y B y el nodo 8 es el ancestro común más reciente de las secuencias C, D y E. El ancestro común más reciente a las cinco secuencias es el nodo 6 y la rama anterior es la raíz del árbol, por lo que el nodo 6 también se lo llama nodo raíz. Los árboles pueden o no tener determinada la raíz. Cada línea horizontal representa una serie de secuencias ancestrales que al final lleva o bien a un nodo interno que es una secuencia ancestral en común entre dos descendientes, o bien a una secuencia actual. Dos secuencias son más similares si tienen un ancestro común más reciente y menos similares si tienen un ancestro común menos reciente. Por ejemplo, D y E son más similares entre sí que con C. Todos los descendientes que contienen un antepasado común conforman un clado (rectángulos). Por ejemplo, el nodo 7 da origen a un clado (amarillo) y el nodo 8 da origen a otro clado (celeste). Por último, al considerar un grupo de secuencias extantes se dice que provienen de un grupo monofilético, cuando todas las secuencias comparten un ancestro común y todos los descendientes de ese ancestro común están incluidos en el grupos. Por ejemplo, el grupo de secuencias C, D y E es monofilético.

Existen tres tipos principales de árboles que difieren en el significado de la longitud de las ramas: (1) el cladograma, donde cada rama únicamente representa la transición evolutiva entre un nodo ancestral y sus descendientes independientemente de la longitud; (2) el filograma donde la longitud de cada rama es proporcional al número de cambios que existen entre un ancestro y sus descendientes; y (3) el cronograma o árbol ultramétrico donde las ramas representan el tiempo y las secuencias u organismos actuales son equidistantes a la raíz. En la Figura 2.13 el árbol representado es un filograma y sobre cada rama se indica la longitud correspondiente. A la derecha del mismo

se muestra la matriz de distancias utilizada para construir el árbol. Cada elemento de la matriz equivale a la suma de las longitudes de las ramas que separan dos secuencias. Por ejemplo, la distancia entre A y E es 3.03 que se obtiene de sumar las longitudes $0.95+0.29+0.94+0.27+0.58$.

Las representaciones de árboles utilizadas a lo largo de esta tesis se realizaron con el paquete PhyTools versión 0.6-44 (Revell, 2012) de R. La descripción de los formatos y árboles utilizados se encuentra disponible en Apéndice D.

Métodos de construcción filogenética

Existen dos métodos principales para la construcción de un árbol filogenético a partir de secuencias: (1) métodos de distancias, siendo los más conocidos el método de agrupación de pares no pesados con media aritmética (UPGMA) (en inglés, *Unweighted Pair Group Method with Arithmetic Mean*), unión de vecinos (NJ) (en inglés, *Neighbor-Joining*), o su derivado BioNJ (Gascuel, 1996), y el método de mínima evolución; y (2) métodos basados en caracteres, como máxima parsimonia, máxima verosimilitud (en inglés, *maximum likelihood*) y el método de inferencia Bayesiana (Holder y Lewis, 2003).

Los métodos basados en distancias, en primer lugar, calculan las distancias genéticas entre todos los pares de secuencias y resumen esta información en una matriz de distancias. Luego, se utilizan los valores obtenidos para la construcción de un árbol filogenético. Estos métodos asumen que la tasa de cambio, o reloj molecular, de todas las secuencias es la misma y no permiten determinar en qué parte del árbol ocurre un determinado cambio en la secuencia. Otras desventajas de estos métodos es que no analizan hipótesis evolutivas alternativas y no tienen en cuenta si ocurre o no un cambio más de una vez en una determinada posición, subestimando la distancia genética real. Por el contrario, los métodos basados en caracteres tienen en cuenta los cambios que ocurren en cada posición del alineamiento, si ocurre o no un cambio más de una vez en una posición y consideran una tasa evolutiva variable. La tasa evolutiva se obtiene utilizando un modelo evolutivo que es una descripción matemática de la evolución de las secuencias. Existen modelos evolutivos específicos para nucleótidos, aminoácidos y codones. Los modelos tienen dos tipos de parámetros principales: (1) la probabilidad de cambio entre pares de estados posibles, definiendo como estado cada nucleótido, aminoácido o codón y (2) las frecuencias de los estados en el alineamiento. También se pueden utilizar otros parámetros que consideran qué partes del alineamiento evolucionan a tasas distintas. Para modelar la variación de las tasas evolutivas entre las posiciones se utiliza una distribución de probabilidad continua llamada distribución gamma que se especifica con un único parámetro, α. A menor valor de α la variación de la tasa es mayor y a mayores valores de α la variación de la tasa es menor.

Lo ideal es generar y evaluar todas las hipótesis evolutivas posibles. Sin embargo, el número de árboles que se pueden construir crece de forma factorial respecto al número de secuencias actuales. Por ejemplo, el número de árboles posibles para tres secuencias es 3, para seis secuencias 945 y para doce secuencias $1.37\ 10^{10}$. Para la construcción del árbol los modelos basados en caracteres realizan una búsqueda heurística partiendo de un árbol aleatorio o construido por un

método de distancias. Es decir, generan árboles similares mediante pequeñas reorganizaciones locales a partir de un árbol inicial. Los algoritmos más comunes para realizar la búsqueda heurística son el algoritmo de movimientos topológicos de subárboles (SPR) (en inglés, *Subtree Pruning and Regrafting*) y el algoritmo de intercambio de vecinos cercanos (NNIs) (en inglés, *Nearest Neighbor Interchanges*) que se esquematizan en la Figura 2.14.

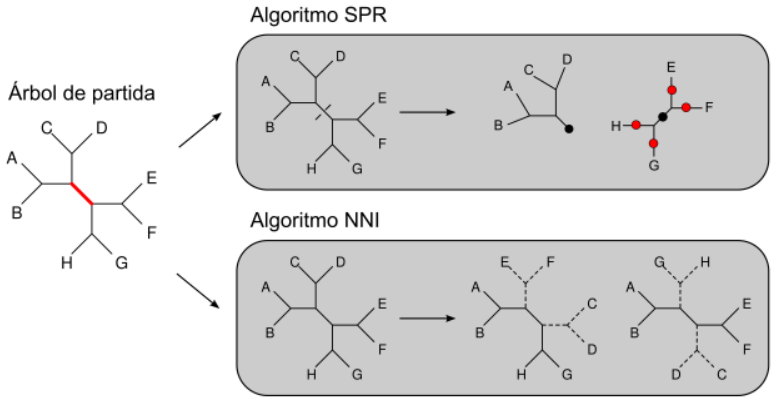


Figura 2.14: Algoritmos de búsqueda heurística. Se esquematizan los algoritmos de búsqueda heurística SPR y NNIs. la línea roja en el árbol principal indica la rama sobre la cual se realizarán los cambios. Para el algoritmo SPR, se señala el punto de corte del árbol inicial y en círculos rojos las cuatro reubicaciones posibles para generar una nueva topología. Para el algoritmo NNI, se señala con líneas punteadas los subárboles y las nuevas posiciones que dan origen a una nueva topología. Figura adaptada de Lemey *et al.* (2009).

Estos algoritmos realizan la reubicación de las ramas en distintos puntos del árbol. El algoritmo de SPR, a partir de un árbol inicial remueve un subárbol y lo ubica en una rama creando un nuevo nodo (Figura 2.14) y el algoritmo de NNI intercambia los subárboles que se encuentran a dos nodos de distancia entre sí (Figura 2.14).

Durante la búsqueda heurística los distintos árboles se aceptan o rechazan utilizando una función o criterio de optimización que depende del modelo evolutivo utilizado y permite elegir el mejor árbol entre todos los generados. Finalmente, la búsqueda continúa hasta que no se encuentran mejoras significativas al crear nuevas topologías. En este trabajo de tesis se utilizó el criterio de optimización de máxima verosimilitud que se explica a continuación.

Máxima verosimilitud. La idea principal del principio de máxima verosimilitud aplicado a filogenia es, dado un árbol y el alineamiento correspondiente, cuál es la topología, longitud de ramas y modelo evolutivo que maximiza la probabilidad de observar las secuencias de dicho alineamiento (Felsenstein, 1981). En otras palabras, la función de verosimilitud es la probabilidad condicional de los datos (alineamiento) dada una hipótesis, donde la hipótesis incluye un modelo de sustitución con ciertos parámetros (θ) y un árbol o topología (τ) con determinada longitud de ramas:

$$L(\tau|\theta) = Pr(datos|\tau,\theta) = Pr(alineamiento \mid árbol, modelo\ evolutivo) \tag{2.33}$$

El objetivo del método es encontrar los parámetros θ y τ que maximizan la probabilidad de explicar el alineamiento.

Dado un alineamiento de largo l y N secuencias, se pueden definir las secuencias como vectores $S_i = (s_i^1 ... \ s_i^l)$ con $i = 1...N$, donde s_i^j es el residuo en la posición j de la secuencia i. La verosimilitud de dos secuencias se calcula como la probabilidad de observar dichas secuencias dada una distancia genética d:

$$L(d) = \prod_{j=1}^{l} \pi_{s_1^j} \, P_{s_1^j s_2^j}\left(\frac{d}{\mu}\right) \tag{2.34}$$

donde s_1^j es el residuo en el sitio $j = 1, ..., l$ en la secuencia 1, $\pi_{s_1^j}$ es la frecuencia del residuo s_1^j en el alineamiento y $P_{s_1^j s_2^j}\left(\frac{d}{\mu}\right)$ la probabilidad de transición del residuo en la secuencia 1 al residuo en la secuencia 2. Esta probabilidad depende de μ que es el número total de sustituciones por unidad de tiempo.

Para calcular la máxima verosimilitud se utiliza un modelo evolutivo que define la probabilidad de cambio entre estados, $P_{s_1^j s_2^j}$. Estos modelos asumen que las posiciones del alineamiento son independientes entre sí, ya que si bien no es del todo verdadero que exista la independencia simplifica los cálculos a realizar. Además, asumen que el proceso de evolución molecular es reversible. Por lo tanto, la verosimilitud del árbol es independiente de la posición de la raíz (Lemey *et al.*, 2009). El número total de sustituciones por unidad de tiempo, μ, puede considerarse que es igual para todas las posiciones o modelarse, por ejemplo, utilizando la distribución Gamma.

Cuando se parte de más de dos secuencias, se calcula la verosimilitud de cada posición. Sea D_j el patrón de residuos en la posición $j = 1, ..., l$ en el alineamiento, el modelo evolutivo, M y el árbol τ con longitud de ramas, la probabilidad de observar el alineamiento, D, dado es el producto de las probabilidades de cada sitio:

$$L(\tau, M, \mu|D) \equiv P[D|\tau, M, \mu] = \prod_{j=1}^{l} Pr[D_j|\tau, M, \mu_i] \tag{2.35}$$

se consideran para cada posición todos los cambios posibles que den lugar al estado actual, incluyendo los más probables y los menos probables, y se suman las probabilidades de cada combinación única de estados ancestrales hipotéticos y longitudes de rama. Una vez calculada la verosimilitud para cada posición, se calcula la verosimilitud total del árbol como el logaritmo natural de la suma de las verosimilitudes de cada posición y se estiman las longitudes de rama, τ, que maximizan la función:

$$log[L(\tau, M, \mu|D)] = log\left[\prod_{j=1}^{l} Pr[D_j|\tau, M, \mu_i]\right] = \sum_{j=1}^{l} log\left[Pr[D_j|\tau, M, \mu_i]\right] \tag{2.36}$$

es decir, se buscan los puntos críticos de la función.

El valor obtenido es el logaritmo del valor de máxima verosimilitud. Esta topología es la más probable pero no significa que sea correcta. Todo árbol filogenético es una hipótesis evolutiva con una probabilidad asociada que se prefiere entre un conjunto de hipótesis alternativas.

2.3.2. Filogenia de *Mastadenovirus*

Los árboles filogenéticos de *Mastadenovirus* existentes en la bibliografía incluyen únicamente una parte de los serotipos utilizados en este trabajo. Los árboles más completos reportados solo incluyen a los serotipos que infectan a primates humanos y no humanos, por lo que se necesitó realizar una filogenia propia que incluyera a todos los serotipos utilizados.

En la Figura 2.15 se resume el procedimiento realizado para la creación de la filogenia.

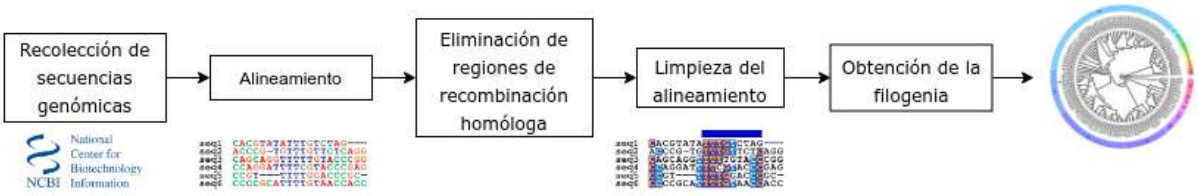


Figura 2.15: Flujo de trabajo para la construcción del árbol filogenético de *Mastadenovirus*.

Se recolectaron las secuencias de 139 genomas completos de *Mastadenovirus* obtenidos de la NCBI (véase Apéndice A). De estas secuencias, 116 corresponden a las secuencias de E1A de nuestra base de datos, 19 secuencias son variantes que se usaron como control y cuatro secuencias son nuevos serotipos. Las secuencias genómicas se alinearon con LAGAN (Brudno *et al.*, 2003). Como secuencia de referencia se utilizó el serotipo Human adenovirus 2 (HAdV2) de la especie *Human mastadenovirus A* (GenBank ID J01917) de la cual se incluyeron 13 fragmentos diferentes como control. Los alineamientos de genomas completos fueron recortados dejando el bloque homólogo de mayor tamaño (posiciones 4089 to 6577), que abarca los genes que codifican para la proteína IVa2 y ADN polimerasa. Este bloque se encuentra en todas las secuencias y no incluye sitios con alta frecuencia de recombinación, es decir, no incluye la región codificante de la base del pentón, hexón y fibra (Robinson *et al.*, 2011). Este alineamiento está disponible en el Apéndice B.

Limpieza del Alineamiento. Para limpiar el alineamiento de regiones ricas en sitios vacíos y alineadas de manera no confiable se utilizó el programa GBlocks (Castresana, 2000). Con el objetivo de identificar bloques de secuencia altamente conservados, GBlocks evalúa posiciones y bloques de posiciones usando cinco umbrales (IS, FS, CP, BL1 y BL2). En los cinco casos, se usó el valor por defecto para el umbral. El umbral IS se define como el 50 % de las secuencias más uno utilizadas en el alineamiento. En este caso es 77. El umbral FS se define como el 85 % de las secuencias utilizadas en el alineamiento, siendo 130 en este caso. El umbral CP indica el mínimo valor de posiciones no conservadas contiguas necesarias para eliminar un bloque no conservado,

siendo el valor por defecto ocho. BL1 y BL2 indican valores mínimos de posiciones conservadas o altamente conservadas contiguas para eliminar un bloque pequeño conservado en el caso de BL1 o una región pequeña conservada en el caso de BL2. Los valores por defecto son 15 y 10, respectivamente. En la Figura 2.16 se describe el algoritmo que usa GBlocks.

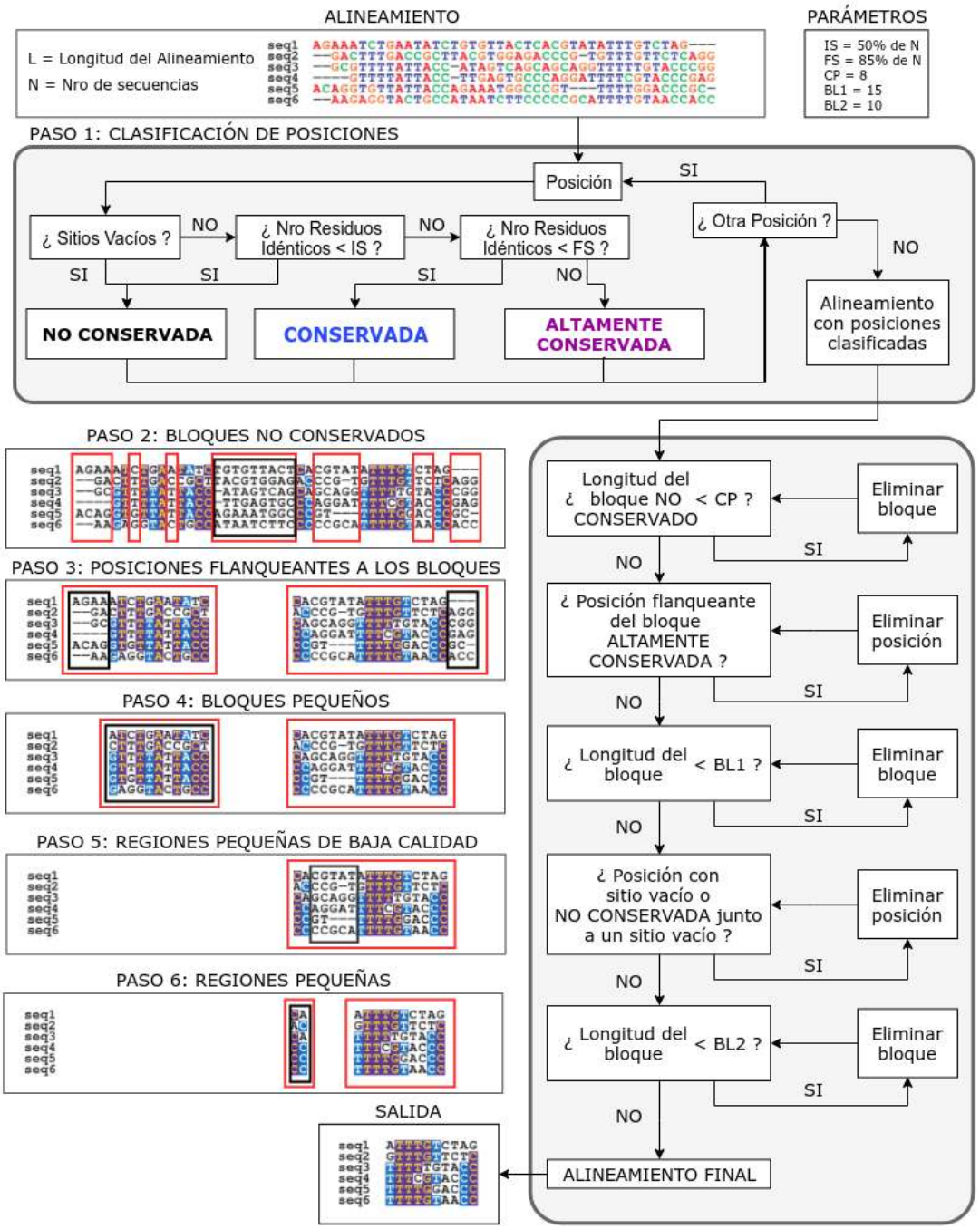


Figura 2.16: Algoritmo Gblocks para la limpieza del alineamiento. Se muestra en forma de flujo el funcionamiento del algoritmo Gblocks para la eliminación de bloques no conservados en el alineamiento junto con un alineamiento ficticio de ejemplo. En los alineamientos con posiciones clasificadas se señalan las posiciones altamente conservadas, conservadas y no conservadas con fondo violeta, azul y blanco y letras amarillas, blancas y negras, respectivamente.

Brevemente, el algoritmo GBlocks consta de seis pasos: (1) Clasificación de las posiciones en no conservadas cuando presentan sitios vacíos o la cantidad de residuos idénticos es menor a IS, conservadas si es menor a FS y en caso contrario se clasifican en altamente conservadas. (2) Se identifican los bloques no conservados y se eliminan los de longitud menor a CP. (3) Se eliminan las posiciones no conservadas o conservadas de los extremos de cada bloque de manera consecutiva hasta que en ambos extremos de cada bloque haya posiciones altamente conservadas. (4) Se eliminan bloques conservados de longitud menor a BL1 ya que son regiones pequeñas donde es difícil asegurar la calidad del alineamiento. (5) En los bloques restantes se eliminan las posiciones con sitios vacíos y las posiciones no conservadas adyacentes a posiciones con sitios vacíos, hasta alcanzar una posición conservada. De esta manera, se eliminan regiones pequeñas donde es difícil asegurar la calidad del alineamiento debido a la presencia de sitios vacíos. Por último, (6) se eliminan todos los bloques conservados pequeños con una longitud menor a BL2.

A través de este procedimiento, se obtiene un alineamiento de una región homóloga que incluye bloques lo suficientemente grandes y alineados de manera confiable como para realizar la reconstrucción del árbol filogenético.

Construcción del Árbol. Para la reconstrucción del árbol filogenético se utilizó un alineamiento con 1826 posiciones alineadas seleccionadas por GBlocks y el algoritmo PhyML 3.0 (Guindon *et al.*, 2010), con la opción SPR del algoritmo.

PhyML es un algoritmo basado en máxima verosimilitud y realiza la búsqueda heurística de las distintas topologías de árboles por SPR a partir de un árbol inicial (Figura 2.17). Para generar el árbol inicial (paso 1) se utilizó una variación del algoritmo NJ (Saitou y Nei, 1987) llamada BioNJ (Gascuel, 1996). Brevemente, a partir del alineamiento BioNJ calcula una matriz de distancias a partir de las diferencias observadas en las secuencias. Luego, transforma esta matriz de distancias observadas en una matriz de varianzas de las distancias. Para minimizar la matriz, agrupa de manera iterativa los pares de taxa que poseen el menor valor en la misma y los representa como un nuevo nodo. A partir de la matriz de distancias minimizada construye el árbol filogenético inicial.

A partir de este árbol filogenético inicial se realiza una búsqueda heurística por SPR (paso 2). A cada reubicación posible que genera SPR le asigna una puntuación por parsimonia (paso 3). Las puntuaciones por parsimonia se calculan utilizando el algoritmo de Fitch (Fitch, 1971) cuantificando el número de cambios entre las secuencias a lo largo del árbol necesarios para explicar la filogenia observada (véase Sección 2.3.4). Se ordenan los valores obtenidos y se seleccionan aquellos menores al umbral de parsimonia (PT) elegido por el usuario. Este umbral define hasta qué diferencia en la puntuación de una nueva topología menos parsimoniosa y la actual se acepta para seguir analizando. Por ejemplo, si PT=0, no se aceptan topologías con una puntuación de parsimonia mayor a la actual, solo aquellas que sean más parsimoniosas (menor puntuación). En la construcción de la filogenia de *Mastadenovirus*, se utilizó el valor por defecto, PT=5. Es decir, además de analizar las soluciones más parsimoniosas, se incluyen para analizar las nuevas topologías menos parsimoniosas con una puntuación hasta 5 puntos mayores que la topología actual.

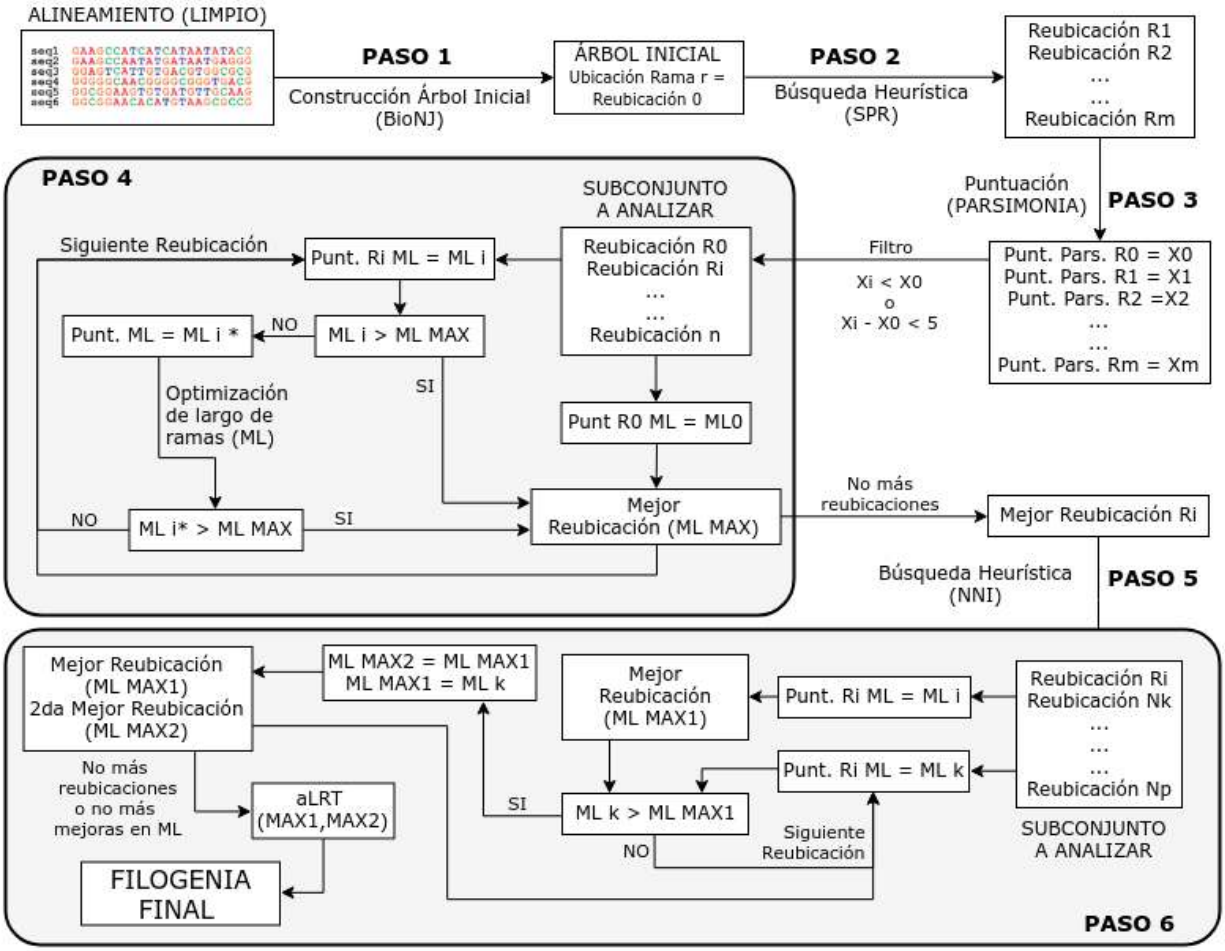


Figura 2.17: Algoritmo PhyML para la construcción de la filogenia de *Mastadenovirus*. Se muestra en forma de flujo el funcionamiento del algoritmo PhyML para la construcción de la filogenia.

El subgrupo de soluciones seleccionado es analizado por la función de máxima verosimilitud como criterio de optimización para la valoración de las nuevas topologías (paso 4, recuadro gris). PhyML primero evalúa la verosimilitud de la nueva topología obtenida por el algoritmo SPR sin ajustar las longitudes de las ramas. Si la nueva reubicación tiene mayor verosimilitud que la reubicación de mayor valor encontrada hasta el momento, entonces pasa a ser la mejor reubicación. Si no, optimiza las longitudes de las ramas, es decir el número de cambios entre los extremos de la rama, hasta maximizar la verosimilitud. De esta manera se evita evaluar la verosimilitud de todo el árbol y solo actualiza la verosimilitud de un número limitado de árboles y regiones del mismo. Una vez que todas las reubicaciones son evaluadas, si hay una reubicación con mayor verosimilitud a la de la filogenia actual, se aplica esa reubicación y se utiliza esa nueva topología para seguir adelante.

Para el ajuste de las longitudes de ramas se utilizó el modelo de sustitución HKY85 (Hasegawa *et al.*, 1985), con la tasa de transición/transversión fijada a 4. Este modelo no asume que las frecuencias de las bases nucleotídicas son iguales y hace diferencia entre transiciones y transversiones. La proporción de sitios invariables y el parámetro α de la distribución Gamma son estimados a partir de los datos. Una vez que se llega a un máximo de verosimilitud, se realiza una búsqueda

heurística por NNI (Guindon y Gascuel, 2003) alrededor de la rama de interés (paso 5). Para cada reconstrucción obtenida por NNI se evalúa la verosimiltud (paso 6). La búsqueda termina cuando no se observa una mejora significativa en la verosimilitud entre dos pasos completos de NNI.

Una vez finalizada la puntuación por verosimilitud, se evalúa la hipótesis evolutiva mediante el contraste contra una historia evolutiva alternativa. La hipótesis evolutiva alternativa es la segunda mejor construcción obtenida por NNI en cada rama. Para cada nodo se evalúa el soporte utilizando un procedimiento similar al de Shimodaira-Hasegawa para la prueba de la tasa de probabilidad aproximada (aLRT) (en inglés, *Approximate Likelihood-Ratio Test*) (Anisimova y Gascuel, 2006). La hipótesis nula de esta prueba es que la diferencia de la verosimilitud entre ambas construcciones es nula. El estadístico aLRT sigue una distribución basada en la distribución de chi-cuadrado y se estima como:

$$aLRT = 2(Ln(L_1) - Ln(L_2)) \tag{2.37}$$

donde $Ln(L_1)$ es el logaritmo de la verosimilitud del árbol con máxima verosimilitud, y $Ln(L_2)$ corresponde a la segunda mejor configuración obtenida por NNI alrededor de la rama de interés. Si la prueba es significativa, se rechaza la hipótesis de igual verosimilitud para ambas construcciones. El soporte de cada rama se expresa como 1 menos el valor p obtenido y varía entre 0 y 1.

Por último, la raíz del árbol se eligió como el punto medio entre los dos serotipos con mayor distancia. Aquellas ramas que contenían variantes fueron colapsadas. Las ramas que no correspondían a serotipos incluidos en nuestra base de datos no fueron consideradas para los análisis y se removieron las terminales correspondientes del árbol.

2.3.3. Filogenia del hospedador de *Mastadenovirus*

El árbol del hospedador de *Mastadenovirus* fue construido a partir de la literatura (Murphy *et al.*, 2001; O'Leary *et al.*, 2013). El género *Mastadenovirus* infecta a mamíferos. En el árbol se incluyeron únicamente aquellos mamíferos que figuran como hospedadores de los serotipos incluidos en nuestra base de datos.

2.3.4. Reconstrucción de estados ancestrales

Los estados ancestrales de una secuencia pueden determinarse a partir de los estados actuales, una filogenia y los métodos de Máxima Parsimonia, Máxima Probabilidad (*Maximum Likelihood*, en inglés) o Estadística Bayesiana. En este trabajo se utilizaron el primer método y una combinación de los dos últimos métodos.

Parsimonia

Este método fue usado en el marco de esta tesis para la reconstrucción de estados ancestrales. Se utilizó el programa Mesquite versión 3.11 que incluye el paquete para la reconstrucción de estados ancestrales (en inglés, *Ancestral State Reconstruction*) (Maddison y Maddison, 2015). El método de parsimonia se explica a continuación.

Algoritmo de parsimonia de Fitch. El principio de parsimonia aplicado a la resolución de problemas establece que cuando existen diversas hipótesis, se debe elegir la que tenga menor cantidad de asunciones. En biología este principio se puede usar para la reconstrucción de árboles filogenéticos y de estados ancestrales, la conexión se basa en la afirmación de que cada instancia de homoplasia, como ser estados de un carácter idénticos que no pueden ser explicados por herencia de un ancestro común, constituyen una hipótesis *ad hoc* y, por lo tanto, el número de estas hipótesis debe ser minimizado. En este trabajo se utilizará el método de parsimonia con el objetivo de definir los estados ancestrales de los motivos lineales considerados como caracteres.

Dado un árbol y unos estados o caracteres actuales, se puede aplicar el principio de parsimonia para encontrar la distribución de estados ancestrales que explique los estados observados en las hojas del árbol y minimice el número total de cambios de un carácter a lo largo de la filogenia. Este método se conoce como máxima parsimonia. Existen numerosos algoritmos que aplican este principio siendo el más común y el utilizado en este trabajo el algoritmo de Fitch (Fitch, 1971). Este algoritmo consiste en dos fases: (1) la fase preliminar y (2) la fase final que se representan en la Figura 2.18.

En la fase preliminar (Figura 2.18, izquierda), la dirección de aplicación del algoritmo es desde las hojas del árbol (círculos verdes), o descendientes, hacia el ancestro más distante (nodo 6, círculo azul). Se crea el estado del carácter ancestral para el ancestro inmediato. Para la creación de los estados del carácter en cada nodo se sigue la regla de que el nodo debe contener todos los estados del carácter comunes a los descendientes inmediatos, es decir el estado del carácter del nodo ancestral es la intersección de los conjuntos de caracteres de los descendientes. Por ejemplo, dado los individuos *s6* y *s7*, ambos presentan como estado de carácter el 3. Para definir el estado del nodo ancestral más reciente que pertenece se hace la intersección entre: $s6 = \{3\} \cap s7 = \{3\} = \{3\}$. Si no tienen estados en común, la intersección es vacía y el nodo incluye a todos los estados del carácter en los descendientes. Es decir, es la unión de los estados del carácter de los descendientes inmediatos. Por ejemplo, dados los individuos s1 y s2, s1 presenta como estado de carácter 0, y s2 el 1. Como $s1 = 0 \cap s2 = 1 = \varnothing$, se define el estado del nodo ancestral inmediato como $s1 = \{0\} \cup s2 = \{1\} = 0, 1$. Este procedimiento se realiza hasta alcanzar al nodo más distante.

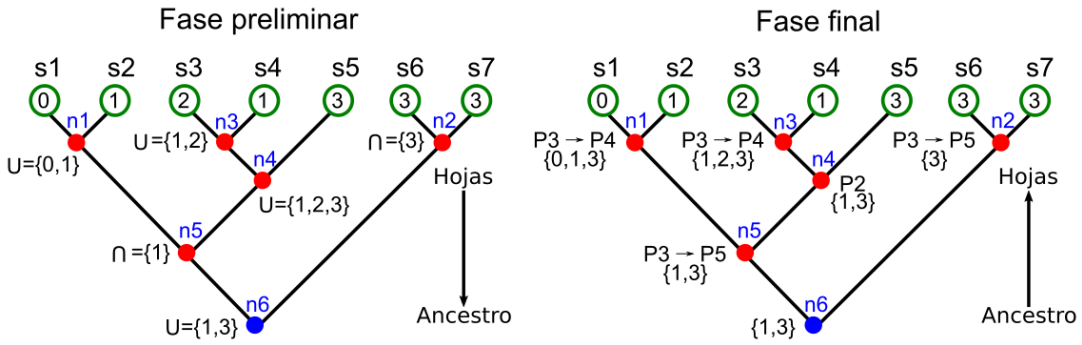


Figura 2.18: Método de Fitch. Se muestra la reconstrucción de los estados ancestrales según el método de Fitch de un caracter con cuatro estados posibles: 0, 1, 2 y 3. *Izquierda.* Fase Preliminar. Se indica al lado de cada nodo interno los estados ancestrales posibles como producto de la unión, ∪, o intersección, ∩, de los estados en los descendientes inmediatos. *Derecha.* Fase Final. Se indica al lado de cada nodo interno los estados ancestrales posibles y los pasos seguidos para adjudicarlos, omitiendo el paso 1 (P1) y paso 6 (P6) comunes a todos. En ambos casos, en la parte superior, se indican siete individuos (s1-s7). En círculos de borde verde se representa las hojas del árbol, es decir, los datos conocidos. Los nodos internos están indicados con círculos rojos y numerados del n1 al n6, con excepción del nodo más ancestral, n6, que está indicado con un círculo azul.

En la fase final (Figura 2.18, derecha), la reconstrucción se realiza desde el nodo más distante (nodo 6, círculo azul) hacia las hojas de los árboles (círculos verdes). El carácter del nodo más distante (nodo 6, círculo azul) queda adjudicado en la fase preliminar y se pasa a los dos descendientes. Cuando aún no se pasó por la fase final, se hace referencia al nodo como nodo preliminar. El algoritmo en esta fase funciona de la siguiente manera:

P1. Si el nodo preliminar contiene un conjunto de estados del carácter que abarca los estados del carácter presentes en el ancestro inmediato, se pasa a P2, si no a P3. Ejemplo: El estado posible del nodo preliminar 5 es 1. El nodo 6 contiene los estados 1 y 3. Por lo tanto, se pasa a P3. En la siguiente ronda, el nodo preliminar 4 tiene los estados 1, 2 y 3. El nodo final 5 tiene los estados 1 y 3. Por lo tanto, se pasa a P2.

P2. En el nodo preliminar se eliminan del conjunto de estados del carácter aquellos estados que no están en el conjunto de estados final del nodo ancestral inmediato y se pasa a P6. Ejemplo: En el nodo preliminar 4 se elimina el estado 2, ausente en el nodo final 5, y se pasa a P6.

P3. Si el conjunto de estados en el nodo preliminar se formó por la unión de los estados de sus descendientes, se pasa a P4. Si se formó por la intersección, se pasa a P5. Ejemplo: El conjunto de estados del nodo preliminar 2 se formó por la intersección, por lo tanto se pasa a P5. El conjunto de estados del nodo preliminar 1 se formó por la unión, por lo tanto se pasa a P4.

P4. Se agrega al conjunto de estados del nodo cualquier estado que esté presente en el nodo ancestral inmediato final y se pasa a P6. Ejemplo: En el nodo preliminar 1 obtenido por unión los estados son 0 y 1. El nodo ancestral inmediato es el nodo final 5 que posee los estados 1 y 3. Por lo tanto, el nodo final 4 posee los estados 0, 1 y 3, y se pasa a P6.

P5. Se agrega al conjunto de estados del nodo los estados que están presentes en el ancestro inmediato y al menos en uno de los dos descendientes inmediatos y se pasa a P6. Ejemplo: En el nodo preliminar 5 el estado es 1. En el nodo ancestral inmediato es el nodo final 6 que posee los estados 1 y 3. Los nodos descendientes inmediatos son los nodos preliminares 1 y 4, que poseen los estados 1 y 0, y 1, 2 y 3. Por lo tanto, se agrega al nodo 5 el estado 3, y se pasa a P6.

P6. Se finaliza con ese nodo y se desciende al siguiente nodo analizado en la fase preliminar, comenzando en P1.

Utilizando este método se puede asignar de manera inequívoca a cada nodo un estado de carácter. Cuando existen diferencias entre el ancestro y el descendiente inmediato, se considera que en algún lugar de la rama ocurrió un evento de cambio del estado del carácter. Sin embargo, en algunos casos no es posible definir el estado del carácter del nodo, quedando *indefinido* y considerando que ambas posibilidades son igualmente probables.

Método empírico de Bayes

Este método fue usado en el marco de esta tesis para la reconstrucción de secuencias ancestrales. Se utilizó el paquete PAML4 versión 4.4 (Yang, 2007) con el modelo evolutivo de WAG (Whelan y Goldman, 2001) e incorporando la distribución Gamma para tasas de reemplazo variables en los distintos sitios (véase Sección 2.3.1). El modelo evolutivo de WAG es una matriz donde cada elemento de la matriz está dada por la probabilidad de cambio entre cada par de aminoácidos. Esta matriz fue construida a partir de 3905 secuencias de 182 familias de proteínas. El programa CODEML incluido en el paquete PAML permite la reconstrucción de secuencias ancestrales aplicando el método empírico de Bayes (Yang *et al.*, 1995) que se explica a continuación.

Dadas las secuencias de aminoácidos extantes, el árbol filogenético del organismo y dada una tasa de sustitución variable para cada posición modelada por la distribución de probabilidad continua Gamma. El objetivo es poder asignar un aminoácido cada posición de las secuencias situadas en los nodos internos del árbol.

Una posición de una secuencia en un alineamiento de n secuencias puede describirse como $x_i = (x_1, x_2, x_3, ..., x_n)$ donde x_i es el aminoácido en la secuencia extante i en el sitio que está siendo analizado como se muestra en la Figura 2.19 para el sitio m extante. De la misma forma, esa posición para las n secuencias ancestrales de los n nodos internos puede describirse como $y_j = (y_1, y_2, y_3, ..., y_n)$ donde y_j es el aminoácido asignado a ese mismo sitio en el nodo interno j del árbol como se muestran en la Figura 2.19 para el sitio m ancestral. El objetivo es estimar y cuando se conoce x.

Para esto, se puede construir una distribución de probabilidades de los distintos aminoácidos en cada posición de cada secuencia ancestral y asignar a cada sitio el aminoácido que maximiza la probabilidad. Para estimar las probabilidades se utiliza un modelo empírico de sustitución de aminoácidos, en este caso se utiliza el modelo evolutivo WAG (Whelan y Goldman,

2001). Los parámetros del modelo son las longitudes de las ramas, que se pueden escribir como $\sigma = (t_1, t_2, t_3, ..., t_l)$.

Se puede calcular la probabilidad de los datos observados, es decir, las secuencias extantes, x, como la suma de todas las posibilidades de y, es decir:

$$f(x;\sigma) = \sum_{y} f(y) f(x \mid y;\sigma) \tag{2.38}$$

donde $f(y)$ representa la probabilidad *a priori* de y y $f(x \mid y;\sigma)$ es la probabilidad condicional de obtener x dado y. $f(x;\sigma)$ significa que f es una función de x con los parámetros σ. Utilizando la Ecuación 2.38 en la construcción que se muestra en la Figura 2.19 para el sitio m:

$$f(x;\sigma) = \sum_{y5}\sum_{y6}\sum_{y7}[\pi_{y_5} P_{y_5 y_6} t_6 \; P_{y_6 y_7} t_4 \; \times P_{y_5 x_4} t_5 \; \times P_{y_6 x_3} t_3 \; \times P_{y_7 x_1} t_1 \; P_{y_7 x_2} t_2] \tag{2.39}$$

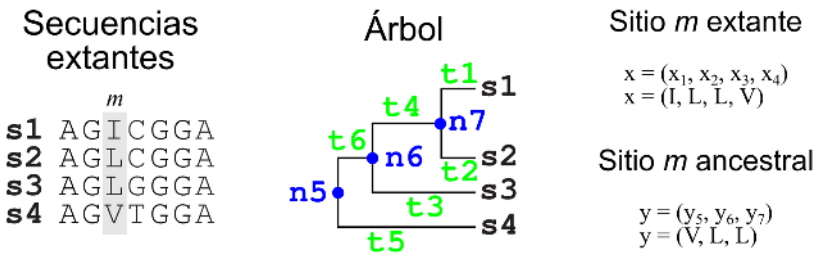


Figura 2.19: Ejemplo para el cálculo de probabilidades *a priori* y *posteriori*.

El resultado de lo que están contenido entre los corchetes es la probabilidad de observar el aminoácido y_5 en el nodo 5, y es la frecuencia de equilibrio π_{y_7} del aminoácido y_7 multiplicado por las probabilidades de transición de las seis ramas del árbol. σ es estimado utilizando el método de máxima verosimilitud (véase Sección 2.3.1) y es el producto de $f(x, \sigma)$ sobre todos los sitios asumiendo que los sitios son independientes.

Cuando el objetivo es la reconstrucción de los sitios ancestrales y, se estudia la probabilidad condicional de cada sitio y dado x, es decir:

$$f(y \mid x;\sigma) = \frac{f(y) f(x \mid y;\sigma)}{f(x;\sigma)} \tag{2.40}$$

como el aminoácido y es una variable discreta, se usa el aminoácido y estimado que maximiza la probabilidad condicional, reemplazando σ por el estimado por máxima verosimilitud. $f(y \mid x;\sigma)$ representa la probabilidad *a posteriori* y es una medida de cuan buena es la reconstrucción de los sitios cuando los datos de entrada son x, siendo esta la interpretación bayesiana.

Cuando se estima la probabilidad de un aminoácido en un sitio no se incluyen aquellos aminoácidos que no están presentes en ese sitio en las secuencias extantes, ya que dicha probabilidad es muy baja y el cálculo incluyendo todos los aminoácidos posibles implicaría una recons-

trucción de un número de secuencias que crece exponencialmente con el número de nodos internos en el árbol.

Finalmente, la probabilidad posterior de la asignación de un aminoácido a un nodo determinado es la suma de la contribución de la probabilidad de los datos observados para ese sitio sobre todas las reconstrucciones que asignan ese mismo aminoácido a ese mismo nodo, eligiendo entre los 20 aminoácidos posibles el que tiene la mejor probabilidad posterior.

2.3.5. Cofilogenia entre parásito y hospedador

Según la regla de Fahrenholz, la estrecha relación existente entre hospedadores y parásitos tiene como consecuencia que la filogenia del parásito refleje la filogenia del hospedador. En este escenario hay coespeciación de hospedador y parásito y se asume que no existe una dispersión de parásitos entre hospedadores no relacionados. La realidad en la mayoría de los estudios es que la relación entre la filogenia del parásito y la filogenia del hospedador es mucho más compleja y la primera no siempre refleja la segunda. Muchas veces la relación no es uno a uno y existen muchas especies de hospedador por parásito, o distintos hospedadores que comparten especies de parásitos. Esto indica que pueden existir otros eventos evolutivos distintos de la coespeciación, que llevan a los distintos patrones filogenéticos. Los numerosos métodos para realizar estudios de cofilogenia se pueden clasificar en dos categorías principales: métodos de ajuste global y métodos basados en eventos (Desdevises, 2007).

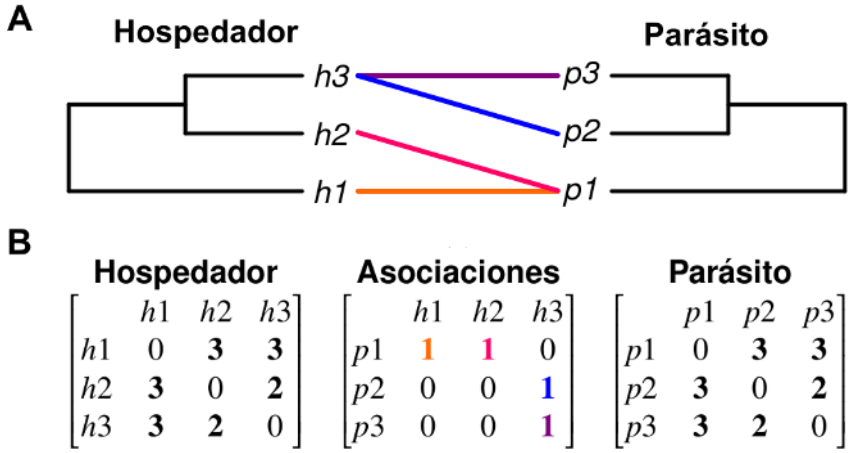
Métodos de ajuste global

El objetivo principal de los métodos de ajuste global es cuantificar el grado de congruencia entre dos topologías dadas e identificar la contribución de las asociaciones actuales parásito-hospedador a la estructura codivergente. De esta manera se puede evaluar la importancia de la coevolución en el sistema estudiado. El algoritmo ParaFit evalúa la congruencia global de las dos filogenias y la contribución de cada asociación hospedador-parásito a dicha congruencia usando matrices de distancias (Legendre *et al.*, 2002). En este trabajo de tesis, se utilizó este algoritmo que está incluido en el programa CopyCat versión 2.04 (Meier-Kolthoff *et al.*, 2007).

Congruencia global. Para poder evaluar la congruencia global es necesario combinar tres tipos de información (Figura 2.20): (1) la filogenia del parásito (Figura 2.20A, derecha), (2) la filogenia del hospedador (Figura 2.20A, izquierda) y (3) las asociaciones hospedador-parásito observadas (Figura 2.20A, centro). Para esto, se describen las filogenias como matrices de distancia (Figura 2.20B).

En este trabajo, el objetivo es evaluar la congruencia entre las topologías del árbol, por lo que se considera que todas las ramas tienen longitud 1. Las dimensiones de la matriz del hospedador son $m \times m$, donde m es el número de hospedadores presentes en la filogenia del hospedador. Las dimensiones de la matriz del parásito son $n \times n$, donde n es el número de parásitos en la filogenia del parásito. También se expresan en forma de matriz las asociaciones hospedador-parásito observadas.

Las dimensiones de esta matriz son n x m, es decir, tendrá tantas filas como hospedadores haya y tantas columnas como parásitos haya, y se utiliza un 1 o un 0 para indicar si se observó o no, respectivamente, una asociación hospedador-parásito (Figura 2.20B, centro).



Hospedador	h1	h2	h3
h1	0	**3**	**3**
h2	**3**	0	**2**
h3	**3**	**2**	0

Asociaciones	h1	h2	h3
p1	**1**	**1**	0
p2	0	0	**1**
p3	0	0	**1**

Parásito	p1	p2	p3
p1	0	**3**	**3**
p2	**3**	0	**2**
p3	**3**	**2**	0

Figura 2.20: Representación de los árboles filogenéticos y asociaciones ente hospedador y parásito para analizar por el método global la coespeciación. (a) Árboles filogenéticos y asociaciones. *Izquierda.* Árbol filogenético del Hospedador con tres especies *h1*, *h2* y *h3*. *Derecha.* Árbol filogenético del Parásito con tres especies *p1*, *p2* y *p3*. *Centro.* Asociaciones existentes entre hospedador y parásito. (b) Matrices de distancia y asociación. *Izquierda.* Matriz de distancia correspondiente al árbol filogenético del hospedador en (a). *Derecha.* Matriz de distancia correspondiente al árbol filogenético del parásito en (a). *Centro.* Matriz de Asociación Hospedador-Parásito. En las filas, *i*, se indican las especies del parásito y en las columnas, *j*, las especies del hospedador. Un 0 en la posición *ij* indica que el parásito *i* no infecta al hospedador *j*, mientras que un 1 indica lo contrario. Los colores utilizados para resaltar los 1 en la matriz de asociación se corresponden con las asociaciones mostradas en (a).

Las matrices de distancia del hospedador y el parásito son transformadas mediante un análisis de coordenadas principales (Buneman, 1974) en matrices que tendrán m o n filas respectivamente y, a lo sumo, $m - 1$ o $n - 1$ columnas. La matriz de asociación hospedador-parásito entonces puede ser descrita por una nueva matriz D, que incluye la filogenia del hospedador y del parásito según:

$$D = H\ A'\ P \tag{2.41}$$

donde H es la matriz del hospedador transpuesta luego del análisis de coordenadas principales, la matrix A' es la matriz de asociación de la Figura 2.20B y P es la matriz del parásito transformada mediante el análisis de coordenadas principales.

La matriz D es el punto de partida para realizar la prueba de congruencia global. La hipótesis nula de esta prueba es:

> H_0 *Global = Dadas la filogenia del hospedador y del parásito, y las asociaciones observadas, la evolución de los dos grupos fue independiente*

El estadístico a utilizar se define como la suma de los cuadrados de los elementos de la matriz,

que equivale a la traza de la multiplicación de la matriz transpuesta D consigo misma:

$$ParaFitGlobal = tr(D'D) = \sum_{ij} d_{ij}^2 \tag{2.42}$$

La distribución del estadístico se construye mediante la aleatorización de la matriz de asociación A, ya que es el único factor supuestamente aleatorio.

Contribución individual. ParaFit analiza además la contribución de cada asociación parásito-hospedador al ajuste global. La hipótesis nula de esta prueba es:

> H_0 *Individual = Cualquier contribución de cada asociación hospedador-parásito individual al ajuste global no es distinta de la obtenida al azar, y por ende podría ser omitida.*

Para esta prueba se utilizan dos estadísticos, *ParaFitLink1* y *ParaFitLink2*, que se basan en la idea de que el valor del estadístico global disminuye si retiramos una asociación significativa de la matriz de asociación A. El estadístico *ParaFitLink1* define la contribución de cada asociación parásito-hospedador al ajuste global observado. El estadístico *ParaFitLink2* define cuanto aporta cada asociación a la congruencia máxima.

Dada una asociación k entre parásito y hospedador, reemplazamos en la matriz A el 1 que representa esa asociación por un 0 y llamamos $A(k)$ a la nueva matriz. Construimos una nueva matriz $D(k)$ a partir de las matrices H, P y $A(k)$, según:

$$D(k) = H\ A(k)'\ P \tag{2.43}$$

y calculamos de manera similar a la Ecuación 2.42 el valor de $t(k)$, que sería la contribución de todas las asociaciones a la congruencia global, menos la asociación k.

$$t(k) = \sum_{ij} d_{ij}^2 \tag{2.44}$$

Finalmente, en base a lo dicho antes, el estadístico para evaluar la contribución de la asociación k a la congruencia global se calcula como:

$$ParaFitLink1(k) = ParaFitGlobal - t(k) \tag{2.45}$$

El segundo estadístico está dado por:

$$ParaFitLink2(k) = \frac{ParaFitLink1(k)}{t_{Max} - ParaFitGlobal} \tag{2.46}$$

El numerador corresponde al estadístico de la Ecuación 2.45. El denominador mide la diferencia entre los datos y el modelo estimado y es análogo a la suma de los cuadrados de los residuales. En este caso, *ParaFitGlobal* se corresponde con los datos y t_{Max} con el modelo estimado. En

el modelo estimado, los árboles filogenéticos del hospedador y parásito son completamente congruentes, es decir, la congruencia es la máxima posible. Esta situación da el valor máximo de traza en la matriz D, que en ese caso es la suma de los cuadrados de los autovalores de las coordenadas principales que se encuentran en las matrices H o P. Por lo tanto:

$$t_{Max} = max\left(\sum \lambda_H^2 , \sum \lambda_P^2\right) \tag{2.47}$$

donde λ_H y λ_P son los autovalores de las matrices H y P respectivamente.

La distribución de ambos estadísticos se construye de la misma manera que antes, es decir, mediante la aleatorización de la matriz de asociación *A(k)*. Cuando los resultados de la prueba global y de ambos estadísticos individuales son significativos, la prueba detecta una asociación significativa. Es decir, existe una coespeciación global y dicha asociación aporta de manera significativa a la coespeciación.

Métodos basados en eventos

Los métodos basados en eventos buscan encontrar la historia cofilogenética más probable entre parásito y hospedador. Para esto se requiere reconstruir la historia de las asociaciones entre parásito y hospedador a partir de las asociaciones actuales, ϕ, observadas entre las hojas del árbol del hospedador, H, y las del parásito, P, dado que no existen datos sobre las asociaciones ancestrales (Ronquist, 1995). El problema de la reconstrucción de la cofilogenia no es trivial y pertenece al conjunto de problemas NP-Completo (Ovadia *et al.*, 2011). Esto implica que aún no se ha encontrado un algoritmo que lo resuelva en tiempo polinomialmente proporcional al tamaño de la entrada (H y P) y es probable que no exista tal algoritmo que encuentre una solución óptima. Una manera de resolver estos problemas es utilizando métodos heurísticos que construyan una solución aproximada, aunque no sea óptima. Este problema está aún en estudio y existen numerosos algoritmos computacionales que aproximan a una solución. Dicha reconciliación de la topología de los árboles de parásito y hospedador se encuentra solucionando las incongruencias mediante eventos evolutivos. La mayoría de los algoritmos consideran cuatro eventos evolutivos (Page, 1994): (A) Coespeciación, (B) duplicación, (C) duplicación con cambio de hospedador y (D) extinción parcial. Cada uno de los eventos se explican a continuación y se detallan en la Figura 2.21 donde están representados de tres maneras diferentes: (1) enfrentada donde se colocan enfrentados el árbol del hospedador y el árbol del parásito (Figura 2.21, primera columna), se grafican las asociaciones entre ellos y se marcan en los nodos de cada uno de los árboles los distintos eventos evolutivos que se observan en cada árbol como ser especiación (círculos blancos con borde rojo) o duplicación (círculos negros) o una rama truncada cuando ocurre una extinción parcial (cruz); (2) superpuesta se superpone el árbol del parásito sobre el árbol del hospedador y se marcan en cada uno de los nodos y ramas del árbol del parásito los eventos evolutivos correspondientes que reconcilian ambas topologías, como ser coespeciación (círculos blancos con borde rojo), duplicación (círculos negros), cambio de hospedador (triángulo azul) y extinción parcial (línea punteada);

y (3) proyección donde se marcan los eventos evolutivos sobre el árbol del parásito con su hospedador correspondiente al lado, como ser coespeciación (círculos blancos con borde rojo y rama roja), duplicación (círculos negros y rama negra), cambio de hospedador (triángulo azul y rama azul) y extinción parcial (línea punteada). Esta última es la representación utilizada a lo largo de esta tesis. A continuación se describen los eventos evolutivos y la interpretación correspondiente de cada una de las representaciones presentadas en la Figura 2.21.

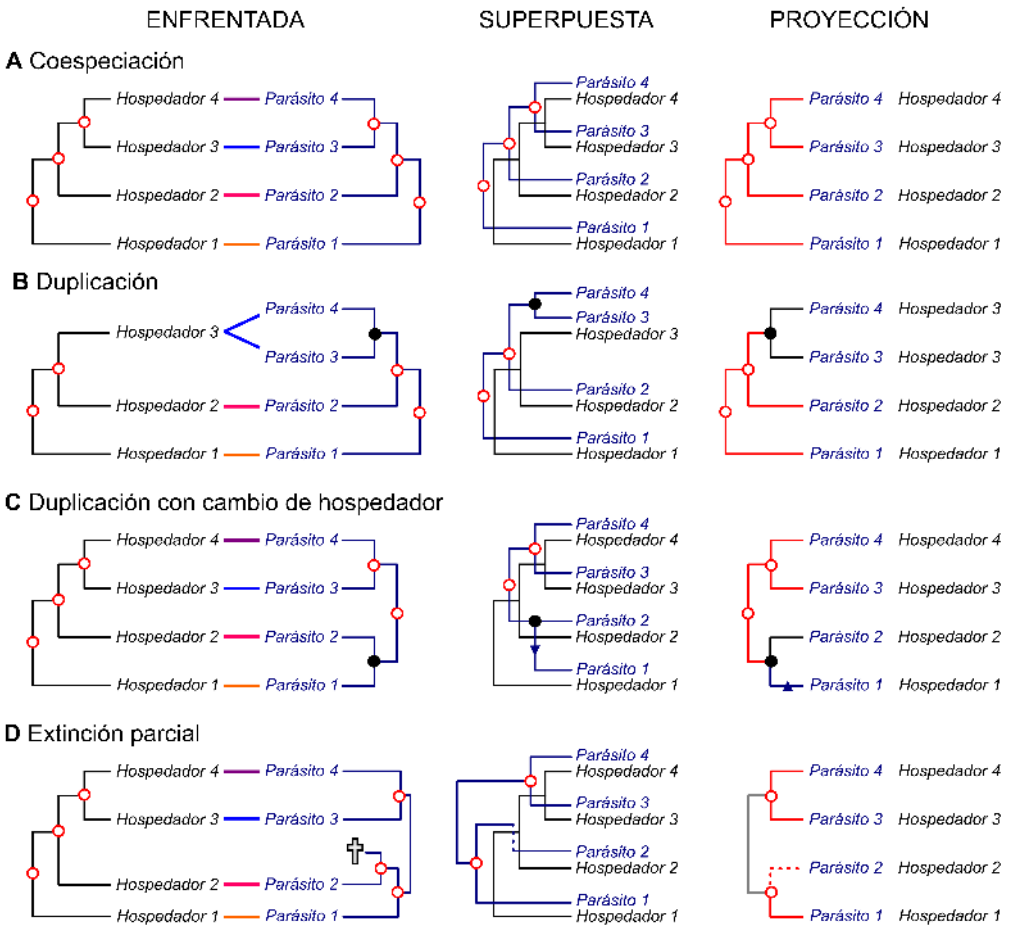


Figura 2.21: Representación de los eventos evolutivos utilizados por los métodos basados en eventos. Se representan los cuatro eventos evolutivos necesarios para la reconciliación de los árboles del parásito y hospedador, (A) coespeciación, (B) duplicación, (C) duplicación con cambio de hospedador y (D) extinción parcial. En la primera columna se muestra a la representación enfrentada, con el árbol del hospeador a izquierda, el árbol del parásito a la derecha y las asociaciones entre ellos en el centro. En la segunda columna se muestra la representación superpuesta del árbol del parásito (en azul) y el árbol del hospedador (en negro). En la tercera columna se muestra la proyección de los eventos evolutivos en el árbol del parásito. La línea gris en (D) indica que no superpone con el árbol del hospedador.

A. Coespeciación: Es un evento de especiación en el árbol del parásito que está asociado a un evento de especiación en el árbol del hospedador (Figura 2.21A). Es decir, parásito y hospedador especian de manera casi simultánea. Implica la existencia de un parásito por hospedador. En la representación enfrentada se puede ver que cada vez que ocurre una especiación en el hospedador lo mismo ocurre en el parásito (círculos blancos con borde rojo). Cuando ocu-

rren únicamente eventos de coespeciación entre parásito y hospedador, en la representación superpuesta se observa que los árboles del parásito (azul) y hospedador (negro) son totalmente congruentes. En la proyección se observa los tres eventos de coespeciación en el árbol del parásito marcados sobre las ramas (círculos blancos con línea roja y ramas rojas). La decisión de considerar el evento sobre las ramas se basa en asumir que el proceso de especiación no es inmediato o posible de definir en un nodo. Se considera que alguno de los cambios que se producen en el genoma del parásito a lo largo de la rama producen un cambio en el fenotipo que lleva a la especiación del parásito.

B. Duplicación: Es un evento de especiación en el árbol del parásito, tras el que ambos descendientes continúan asociados al mismo hospedador (Figura 2.21B). Es decir, el parásito especia de manera independiente del hospedador. Implica la existencia de más de un parásito para un mismo hospedador. En la representación enfrentada se observa que en la base de cada uno de los árboles ocurren dos eventos de coespeciación (círculos blancos con borde rojo), mientras que los parásitos 3 y 4 divergen (círculo negro), este evento de divergencia no está reflejado en el árbol del hospedador porque ambos parásitos están asociados al hospedador 3. En la representación superpuesta se observa en el árbol del parásito (azul) que las ramas que dan origen a los parásitos 3 y 4 se superponen bien en el árbol del hospedador (negro) con la rama que da origen al hospedador 3. En la proyección, se señala el inicio del evento de duplicación en el nodo (círculo negro) y sobre las ramas (ramas negras) porque no es posible definir en qué momento exacto ocurre la duplicación que lleva a la divergencia de los parásitos 3 y 4.

C. Duplicación con cambio de hospedador: Es un evento de duplicación en el árbol del parásito, tras el que uno de los dos descendientes del parásito sigue la rama del hospedador original mientras que el otro descendiente cambia a una rama distinta en el árbol del hospedador (Figura 2.21D). En la representación enfrentada se puede observar que en el árbol del parásito hay una divergencia entre el parásito 1 y 2 (círculo negro) que no ocurre en el árbol del hospedador. Este evento es mucho más fácil de observar en la superposición de los dos árboles, donde se observa que en el árbol del parásito (azul) las ramas de los parásitos 2, 3 y 4 coinciden en el árbol del hospedador (negro) con las ramas de sus respectivos hospedadores mientras que la rama del parásito 1 necesita ser reconciliada introduciendo el evento evolutivo de cambio de hospedador (triángulo azul). En la proyección se representan el evento de duplicación que ocurre en algún momento entre el nodo y el origen del parásito 2 (círculo negro y rama negra), y el evento de cambio de hospedador que ocurre en algún momento luego de la duplicación y origen del parásito 1 (triángulo azul y rama azul).

D. Extinción parcial: Es un evento de especiación en el árbol del hospedador que no se ve reflejado en el árbol del parásito (Figura 2.21C). Es decir, el hospedador especia de manera independiente y el parásito sigue una de las ramas. En la representación enfrentada se puede observar los tres eventos de especiación que coinciden entre el árbol del parásito y el árbol del hospedador. Sin embargo, una de las ramas a las cuales da origen el evento de especiación que origina el

parásito 2 se extingue. Este evento evolutivo es mucho más fácil de comprender observando la superposición del árbol del parásito sobre el árbol del hospedador. Cuando ocurre una especiación en el árbol del hospedador (negro) que da origen a los hospedadores 2 y 3, en el árbol del parásito (azul) está presente la rama del parásito 2 asociada al hospedador 2 pero la otra rama no está (línea punteada). En la proyección sobre el árbol del parásito se observa que existe una extinción parcial entre el evento de coespeciación y el origen del parásito 2 (línea punteada).

Es importante resaltar que la topología de los árboles del parásito y las asociaciones entre los parásitos y hospedadores de los ejemplos utilizados para los eventos duplicación con cambio de hospedador y extinción parcial (Figura 2.21D y 2.21C) son iguales. Se eligieron para ejemplificar como una misma topología puede reconciliarse utilizando distintos eventos evolutivos.

Uno de los objetivos del presente trabajo fue poner a punto un protocolo de análisis de cofilogenia parásito-huésped en el laboratorio. Se mencionan en adelante los algoritmos considerados y las dificultades encontradas en cada caso.

El algoritmo TreeFitter (Ronquist, 1995) realiza la reconciliación de árboles utilizando parsimonia. Se asigna un costo a cada uno de los cuatro eventos y se busca la reconstrucción óptima que minimiza el costo global. Sin embargo, la versión más utilizada de TreeFitter 1.1 solo puede utilizarse en un número limitado de plataformas, algunas de las cuales ya no están disponibles, y no se pudo acceder a dicha versión en el sitio indicado en la literatura (`http://www.ebc.uu.se/systzoo/research/treefitter/treefitter.html`) ni conseguirla luego de varios intentos de comunicación con los autores. En el presente trabajo se consiguió e intentó utilizar la versión 1.3b de TreeFitter en un entorno de Linux-Ubuntu versión 12.04 con un procesador i5 y 8 Gb de RAM. TreeFitter 1.3b, disponible en `https://sourceforge.net/p/treefitter/wiki/Home/`, pero presentó errores en la lectura de los árboles no permitiendo ingresar árboles de gran tamaño (Figura 2.22, entrada 4) ni repetir el ingreso de alguno de los árboles que se pudieron ingresar previamente (entrada 3 y entrada 7). El error fue comunicado a los autores del programa sin respuesta y consultado con el grupo de trabajo del Dr. Bravo (Bravo *et al.*, 2010) quienes pudieron reproducir los mismos errores.

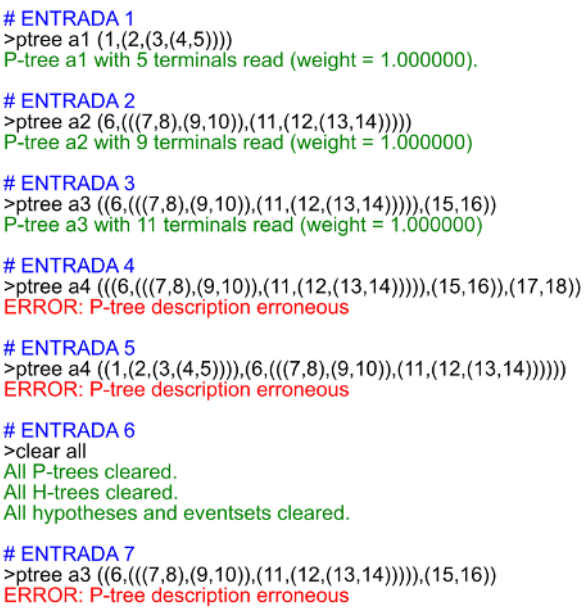


Figura 2.22: Error en TreeFitter versión 1.3b. Texto correspondiente a la terminal utilizando TreeFitter versión 1.3b. En negro se indican los comandos ingresados. Se indican las salidas del programa TreeFitter cuando el árbol o comando ingresado es correcto (verde) o cuando TreeFitter lo considera incorrecto (rojo). En azul se indican comentarios que no son ingresados en la línea de comando.

TreeMap utiliza un algoritmo denominado Jungles (Charleston y Robertson, 2002), cuyo tiempo de resolución crece exponencialmente con el número de hojas de los árboles. En este trabajo se intentó utilizar la versión 3 disponible en `https://sites.google.com/site/cophylogeny/software` en un entorno de Linux-Ubuntu 12.04 con un procesador i5 y 8 Gb de RAM. Luego de una semana de corrida no se obtuvieron resultados cuando el programa se cerró abruptamente sin producir resultados. Considerando que esto podía ser un problema debido al gran tamaño del árbol, se utilizó un árbol más pequeño colapsando aquellas ramas del árbol del parásito que correspondían a grupos monofiléticos que infectaban un mismo hospedador, reduciendo considerablemente el tamaño del árbol del parásito. Nuevamente, luego de una semana de corrida, el programa se cerró abruptamente sin producir resultados.

Finalmente, se tuvo éxito con el software Jane versión 4.0 (Conow *et al.*, 2010) disponible en `https://www.cs.hmc.edu/~hadas/jane/`. Jane sortea la complejidad computacional de la reconstrucción de la cofilogenia mediante un algoritmo de programación dinámica partiendo de una búsqueda heurística. Dada una secuencia temporal de los nodos en el árbol del hospedador, dicho algoritmo encuentra soluciones pareto óptimas en tiempos polinomiales. Dada una asignación inicial de eventos a cada una de las ramas, un cambio hacia una nueva asignación que al menos mejora la reconciliación de una de las ramas sin hacer que empeore la reconciliación del resto de las ramas se la denomina mejora de Pareto. Una solución es pareto óptima si un cambio en el número de eventos o la ubicación de los mismos no produce una mejora en la reconciliación de las filogenias. En concreto, Jane parte de una búsqueda heurística y combina un algoritmo genético y un algoritmo de cruzamientos para generar un subconjunto de soluciones posibles. Los costos

de cada evento evolutivo pueden ser elegidos por el usuario, así como el tamaño poblacional, el número de generaciones, la tasa de mutación de nodos y la fuerza de selección.

Algoritmo genético. El árbol del parásito se toma como fijo. En primer lugar, Jane calcula un conjunto de secuencias temporales para los nodos en el árbol del hospedador, es decir, determina el orden temporal posible para cada evento de divergencia. El cojunto de secuencias temporales es de tamaño S (análogo al tamaño poblacional). Para cada una de las secuencias temporales se resuelve la reconstrucción de manera óptima y se calcula su costo, o en términos genéticos, el *fitness*. A continuación, se le asigna una probabilidad a cada secuencia temporal, pesada de manera exponencial según el *fitness*. Se eligen dos secuencias temporales (τ_1, τ_2) al azar, con repetición. Luego, se construye una nueva secuencia temporal, τ_{new} a partir de τ_1 y τ_2 con elementos de cada uno, utilizando el algoritmo de cruzamiento descrito más abajo. Estos pasos se repiten según las iteraciones (o generaciones) elegidas por el usuario. Finalmente, se reporta el mejor grupo de soluciones obtenidas en la última iteración.

Algoritmo de cruzamiento. Para crear una nueva secuencia temporal, τ_{new}, se seleccionan al azar dos secuencias temporales, τ_1 y τ_2. A partir del árbol del hospedador se selecciona al azar un sub-árbol, T. Luego se construye τ_{new} seleccionando el tiempo relativo de los nodos internos a partir de los tiempos combinados de τ_1 y τ_2. Para esto se listan en orden temporal los nodos de τ_1 sin considerar el sub-árbol T, los nodos de τ_2 solo considerando el sub-árbol T y se crea una lista vacía para τ_{new}. Para asignar los tiempos en τ_{new}, se respeta el orden relativo de los tiempos de τ_1 y τ_2, siempre y cuando los nodos parentales ya hayan sido asignados. Cuando dos nodos pueden ser asignados a un mismo tiempo, el algoritmo elije al candidato cuyo tiempo original es más cercano al tiempo bajo consideración. Si la distancia es la misma, se elije uno al azar. Para obtener variaciones adicionales, se utilizan *mutaciones al azar* en algunas secuencias temporales. Es decir, se cambia el orden de dos nodos que ocurren en tiempos consecutivos y no tienen relación ancestro-descendiente.

Parámetros. El *algoritmo genético* tiene los siguientes parámetros que pueden ser seleccionados por el usuario:

- **Costos:** Son específicos de cada evento evolutivo. Los costos utilizados para cada evento en este trabajo se discuten en la sección de resultados.
- **Tasa de mutación:** Determina con qué frecuencia se mutan los tiempos de nodos internos. Este parámetro puede variar entre 0 y 1, donde 0 significa que nunca se realiza una mutación y 1 implica una mutación luego de cada iteración. Se utilizó el valor por defecto, 0.8.
- **Fuerza de Selección:** Determina cuánto se confía en la función de *fitness*. Varía entre 0 y 1, donde 0 implica que la elección de las secuencias temporales no está pesada por la función

de fitness y 1 significa que siempre se eligen las mejores secuencias temporales de cada iteración. Se utilizó el valor por defecto, 0.6.

- **Tamaño Poblacional:** Es el número de soluciones diferentes a ser consideradas en cada iteración del algoritmo. Se utilizó el valor por defecto, 100.
- **Número de Generaciones:** Es el número de iteraciones realizadas por el algoritmo. Se utilizó el valor por defecto, 100.

Reporte de soluciones. Jane brinda un conjunto de soluciones que comparten el mismo costo total calculado a partir de los costos individuales asignados a cada evento evolutivo. Las soluciones están agrupadas en soluciones isomórficas, esto es un conjunto de soluciones que poseen los mismos eventos evolutivos, es decir son técnicamente iguales, pero difieren en el espacio temporal ya que cambia la longitud de las ramas de los árboles. Para este trabajo de tesis por lo tanto se utiliza una solución por cada conjunto de soluciones isomórficas. Jane asigna un valor de soporte a cada evento evolutivo que aparece en una determinada posición en el árbol del parásito. Este valor de soporte se calcula como el porcentaje de ver ese mismo evento en esa misma posición entre todas las soluciones encontradas. Por lo tanto, se seleccionaron aquellas soluciones con los valores de soporte más altos en la mayoría de los eventos.

Problemas encontrados en la ejecución de Jane. El primer problema que presenta la utilización de Jane es el tiempo de ejecución. Para solucionar este problema se utilizó un árbol de *Mastadenovirus* más pequeño con las ramas que corresponden a grupos monofiléticos que infectan a un mismo hospedador colapsadas y se hace referencia al mismo como árbol colapsado. Jane se puede correr en modo gráfico o por línea de comandos. Sin embargo, la línea de comandos proporciona una única solución y no el subconjunto de soluciones obtenido. El segundo problema es que si bien el modo gráfico ofrece la opción de guardar cada una de las soluciones de manera individual en modo texto, su implementación no es correcta y recupera en cada caso la misma solución con los costos por defectos que ofrece la ejecución por línea de comando. Este problema fue reportado al Dr. Hadas director del grupo de trabajo autor de Jane (Conow *et al.*, 2010). Por lo tanto, se guardó cada una de las soluciones en formato gráfico y se tradujo cada una a modo texto.

Cada una de los soluciones se representó como una proyección sobre el árbol de *Mastadenovirus*, asumiendo que dentro de las ramas colapsadas el único evento que ocurría era el de duplicación. De esta manera se obtiene para cada rama un evento evolutivo asociado.

2.4. Correlación entre rasgos moleculares y fenotípicos

Los motivos lineales reportados para cada proteína y los rasgos fenotípicos de cada secuencia, tropismo, patología y hospedador, así como los blancos proteicos, fueron obtenidos y curados manualmente de la literatura disponible.

2.4.1. Búsqueda de blancos proteicos

La búsqueda de blancos proteicos para la proteína E1A se realizó en noviembre 2013 utilizando la siguiente combinación de palabras claves:

```
(\binding"[All Fields] OR (\interaction"[All
Fields] OR (\target"[All Fields] OR (\partner"[All
Fields])))) AND ((e1a[All Fields] OR (\early
protein"[All Fields]))) AND (\adenoviridae"[All
Fields] OR (\adenovirus"[All Fields]))
```

Esta búsqueda arrojo más de 1600 resultados. Se realizó una limpieza en base a la lectura de los resúmenes de cada artículo y los resultados disminuyeron a 290 artículos. Las interacciones reportadas en la literatura fueron clasificadas en base a si la metodología experimental correspondía a evidencia directa de interacción o indirecta de interacción. Los métodos de evidencia de interacción directa, son por ejemplo, complementación biomolecular de fluorescencia (BiFC) (en inglés, *biomolecular fluorescence complementation*), titulación isotérmica calorimetría, titulación en RMN, ensayos de competición, ensayo de inmunoabsorción ligado a enzimas (ELISA), cristalización, coprecipitacion y coinmunoprecipitación *in vitro*, entre otros. Los métodos de interacción que no permiten determinar si la interacción es directa o está mediada por una tercera proteína son, la prueba de doble híbrido en levadura, coprecipitación y coinmunoprecipitación partiendo de un extracto celular, y colocalización, entre otras. De esta manera, el número de artículos disminuyó a 91 y proporcionó un total de 68 blancos proteicos de la proteína E1A.

2.4.2. Recolección de datos fenotípicos

Los datos fenotípicos, patología, tropismo y hospedador se recolectaron para cada secuencia partiendo del paper original dónde es reportada la secuencia.

2.5. Analisis estadísticos

2.5.1. Prueba hipergeométrica de asociación

Una forma de evaluar la asociación entre dos atributos es utilizar una prueba hipergeométrica (Rivals *et al.*, 2007). Dada una población de N individuos, de los cuales $x \leq N$ poseen un atributo A, $y \leq N$ poseen un atributo B, $z \leq min(x, y)$ poseen ambos atributos A y B. La función hipergeométrica permite calcular la probabilidad de obtener z casos de éxito cuando se retiran sin reemplazo un número y de objetos de una población de tamaño N, dado que los casos totales de éxito de la población son x. Si existe una asociación positiva, el valor p se define como la suma de la probabilidad de tener z o más casos de éxito.

Este prueba se aplicó para determinar qué asociaciones eran significativas entre las co-ocurrencias de pares de motivos, motivo-hospedador y motivo-tropismo, y entre las asociaciones

de eventos de aparición/desaparición de motivos y entre las asociaciones de eventos de aparición/desaparición de motivos y eventos evolutivos.

2.5.2. Corrección de Benjamini-Hochberg para comparaciones múltiples

Si se considera un nivel de significancia, o valor crítico, de 0.05 eso implica que existe un 5 % de probabilidad de rechazar la hipótesis nula siendo verdadera. Esta también se conoce como probabilidad de falso positivo (error tipo I), es decir, estamos definiendo la proporción de falsos positivos que toleramos. Por ejemplo, un valor de 0.05 de significancia, significa que de 20 pruebas significativas una podría ser un falso positivo.

Cuando en un análisis estadístico se comparan múltiples características de una misma muestra, a medida que se aumenta el número de características o atributos asociados entre sí a comparar aumenta la probabilidad de observar diferencias significativas simplemente por un error de muestreo o azar. Es decir, aumenta el error de tipo I. Si se realizan k comparaciones independientes, la probabilidad de que al menos ocurra un falso positivo (FWER) (en inglés, *family-wise error rate*) viene dada por la ecuación:

$$\text{FWER} = 1 - (1 - \alpha)^k \tag{2.48}$$

El método de Bonferroni controla la probabilidad de tener al menos un falso positivo. En lugar de utilizar un valor crítico de 0.05, se utiliza el valor crítico elegido penalizado por el número de comparaciones realizadas. Se obtiene por lo tanto un valor crítico menor aumentado la probabilidad de no rechazar la hipótesis nula siendo falsa en la población. Es decir, este método aumenta la probabilidad de obtener falsos negativos o error de tipo II.

Otra forma de realizar esto es controlando la tasa de descubrimientos de falsos positivos (FDR) (en inglés, *false discovery rate*) en lugar de la FWER. La FDR es la proporción esperada de falsos positivos entre las pruebas significativas. El objetivo de este control es que el número de falsos positivos no supere determinado valor. Esto es una prueba menos conservativa que el control de FWER. Uno de los métodos que controla la FDR es la corrección de Benjamini-Hochberg para comparaciones múltiples (Benjamini y Hochberg, 1995).

Esta corrección consiste básicamente en corregir los valores p por el número total de comparaciones realizadas y el orden relativo de los valores p obtenidos. Brevemente, los valores p obtenidos para cada comparación individual son ordenados de manera creciente y se le asigna un índice, $i = 1, 2, ..., n$. Luego, el valor p es corregido como:

$$p_i^* = p\frac{m}{i} \tag{2.49}$$

donde p es el valor p individual y m es el número total de comparaciones realizadas. En este trabajo, el número de comparaciones realizadas se indica en cada uno de los casos donde la corrección es aplicada.

Esta corrección fue aplicada cuando se obtuvieron valores p significativos en los análisis estadísticos de asociación incluyendo asociación entre motivos, motivo-tropismo y motivo-hospedador, y las asociaciones entre eventos de aparición/desaparición de motivos y eventos de aparición/desaparición de motivos con eventos evolutivos. De igual manera se utilizó esta corrección en la prueba de permutación realizada para la comparación de la conservación de aminoácidos correspondientes a distintos grupos de las proteínas E1A y E7.

Capítulo 3

Proteína E7

En este capítulo describo y discuto brevemente los resultados obtenidos para la proteína E7 de la familia *Papillomaviridae*.

Los resultados de este capítulo los realicé al comienzo de mi doctorado y se analizaron en conjunto con la Dra. Lucía Chemes del Instituto de Investigaciones Biotecnológicas de la Universidad Nacional de San Martín. Partiendo de resultados previos de la proteína E7 de papilomavirus de la Dra. Chemes se obtuvieron los resultados que definieron las bases para el análisis de la proteína E1A de *Mastadenovirus* que se discute en el próximo capítulo. En relación a la proteína E7 únicamente incluyo y discuto aquellos resultados donde mi participación fue relevante. Los análisis de coevolución de secuencia por información mutua fueron realizados en conjunto con la Dra. Cristina Marino-Buslje del laboratorio de Bioinformática Estructural de la Fundación Instituto Leloir. Mi trabajo como autora en este caso fue proporcionar los datos necesarios y la interpretación de los resultados. Los métodos de predicción de contactos entre residuos evolucionaron durante mi doctorado. En nuestro grupo de trabajo, la Dra. Rocío Espada se dedicó a estudiar la implementación del método de información directa. Por lo tanto, utilizamos este nuevo método en colaboración con la Dra. Espada para corroborar los resultados obtenidos por información mutua. A partir del alineamiento del dominio E7C, el Dr. Leonardo G. Alonso del laboratorio de Estructura-Función e Ingeniería de Proteínas de la Fundación Instituto Leloir observó una alta densidad de cisteínas que derivó en su estudio en esta tesis. Por último, todos los resultados obtenidos fueron discutidos en conjunto con el Dr. Gonzalo de Prat-Gay del laboratorio de Estructura-Función e Ingeniería de Proteínas de la Fundación Instituto Leloir.

Los resultados obtenidos derivaron en dos publicaciones:

- Chemes, L. B., Glavina, J., Alonso, L. G., Marino-Buslje, C., de Prat-Gay, G., y Sánchez, I. E. (2012a). *Sequence evolution of the intrinsically disordered and globular domains of a model viral oncoprotein.* PLoS One, 7(10):e47661

- Chemes, L. B., Glavina, J., Faivovich, J., de Prat-Gay, G., y Sánchez, I. E. (2012b). *Evolution of linear motifs within the papillomavirus E7 oncoprotein.* Journal of molecular biology, 422(3):336–46

3.1. Recoleccion de secuencias de la proteína E7 de la familia *Papillomaviridae*

Los resultados obtenidos en la proteína E7 publicados (Chemes *et al.*, 2012a,b) se obtuvieron al inicio de mi trabajo en el laboratorio en el año 2011. En ese momento se creó la base de datos 1 de la proteína E7 de papilomavirus. Numerosos serotipos virales son reportados cada año, en especial para los virus de relevancia clínica. Hacia el final de mi doctorado realizamos un análisis de coevolución de secuencia junto con la Dra. Rocío Espada, para lo cual actualizamos la base de datos 1 y creamos la base de datos 2 de secuencias de la proteína E7 de papilomavirus.

3.1.1. Creación de la base de datos 1

Se recolectaron un total de 224 secuencias proteicas correspondientes a la proteína E7 de la familia *Papillomaviridae* de la base de datos NCBI en junio 2011 (véase Sección 2.1.2), y se las nombró según las convenciones de ICTV establecidas en Bernard *et al.* (2010) (Sección A.1). Se recolectó una secuencia por serotipo para evitar la sobre-representación de los serotipos de mayor importancia clínica. Los serotipos recolectados infectan a un total de 41 especies correspondientes a 14 órdenes de vertebrados amniotas incluyendo tortugas, aves, primates y murciélagos, entre otros (Sección A.3).

De los 30 géneros de la familia *Papillomaviridae* reportados en Zur Hausen (2009), cuatro no contienen un marco de lectura abierto para la proteína E7, *Upsilonpapillomavirus* (serotipos Delphinus delphis papillomavirus 1 y Tursiops truncatus papillomavirus 1, 2 y 3), *Dyodeltapapillomavirus* (Sus scrofa papillomavirus 1), *Omegapapillomavirus* (Ursus maritimus papillomavirus 1) y *Omikronpapillomavirus* (Phocoena spinipinnis papillomavirus 1) (Gottschling *et al.*, 2011a; Rehtanz *et al.*, 2006; Stevens *et al.*, 2008a,b; Van Bressem *et al.*, 2007).

Las 224 secuencias recolectadas están distribuidas en 26 géneros de papilomavirus, reportados en Zur Hausen (2009) y 63 especies, mientras que seis secuencias no pudieron ser clasificadas dentro de ningún género y tres no pudieron ser clasificadas dentro de ninguna especie. En la Figura 3.1 se muestra el número de representantes por especie coloreados según el orden al que pertenece el hospedador que infectan, con excepción de las seis secuencias que no pudieron ser clasificadas que se muestran en negro.

Las seis secuencias que no fueron clasificadas corresponden a los serotipos Bettongia penicillata papillomavirus 1, Equus ferus caballus papillomavirus 3, Mus musculus papillomavirus 1, Zalophus californianus papillomavirus 1 y Ovis aries papillomavirus 3, Bos taurus papillomavirus 7, cuyos hospedadores pertenecen a los órdenes Diprotodontia, Perissodactyla, Rodentia, Carnivora y Artiodactyla.

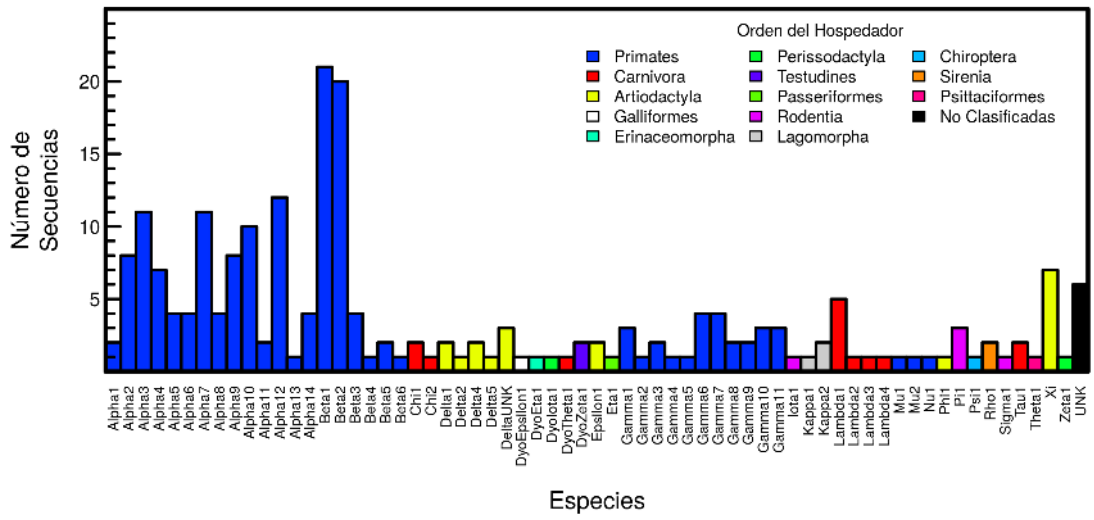


Figura 3.1: Número de secuencias de E7 por especie. La altura de cada barra indica el número total de secuencias (o serotipos) recolectados para cada especie de la familia *Papillomaviridae* y está coloreada según el orden al que pertenece el hospedador de los serotipos correspondientes a cada especie.

En resumen, la base de datos inicial (Sección A.1) es representativa de la familia *Papillomaviridae*, abarcando los 26 géneros que poseen una región codificante para la proteína E7 de papilomavirus y las especies conocidas hasta la fecha de recolección.

3.1.2. Creación de la base de datos 2

Actualmente la familia *Papillomaviridae* está dividida en 2 subfamilias: (1) *Firstpapillomavirinae* y (2) *Secondpapillomavirinae* (Van Doorslaer *et al.*, 2018). La subfamilia *Firstpapillomavirinae* consiste en 52 géneros. Además de los 30 géneros discutidos antes, se adicionaron 17 géneros nuevos con una única especie representante. La subfamilia *Secondpapillomavirinae* contiene un único género y una única especie con un único serotipo representante, Sparus aurata papillomavirus 1. Estos datos se resumen en la Tabla 3.1.

En abril 2017 se realizó una actualización de la base de datos, incorporando 121 secuencias a nuestra base de datos bajadas de NCBI y nombradas según la taxonomía actual de ICTV, que sigue los lineamientos establecidos en Bernard *et al.* (2010). En primer lugar, pudimos clasificar las secuencias que no poseían clasificación al inicio del trabajo. Las seis que no tenían el género identificado son ahora *Dyokappapapillomavirus* (Ovis aries papillomavirus 3), *Dyolambdapapillomavirus* (Bettongia penicillata papillomavirus 1), *Dyonupapillomavirus* (Zalophus californianus papillomavirus 1), *Dyorhopapillomavirus* (Equus caballus papillomavirus 3) y *Dyoxipapillomavirus* (Bos taurus papillomavirus 7). Pudimos clasificar dos de las tres secuencias del género *Deltapapillomavirus* dentro de las especies *Deltapapillomavirus 4* (Bos taurus papillomavirus) y *Deltapapillomavirus 6* (Camelus dromedarius papillomavirus 1). La única secuencia restante no tiene especie asignada (Camelus dromedarius papillomavirus 2). En segundo lugar, obtuvimos representantes para 15 de los 17 nuevos géneros (Sección A.2). Sólo los serotipos correspondientes a

los géneros *Dyopipapillomavirus* (Phocoena phocoena papillomavirus 4), *Treisiotapapillomavirus* (Myotis ricketti papillomavirus 1) (Gottschling *et al.*, 2011b; Wu *et al.*, 2012) y el único serotipo de la subfamilia *Secondpapillomavirinae* (Sparus aurata papillomavirus 1) no poseen una región codificante para la proteína E7 (López-Bueno *et al.*, 2016). El serotipo Sparus aurata papillomavirus 1 es el único representante de la familia *Papillomaviridae* que infecta a un vertebrado anamniota (López-Bueno *et al.*, 2016).

Subfamilia	Género	Especie
Firstpapillomavirinae	*Dyochipapillomavirus*	*Dyochipapillomavirus 1*
	Dyomupapillomavirus	*Dyomupapillomavirus 1*
	Dyoomegapapillomavirus	*Dyoomegapapillomavirus 1*
	Dyoomikronpapillomavirus	*Dyoomikronpapillomavirus 1*
	Dyophipapillomavirus	*Dyophipapillomavirus 1*
	Dyopipapillomavirus	*Dyopipapillomavirus 1*
	Dyopsipapillomavirus	*Dyopsipapillomavirus 1*
	Dyosigmapapillomavirus	*Dyosigmapapillomavirus 1*
	Dyotaupapillomavirus	*Dyotaupapillomavirus 1*
	Dyoupsilonpapillomavirus	*Dyoupsilonpapillomavirus 1*
	Treisdeltapapillomavirus	*Treisdeltapapillomavirus 1*
	Treisepsilonpapillomavirus	*Treisepsilonpapillomavirus 1*
	Treisetapapillomavirus	*Treisetapapillomavirus 1*
	Treisiotapapillomavirus	*Treisiotapapillomavirus 1*
	Treiskappapapillomavirus	*Treiskappapapillomavirus 1*
	Treisthetapapillomavirus	*Treisthetapapillomavirus 1*
	Treiszetapapillomavirus	*Treiszetapapillomavirus 1*
Secondpapillomavirinae	*Alefpapillomavirus*	*Alefpapillomavirus 1*

Tabla 3.1: Nuevos géneros y especies de la familia *Papillomaviridae*.

La base de datos 2 posee un alto número de secuencias de la proteína E7 que son representativas de la familia *Papillomaviridae*, con representantes para los 41 géneros y las especies correspondientes que poseen una región codificante para la proteína E7 de papilomavirus.

3.2. Definición de expresiones regulares de motivos lineales de la proteína E7

Como se discutió previamente en la introducción (véase Sección 1.3.5), el dominio desordenado de E7 está conformado por las regiones conservadas CR1 y CR2. Estas regiones contienen múltiples motivos lineales que median interacciones proteína-proteína y que son responsables, por lo menos en parte, del alto número de blancos proteicos reportados para E7 (véase Sección 1.2). La Dra. Lucía Chemes realizó una búsqueda bibliográfica que reveló más de 100 blancos proteicos (Figura 3.2), que interactúan con el dominio E7N desordenado, con el dominio E7C globular o con más de una región. Por ejemplo, los motivos lineales de la región CR2, el motivo LxCxE, el

motivo de fosforilación de CKII y la región acídica cooperan en la unión a la pRb (Chemes *et al.*, 2011, 2010) que interactúa además con la región CR1 y el dominio E7C.

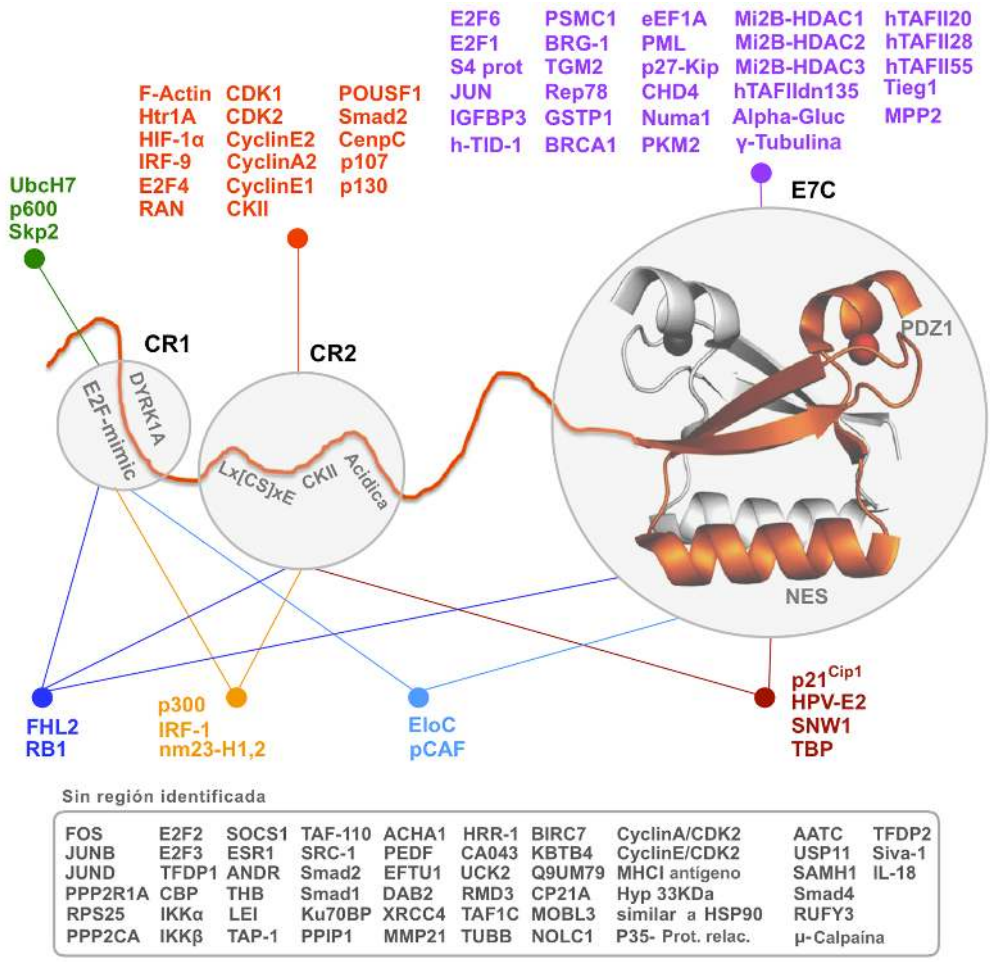


Figura 3.2: Representación esquemática de los dominios de E7 y blancos proteicos. El homodímero que forma el dominio E7C está representado utilizando la estructura obtenida por RMN del dominio E7C de la proteína E7 de HPV45 (PDB ID: 2F8B) (Ohlenschläger *et al.*, 2006) en forma de cintas. Los átomos de zinc asociados están representados como esferas. El dominio E7N está representado para uno de los monómeros de E7 en naranja como una cinta extendida. Se muestran además las ubicaciones aproximadas de las regiones CR1 y CR2, y los motivos pRb_ABGroove (señalado como E2F-Mimic), DYRK1A, LxCxE, CKII, región acídica, NES y PDZ. Los blancos proteicos de E7 que tienen las regiones blanco en la proteína E7 identificadas están agrupados según sus sitios de interacción. Aquellos blancos proteicos que no tiene los sitios de interacción identificados se muestran en gris. Figura adaptada de Chemes *et al.* (2012a).

Se realizó una búsqueda bibliográfica que reveló numerosos motivos para la proteína E7 de determinados serotipos como ser HPV16 perteneciente a la especie *Alphapapillomavirus 9*. Para uno de los nueve motivos, el pRb_ABGroove se recolectó toda la evidencia experimental disponible para ampliar su definición. Dos de los nueve motivos, la región acídica y las posiciones ricas en cisteínas, son propuestos por primera vez. En la Tabla 3.2 se presenta un resumen de las expresiones regulares utilizadas.

Dominio	Motivo	Expresión Regular	% Secuencias	Referencias	Serotipo
CR1	DYRK1A	`[ST]P[ST]`	4.8	Liang *et al.* (2008)	HPV16
	pRb_ABGroove	`[IVLA].[NQDE][IVLFMYA][IVLFMYA]`	70.5	Shan *et al.* (1996)	-
		`[IVLA]{0,1}[AHKTNQDES]`		Lee *et al.* (2002)	-
				Xiao *et al.* (2003)	-
				Liu y Marmorstein (2007)	-
				Burke *et al.* (2010)	-
				Chemes *et al.* (2010)	HPV16
CR2	LxCxE	`[IL].[CS].[DE]`	89.9	Lee *et al.* (1998)	HPV16
				Wang *et al.* (2010)	CPV2
				Chemes *et al.* (2010)	HPV16
	Region Acídica	`Carga Neta ≤ -4`	95.2	Chemes *et al.* (2010)	-
	CKII	`[ST]..[DE]`	71	Smotkin y Wettstein (1987)	HPV16
				Barbosa *et al.* (1990)	HPV16
CR3	Posiciones Ricas en Cisteínas	`% Cys ≥ 5.9`	68.9	-	-
	Motivo de Unión a Zinc	CxxC	100	Ohlenschläger *et al.* (2006)	HPV45
				Liu *et al.* (2006)	HPV1a
	NES	`[ILVM]..[ILVMF]...[ILMF]...[ILVMF].[ILVMF]`	90	Knapp *et al.* (2009)	HPV16
				Güttler *et al.* (2010)	-
	PDZ	`[ST].[VIL]`	2.28	Tomaić *et al.* (2009)	RhPV1
				Gould *et al.* (2010)	-

Tabla 3.2: Definición de los motivos de la proteína E7 de papilomavirus. Se indican las expresiones regulares o el criterio establecido y el porcentaje de secuencias de la proteína E7 de papilomavirus que poseen cada motivo. El porcentaje de secuencias que contienen cada motivo está indicado en la cuarta columna. La bibliografía correspondiente y el serotipo viral al que corresponde la proteína E7 utilizada en la bibliografía están indicadas en la quinta y sexta columna respectivamente.

En las secciones siguientes se describe cómo se obtuvieron las expresiones regulares de los motivos utilizados.

3.2.1. Motivos lineales en el dominio CR1 de E7

Sitio de fosforilación de la quinasa 1A de especificidad dual regulada por fosforilación de tirosina. La expresión regular del motivo DYRK1A es una ampliación conservativa del sitio de fosforilación determinado por Liang *et al.* (2008) distinto de la expresión del sitio canónico de esta quinasa `RP.[ST]P` (Himpel *et al.*, 2000).

Expresión regular:

`[ST]P[ST]`

Motivo de unión al bolsillo AB de la proteína retinoblastoma. La expresión regular del motivo pRB_ABGroove se dedujo a partir de los complejos estructurales presentes y las instancias conocidas del motivo. La proteína pRb consiste de un dominio N-terminal estructurado (posiciones 52-354) conectado por una región no estructurada llamada región conectora interdominio (RbIDL) (en inglés, *Retinoblastoma Interdomain linker*, posiciones 355-379) a un dominio central estructurado llamado bolsillo (en inglés, *pocket*, posiciones 380-787) y un dominio C-terminal desordenado (posiciones 788-928) (véase Sección 1.5.1). El dominio bolsillo consiste en dos subdominios, A y B, que están conectados por una región no estructurada (RbPL) (en inglés, *Retinoblastoma pocket*

linker, posiciones 592-624). Entre las regiones internas de los subdominios A y B se forma el bolsillo AB (en inglés, *AB groove* en inglés). Burke *et al.* (2010) demuestran que la región RbPL se une a este bolsillo inhibiendo la unión a E2F-1.

Las instancias conocidas del motivo, es decir, las proteínas que se conoce que poseen un fragmento identificado que interactúa con esta región, son cinco celulares y dos virales. Para establecer la expresión regular se recolectó la información de los distintos experimentos relevantes que involucran estas instancias haciendo enfoque en la región común a todas las instancias. Para la generación de la expresión regular se tuvo en cuenta aquellas posiciones que formaban contactos con pRb en las estructuras y aquellas cuya mutación impedían o disminuían de manera significativa la unión a pRb respecto al péptido silvestre (en inglés *wild-type*), o inhibían la activación en los ensayos de transcripción o no desplazaban a E2F en los ensayos de competencia de igual manera que el péptido silvestre. Los datos experimentales en base a los cuales se definió la expresión regular se resumen en la Tabla 3.3, indicando la observación realizada en cada experimento para cada una de las posiciones.

Las instancias celulares corresponden a los factores de transcripción E2F 1-5 que pertenecen a la familia de factores de transcripción E2F y a la misma proteína pRb. La familia E2F tiene un total de 8 miembros, E2F 1-8 (véase Sección 1.5.1). Las proteínas E2F 6-8 carecen del dominio de transactivación y unión a pRb (TA-PB). Cada una de las proteínas E2F que poseen el dominio TA-PB interacciona con miembros específicos de la familia de las proteínas bolsillo. E2F-1, E2F-2 y E2F-3 interaccionan casi exclusivamente con pRb, E2F-5 interacciona con p130 mientras que E2F-4 se asocia con pRb, p130 y p107 (Lee *et al.*, 2002). Por lo tanto, se incluyeron en el análisis únicamente los miembros E2F 1-5.

Shan *et al.* (1996) y Lee *et al.* (2002) realizan experimentos de mutagénesis en E2F-1. En Shan *et al.* (1996) utilizan un péptido de 18 aminoácidos E2F-1 (posiciones 409-426) correspondiente a la región de unión a pRb y realizan experimentos de mutagénesis combinados con doble híbrido para evaluar la unión a pRb, con el péptido mutante aislado y con el péptido mutante inmerso en el contexto de secuencia de E2F-1 correspondiente al dominio de transactivación (posiciones 284-437). A partir de los resultados observan que las posiciones 423, 424 y 425 de E2F-1 son importantes en la interacción con pRb. Lee *et al.* (2002) realizan ensayos de activación transcripcional utilizando la técnica del gen reportero con E2F-1, observando que las posiciones 421, 424 y 425 de E2F-1 cumplen un rol importante en la interacción con pRb. En Xiao *et al.* (2003) determinan la estructura de pRb unido a un péptido de 18 aminoácidos de E2F-1 (posiciones 409-426) (PDB ID: 1O9K). Entre sus observaciones, determinan que el residuo 424 establece un contacto con pRb.

Posición en E2F-1	Proteína	Mutación	Experimento y Resultado	Referencia
I421	E2F-1	TM: I421A, L424A, F425A	Inhiben activación transcripcional	Lee *et al.* (2002)
	E2F-1	TM: Y411A, F413A, I421A	Inhiben activación transcripcional	Lee *et al.* (2002)
	E2F-1	MM: Y411A, F413A, I421A, L424A, F425A	Inhiben activación transcripcional	Lee *et al.* (2002)
	E1A	L43A	Co-P: No purifica con pRb ELISA: No desplaza E2F-1 Afinidad con pRb por ITC: Disminuye K_d wt = 9μM vs. K_d mutante > 100μM	Liu y Marmorstein (2007)
	E1A		Estructura: Contacto con pRb y enterrado	Liu y Marmorstein (2007)
R422	E1A	H44Y, H44A, H44Q	Co-P: Purifica parcialmente con pRb ELISA: Desplazo parcial de E2F-1 (H44Y y H44A) Afinidad con pRb por ITC: Disminuye. K_d wt = 9μM vs. K_d mutante H44Y = 20μM K_d mutante H44A y H44Q > 100μM	Liu y Marmorstein (2007)
	E1A		Estructura: Contacto con pRb y enterrado	Liu y Marmorstein (2007)
D423	E2F-1	D423G	Doble híbrido: Inhibe interacción con pRb	Shan *et al.* (1996)
	pRb	D604A	Afinidad con E2F-1 por ITC: Aumenta. K_d wt pRb no fosforilado 0.045 ± 0.007μM K_d wt pRb fosforilado 0.7 ± 0.4μM vs. K_d mut pRb fosforilado 0.4 ± 0.2μM	Burke *et al.* (2010)
L424	E2F-1	L424P	Doble híbrido: Inhibe interacción con pRb	Shan *et al.* (1996)
	E2F-1	DM: L424A, F425A	Inhiben activación transcripcional	Lee *et al.* (2002)
	E2F-1	TM: I421A, L424A, F425A	Inhiben activación transcripcional	Lee *et al.* (2002)
	E2F-1	MM: Y411A, F413A, I421A, L424A, F425A	Inhiben activación transcripcional	Lee *et al.* (2002)
	E2F-2		Estructura: Contacto con pRb y enterrado	Lee *et al.* (2002)
	E2F-1		Estructura: Contacto con pRb y enterrado	Xiao *et al.* (2003)
	E1A	L46A	Co-P: Purifica parcialmente con pRb ELISA: Desplazo parcial de E2F-1 Afinidad con pRb por ITC: Disminuye K_d wt = 9μM vs. K_d mutante = > 100μM	Liu y Marmorstein (2007)
	E1A		Estructura: Contacto con pRb y enterrado	Liu y Marmorstein (2007)
F425	E2F-1	F425A, F425S	Doble híbrido: Inhibe interacción con pRb	Shan *et al.* (1996)
	E2F-1	DM: L424A, F425A	Inhiben activación transcripcional	Lee *et al.* (2002)
	E2F-1	TM: I421A, L424A, F425A	Inhiben activación transcripcional	Lee *et al.* (2002)
	E2F-1	MM: Y411A, F413A, I421A, L424A, F425A	Inhiben activación transcripcional	Lee *et al.* (2002)
	E2F-2		Estructura: Contacto con pRb y enterrado	Lee *et al.* (2002)
	E2F-1		Estructura: Contacto con pRb y enterrado	Xiao *et al.* (2003)
	pRb	Y606A	Afinidad con E2F-1 por ITC: Aumenta. K_d wt pRb no fosforilado 0.045 ± 0.007μM K_d wt pRb fosforilado 0.7 ± 0.4μM vs. K_d mut pRb fosforilado 0.22 ± 0.09μM	Burke *et al.* (2010)
	E1A	Y47G, Y47R	Co-P: No purifica con pRb Afinidad con pRb por ITC: Disminuye K_d wt = 9μM vs. K_d mutante = > 100μM ELISA: Desplazo parcial de E2F-1 (Y47G)	Liu y Marmorstein (2007)
	E1A		Estructura: Forma contacto con pRb	Liu y Marmorstein (2007)

Tabla 3.3: Datos experimentales para el desarrollo de la expresión regular del motivo pRb_ABGroove. En la primera columna se indica la posición de referencia en la proteína E2F-1. En la segunda columna se indica la proteína con la cuál se realizó el experimento. En la tercera columna se indica cuando corresponde la mutación realizada sobre la proteína. Se señalan además cuando es una doble, triple o múltiple mutación (DM,TM o MM, respectivamente). En la cuarta columna se indica el experimento realizado y el resultado observado. En la quinta columna se indica la referencia correspondiente. Las abreviaturas utilizadas corresponden a: calorimetría de titulación isotérmica (ITC), coprecipitación (Co-P), ensayo de inmunoabsorción ligado a enzimas (ELISA) y péptido silvestre (wt).

Lee *et al.* (2002) realizan experimentos de calorimetría de titulación isotérmica (ITC) (en inglés, *Isothermal Titration Calorimetry*) donde observan que E2F-2 es más afín

($K_d = 0.19 \pm 0.04\mu$M) que E2F-5 ($K_d = 0.69 \pm 0.01\mu$M) por pRb. Además, determinan la estructura de pRb unido a un péptido de 18 aminoácidos de E2F-2 (posiciones 410-427) (PDB ID: 1N4M). En la estructura observan un cierto número de posiciones de E2F-2 que tienen distintos residuos en E2F-5, que explicarían la diferencia en afinidad. En particular, observan que la posición I422 (421 en E2F-1) tiene un efecto de empaquetamiento en un núcleo hidrofóbico en la superficie de contacto entre pRb y E2F-2. Este efecto podría verse disminuido por la presencia de una valina en la posición homóloga 335 de E2F-5. El grupo hidroxilo de la serina en la posición 423 de E2F-2 (422 en E2F-1) está expuesto y cercano a un oxígeno del residuo E464 en pRb. El cambio a C336 en E2F-5 no sería favorable en ese contexto. Por último, observan que los residuos en las posiciones 425 y 426 de E2F-2 (424 y 425 en E2F-1) establecen contactos con pRb.

Por último, Burke *et al.* (2010) estudia la regulación por fosforilación de pRb. Cuando pRb está fosforilada se inhibe la asociación con el factor de transcripción E2F (K_d pRb defosforilada = $0.3 \pm 0.2\mu$M vs. K_d pRb fosforilada = $13 \pm 3\mu$M). Además, realizan experimentos de mutagénesis, fosforilación e ITC para medir la constante de afinidad entre pRb, fosforilada o desfosforilada, y el dominio de transactivación de E2F-1. Observan que la mutación del sitio de fosforilación S608 en el fragmento RbPL (426 en E2F-1) disminuye la afinidad del dominio de transactivación de E2F-1 por pRb. La afinidad resultante era similar a la afinidad por la pRb silvestre defosforilada ($K_d = 0.15 \pm 0.01\mu M$). Mediante ensayos de RMN determinan que la fosforilación de la S608 de RbPL induce la unión de RbPL al bolsillo AB y a su vez desplaza a E2F-1. Además, mediante ensayos de mutagénesis en pRb 380-787 fosforilado combinados con ITC determinan que las posiciones 604 y 606 de Rb-PL son relevantes para la interacción con el bolsillo AB. Observan que la mutación L607A en pRb aumenta la afinidad por E2F-1 cuando pRb está fosforilada ($K_d$wt pRb no fosforilado $0.045 \pm 0.007\mu$M, $K_d$wt pRb fosforilado $0.7 \pm 0.4\mu$M vs. $K_d$mut pRb fosforilado $0.3 \pm 0.2\mu$M). Esta posición no se encuentra en las otras instancias con excepción de la proteína E7 de papilomavirus (Figura 3.4)

Además de las instancias celulares existen dos instancias virales, la proteína E1A de adenovirus y la proteína E7 de HPV16. Ambas proteínas se unen a la proteína pRb a través del dominio CR1 y CR2 (Chemes *et al.*, 2010; Liu y Marmorstein, 2007). El dominio CR1 interacciona con el bolsillo formado entre los dominios A y B de pRb, mientras que el dominio CR2 se une al dominio B de pRb.

Liu y Marmorstein (2007) determinan la estructura de pRb unido a un péptido de 10 aminoácidos de E1A HAdV5 (posiciones 40 a 49) (PDB ID: 2R7G). Entre sus observaciones, determinan que las posiciones 43, 44, 46 y 47 de E1A (421, 424 y 425 en E2F-1) establecen contactos con pRb. Además, evalúan estas posiciones por distintos métodos. Realizan experimentos de mutagénesis en combinación con experimentos de coprecipitación (en inglés, *pull-down*) mediante los cuales determinan que los residuos en estas posiciones median la interacción con pRb. Por experimentos de ITC determinan que la mutación de estos residuos en el contexto del dominio CR1 de E1A disminuye la afinidad por pRb. Por último, utilizan estas mutantes para medir el la competencia entre E1A y E2F-1 por pRb utilizando el ensayo de inmunoabsorción ligado a enzimas (ELISA)

(en inglés, *Enzyme-Linked ImmunoSorbent Assay*) para medir el desplazamiento de E2F-1 de pRb. En el ensayo se observa que las mutantes no desplazan a E2F-1 o lo hacen en menor medida que el péptido silvestre.

En segundo lugar, se observó la ubicación de dichas posiciones en los complejos estructurales. En las tres estructuras se observa que el extremo C-terminal de los péptidos utilizados forma una estructura de hélice (Figura 3.3). Esta misma observación fue realizada para la proteína E7 de papilomavirus por dicroísmo circular (Chemes *et al.*, 2010).

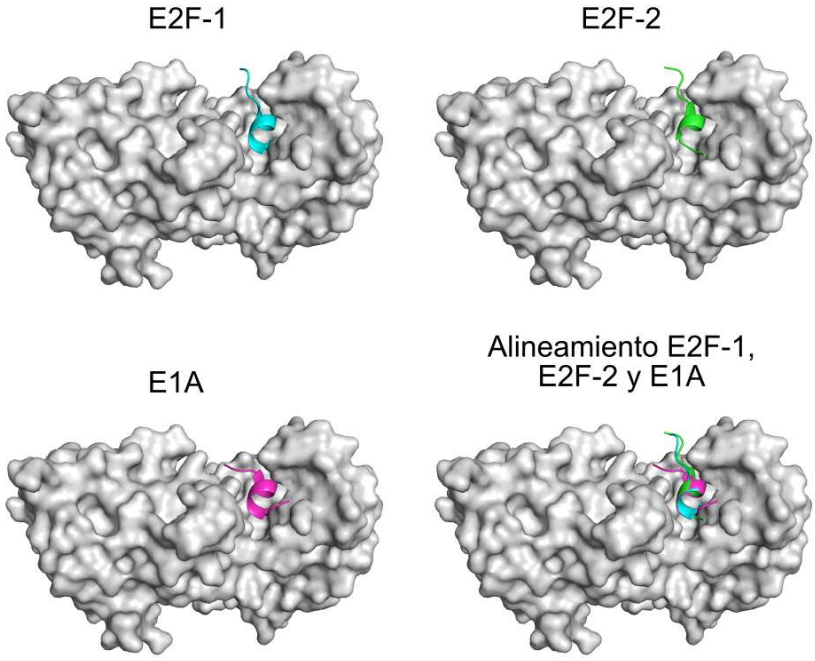


Figura 3.3: Estructuras para la determinación de la expresión regular del motivo pRb_ABGroove. Se muestran las estructuras correspondientes a los complejos E2F1-Rb (PDB ID: 1O9K), E2F2-RB (PDB ID: 1N4M) y E1A-Rb (PDB ID: 2R7G) y el alineamiento de los tres fragmentos de E2F-1, E2F-2 y E1A utilizando la estructura de pRb (representación en superficie gris) del complejo E1A-Rb como base. Los tres fragmentos, E2F-1, E2F-2 y E1A, están representados en forma de cintas celeste, verde y rosa respectivamente.

En tercer lugar, se realizó un alineamiento de las siete instancias conocidas (Figura 3.4) que incluye las posiciones estudiadas experimentalmente en tres instancias celulares y una instancia viral. En el alineamiento se puede observar que la Ser 608 que se fosforila en pRb alinea con una posición conservada correspondiente a un aspártico en las otras instancias.

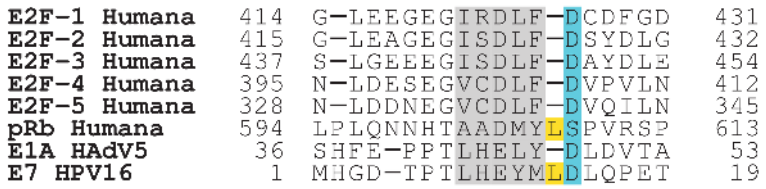


Figura 3.4: Alineamiento de las instancias celulares utilizadas para la determinación de la expresión regular del motivo pRb_ABGroove. Se indican las posiciones analizadas en los experimentos de la Tabla 3.3 (gris), la posición única a las instancias de las proteínas pRb y E7 (amarillo) y la posición correspondiente a la serina 608 fosforilada de pRb (celeste).

En base a sus resultados, Shan *et al.* (1996) proponen la expresión regular `Y.......E...DLF` y en Xiao *et al.* (2003) proponen `[LI]..L[YF]`. En este trabajo de tesis, se amplía la expresión regular con mutaciones conservativas observadas en el alineamiento de la proteína E7 (Figura 3.4) en las posiciones correspondientes y teniendo en cuenta los datos experimentales y estructurales analizados. La expresión regular es:

```
[IVLA].[NQDE][IVLFMYA][IVLFMYA][IVLA]{0,1}[AHKTNQDES]
```

3.2.2. Motivos lineales en el dominio CR2 de E7

Motivo de unión a la proteína retinoblastoma. Definimos la expresión regular del motivo LxCxE según (Dinkel *et al.*, 2014). La expresión regular es:

```
[IL].[CS].[DE]
```

Sitio de fosforilación de la quinasa de caseina II. Definimos la expresión regular del motivo CKII según (Allende y Allende, 1995). La expresión regular es:

```
[ST]..[DE]
```

Región acídica. Analizamos la carga presente en las secuencias de las proteínas E7 en la región comprendida entre el motivo LxCxE y el dominio E7C, considerando que los aminoácidos ácidos, aspártico y glutámico aportan 1 carga negativa cada uno, y que los aminoácidos básicos, arginina y lisina, aportan una carga positiva cada uno. El 87 % de las proteínas E7 posee una carga neta menor o igual a -4 en esa región (Figura 3.5).

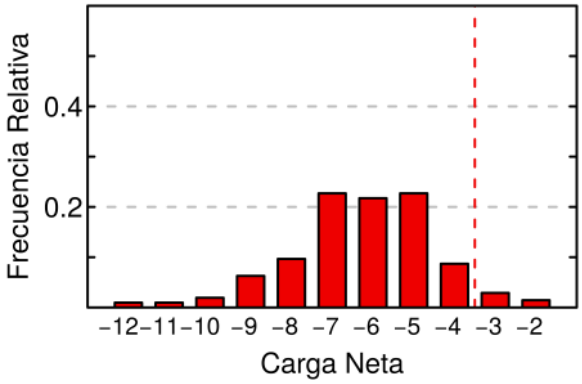


Figura 3.5: Distribución de la carga neta en la región acídca por proteína para E7. Se muestra la carga neta en la región comprendida entre el motivo LxCxE y el dominio E7C.

Definimos arbitrariamente que una región acídica está presente en una proteína de E7 si la carga neta en esa región es menor o igual a -4.

3.2.3. Motivos lineales en el dominio CR3 de E7

Posiciones ricas en Cisteína. Una inspección visual del alineamiento del dominio E7C en la vecindad del motivo CxxC de unión a Zinc sugirió la presencia de un número inusualmente elevado de cisteínas. 10 posiciones presentaban al menos un 5.9 % de cisteinas. Este valor es 4 veces mayor que el promedio de cisteínas en Uniprot, 1,36 % en 2012 y 1,20 actualmente (Consortium, 2017). Por lo tanto, definimos la presencia de posiciones ricas en cisteína en el dominio de CR3 de E7 como aquellas posiciones alineadas confiablemente con una abundancia de cisteínas mayor a 5.9 (Chemes *et al.*, 2012b).

Señal de Exportación Nuclear. La expresión regular de este motivo fue definida según Güttler *et al.* (2010). La expresión regular es:

```
[ILVM]..[ILVMF]...[ILMF]...[ILVMF].[ILVMF]
```

Motivo de unión a los dominios PDZ. Definimos el motivo de unión a los dominios PDZ tipo 1 según Gould *et al.* (2010). La expresión regular es:

```
[ST].[VIL]$
```

donde $ indica que el motivo se encuentra en el extremo C-terminal de la proteína.

3.3. Alineamiento múltiple de secuencias y Dominios de E7

Experimentos de dicroísmo circular, RMN, y cristalización en la proteína E7 de HPV16, HPV1a y HPV45 permitieron determinar que la proteína E7 contiene un dominio N-terminal desordenado (E7N), posiciones 1-40 en HPV16, formado por las regiones conservadas CR1 y CR2, y

un dominio C-terminal globular (E7C), posiciones 51-98 en HPV16, formado por la región conservada CR3 que forma un homodímero (Alonso *et al.*, 2002; García-Alai *et al.*, 2007; Liu y Marmorstein, 2007; Ohlenschläger *et al.*, 2006) que coordina dos átomos de Zinc.

De los 224 serotipos de papilomavirus que contenían al menos una región codificante para la proteína E7, 17 secuencias de E7 poseían un dominio N-terminal no homólogo a los restantes dominios de E7. Estas secuencias incluyen secuencias del género *Deltapapillomavirus*, *Epsilonpapillomavirus*, *Etapapillomavirus*, *Thetapapillomavirus*, *Dyoepsilonpapillomavirus*, *Dyozetapapillomavirus* y una de las secuencias no clasificadas (BPV7). 5 secuencias de E7 de los serotipos de papilomavirus que infectan a los quelónidos (Caretta caretta papillomavirus 1, Chelonias midas papillomavirus 1) y aves (Fringilla coelebs papillomavirus, Psittacus erithacus papillomavirus y Francolinus leucoscepus papillomavirus 1) presentan una deleción de 5 a 6 residuos en el dominio C-terminal que corresponden a la hélice α principal, necesarios para mantener la estructura globular conocida. Por lo tanto, no se las incluyó en el alineamiento.

Las secuencias restantes se utilizaron para construir dos alineamientos por separado. Un alineamiento correspondiente al dominio N-terminal de la proteína E7 (E7N) con 207 secuencias y un alineamiento correspondiente al dominio C-terminal de la proteína E7 con 219 secuencias (Sección B.1.1).

Para realizar el alineamiento se utilicó el software MUSCLE con los valores por defecto. El alineamiento se curó manualmente en base a los motivos descriptos en la Sección 3.2. El alineamiento de E7N final tiene una longitud de 124 posiciones y el alineamiento del E7C final tiene una longitud de 69 posiciones.

En el alineamiento del E7N se pueden distinguir claramente dos tipos de regiones. Por un lado se observan regiones que están definidas por bloques de posiciones que presentan menos de 30 % de sitios vacíos y fácilmente alineables, ya que los motivos funcionales conocidos (Sección 3.2) se encuentran dentro de estas regiones. Por otro lado, la mayoría de las posiciones por fuera de estos bloques presentaban más del 30 % de sitios vacíos y no podían ser alineados confiablemente ya que no se reconocían posiciones homólogas entre las distintas secuencias. A estas regiones pobremente alineadas se las denominó conectores (en inglés, *linkers*). Por lo tanto, se construyó un nuevo alineamiento eliminando las posiciones con alto porcentaje de sitios vacíos. Para esto se calculó la abundancia de sitios vacíos por posición y se eliminó un total de 84 posiciones que tenían más del 30 % de sitios vacíos para el E7N (Sección B.1.1). En el alineamiento del dominio E7C (Sección B.1.1) se observó que el alineamiento era de muy alta calidad, con un pequeño porcentaje de sitios vacíos en posiciones correspondientes a la estructura globular homodimérica determinada para HPV1a y HPV45 (Liu y Marmorstein, 2007; Ohlenschläger *et al.*, 2006). Únicamente había posiciones con un alto porcentaje de sitios vacíos en el extremo C-terminal. Por lo tanto, se construyó un nuevo alineamiento eliminando las posiciones con alto porcentaje de sitios vacíos. Para esto se calculó la abundancia de sitios vacíos por posición y se eliminaron 21 posiciones que tenían más del 30 % de sitios vacíos en el E7C (Sección B.1.1).

3.3.1. Regiones conectoras en la proteína E7 de papilomavirus

Para estudiar las regiones variables dentro del dominio E7N y en el extremo C-terminal del dominio E7C se representaron como un histograma de la longitud observada y la composición promedio de residuos. Los resultados se muestran en la Figura 3.6. No se observó ninguna región rica en sitios vacíos entre los motivos DYRK1A y pRb_ABGroove.

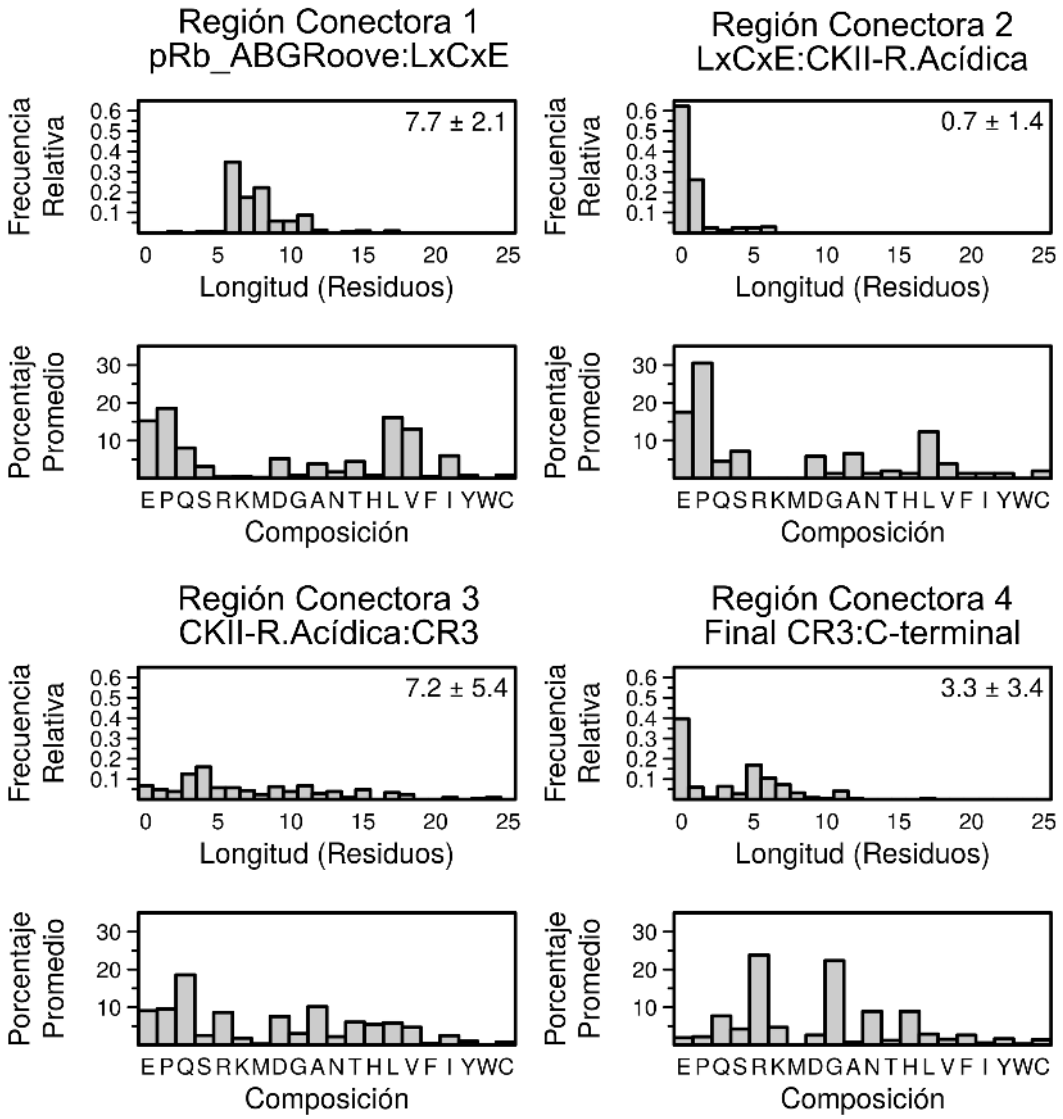


Figura 3.6: Distribución de longitudes y composición de aminoácidos de las regiones conectoras. Para cada región conectora, se muestra en el panel superior como histogramas las longitudes de la región correspondiente y en el panel inferior la abundancia promedio de aminoácidos. Los aminoácidos están ordenados según la tendencia decreciente a aparecer en regiones desordenadas (Brown *et al.*, 2010). Las regiones conectoras 1, 2 y 3 conectan las posiciones 14 y 21, 28 y 30 y 40 y 51 del logo de secuencias (Figura 3.9) respectivamente, mientras que la región conectora 4 se encuentra a continuación de la posición 98.

Observamos una región conectora entre el motivo pRb_ABGroove y el motivo LxCxE con una longitud de 7.4 ± 2.1 residuos y una composición rica en los aminoácidos prolina, valina, leucina

y glutámico, esta región se denominó Región conectora 1. Sólo 4 posiciones dentro de esta región no presentaban un alto porcentaje de sitios vacíos. Identificamos una segunda región conectora dentro del CR2, entre el motivo LxCxE y el fragmento que contiene los motivos CKII y la región acídica. Esta región denominada región conectora 2 posee una longitud corta y restringida (0.7±1.4 residuos) y está enriquecida en prolinas. La última región del dominio E7N es la región conectora entre el fragmento que contiene los motivos CKII y la región acídica y el dominio E7C. Esta región denominada región conectora 3, de las 4 regiones analizadas es la región de mayor longitud y variabilidad (7.2 ± 5.4 residuos). Su composición está enriquecida en glutamina. Por último, la región conectora 4 corresponde al extremo C-terminal del dominio E7C. Se observó una longitud variable de 3.3 ± 3.4 residuos y su composición está enriquecida en los aminoácidos arginina y glicina.

En resumen, se construyeron dos alineamientos múltiples de secuencia con posiciones alineadas confiablemente y se identificaron 4 regiones conectoras con posiciones con alto porcentaje de sitios vacíos. Es importante resaltar que la región conectora 2, que es la que presenta más restricción en longitud de secuencia, conecta los motivos LxCxE, CKII y la región acídica identificados como motivos asociados.

3.4. Abundancia y distribución por especie de los motivos lineales de E7.

Se estudió la abundancia de cada motivo lineal entre las secuencias de las proteínas de E7 (Tabla 3.2). Para ello, se determinó la presencia y ausencia de cada motivo en cada secuencia de E7, restringiendo la búsqueda a la región de la proteína donde el motivo fue descripto originalmente. Los motivos de interacción proteína-proteína pRb_ABGroove, LxCxE, el sitio de fosforilación de CKII y las posiciones ricas en cisteínas estaban en un porcentaje sustancial de las secuencias (70 a 82 %). Sólo 2 motivos de interacción proteína-proteína, el sitio de fosforilación DYRK1A y el motivo de unión a dominios PDZ estuvieron presentes en menos del 10 % de las secuencias de E7.

Se cuantificó el número de sitios CKII por secuencia en la región posterior al motivo LxCxE y anterior al dominio E7C (Figura 3.7). El 70 % de las secuencias tenían al menos 1 sitio CKII. Dos tercios de ellas tenían al menos 2 sitios de fosforilación y únicamente dos secuencias tenían 4 sitios de fosforilación.

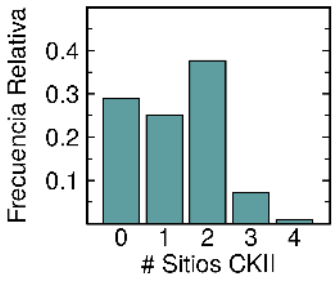


Figura 3.7: Número de sitios de fosforilación CKII en el dominio E7N. Distribución del número de sitios CKII en la región que incluye al sitio CKII de cada secuencia de la proteína E7 de papilomavirus.

En el dominio E7C, el motivo de unión a zinc, CxxC, estaba presente en el 100 % de las secuencias. Se identificaron un total de diez posiciones ricas en cisteínas. El 68 % de las secuencias tenían al menos una cisteína en una de las posiciones ricas en cisteínas (Tabla 3.2 y Figura 3.8). De estas secuencias, dos tercios tenían una cisteína extra y un tercio tenía dos posiciones ricas en cisteínas, y solamente un 3 % tenía tres cisteínas extra. Estas posiciones ricas en cisteínas pueden ser clasificadas en dos grupos. Un grupo son posiciones ricas en cisteínas que se encuentran cercanas en secuencia y espacio al motivo CxxC de cada monómero de E7C (residuos azules, centro y derecha, Figura 3.8). Este grupo incluye las posiciones 56, 57, 59, 60, 63 y 98 de HPV16, y se encuentran en el 6.8 %, 10.5 %, 18.7 %, 9.4 %, 5.9 % y 9.7 % de las secuencias, respectivamente. El segundo grupo son posiciones ricas en cisteínas que se encuentran más alejadas en secuencia del motivo CxxC. Sin embargo, el conjunto de posiciones de uno de los monómeros de E7C se encuentra cercano en el espacio al motivo CxxC de la otra molécula del homodímero (residuos verdes, centro y derecha, Figura 3.8). Este grupo incluye las posiciones posiciones 51, 68, 69 y 71 de HPV16, los cuales se encuentran en el 8.3 %, 9.1 %, 6.8 % y 20.1 % de las secuencias, respectivamente.

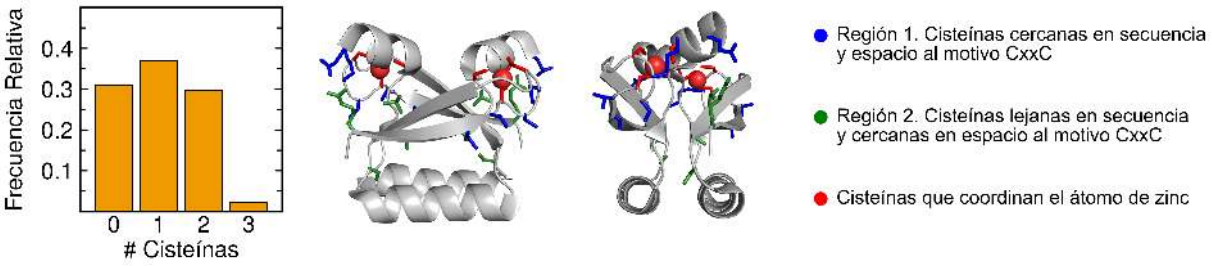


Figura 3.8: Regiones ricas en cisteínas en el dominio E7C. *Izquierda.* Distribución del número de cisteínas en las regiones ricas en cisteínas del dominio E7C de cada secuencia. *Centro y Derecha.* Representación en forma de cintas de la estructura del dominio E7C de HPV45 (PDB ID: 2F8B, (Ohlenschläger *et al.*, 2006)), con el residuo que se encuentra en las posiciones ricas en cisteínas de la región 1 (azul) y 2 (verde) representado en forma de palillos, junto con las cisteínas que coordinan el zinc (rojo). Es importante resaltar que para muchas de las posiciones ricas en cisteínas el residuo presente en HPV45 no es una cisteína. Los átomos de zinc se encuentran representados como esferas rojas. La representación de la derecha tiene una rotación de 90 grados en el eje y respecto de la del centro.

La abundancia y conservación de cada motivo lineal de la proteína E7 es específica de cada motivo, sugiriendo que su distribución no es homogénea entre los distintos serotipos de la familia *Papillomaviridae*.

3.5. Conservación de secuencia en E7

La organización de los dominios está representada en forma de logos de secuencia (Schneider *et al.*, 1986) Figura 3.9 construidos a partir de los alineamientos del E7N y del E7C sin sitios vacíos (Sección B.1). En un logo de secuencia cada posición del alineamiento está representada como una columna de letras. La altura de cada columna, $IC(l)$ mide la conservación correspondiente a esa posición en bits (véase Sección 2.1.9). Los aminoácidos presentes en cada posición

están representados como letras en la columna y las alturas son proporcionales a las abundancias relativas. Los logos de secuencia permiten, por lo tanto, medir la conservación de secuencia a nivel de residuo y a nivel de dominio (véase Sección 2.1.9 y Sección 2.1.8). En contraste a lo esperado para una proteína con un alto grado de desorden intrínseco, ~ 58 % de las posiciones del dominio E7N (19 de 33) muestran una conservación media a alta ($IC(l) > 2.0$).

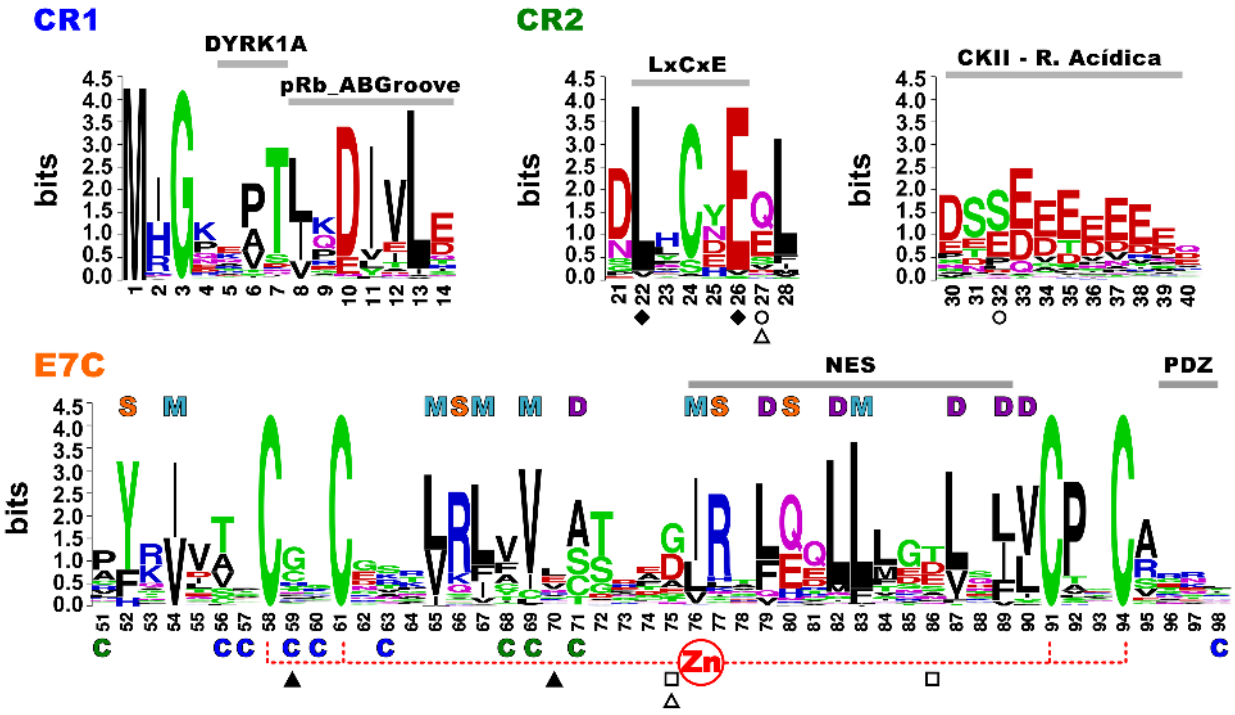


Figura 3.9: Conservación de secuencia de la proteína E7 de papilomavirus. Las posiciones que pudieron alinearse con confianza están representadas como logos de secuencia. De *arriba* a *abajo* Estructura de dominios y regiones de la proteína E7. Dominio N-terminal (33 posiciones de 207 secuencias), incluyendo las regiones CR1 y CR2, y Dominio C-terminal, E7C (48 posiciones de 219 secuencias). La numeración de secuencia se construyó a partir de la proteína E7 de HPV16. La posición de los motivos funcionales conocidos se indica en la parte superior de los logos (líneas grises). Las posiciones correspondientes a los residuos conservados que integran el núcleo hidrofóbico de cada monónmero (M, celeste), la superficie de dimerización (D, violeta) y expuestos en superficie (S, naranja) se indican en la parte superior del E7C. Para el dominio E7C se muestra la coordinación del zinc por cuatro cisteínas (línea roja punteada) y las diez posiciones ricas en cisteínas (letras C azules y verdes). Los residuos identificados en el análisis de coevolución por información mutua están señalados en la parte inferior de cada logo (rombos y triángulos negros, círculos, triángulos y cuadrados blancos).

Las regiones más conservadas de E7N corresponden con sitios funcionales conocidos como los motivos pRb_ABGroove, DYRK1A en el CR1 y los motivos LxCxE, CKII y la región acídica en el CR2. Además de la metionina N-terminal, la posición con la mayor conservación ($IC(l) > 4$) es la glicina en la posición 3. El sitio de fosforilación de DYRK1A (posiciones 5-7, Figura 3.9) está parcialmente conservado. Sólo dos de las tres posiciones poseen un $IC(l) > 2.0$. Cinco de las siete posiciones del motivo pRb_ABGroove (posiciones 8-14, Figura 3.9) mostraron una conservación media a alta ($IC(l) > 2.0$). Dentro del motivo LxCxE (posiciones 22-26, Figura 3.9) dos posiciones extra, la 21 y la 28 mostraron alto nivel de conservación.

En el dominio E7C, el $\sim$ 48 % de las posiciones (23 de 48) estaban altamente conservadas ($IC(l) > 2.0$) y se observan cuatro grupos de estos residuos altamente conservados en la estructura del homodímero del dominio E7C. El primer grupo corresponde al motivo de coordinación del zinc, conformado por las cuatro cisteínas y una prolina (posiciones 58, 61, 91, 92 y 94, Figura 3.9 y Figura 3.10). Un segundo grupo de seis residuos constituye el núcleo hidrofóbico del monómero (posiciones 54, 65, 67, 69, 76 y 83, M celeste, Figura 3.9 y Figura 3.10). Un tercer grupo de seis residuos conforman la superficie hidrofóbica de dimerización (posiciones 71, 79, 82, 87, 89 y 90, D violeta, Figura 3.9 y Figura 3.10). Por último, el cuarto grupo incluye cuatro residuos que se encuentran en la superficie (posiciones 52, 66, 77 y 80, S naranja, Figura 3.9 y Figura 3.10).

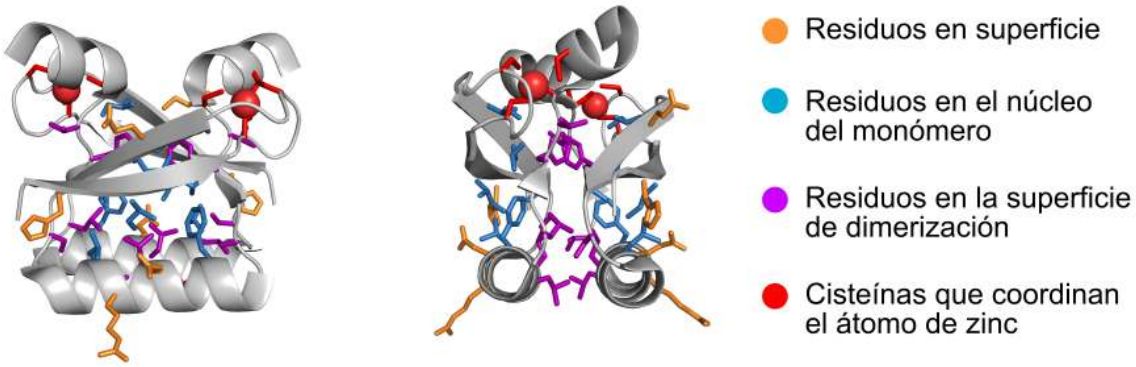


Figura 3.10: Grupo de residuos con conservación alta en el dominio E7C. Las orientaciones *izquierda* y *derecha* difieren en una rotación de 90 grados en el eje y. Se muestran en representación de palillos los residuos involucrados en el núcleo hidrofóbico de cada monómero (celeste), en la superficie de dimerización (violeta) y superficie expuesta (naranja) y las cisteínas que coordinan el zinc (rojo). El código de colores es el mismo al utilizado en la Figura 3.9. El átomo de zinc está representado por una esfera de color rojo. El esqueleto de la proteína se muestra en representación de cintas y color gris (PDB ID: 2F8B) (Ohlenschläger *et al.*, 2006).

Se realizó una comparación entre el grado de conservación promedio del dominio desordenado, E7N, y el dominio ordenado E7C de la proteína E7. En la Figura 3.11 se observa la conservación de las posiciones dentro de cada dominio como histogramas. El eje x indica los intervalos de $IC(l)$ considerados y el eje y indica la frecuencia relativa de las posiciones en cada dominio para cada intervalo de $IC(l)$.

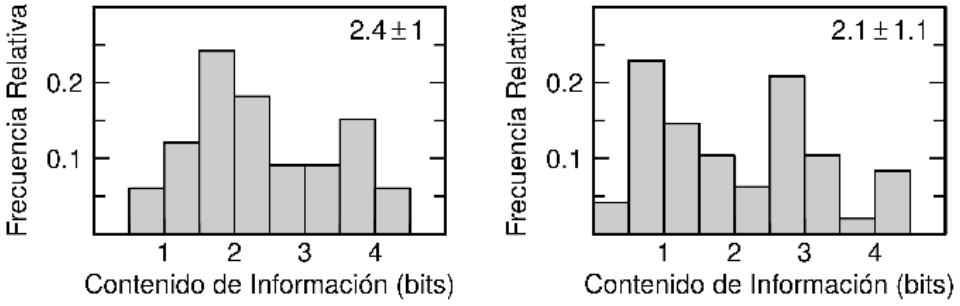


Figura 3.11: Conservación de secuencia de los dominios de la proteína E7. Se muestra en forma de histogramas la distribución del contenido de información de cada dominio de la proteína E7. En la parte superior del histograma se indica el valor promedio y el desvío estándar.

La conservación promedio del dominio desordenado E7N y el dominio ordenado E7C fue muy similar (2.4 ± 1.0 bits y 2.1 ± 1.2 bits, respectivamente) y no se observaron diferencias significativas (valor $p > 0.05$) según las pruebas de permutación (véase Sección 2.1.11).

En resumen, el grado de conservación general de la proteína E7 es alto y el dominio desordenado está tan conservado como el dominio ordenado. En el dominio desordenado, la mayoría de las posiciones más conservadas están asociadas a sitios funcionales, mientras que en el dominio ordenado las posiciones más conservadas están relacionadas con residuos que cumplen un rol estructural.

3.6. Desorden intrínseco en la proteína E7

Estudios experimentales en la proteína E7 de HPV16, HPV45, y HPV1a muestran que el dominio N-terminal de E7 (E7N) es intrínsecamente desordenado (García-Alai *et al.*, 2007), mientras que el dominio C-terminal (E7C) posee una estructura globular, forma un dímero en solución y cada monómero de E7 coordina mediante un motivo CxxC un átomo de zinc (Alonso *et al.*, 2002; Liu *et al.*, 2006; Ohlenschläger *et al.*, 2006). Sin embargo, cambios en la temperatura y el pH inducen transiciones estructurales indicando la adquisición de elementos de estructura secundaria en el dominio E7N (García-Alai *et al.*, 2007; Noval *et al.*, 2013). Por otro lado, el análisis computacional de cuatro secuencias prototípicas de E7, HPV16, HPV18 (de alto riesgo), HPV6 y HPV11 de bajo riesgo (Uversky *et al.*, 2006) reveló que el patrón de desorden de los serotipos de alto riesgo es diferente del observado en los serotipos de bajo riesgo.

Para evaluar la generalidad de este resultado se estudió la tendencia al desorden en la base de datos 1 de 207 secuencias para el dominio E7N y 219 secuencias para el dominio E7C con el algoritmo IUPred (Dosztányi *et al.*, 2005a) (véase Sección 2.2.3). El valor de IUPred toma valores entre 0 y 1. Un valor superior a 0.5 indica desorden y un valor menor a 0.5 indica orden (Daughdrill *et al.*, 2011; Dosztányi *et al.*, 2005a). Estos resultados están representados como un gráfico de puntos (Figura 3.12) donde el eje x indica la posición de secuencia y el eje y indica el valor promedio de IUPred y el desvío estándar para las posiciones alineadas de manera confiable.

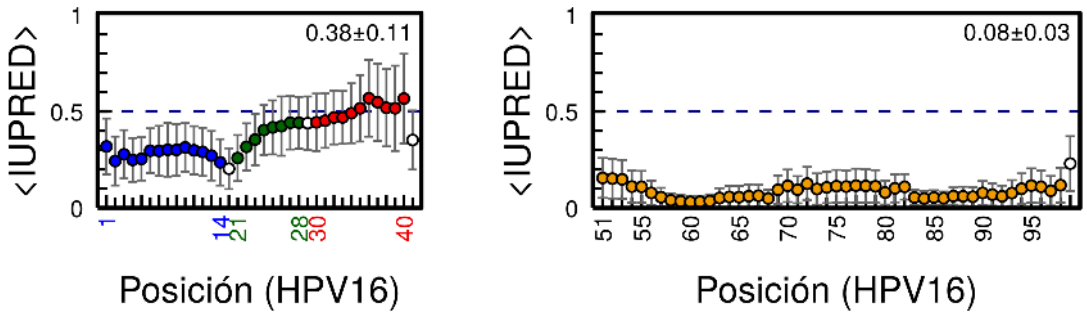


Figura 3.12: Predicción del desorden intrínseco para cada dominio de la proteína E7. Se muestra el índice de IUPred de cada dominio de la proteína E7 para aquellas posiciones que pudieron ser alineadas de manera confiable. La numeración de la secuencia se origina en la proteína E7 de HPV16. *Izquierda*. Predicción de desorden intrínseco del dominio E7N para las regiones CR1 (azul) y CR2. En este caso se indica el motivo LxCxE (verde) y la región que contiene el motivo CKII y la región acídica (rojo). En blanco se indica el promedio de todas las posiciones que comprenden la región conectora 1, 2 y 3 y su ubicación relativa entre el dominio CR1 y el motivo LxCxE, entre este y la región que contiene el motivo CKII y la región acídica, y entre esta última y el CR3, respectivamente. En el eje x se indican las posiciones que corresponden al inicio y final de cada región en la proteína E7 de HPV16. *Derecha*. Predicción de desorden intrínseco del dominio E7C. En blanco se indica el promedio de todas las posiciones que comprenden la región conectora 4 y su ubicación relativa en el dominio CR3. La numeración se corresponde con la proteína E7 de HPV16. El valor promedio y desvío estándar se muestran en la parte superior de cada gráfico.

Los valores promedio de IUPred de los dominios E7N (0.38 ± 0.11) y E7C (0.08 ± 0.03) sugieren que estos dominios no son desordenados (valor $p < 0.01$). Estos resultados concuerdan con el hecho de que la mayoría de las posiciones del dominio E7N (81.82 %) y el todas las posiciones del dominio E7C muestran un valor promedio de IUPred inferior a 0.5.

El extremo N-terminal del dominio E7N presenta una valor promedio de IUPred menor que el extremo C-terminal del dominio E7N. Esto puede deberse a la presencia de una hélice propuesta en pRB_ABGroove (Chemes *et al.*, 2010).

En resumen, los resultados confirman que el dominio E7C de las proteínas E7 en la base de datos abarcativa es predicho globular lo que se condice con los resultados experimentales. Sin embargo, los resultados experimentales obtenidos para las proteínas E7 de unos pocos serotipos virales no parecen ser representativos para el dominio E7N del conjunto de proteínas E7 en la base de datos abarcativa siendo el dominio E7N parcialmente ordenado. Estas diferencias entre los resultados predichos y experimentales pueden deberse a la tendencia a adquirir elementos de estructura secundaria en determinadas condiciones observada en el dominio E7N (García-Alai *et al.*, 2007; Noval *et al.*, 2013).

3.6.1. Desorden y conservación de secuencia en la proteína E7

Es comúnmente esperado que la conservación de los dominios intrínsecamente desordenados sea menor a la conservación de los dominios globulares debido a que estos deben mantener una estructura tridimensional estable (Daughdrill *et al.*, 2011; Toth-Petroczy *et al.*, 2008).

Para evaluar esta hipótesis en la proteína E7 de papilomavirus se comparó la conservación

promedio de cada dominio (Figura 3.11) por posición con el valor promedio de desorden por posición (Figura 3.12) para aquellas posiciones que se pueden alinear de manera confiable.

Los resultados se muestran en la Figura 3.13. En el eje *x* se grafica el valor promedio de IUPred por posición y en el eje *y* el contenido de información por posición. Estos resultados muestran que no hay correlación entre ambas mediciones.

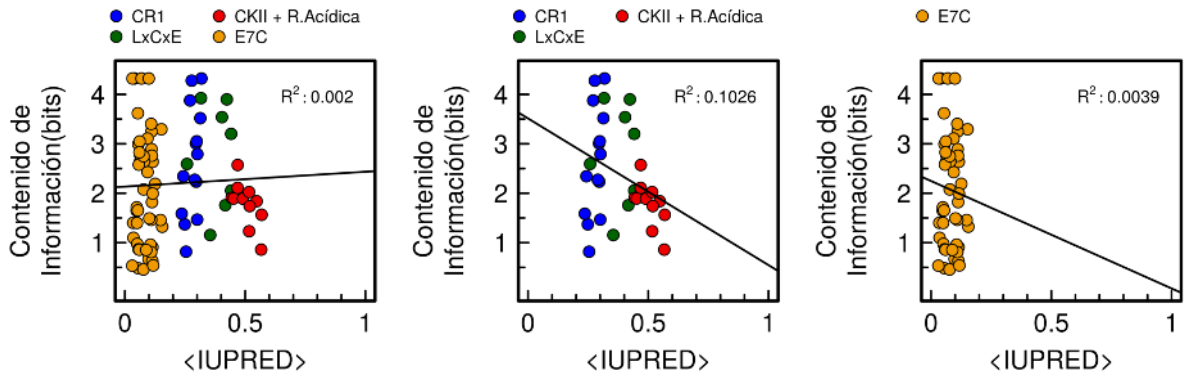


Figura 3.13: Comparación entre conservación de secuencia y desorden de los dominios de la proteína E7. En el eje *y* se indica el valor por posición del contenido de información en bits. En el eje *x* se indica el valor promedio de IUPred por posición. En la parte inferior derecha del gráfico se indica el cuadrado del coeficiente de pearson.

Por lo tanto, la conservación de secuencia de los dominios de E7 no está únicamente dictada por el grado de desorden intrínseco.

3.7. Identificación de un posible sitio de interacción en el dominio E7C

La superficie del dominio E7C fue identificada como sitio de interacción de un péptido no estructurado de la proteína p21 del hospedador (Ohlenschläger *et al.*, 2006), del dominio RbC no estructurado (Liu *et al.*, 2006) y de un dominio no estructurado dentro de la proteína Mi2b (Brehm *et al.*, 1999). Estos resultados sugieren que el dominio E7C podría unir motivos lineales contenidos dentro de dominios desordenados de sus blancos proteicos (Figura 3.2).

Una forma de determinar una posible región en la superficie del dominio E7C que pueda cumplir con esa función es identificar las regiones conservadas sobre ella. Para esto, se representa la conservación de secuencia (medida como el contenido de información por posición) sobre la estructura del homodímero de E7C representado como superficie Figura 3.14.

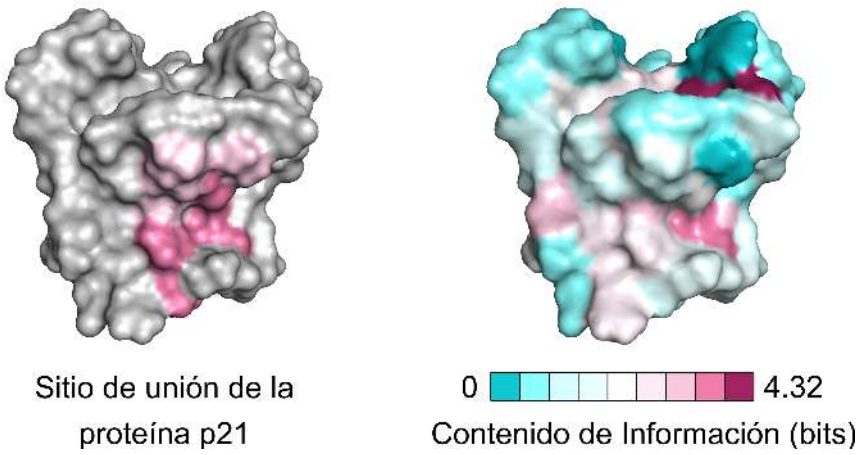


Figura 3.14: Posible sitio de unión de motivos lineales en el dominio E7C. *Izquierda.* Sitio de unión de la proteína p21 (Ohlenschläger *et al.*, 2006) representada en la superficie de la estructura del dominio E7C de HPV45 (PDB ID: 2F8B, (Ohlenschläger *et al.*, 2006)). En rosa se muestran las amidas que son perturbadas fuertemente (rosa oscuro) o moderadamente (rosa claro) al unirse p21. *Derecha.* Conservación de secuencia en la superficie del dominio E7C representada en la superficie de la estructura del dominio E7C de HPV45 (PDB ID: 2F8B, (Ohlenschläger *et al.*, 2006)). Se midió la conservación de cada posición como el contenido información y se lo clasificó en 9 categorías (a intervalos de 0.5 bits desde 0 a 4, con un intervalo extra desde 4 a 4.32)

Se puede observar una región moderadamente conservada en la superficie del dominio E7C (derecha, Figura 3.14) que incluye, entre otros, a los cuatro residuos que se encuentran en la superficie (posiciones 52, 66, 77 y 80, S naranja, Figura 3.9 y Figura 3.10). Esta región se superpone parcialmente con el sitio de unión de p21 (izquierda, Figura 3.14).

Estos resultados sugieren que una fracción significativa de los dominios E7C podrían interactuar con la proteína p21 en este sitio y que, probablemente, este sea el sitio de interacción de otros blancos proteicos.

3.8. Coevolución de secuencia en la proteína E7

Las señales obtenidas por análisis de coevolución permiten predecir pares de residuos cercanos en el espacio dentro de una proteína o entre proteínas, globulares o desordenadas (Cheng *et al.*, 2014; Morcos *et al.*, 2011; Toth-Petroczy *et al.*, 2016).

Para estudiar si existen pares de residuo coevolucionando en la proteína E7 de papilomavirus se utilizaron dos métodos diferentes ya que a lo largo del desarrollo de esta tesis hubo un cambio en los métodos para predecir pares de residuos en contacto. El primer método utilizado es el método de información mutua (véase Sección 2.2.5) y se realizó utilizando la base de datos 1 de la proteína E7 de papilomavirus. Este análisis se realizó en conjunto con la Dra. Cristina Marino-Buslje. La autora de esta tesis proporcionó los datos y analizó los resultados. El segundo método utilizado es el método de información directa (véase Sección 2.2.5). Este análisis se realizó en colaboración con la Dra. Rocío Espada, integrante del laboratorio de Fisiología de Proteínas quien participó en el desarrollo del experimento y análisis de resultados junto con la autora de esta tesis. A diferencia

del método de información mutua, el método de información directa permite diferenciar las señales de coevolución directas de las indirectas (véase Sección 2.2.5).

3.8.1. Análisis de coevolución por información mutua en la proteína E7

Utilizando la base de datos 1 de E7 (véase Sección A.1), se identificaron un total de cinco pares de residuos coevolucionando utilizando el algoritmo basado en información mutua (véase Sección 2.2.5). Dos de los cinco pares, posiciones 22-26 y 27-32, se encuentran dentro del dominio E7N (rombos negros y círculos blancos, Figura 3.9) y otros dos pares, posiciones 59-70 y 75-86, se encuentran dentro del dominio E7C (triángulos negros y cuadrados blancos, Figura 3.9 y Figura 3.15). Uno de los cinco pares de residuos ocurre entre un residuo del dominio E7N, posición 27, y un residuo de la superficie del dominio E7C, posición 75 (triángulos blancos, Figura 3.9).

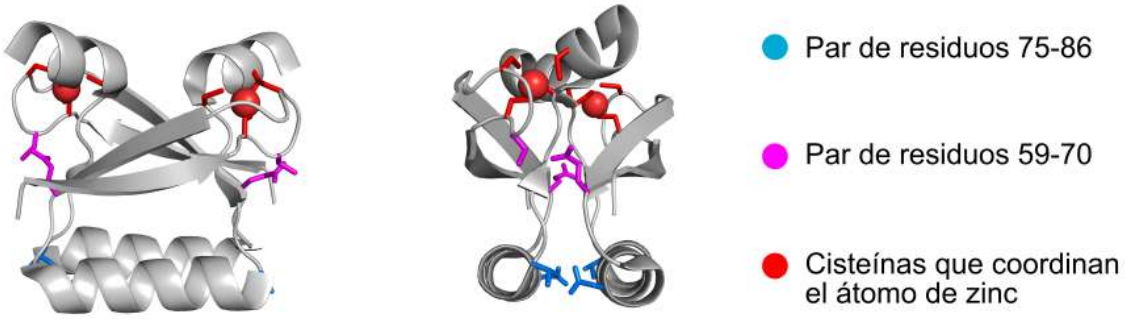


Figura 3.15: Representación de pares de residuos que coevolucionan en E7C. Las orientaciones *izquierda* y *derecha* difieren en una rotación de 90 grados en el eje *y*. Los pares de residuos identificados con alto valor de información mutua en el dominio E7C están representados como palillos. Se muestran los pares de residuos correspondientes a las posiciones 75 y 86 (celeste, cuadrados blancos en la Figura 3.9) y los correspondientes a las posiciones 59 y 70 (rosa, triángulos negros en la Figura 3.9) y las cisteínas que coordinan el zinc (rojo). El átomo de zinc está representado por una esfera de color rojo. El esqueleto de la proteína se muestra en representación de cintas y color gris (PDB ID: 2F8B) (Ohlenschläger *et al.*, 2006).

Estos resultados sugieren en primer lugar que la coevolución de residuos ocurre tanto dentro de secuencias con una estructura globular definida o sin ella y en segundo lugar que puede existir alguna actividad relevante de la proteína que involucre el contacto físico entre los dominios sugiriendo que la evolución de los dos dominios de E7 no ha sido completamente independiente.

3.8.2. Análisis de coevolución por información directa en la proteína E7

Utilizando la base de datos 2 de la proteína E7 de papilomavirus (véase Sección A.2) identificamos un total de 44 pares de residuos que poseían un valor Z≥3, lo que representa un 1.26 % de los pares totales. Estos datos están representados como un mapa de contactos predichos en la Figura 3.16. De los 44 residuos identificados, 23 (~56 %) se encuentran a una distancia de cinco

o menos residuos. 12 (~28 %) están separados por seis a 30 residuos. Los nueve residuos (~21 %) restantes estaban a una distancia de 31 o más.

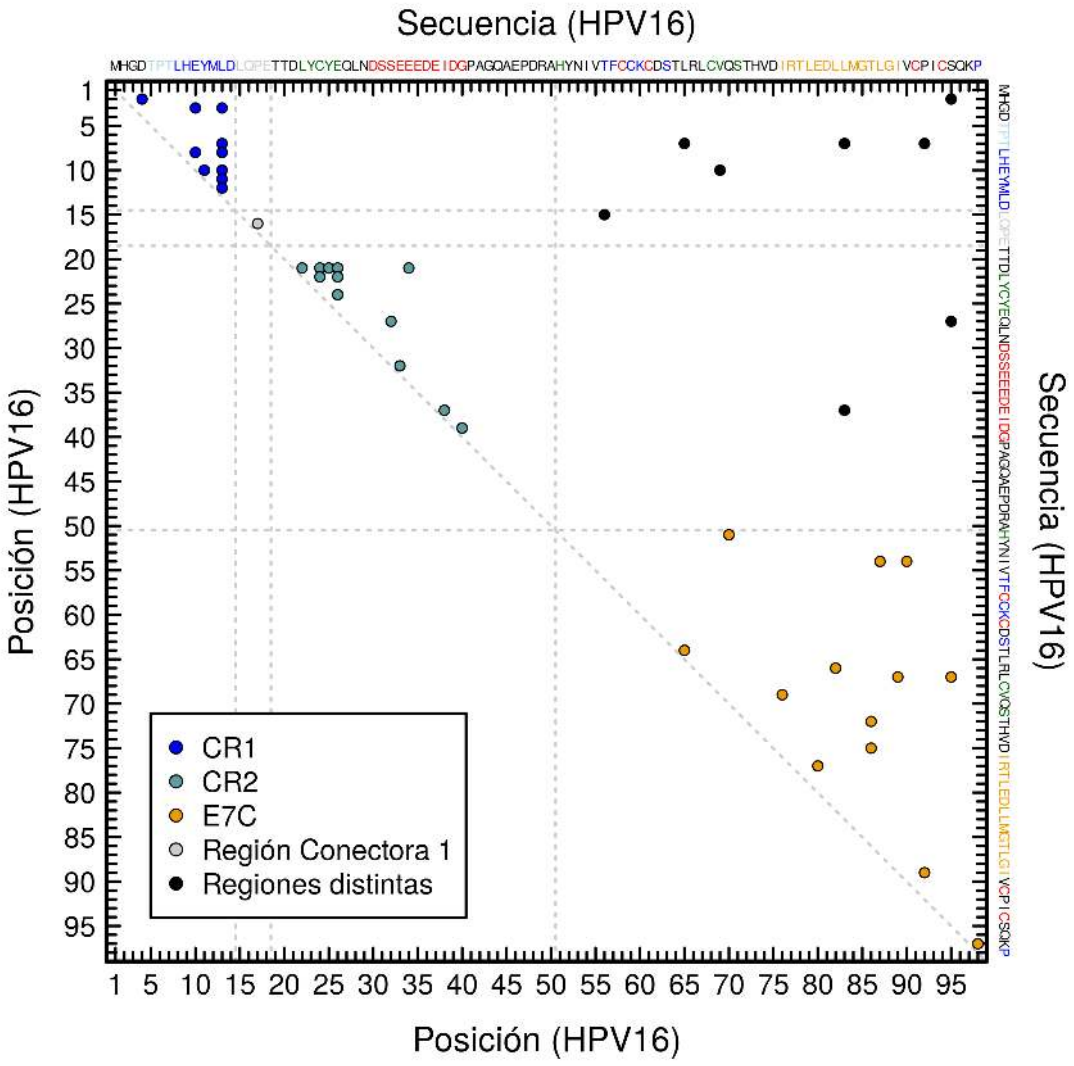


Figura 3.16: Predicción del mapa de contactos de la proteína E7 utilizando información directa. Los pares de residuos con alto valor de información directa están representados en forma de puntos. Los contactos dentro del mismo dominio o misma región están señalados utilizando el siguiente código de colores: dentro del dominio CR1 están indicados en azul, dentro del dominio CR2 están indicados en verde, dentro del dominio E7C en naranja y dentro de la región conectora en gris. Los contactos entre distintos dominios o regiones están indicados en negro. Los distintos dominios están separados por líneas punteadas. La numeración de secuencia utilizada se basa en la proteína E7 de HPV16. En la parte superior y derecha del gráfico se indica la secuencia de HPV16. Los residuos correspondientes a los motivos están resaltadas con colores: En el CR1, DYRK1A en celeste y pRB_ABGroove en azul, en el CR2 el LxCxE en verde y el CKII y la región acídica en rojo, en el E7C las cisteínas que coordinan el zinc en rojo, las posiciones ricas en cisteínas lejanas en secuencia al sitio de coordinación del zinc en verde y las cercanas en secuencia en azul, y la NES en naranja. Para una mejor visualización, las predicciones de contacto se muestran únicamente en la parte superior de la diagonal.

Los pares de residuos que coevolucionan dentro de un mismo dominio (35 pares, 81 %) están coloreados por dominio en la Figura 3.16. Dentro del dominio E7N, coevolucionan 22 pares. Diez

pares de residuos coevolucionan dentro del CR1 y doce dentro del CR2. Dentro del CR1, seis de los diez pares identificados ocurren entre residuos que pertenecen al motivo pRb_ABGroove. Dentro del CR2, tres de los diez pares identificados ocurren entre residuos que pertenecen al motivo LxCxE, cuatro ocurren entre la posición acídica anterior al LxCxE (posición 21, Figura 3.9) y un residuo del LxCxE y los tres pares restantes ocurren entre residuos que pertenecen a la región que contiene el motivo CKII y la región acídica. Un único par coevoluciona dentro de la región conectora 1 (gris, Figura 3.16). 13 pares de residuos coevolucionan dentro del E7C. Seis de los 13 pares identificados poseen sus carbonos alfa a una distancia menor a 9Å dentro de la misma cadena y cuatro de los 13 pares identificados poseen sus carbonos alfa a una distancia menor a 9Å entre las distintas cadenas. Diez de los trece pares identificados están en contacto en la estructura conocida, sugiriendo que las señales de coevolución obtenidas dentro del dominio E7N se deben a la formación de contactos dentro del mismo.

Los pares de residuos que coevolucionan entre distintos dominios (ocho pares, 19 %) están coloreados en negro en la Figura 3.16. Siete de los ocho pares identificados ocurren con alguno de los cinco residuos que se encuentran en la superficie del E7C y se muestran en azul en la Figura 3.17. Estas señales de coevolución sugieren que los dominios E7N y E7C no son dos unidades estructurales independientes, sino que existe un número importante de residuos que están en contacto entre los distintos dominios.

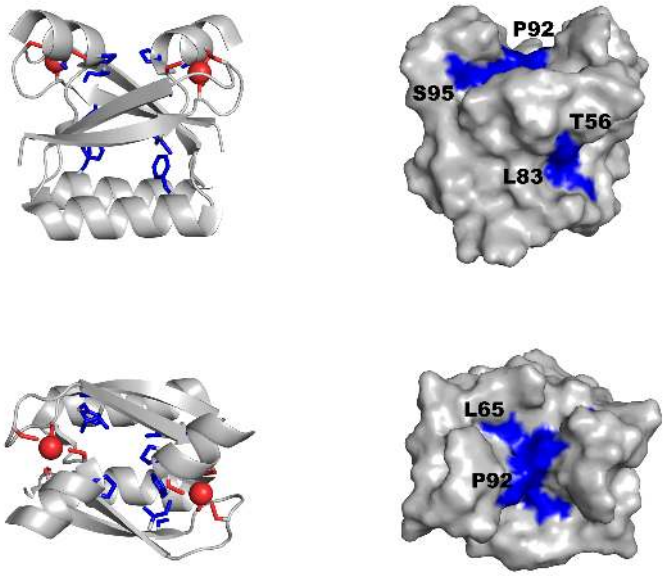


Figura 3.17: Región predicha de interacción del dominio E7C con el dominio E7N. *Arriba.* Vista frontal de la estructura del homodímero E7C (PDB ID: 2F8B) (Ohlenschläger *et al.*, 2006) representada en forma de cintas (*izquierda*) y superficie (*derecha*). *Abajo.* Vista superior, girada 90 grados en el eje *x* respecto del panel superior, de la estructura del homodímero E7C (PDB ID: 2F8B) (Ohlenschläger *et al.*, 2006) representada en forma de cintas (*izquierda*) y superficie (*derecha*). En ambos casos los residuos que conforman pares que interactúan con el dominio E7N están representados en forma de palillos (derecha) y coloreados en azul (derecha e izquierda). Se indican los respectivos residuos y posiciones de la proteína E7 de HPV16.

Cinco pares ocurren entre el CR1 y el E7C, de los cuales cuatro ocurren con algún residuo en la superficie del E7C. De estos cuatro, tres están formados por la posición 7 que pertenece al motivo DYRK1A y tres posiciones del E7C que se encuentran en la superficie (posiciones 65, 83 y 92, Figura 3.9). El par restante ocurre entre la posición 2 y la posición 95 del E7C. Un par ocurre entre una posición del motivo pRb_ABGroove (posición 10, Figura 3.9) y una posición rica en cisteínas en el E7C (posición 69, Figura 3.9) que no se encuentra en la superficie. Dos pares están formados por dos residuos del CR2, el residuo posterior al motivo LxCxE (posición 27, Figura 3.9) y un residuo de la región acídica (posición 37, Figura 3.9) con dos residuos del dominio E7C que se encuentran en la superficie, las posiciones 95 y 83 respectivamente. El último par está formado entre una posición que pertenece a la región conectora 1 y una posición rica en cisteínas (posición 56, Figura 3.9).

En resumen, la mayoría de los contactos predichos ocurren entre un sitio funcional del dominio E7N y residuos en la superficie del dominio E7C, incluyendo residuos en el bolsillo conservado del dominio E7C. Sugiriendo que existe un acoplamiento funcional entre ambos dominios. De los cinco pares identificados por información mutua, tres fueron también identificados por información directa. Esta puede deberse o bien a diferencias en los algoritmos, o bien a la alta sensibilidad que presentan los métodos utilizados en el número de secuencias utilizadas. Sin embargo, ambos métodos identifican residuos dentro del dominio E7C que están en contacto en la estructura, indicando que son buenos predictores de la formación de contactos y sugiriendo que los dominios de la proteína E7 no evolucionan de manera independiente.

3.9. Conclusiones de E7

Utilizando dos bases de datos con alto número de secuencias de la proteína E7 representativas de la familia *Papillomaviridae* se realizó un análisis sistemático y profundo de la variabilidad de E7 a nivel de secuencia.

Se identificaron un total de nueve motivos lineales en la proteína E7 (véase Sección 3.2), cuya distribución y abundancia observamos que es motivo-específica (Tabla 3.2).

A partir del alineamiento de secuencias se determinó la existencia de regiones conectoras entre regiones conservadas con variabilidades de longitud y composición de aminoácidos propias de cada una. Utilizando dos métodos de coevolución en secuencia diferentes se identificaron relaciones entre residuos, dentro del mismo dominio o región, y entre dominios (véase Sección 3.8). Para la proteína E7, estas herramientas resultaron útiles para identificar motivos lineales funcionalmente acoplados dentro del dominio desordenado. Por ejemplo, la variabilidad de longitud de la región conectora 2, que conecta el motivo LxCxE con la región que contiene el motivo de fosforilación CKII y la región acídica, está más restringida (Figura 3.6), y por el análisis de información mutua se identificaron residuos cercanos al motivo LxCxE que coevolucionan con residuos en la región acídica (Figura 3.9). Estos resultados además indican que el dominio E7N es un dominio propiamente dicho y no un simple conjunto de motivos lineales independientes (García-Alai *et al.*, 2007).

Contrario a lo esperado para un dominio desordenado, las posiciones alineadas de manera confiable en el dominio E7N están en promedio tan conservadas como el dominio globular E7C y el número de pares de residuos que coevolucionan identificados por ambos métodos es similar para ambos dominios (Figura 3.11, Figura 3.16). Esto podría deberse a la alta densidad funcional en el dominio E7N, algo característico de muchas proteínas virales (Davey *et al.*, 2011b). Además, el valor promedio de desorden predicho por IUPred para el dominio E7N no es el esperado para un dominio intrínsecamente desordenado. El grado de conservación y esta predicción pueden deberse a la modulación del conjunto de conformaciones del dominio E7N.

En el dominio E7C se observaron, en primer lugar, posiciones ricas en cisteínas ubicadas en la superficie del homodímero y cercanas en el espacio a las cisteínas que coordinan el zinc (Figura 3.8), sugiriendo una nueva funcionalidad para el dominio E7C y, en segundo lugar, se propuso una superficie de interacción posible para motivos lineales de proteínas del hospedador (Figura 3.14) que contiene dos residuos que interactúan con el dominio E7N (Figura 3.17).

Por último, las señales de coevolución obtenidas entre ambos dominios sugieren que los dominios E7N y E7C no son dos módulos estructurales independientes sino que evolucionan y funcionan de manera coordinada.

Capítulo 4

Proteína E1A

En este capítulo describo y discuto los resultados obtenidos para la proteína E1A de la familia *Adenoviridae*.

El Dr. Ernesto A. Román del Instituto de Química y Fisicoquímica Biológicas de la Facultad de Farmacia y Bioquímica realizó la obtención del modelo estructural del dominio CR3 de la proteína E1A. El Dr. Ricardo Rodríguez de la Vega del Laboratorio de Ecología, Sistemática y Evolución de la Universidad de París construyó la filogenia del género *Mastadenovirus*. La reconstrucción de secuencias ancestrales la realizó la Dra. Valeria A. Risso del departamento de Física Química de la Universidad de Granada. La búsqueda de motivos en las secuencias ancestrales la realizó César Leonetti, estudiante de Bioinformática de Universidad Argentina de la Empresa. En este caso, mi rol como autora de esta tesis fue el aporte de los datos iniciales y el análisis de los resultados, incluyendo la validación del modelo estructural. Los análisis de información directa se realizaron en colaboración con la Dra. Rocío Espada del Laboratorio de Fisiología de Proteínas. La reconstrucción de motivos por el método de parsimonia se discutió en conjunto con el Dr. Julián Faivovich de la División de Herpetología del Museo Argentino Bernardino Rivadavia. Los resultados de este capítulo se analizaron en conjunto con la Dra. Lucía Chemes del Instituto de Investigaciones Biotecnológicas de la Universidad Nacional de San Martín y el Dr. Gonzalo de Prat-Gay del laboratorio de Estructura-Función e Ingeniería de Proteínas de la Fundación Instituto Leloir. Los resultados de este capítulo derivaron en dos publicaciones:

- Glavina, J., Román, E. A., Espada, R., de Prat-Gay, G., Chemes, L. B., y Sánchez, I. E. (2018). *Interplay between sequence, structure and linear motifs in the adenovirus E1A hub protein*. Virology, 525(May):117–131

- Dinkel, H., Van Roey, K., Michael, S., Davey, N. E., Weatheritt, R. J., Born, D., Speck, T., Krüger, D., Grebnev, G., Kuban, M., Strumillo, M., Uyar, B., Budd, A., Altenberg, B., Seiler, M., Chemes, L. B., Glavina, J., Sánchez, I. E., Diella, F., y Gibson, T. J. (2014). *The eukaryotic linear motif resource ELM: 10 years and counting.* Nucleic acids research, 42(Database issue):D259–66

y un trabajo en preparación:

- Glavina, J., Rodríguez de la Vega, R., Risso V.A., Faivovich, J., de Prat-Gay, G., Chemes, L. B., y Sánchez, I. E. *E1A linear motifs contribute to adaptive Mastadenovirus evolution*.

4.1. Recolección de secuencias de la proteína E1A de la familia *Adenoviridae*

De los cinco géneros que conforman la familia *Adenoviridae* únicamente el género *Mastadenovirus* contiene un marco de lectura abierto para la proteína E1A. En total, se recolectaron 116 secuencias proteicas correspondientes a la proteína E1A de la base de datos NCBI en diciembre 2012 y se las nombró acorde a los lineamientos de la ICTV (Harrach *et al.*, 2012) (Sección A.4). Se recolectó una secuencia por serotipo para evitar la sobrerrepresentación de los serotipos de mayor importancia clínica. Los serotipos recolectados infectan a un total de 18 especies de mamíferos miembros de siete órdenes (Sección A.6) incluyendo vaca, ovejas, chanchos, murciélagos, perros, caballos, primates y roedores. En la Figura 4.1 se muestra el número de representantes por especie.

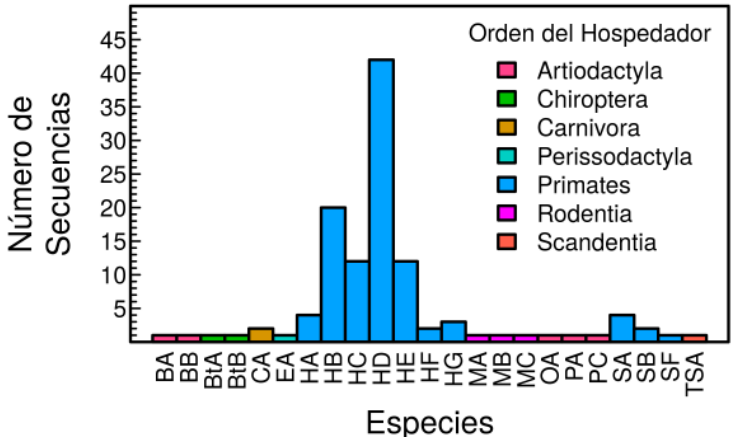


Figura 4.1: Número de secuencias de E1A por especie. La altura de cada barra indica el número total de secuencias (o serotipos) recolectados para cada especie del género *Mastadenovirus* y está coloreada según al orden al que pertenece el hospedador al que infecta los serotipos correspondientes a cada especie.

Por lo tanto, la base de datos posee un alto número de secuencias de la proteína E1A que son representativas del género *Mastadenovirus*.

4.2. Definición de expresiones regulares de motivos lineales de la proteína E1A

Como se discutió previamente en la introducción (véase Sección 1.4.5), la actividad biológica de la proteína E1A involucra la formación de interacciones proteína-proteína. Una búsqueda bibliográfica (véase Sección 2.4.1) reveló más de 50 blancos proteicos (Figura 4.2). Algunos de los blancos proteicos de E1A más estudiados son la proteína retinoblastoma (pRb), p300 y la proteína de unión a la proteína CREB (CBP) (en inglés, *CREB Binding Protein*) (Boyd *et al.*, 2002; Ferreon *et al.*, 2009; Whyte *et al.*, 1988a). Muchas de estas interacciones son mediadas por motivos lineales (véase Sección 1.2). Por ejemplo, el motivo LxCxE media la interacción del dominio CR2 de E1A con la proteína pRb.

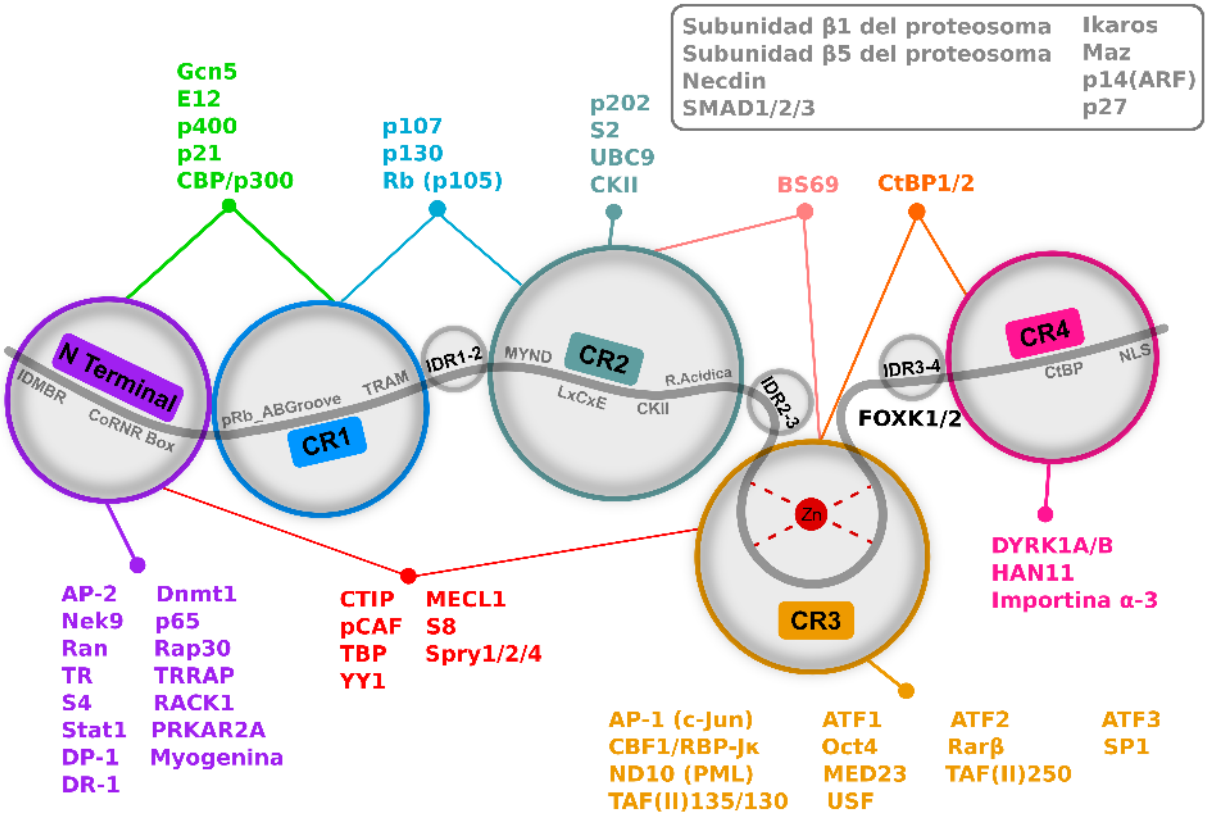


Figura 4.2: Representación esquemática de los dominios de E1A e interacciones. La ubicación aproximada de cada región conservada está marcada por círculos coloreados (Dominio N-terminal, violeta; CR1, azul; CR2, verde; CR3, amarillo, CR4, rosa). El átomo de zinc asociado al CR3 está representado por una esfera roja. Se muestra además la ubicación aproximada de los motivos lineales CoRNR Box, pRB_ABGroove, TRAM, MYND, LxCxE, CKII, Región acídica, CtBP, NLS. Las regiones inter dominio IDR12, IDR23 e IDR34 están indicadas con círculos grises. El tamaño de cada círculo no está relacionado con el tamaño del dominio o región. Los blancos proteicos de cada dominio de la proteína E1A están agrupados según los dominios o regiones blanco como se reporta en la literatura, los blancos proteicos no mapeados se muestran en gris (Sección F.1).

Se realizó una búsqueda bibliográfica que reveló numerosos motivos lineales putativos para la proteína E1A. Dichos motivos putativos se caracterizaron en proteínas E1A de determinados serotipos de sólo unas pocas especies del género *Mastadenovirus*, como HAdV12 de la especie *Human adenovirus A* o HAdV2 y HAdV5 de la especie *Human adenovirus C*. En el presente trabajo se incluyeron solamente motivos lineales con un fuerte respaldo experimental. Por lo tanto algunos motivos reportados en la literatura no se utilizaron en este trabajo y en otros casos se redefinió la expresión regular. Para dos de los doce motivos, la región intrínsecamente desordenada de unión múltiple (IDMBR) (en inglés, *Intrinsically Disordered Multiple Binding Region*) y el motivo de unión a la región TRAM de la CBP, se combinó toda la evidencia experimental disponible para crear una definición. Para uno de los doce motivos, el motivo de la CoRNR Box, se amplió la definición reportada en la literatura. En la Tabla 4.1 se presenta un resumen de las expresiones regulares utilizadas.

Dominio	Motivo	Expresión Regular	% Secuencias	Referencias	Serotipo
N-terminal	IDMBR	I.....T.....LL..L....L.D | C .....LL..L....L | R...C....[IL][ST].....LL..L[IL]	6	Boyd *et al.* (2002) Rasti *et al.* (2005)	HAdV5 HAdV5
	CoRNRBox	[ILF][^P][^P][ILVFY][ILV] [^P][^P][^P][ILVYFHM]	89	Meng *et al.* (2005) Phelan *et al.* (2010)	HAdV5 -
CR1	pRb_ABGroove	[IVLA].[NQDE][IVLFMYA][IVLFMYA] [AHKTNQDES]	95	Chemes *et al.* (2012b) Ikeda y Nevins (1993) Dyson *et al.* (1992a) Whyte *et al.* (1988b)	- HAdV5 HAdV5 HAdV5
	TRAM-CBP	[DE].[NQ][DE][^PG]AV[^PG] [NQDEST][ILMVF]F....[MIL][^PG] A[AV][^PG]..[IVLF]	9	Ferreon *et al.* (2009)	HAdV5
CR2	MYND	P.L.P	52	Hateboer *et al.* (1995) Ansieau y Leutz (2002) Isobe *et al.* (2006) Dinkel *et al.* (2014)	- - - -
	LxCxE	[IL].C.[DE]	97	Whyte *et al.* (1988a) Dyson *et al.* (1992a) Ikeda y Nevins (1993) Corbeil y Branton (1994) Dinkel *et al.* (2014)	HAdV5 HAdV5 HAdV5 - -
	Region Acídica	Carga Neta ≤ −4	94	Chemes *et al.* (2012b)	-
	CKII	[ST]..[DE]	97	Allende y Allende (1995) Whalen *et al.* (1996)	- -
CR3	Posiciones Ricas en Cisteínas	% Cys ≥ 5.9	42	Chemes *et al.* (2012a) Chemes *et al.* (2014)	- -
	Motivo de Unión a Zinc	CxxC	100	Culp *et al.* (1988)	HAdV5
CR4	NLS	[^DE]K[RK][KRP][KR][^DE]	74	Lyons *et al.* (1987) Köhler *et al.* (2001) Madison *et al.* (2002) Dinkel *et al.* (2014)	HAdV5 - HAdV5/2 -
	CtBP	P.DLS	75	Boyd *et al.* (1993) Schaeper *et al.* (1995) Molloy *et al.* (2007, 2006, 1998) Cohen *et al.* (2013) Dinkel *et al.* (2014)	HAdV5/2 HAdV5/2 HAdV12 HAdV5 -

Tabla 4.1: Definición de los motivos de la proteína E1A. Se indican las expresiones regulares o el criterio establecido y el porcentaje de secuencias de la proteína E1A de adenovirus que poseen cada motivo. El porcentaje de secuencias que contienen cada motivo está indicado en la cuarta columna. La bibliografía correspondiente y el serotipo viral al que corresponde la proteína E1A utilizada en la bibliografía están indicadas en la quinta y sexta columna respectivamente.

4.2.1. Motivos lineales en el dominio N-terminal de E1A.

Región de unión múltiple intrínsecamente desordenada. La expresión regular se definió según los experimentos de interacción y mutación secuencial por alanina realizados en la proteína E1A de HAdV5 en Rasti *et al.* (2005). Brevemente, realizan análisis de interacción proteína-proteína *in vitro* e *in vivo* combinando experimentos de mutagénesis, coprecipitación y capacidad de transformación. Para estudiar la interacción *in vitro* realizan mutaciones a alanina o glicina en 22 de las 40 posiciones del dominio N-terminal de la proteína E1A de HAdV5 de 243 residuos (243R) y de-

terminan la capacidad de unión por coprecipitación de las proteínas E1A mutantes a las proteínas: las acetiltransferasas CBP, p300, P/CAF y Gcn5, y a TBP, S8 y Ran. Los resultados obtenidos por Rasti *et al.* (2005) se muestran en la Tabla 4.2.

		Interacción *in vitro* con:								Interacción *in vivo* con:			
Pos.	Mut.	CBP	p300	P/CAF	hGCN5	S4/s8	TBP	Ran	Transformación Relativa a wt	CBP/p300	S8	TBP	Exp.Reg.
	wt	4	4	4	4	4	4	4	100	4	4	4	
2*	R → G	3	3	3	3	4	2	4	11.5	0	2	1	R
5*	I → G	3	2	0	1	2	2	2	4.5	0	2	4	I
6*	C → A	1	1	1	2	1	2	2	40	2	3	1	C
7	H → A	4	4	4	4	4	4	4	98.7	4	3	4	
8	G → A	4	4	4	4	4	4	4	80.4				
10	V → A	3	3	1	2	3	3	4	83				
11*	I → A	1	1	0	0	1	1	4	35.5	3	3	2	[IL]
12*	T → A	4	2	1	1	2	2	2	56	3	1	2	T|[ST]
14	E → A	2	2	4	3	4	4	2	70.5	4	4	4	
16	A → G	1	1	0	1	3	4	3	30.9				
18	S → G	2	2	2	2	3	4	3	112	3	3	3	
19*	L → A	1	1	0	0	1	1	1	38	1	1	2	L
20*	L → A	0	1	0	0	0	1	1	46	1	0	0	L
21	D → A	3	3	4	4	4	4	1	57				.
23*	L → A	0	0	0	0	1	1	1	38	1	0	0	L
24*	I → A	0	0	0	0	1	1	1	37.6	3	4	1	[IL]
25	E → A	4	4	4	3	3	2	2	112				
26	E → A	2	2	4	2	3	4	1	121				
27	V → A	3	3	1	3	3	4	3	74.7	4	3	4	
28*	L → A	0	0	0	0	0	3	2	45	2	0	4	L
29	A → G	4	2	4	3	4	4	4	97				
30*	D → A	4	4	4	2	1	4	2	43.5	4	1	2	D

Tabla 4.2: Mutagénesis secuencial en el dominio N-terminal de E1A. 0: No se detectó interacción. **1:** La unión es entre 0 y 25 % respecto a la proteína silvestre (wt). **2:** La unión es entre 25 y 50 % respecto a la proteína silvestre. **3:** La unión es entre 50 y 75 %. **4:** La unión es entre 75 y 100 % respecto a la proteína silvestre. En la primera columna se indica con un asterisco, "*", las posiciones fijas incorporadas en la expresión regular. En la última columna se indica el aminoácido considerado para cada posición fija. Tabla adaptada de Rasti *et al.* (2005).

Estos resultados muestran que si bien hay puntos de contacto en común en esta región, para cada proteína hay sitios de unión específicos, por lo tanto, la creación de una única expresión regular que abarque a todas no es sencillo de hacer. Por lo tanto, para definir esta expresión regular se incluyeron en la definición aquellas posiciones que al ser mutadas disminuyen la unión o actividad al menos un 50 %, *in vitro* e *in vivo*, para las proteínas CBP/p300, S8 y TBP. Luego, se incluyeron sustituciones conservativas que se encontraron en el grupo de secuencias de la proteína E1A. La expresión regular es:

```
I......T......LL..L....L.D |
  C............LL..L....L |
R...C....[IL][ST]......LL..L[IL]
```

Motivo de unión a receptores nucleares. Se modificó la definición de este motivo a partir de la propuesta por Phelan *et al.* (2010), `[IL]..[IV]I...[LFY]`. La expresión regular original se obtuvo a partir de una estructura cristalina del motivo CoRNR Box del Co-Represor nuclear NCoR unido a Rev-erb α (PDB ID: 3N00). En la modificación se incluyeron sustituciones conservativas de aminoácidos que se encontraron en el grupo de secuencias de E1A y la prohibición de prolina a las posiciones comodín, ya que este motivo tiende a formar hélices α. La expresión regular es:

```
[ILF][^P][^P][ILVFY][ILV][^P][^P][^P][ILVYFHM]
```

4.2.2. Motivos lineales en el dominio CR1 de E1A

Motivo de unión al bolsillo AB de la proteína retinoblastoma. Se utilizó la expresión regular del motivo pRB_ABGroove definida previamente para la proteína E7 de papilomavirus (véase Sección 3.4) (Chemes *et al.*, 2012b) sin incluir la sexta posición de la definición del motivo [IVLA]{0,1}. La expresión regular es:

```
[IVLA].[NQDE][IVLFMYA][IVLFMYA][IVLA][AHKTNQDES]
```

Motivo de unión a la región TRAM de la proteína de unión a CREB. Para este motivo se realizó una nueva definición de la expresión regular a partir de la estructura determinada por RMN de E1A unido a CBP (PDB ID: 2KJE) (Ferreon *et al.*, 2009). O'Connor *et al.* (1999) definen la expresión regular `F.[DE]...L` a partir de un experimento de doble mutación por alanina de las posiciones fijas donde se observa una disminución de la afinidad y el alineamiento de una región de las proteínas E2F y p53 que unen a la misma región TRAM de la proteína CBP. Sin embargo, las mutaciones por alanina individuales de cada posición no disminuyen la afinidad (O'Connor *et al.*, 1999).

Ferreon *et al.* (2009) observan en la determinación de la estructura del complejo (Figura 4.3) que el péptido de E1A está ordenado entre los residuos 54 a 80 con excepción de una región móvil entre los residuos 68 a 70. La construcción es una proteína de fusión en el N-terminal de E1A, por lo que los residuos 50 a 53 son un artefacto de esta fusión. Además, observan que el extremo C-terminal no interactúa fuertemente con CBP y que la interacción está mediada principalmente por una red de interacciones hidrofóbicas. Observan dos grupos de interacciones hidrofóbicas. El primer grupo está compuesto por los residuos 59-66 que conforman una hélice α que está enterrada en un bolsillo hidrofóbico de la superficie de CBP. El segundo grupo está formado por los residuos hidrofóbicas en las posiciones 71-74 y 78 que conforman una segunda hélice α que se empaqueta con un grupo de residuos hidrofóbicos de CBP. Por último, observan que mientras los residuos 80 y 82 podrían contribuir a la unión, la fenilalanina en la posición 83 interactúa débilmente con CBP.

Para la creación de la expresión regular, en primer lugar se descartaron los residuos que presentaban alta movilidad en la estructura de RMN (PDB ID: 2KJE) (Ferreon *et al.*, 2009). Para esto se utilizó MOBI (Martin *et al.*, 2010) utilizando un umbral de 0.95. MOBI determina la mobilidad de un residuo en base a la superposición estructural y los cambios en la conformación local. A partir

de la superposición estructural de todos los modelos de RMN calcula la RMSD y la distancia d entre los C_α y las re-escala según:

$$SD = \frac{1}{1 + \left(\frac{d}{d_0}\right)^2} \tag{4.1}$$

donde d_0 es un factor de normalización. Luego, calcula la media y desvío estándar de las distancias escaladas. El valor por defecto de d_0 es 0.85, pero es sugerido utilizar 0.95 para modelos que son parecidos como en este caso (Figura 4.3A). Los residuos 84 a 91 poseen un valor de RMSD mayor a 2 Å. Los residuos 68 a 70 y 79 a 91 son considerados móviles según MOBI. Por lo tanto, no se incluyeron en la definición del motivo el conjunto de residuos 68-70 y 79-91 (Figura 4.3B). Esto sugiere que la región de interacción más importante se encuentra entre los residuos 56 a 78.

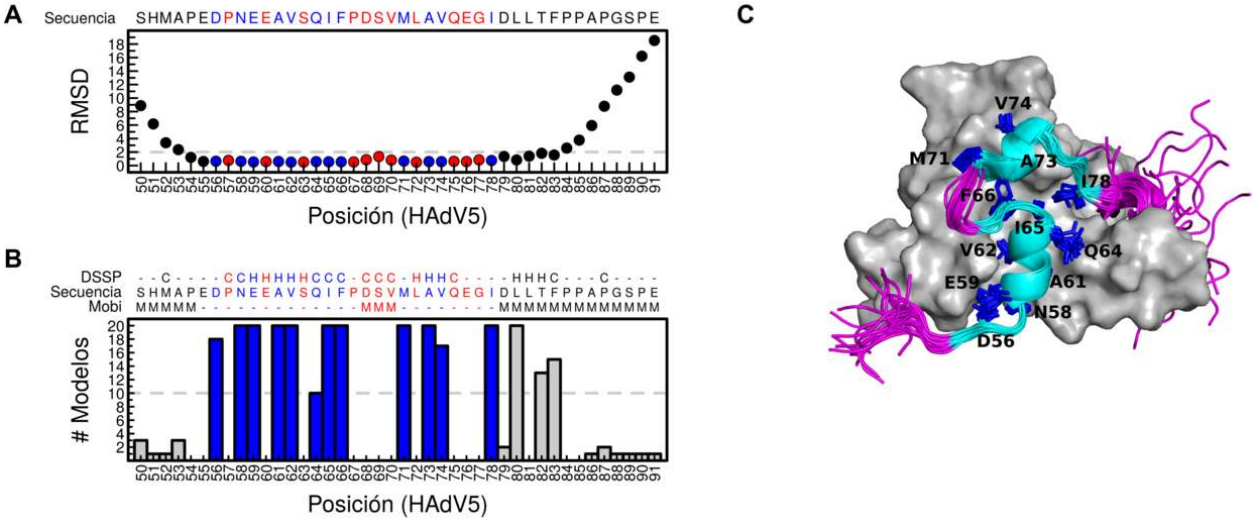


Figura 4.3: Definición del motivo TRAM de unión a la proteína CBP. **(A)** RMSD y **(B)** Frecuencias de contacto de la región de E1A unida a CBP (PDB ID: 2KJE) (Ferreon *et al.*, 2009). Las posiciones fijas y comodín están indicadas en azul y rojo, respectivamente. En la parte superior de ambos gráficos se indica la secuencia correspondiente a la proteína E1A de HAdV5. Para el gráfico de frecuencias se muestra la estructura secundaria obtenida por DSSP a partir del PDB y la mobilidad calculada según MOBI (Martin *et al.*, 2010). Las líneas punteadas indican el umbral de 2 Å en el caso de RMSD y el umbral del 50 % de los modelos de RMSD en el caso de las frecuencias de contacto. **(C)** Complejo entre el fragmento de 50-91 de E1A de HAdV5 unido a CBP (PDB ID: 2KJE) (Ferreon *et al.*, 2009). Los 20 modelos de E1A están representados en forma de cintas. Las posiciones fijas en palillos están indicadas en azul y con su nombre y posición. Las partes móviles están indicadas en magenta. Las partes no móviles en cyan. La proteína CBP está representada como superficie en un modelo.

En segundo lugar, se determinó cuales eran los residuos de E1A que estaban a una distancia de contacto de 6 Å evaluada a partir del centro de masa del residuo a la proteína CBP y que establecían dicho contacto con 6 o más residuos distintos de la proteína CBP en al menos el 50 % de los modelos (Figura 4.3B). Este análisis arrojó un total de doce residuos (posiciones 56, 58, 59, 61, 62, 64, 65, 66, 71, 73, 74 y 78). Estos resultados sugieren que estos doce residuos median la interacción entre E1A y CBP y, por lo tanto, se eligieron para las posiciones fijas de la expresión

regular, mientras que las posiciones 57, 60, 63, 67-70, 72 y 75-77, se eligieron como posiciones comodín por presentar bajo número de contactos en todos los modelos. Los resultados obtenidos por MOBI y el análisis de contactos concuerdan con las observaciones experimentales realizadas por Ferreon *et al.* (2009), con excepción de los residuos 80, 82 y 83. Si bien estos residuos forman contactos en la mayoría de los modelos, MOBI sugiere que son móviles y por lo tanto no se incluyeron en la expresión regular.

En tercer lugar, se incluyeron sustituciones conservativas en las posiciones fijas a partir de lo observado en el conjunto de secuencias de E1A.

Por último, se incluyeron restricciones a las posiciones comodín ya que el motivo se encuentra en una hélice α excluyendo los aminoácidos prolina y glicina (Figura 4.3C). La expresión regular es:

```
[DE].[NQ][DE][^PG]AV[^PG][NQDEST][ILMVF]F....[MIL][^PG]
A[AV][^PG]..[IVLF]
```

4.2.3. Motivos lineales en el dominio CR2 de E1A

Motivo de unión al dominio MYND. La definición de este motivo se corresponde con la expresión regular definida en ELMdb (Gouw *et al.*, 2018). La expresión regular es:

```
P.L.P
```

Motivo de unión a la proteína retinoblastoma. La expresión regular del motivo lineal LxCxE se definió según Dinkel *et al.* (2014). La expresión regular es:

```
[IL].C.[DE]
```

Sitio de fosforilación de la quinasa de caseina II. La expresión regular del motivo CKII se definió según Allende y Allende (1995). La expresión regular es:

```
[ST]..[DE]
```

Región Acídica. Se consideraron las posiciones de secuencia entre la última posición del motivo LxCxE y la última posición del dominio CR2. Al igual que en E7 (véase Sección 3.4), se consideró que esta región estaba presente si la carga neta de la región era menor o igual a -4. (Chemes *et al.*, 2012b).

4.2.4. Motivos lineales en el dominio CR3 de E1A

Posiciones ricas en Cisteína. La presencia de posiciones ricas en cisteína en el dominio de CR3 de E1A se definió como aquellas posiciones alineadas de manera confiable con una abundancia de cisteínas mayor a 5.9 % (Chemes *et al.*, 2012b).

4.2.5. Motivos lineales en el dominio CR4 de E1A

Motivo de unión a la proteína de unión al C-terminal. La expresión regular del motivo de unión a CtBP se corresponde con la definición de ELMdb (Dinkel *et al.*, 2014). La expresión regular es:

```
P.DLS
```

Señal de localización nuclear. La definición de la NLS es acorde a la versión clásica básicamente cargada de núcleo fuerte de la variante monopartita de la señal de localización nuclear (TRG_NLS_MonoCore_2) (Dinkel *et al.*, 2014). La expresión regular para este motivo es:

```
[^DE]((K[RK])|(RK))[KRP][KR][^DE]
```

o sea que puede ser:

```
[^DE]K[RK][KRP][KR][^DE] o [^DE]RK[KRP][KR][^DE]
```

Se realizó la búsqueda de ambas expresiones regulares y se encontró únicamente la primera en las secuencias de E1A. Por lo tanto, la expresión regular es:

```
[^DE]K[RK][KRP][KR][^DE]
```

4.3. Alineamiento múltiple de secuencias de E1A

La proteína E1A de HAdV5 tiene un largo de 289 residuos y pertenece a las de mayor longitud dentro de las proteínas E1A estudiadas en este trabajo. Si bien la distribución de longitudes dentro de cada especie es homogénea (Figura 4.4, izquierda), las longitudes dentro del género *Mastadenovirus* están distribuidas entre 170 y 289 residuos (Figura 4.4, derecha) siendo las longitudes menores a 250 aminoácidos las menos frecuentes.

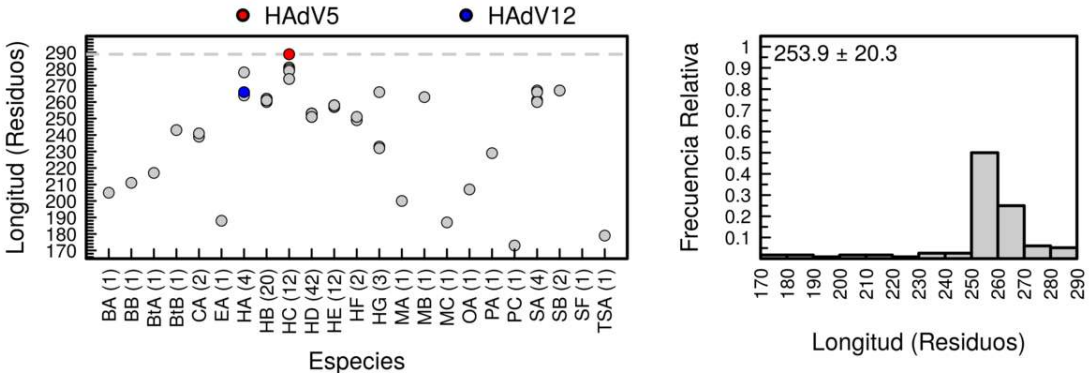


Figura 4.4: Distribución de longitudes de la proteína E1A por especie y dentro del género. Se representa a la *izquierda* la distribución de longitudes de la proteína E1A por especie. En el eje *x* se muestra las especies abreviadas (Sección A.4) y el número de serotipos incluidos entre paréntesis. Las longitudes de las proteínas E1A de HAdV5 y HAdV12 se indican en rojo y azul respectivamente. A la *derecha* se representa la distribución de longitudes dentro del género. La media y el desvío estándar se muestran en la parte superior.

Para realizar el alineamiento se utilizó el software MUSCLE con los valores por defecto. El alineamiento se curó manualmente en base a los motivos descriptos en la Sección 4.2. El alineamiento

final tiene una longitud de 409 posiciones (Sección B.2). Se construyó un segundo alineamiento eliminando las posiciones con un porcentaje de sitios vacíos mayor al 30 %. Para esto se calculó la abundancia de sitios vacíos por posición y se eliminaron un total de 163 posiciones, obteniendo un alineamiento final de 246 residuos (Sección B.2).

El alineamiento inicial no presenta regiones pobremente alineadas y todos los motivos conocidos son continuos en secuencia. Esto es una diferencia a los resultados observados en la proteína E7 de papilomavirus (véase Sección 3.3) funcionalmente relacionada (Chemes *et al.*, 2012a), donde los motivos conocidos y los sitios funcionales se encontraban en algunos casos separados por segmentos de secuencia poco conservados y pobremente alineados, descriptos como conectores (en inglés, *linkers*).

4.3.1. Dominios funcionales conocidos de E1A

Los experimentos en biología molecular y análisis de secuencia en las proteínas E1A de los serotipos prototípicos llevaron a la definición de cinco dominios funcionales conservados. Estos son llamados dominio N-terminal, y CR1 a CR4 (Avvakumov *et al.*, 2002; Kimelman *et al.*, 1985; Subramanian *et al.*, 1988). Los límites de estos dominios se establecieron utilizando 34 secuencias de *Mastadenovirus* que infectan a humanos (Avvakumov *et al.*, 2004). Por lo tanto, se revisó la definición y los límites de los dominios de E1A utilizando la base de datos construida (Sección A.4) que es más amplia y más actualizada.

Todos los dominios reportados se reconocieron fácilmente en el alineamiento múltiple de secuencias. Los límites de los dominios se redefinieron (Tabla 4.3) considerando conservación de secuencia, porcentaje de sitios vacíos por posición y motivos funcionales conocidos. Para facilitar la comunicación de los resultados se estableció una numeración de secuencia sistemática que hace referencia a la proteína prototípica E1A de HAdV5. Aquellas posiciones que no poseen una posición homóloga en HAdV5 se identifican con la última posición presente en la secuencia de E1A de HAdV5 y una letra minúscula en orden alfabético.

Dominio	**Avvakumov *et al.* (2004)**		**Este trabajo**	
	Inicio	**Final**	**Inicio**	**Final**
N-terminal	1	41	1	41
CR1	42	72	42	84
CR2	115	137	110	139
CR3	144	191	140	192
CR4	240	288	231	289

Tabla 4.3: Redefinición de los límites de los dominios de la proteína E1A. La numeración utilizada tiene origen en la proteína E1A de HAdV5.

Por ejemplo, el comienzo del dominio CR2 según Avvakumov *et al.* (2004) es en la posición 115, en este trabajo se expandió para incluir el motivo de unión al dominio MYND a la posición 110.

Dentro de cada especie la distribución de longitudes de los dominios es homogénea (Figura 4.5, izquierda).

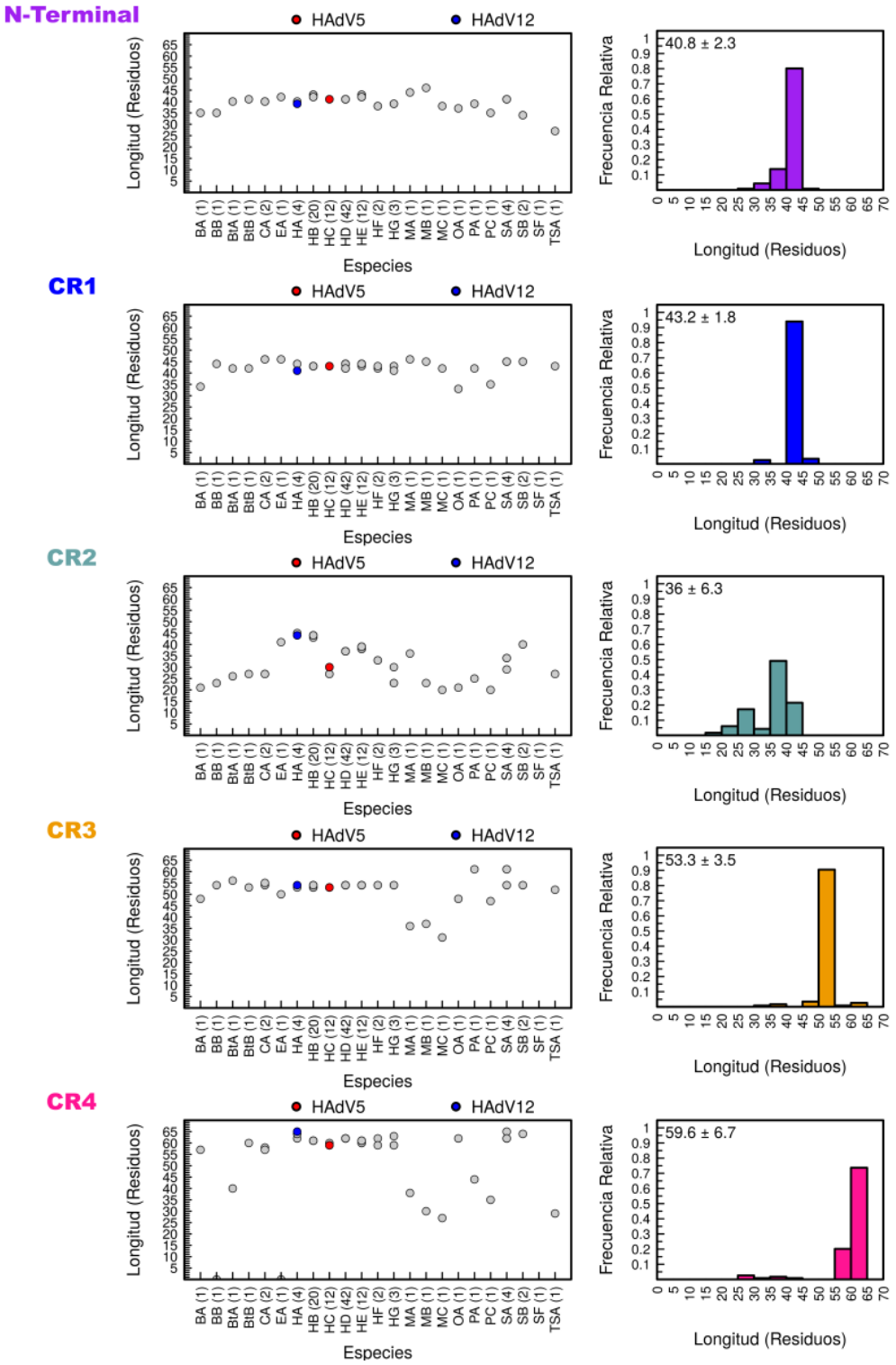


Figura 4.5: Distribución de longitudes de los dominios de la proteína E1A por especie y dentro del género. Para cada dominio de la proteína E1A: *Izquierda*. Distribución de longitudes por especie. En el eje *x* se muestran las especies abreviadas (Sección A.4) y el número de serotipos entre paréntesis. Las longitudes de los dominios de las proteínas E1A de HAdV5 y HAdV12 se indican en rojo y azul, respectivamente. *Derecha*. Distribución de longitudes dentro del género. La media y el desvío estándar se muestran en la parte superior. La distribución del CR4 no incluye los serotipos BAdV3 y EAdV1.

Entre las especies *Human adenovirus A* a *F*, la longitud de los dominios se mantiene con excepción del dominio CR2. Para el resto de las especies las longitudes son más variables para todos los dominios con excepción del dominio N-terminal y CR1. Se puede observar que las proteínas E1A de BAdV3 y EAdV1 carecen de un dominio CR4. Las distribuciones de longitud de los distintos dominios son homogéneas, con diferencias de unos pocos residuos (Figura 4.5, derecha), siendo los dominios CR2 y CR4 los más variables.

Identidad de secuencia. Una inspección visual del alineamiento permitió además observar que las proteínas de los serotipos que infectan primates son mucho más similares entre sí que las proteínas de serotipos que infectan otros órdenes. Para cuantificar esta variabilidad se calculó el porcentaje de identidad de secuencia para el alineamiento entero y para el alineamiento de cada dominio. En la Figura 4.6 se observa que el porcentaje de identidad entre pares de secuencias de E1A se pueden separar en tres grandes grupos. El primer grupo comparte un porcentaje de identidad menor al 25 % y corresponde a los pares de secuencias de E1A de serotipos que pertenecen a distintas especies y que alguno de los dos puede o no infectar a primates (Figura 4.6, barras rojas). Por ejemplo, en este grupo se compara la secuencia E1A de HAdV5 de la especie *Human adenovirus C* y la secuencia de E1A de CAdV1 de la especie *Carnivora adenovirus A*. El segundo grupo es el más abundante, comparte entre un 25 % y un 60 % de identidad y corresponde a pares de secuencias de E1A de serotipos de distinta especie que infectan a primates (Figura 4.6, barras azules). Por ejemplo, en este grupo se compara la secuencia E1A de HAdV5 de la especie *Human adenovirus C* con la secuencia de E1A de HAdV12 de la especie *Human adenovirus A*. El tercer y último grupo es el menos abundante, comparten en su mayoría más del 75 % de identidad, e incluye los pares de secuencia de E1A de serotipos que comparten la misma especie (Figura 4.6, barras verdes). Por ejemplo, en este grupo se compara la secuencia E1A de HAdV5 de la especie *Human adenovirus C* con la secuencia de E1A de HAdV2 de la misma especie. Una separación similar se observa para los dominios CR1, CR2 y CR4.

En el dominio N-terminal se observa que el porcentaje de identidad entre secuencias de E1A de serotipos de distinta especie que infectan a primates es menor (Figura 4.6, barras azules). En el dominio CR3 se observa que el porcentaje de identidad entre secuencias de E1A de serotipos de distinta especie que infectan o no a primates es mayor (Figura 4.6, barras rojas y azules).

En base a estos resultados se identificaron los dominios en las distintas secuencias considerando que debían compartir al menos un 20 % de identidad con alguna otra secuencia en el alineamiento completo. Mientras que el dominio CR3 está conservado en las 116 secuencias, se observó que esto no era así para el resto de los dominios. Para el dominio N-terminal se identificaron un total de 113 secuencias que comparten como mínimo un 20 % de identidad con al menos una secuencia en el alineamiento completo y para los dominios CR1, CR2 y CR4, 112, 114 y 107 secuencias respectivamente. Este conjunto de secuencias serán utilizadas en los distintos análisis (véase Sección A.4).

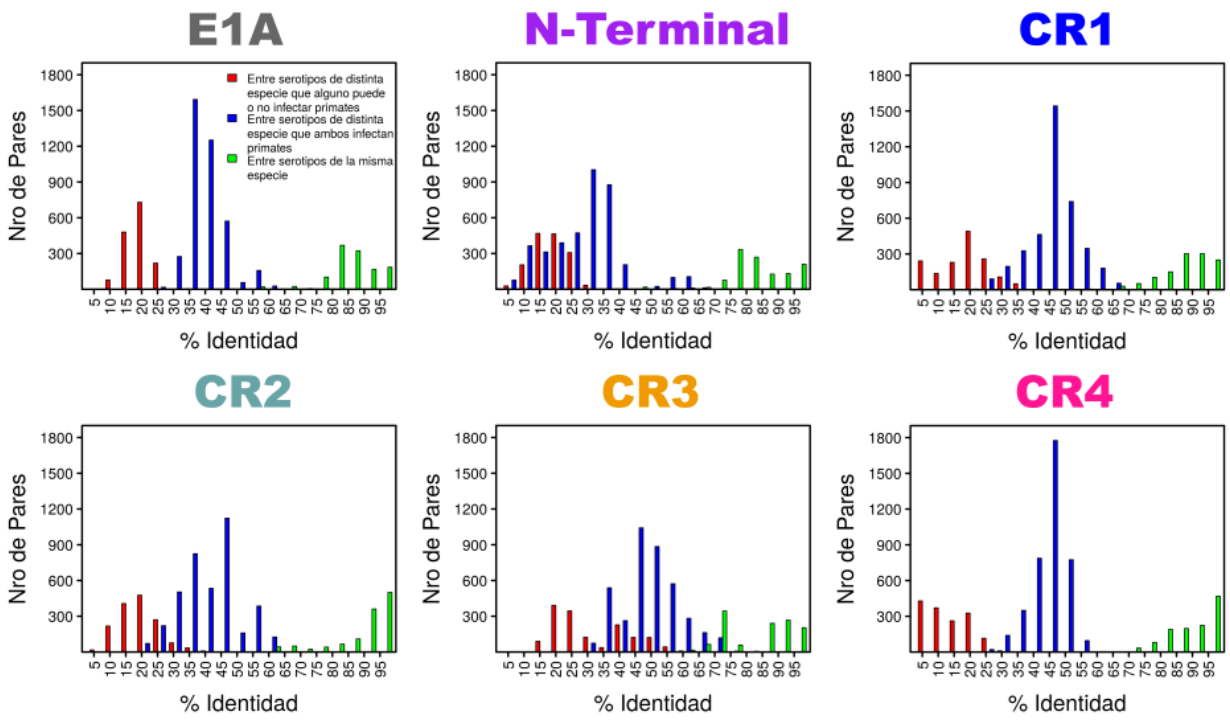


Figura 4.6: Distribución del porcentaje de identidad de secuencia entre pares de secuencias. Se muestra como la distribuciones del porcentaje de identidad de secuencia agrupado para la longitud total de la proteína E1A y para los dominios N-terminal, CR1, CR2, CR3 y CR4. Los pares de secuencia a los cuales se les calculó el porcentaje de identidad son agrupados en 3 grupos: pares de secuencias de E1A de serotipos que pertenecen a distintas especies y que alguno de los dos puede o no infectar a primates (barras rojas), pares de secuencias de E1A de serotipos de distinta especie que infectan a primates (barras azules) y los pares de secuencia de E1A de serotipos que comparten la misma especie (barras verdes). La leyenda utilizada para la proteína E1A es la misma para los dominios.

En resumen, en primer lugar se puede decir que las secuencias de E1A de serotipos que pertenecen a la misma especie comparten más de un 75 % de identidad. En segundo lugar, se confirmó lo observado visualmente en el alineamiento, las secuencias de E1A de serotipos que infectan a primates, efectivamente se parecen más entre sí que las que infectan o no a primates.

4.3.2. Regiones entre dominios funcionales de E1A

Además de los cinco dominios conocidos, se pueden observar en el alineamiento tres bloques de posiciones bien alineadas que no están presentes en todas las secuencias. Se llamó a cada una de estas regiones interdominio y se las denotó con las siglas IDR. A partir de los alineamientos de las regiones entre el CR1, CR2, CR3 y CR4 se definieron las regiones interdominio: IDR12, IDR23 e IDR34. Para la definición de cada región interdominio se consideró la longitud y similitud de secuencia.

Región interdominio 1-2. En la región del alineamiento entre los dominios CR1 y CR2, 57 secuencias no tenían ningún residuo, 42 secuencias tenían menos de diez residuos, 14 secuencias tenían entre 13 y 25 residuos y tres tenían una longitud mayor a 29 residuos. De las 14 secuencias,

doce pertenecen a la especie *Human adenovirus C* y comparten por lo menos un 75 % de identidad. Estas doce secuencias se utilizaron para definir la región interdominio 1-2 con una longitud promedio de 22 ± 4 residuos. Esta región está presente únicamente en los doce serotipos de la especie *Human adenovirus C*, que incluye los serotipos prototípicos HAdV2 y HAdV5. No se encontró literatura relacionada a la funcionalidad de esta región.

Región interdominio 2-3. En la región del alineamiento entre los dominios CR2 y CR3, 16 secuencias no tenían ningún residuo, once secuencias presentaron menos de doce residuos y las 89 secuencias restantes presentaron entre 14 y 25 residuos. El grupo de 89 secuencias comparten al menos un 18 % de identidad y pertenecen a las especies *Human adenovirus A*, *Human adenovirus B*, *Human adenovirus D*, *Human adenovirus E*, *Human adenovirus F*, *Human adenovirus G*, *Porcine adenovirus A*, *Simian adenovirus A*, *Simian adenovirus B* y *Simian adenovirus F*. Estas 89 secuencias se utilizaron para definir la región interdominio 2-3 (IDR23) con una longitud promedio de 16±2 residuos. Esta región se corresponde al espaciador identificado previamente en la proteína E1A de HAdV12 de la especie *Human adenovirus A* (Williams *et al.*, 2004) (véase Sección 1.4.5).

Región interdominio 3-4. En la región del alineamiento entre los dominios CR3 y CR4, 81 secuencias no presentaron residuos, 23 secuencias presentaron menos de 25 residuos y las doce secuencias restantes presentaron entre 32 y 38 residuos. El grupo de doce secuencias pertenecen a la especie *Human adenovirus C* y comparte por lo menos un 75 % de identidad. Con estas secuencias se definió la region interdominio 3-4 (IDR34) con una longitud promedio de 35 ± 3 residuos. Los primeros ocho residuos y los últimos seis residuos de está región pertenecen a las llamadas región auxiliar 1 y región auxiliar 2, respectivamente, caracterizadas previamente en la proteína E1A de HAdV5 (Bondesson *et al.*, 1992).

En conclusión, el análisis abarcativo de las 116 secuencias de E1A confirmó la presencia de los 5 dominios conocidos (N terminal, CR1, CR2, CR3 y CR4). También permitió la actualización de los límites de los dominios y la identificación de tres regiones interdominio presentes en doce (IDR12 e IDR34) y 89 secuencias (IDR23).

4.4. Abundancia y distribución por especie de los motivos lineales de E1A.

Se estudió la abundancia de cada motivo lineal entre las secuencias de las proteínas de E1A. Se determinó la presencia y ausencia de cada motivo restringiendo la búsqueda a la región de la proteína donde el motivo fue descripto originalmente.

Los motivos de interacción proteína-proteína pRb_ABGroove, LxCxE, el sitio de fosforilación de CKII y la región acídica son altamente prevalentes y están presentes entre el 94 y 97 % de las secuencias. Los motivos de interacción proteína-proteína CoRNR Box, MYND, CtBP y la NLS están presentes en un porcentaje sustancial de las secuencias (52 a 89 %). Sólo dos motivos de interacción proteína-proteína, IDMBR y TRAM-CBP, estuvieron presentes en menos del 10 % de

las secuencias de E1A. En el dominio CR3, el motivo de unión a zinc, CxxC, estaba presente en el 100 % de las secuencias y el 42 % tenía al menos una cisteína en una de las tres posiciones ricas en cisteínas (Tabla 4.1 y Figura 4.7).

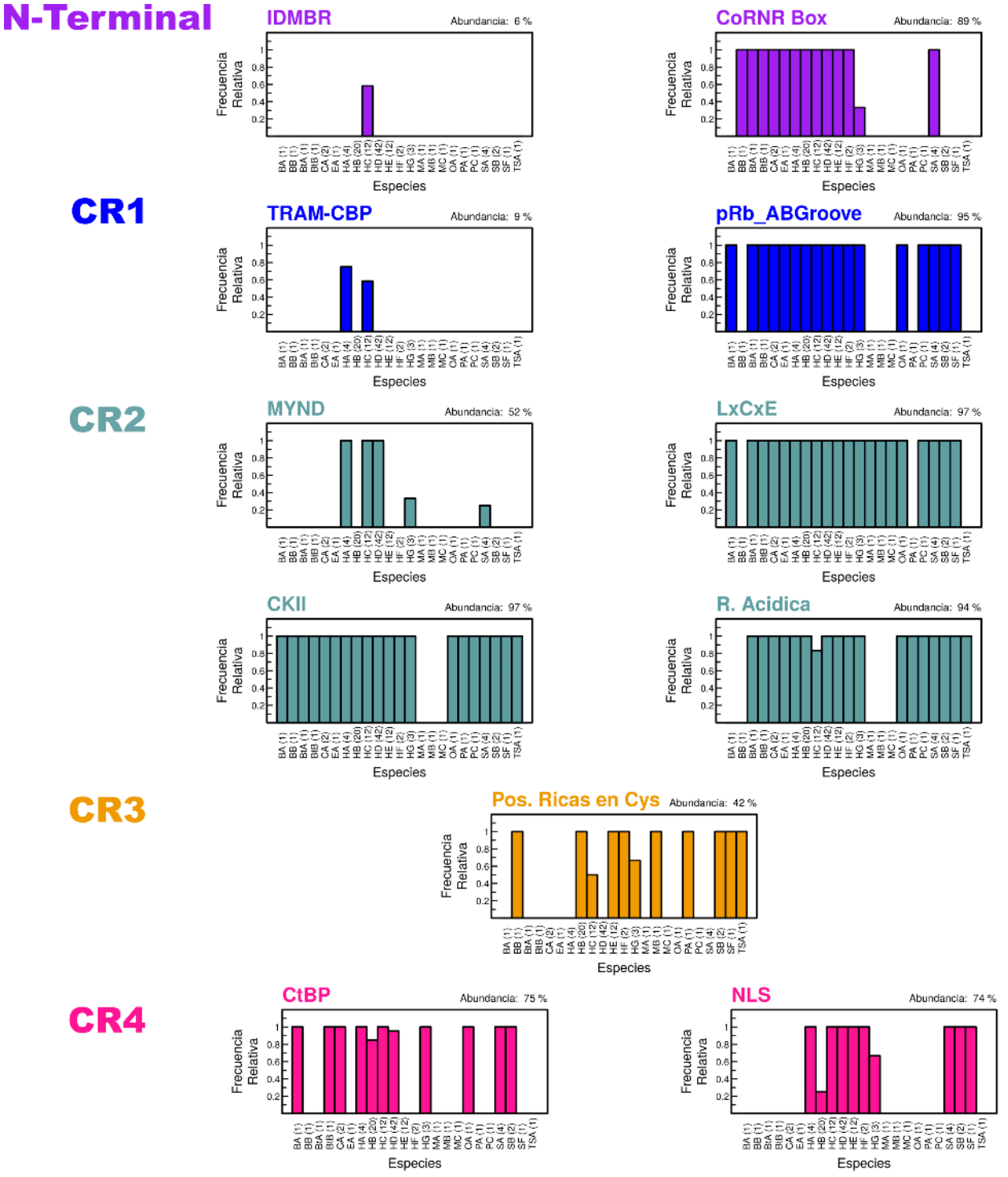


Figura 4.7: Distribución de los motivos por especie. Para cada dominio se muestra la frecuencia relativa de los motivos por especie. El porcentaje de abundancia se indica en la parte superior derecha del histograma. El código de colores es el mismo que se usa en la Figura 4.2.

La distribución de los motivos para cada especie viral no es homogénea (Figura 4.7). Por ejemplo, en el caso de *Human adenovirus C*, los motivos CoRNR Box, pRb_ABGroove, MYND,

LxCxE, CKII, CtBP y NLS se encuentran en el 100 % de los serotipos que la integran. Mientras que los motivos restantes IDMBR, TRAM-CBP, la región acídica y las posiciones ricas en cisteínas se encuentran entre el 50 y 83 % de los serotipos. La distribución de los motivos entre las distintas especies tampoco es homogénea. Si bien los motivos pRb_ABGroove, LxCxE, CKII tienen una distribución similar y están presentes entre el 80 y 100 % de las secuencias de las especies que lo poseen, el resto de los motivos tiene una distribución más variable, estando presentes a veces en el el 30 % de los serotipos de la especie y en otras especies en el 100 % (Figura 4.7). La distribución de los motivos tampoco es homogénea dentro de cada dominio, por ejemplo, el motivo MYND tiene una distribución mucho más variable que los motivos LxCxE, CKII y la región acídica.

Por último, tres de los motivos lineales no tienen una posición o longitud definida (Figura 4.8). Del 97 % de secuencias que poseen el motivo CKII, el ~50 % poseían un único sitio CKII y el resto dos (Figura 4.8A). Del 94 % de secuencias que mostraron una carga neta menor o igual a -4, el ~ 50 % presentaba una carga neta menor o igual a -6 (Figura 4.8B). Respecto a las posiciones ricas en cisteínas, ~ 30 % de las secuencias de E1A presenta dos posiciones y el ~10 % presenta una única posición (Figura 4.8C).

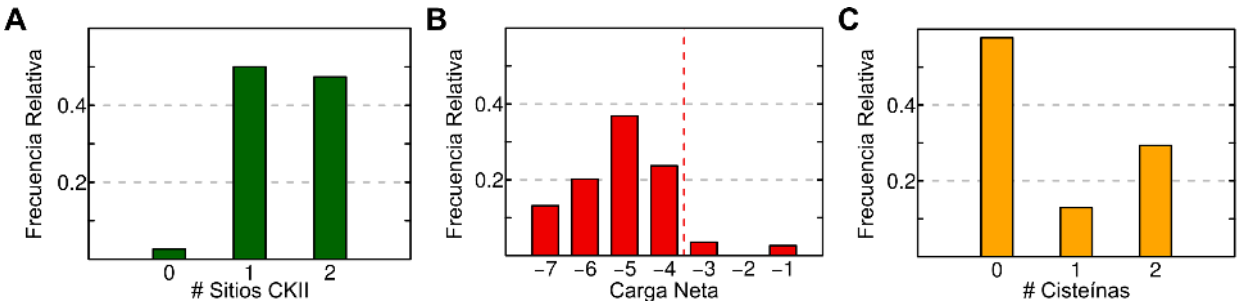


Figura 4.8: Prevalencia de motivos lineales sin posición o logitud fija. (A) Distribución del número de sitios de fosforilación de CKII en las secuencias de E1A. **(B)** Distribución de la carga neta de la región acídica en las secuencias de E1A. La línea punteada roja indica una carga neta de -4. Una carga neta menor o igual indica la presencia del motivo (Chemes *et al.*, 2012b). **(C)** Distribución del número de posiciones ricas en cisteínas en las secuencias de E1A.

En conclusión, la distribución de los motivos lineales de E1A es específica de cada motivo y los motivos lineales de E1A identificados en los serotipos paradigmáticos no son fácilmente extrapolables a otros serotipos.

4.5. Conservación de secuencia en E1A

A partir del alineamiento sin sitios vacíos (Sección B.2) se construyeron logos de secuencia (Schneider *et al.*, 1986) para resumir la organización de dominios y medir la conservación de secuencia a nivel de residuo y a nivel de dominio (Figura 4.9).

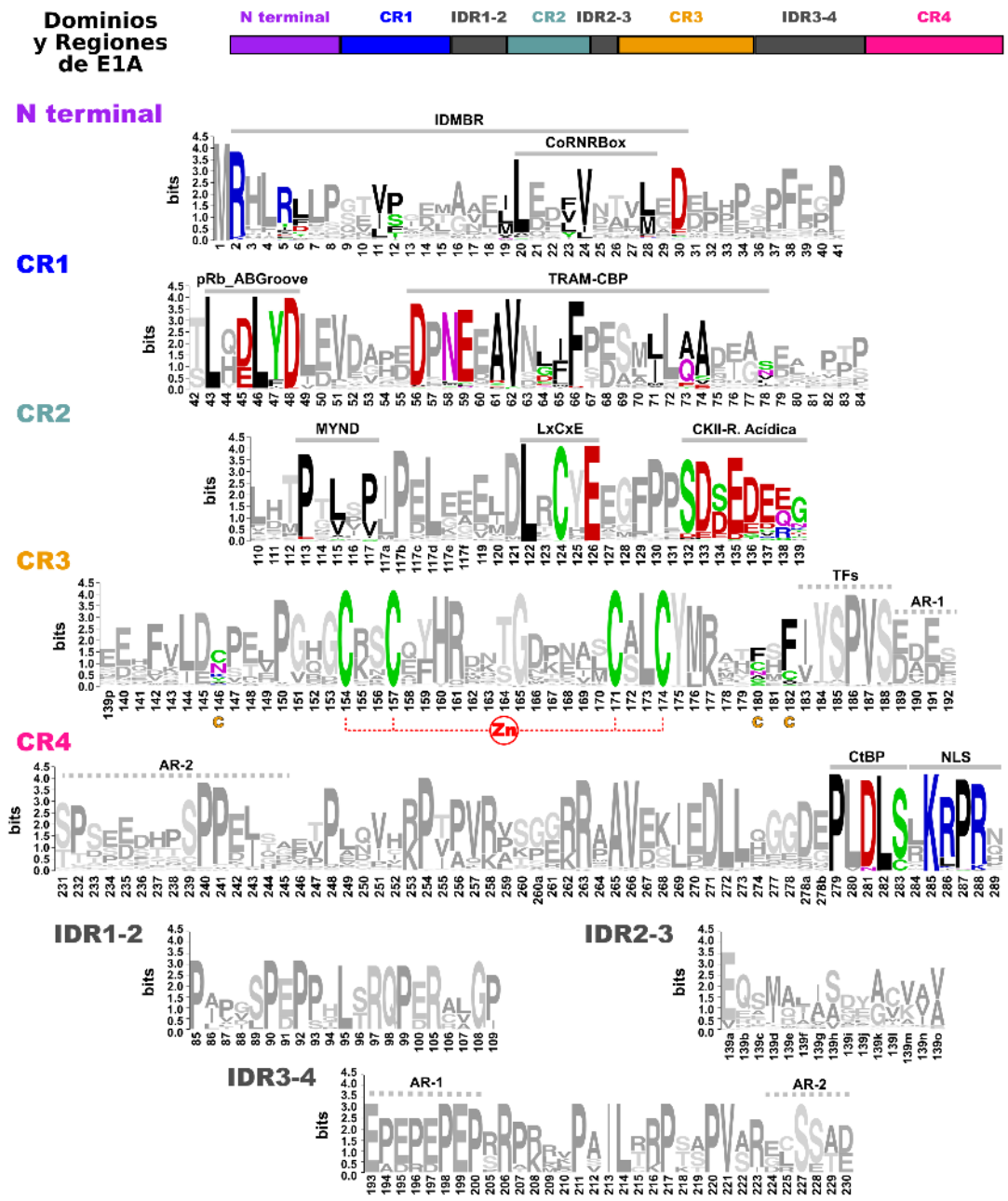


Figura 4.9: Conservación de secuencia de la proteína E1A de *Mastadenovirus*. Las posiciones que pudieron alinearse con confianza están representadas como logos de secuencia. De *arriba* a *abajo* Estructura de dominios y regiones de la proteína E1A. Dominio N-terminal (40 posiciones de 113 secuencias), CR1 (43 posiciones de 112 secuencias), CR2 (35 posiciones de 114 secuencias), CR3 (54 posiciones de 116 secuencias), CR4 (60 posiciones de 107 secuencias), IDR12 (21 posiciones de 12 secuencias) y IDR23 (15 posiciones de 89 secuencias) e IDR34 (33 posiciones de 12 secuencias). La numeración de secuencia se construyó a partir de la proteína HAdV5. Las inserciones en la secuencia de E1A de HAdV5 están indicadas con un número seguido de una letra. La posición de los motivos funcionales conocidos se indica en la parte superior de los logos (líneas grises). Las posiciones fijas de los motivos se muestran en color y el resto de las posiciones en escala de grises. Para el CR3 se muestra la coordinación del zinc por cuatro cisteínas (línea roja punteada), las tres posiciones ricas en cisteínas (letras C amarillas) y el sitio de unión de los factores de transcripción (TFs) (en inglés, *Transcription factors*). El comienzo de la región auxiliar 1 (*AR-1*) (en inglés, *Auxiliary region 1*) en el logo del CR3, el final de la AR-1, el comienzo de la región auxiliar 2 (AR-2) en el logo de la región IDR34 y el final de la región AR-2 en el logo del CR4 están señalados con una línea gris punteada. El código de colores para los dominios es el mismo que en la Figura 4.2.

En un logo de secuencia cada posición del alineamiento está representada como una columna de letras. La altura de cada columna, $IC(l)$ mide la conservación correspondiente a esa posición en bits (Schneider *et al.*, 1986) (Sección 2.1.9). Los aminoácidos presentes en cada posición están representados como letras en la columna y las alturas son proporcionales a las abundancias relativas.

En contraste a lo esperado para una proteína con un alto grado de desorden intrínseco, muchas posiciones de secuencia muestran una conservación media a alta. Comparamos el grado de conservación promedio de los cinco dominios y tres regiones interdominio de E1A. En la Figura 4.10 se muestra la conservación de las posiciones dentro de cada dominio/región como histogramas. El eje x indica los intervalos de $IC(l)$ considerados y el eje y indica la frecuencia relativa de las posiciones en cada dominio para cada intervalo de $IC(l)$.

Los dominios CR1 a CR4 muestran los mayores valores de conservación. Los valores de conservación promedio del CR1 (2.7 ± 0.8 bits), CR2 (2.9 ± 0.7 bits), CR3 (2.8 ± 0.9 bits) y CR4 (2.8 ± 0.8 bits) no son significativamente distintos entre ellos (valor $p^* > 0.05$) según las pruebas de permutación (véase Sección 2.1.11). El dominio N-terminal (2.3 ± 0.7 bits) y las regiones interdominio IDR12 (2.4 ± 0.7 bits), IDR23 (2.2 ± 0.6 bits) e IDR34 (2.5 ± 0.7 bits) mostraron valores de conservación menores. Estos valores promedio de conservación no fueron significativamente distintos entre ellos (valor $p^* > 0.05$). En términos de conservación, estos dos grupos no pudieron ser separados estadísticamente, con excepción del dominio N-terminal cuyo valor de conservación promedio es significativamente menor que el valor promedio de conservación del CR1, CR2 y CR4 (valor $p^* < 0.05$).

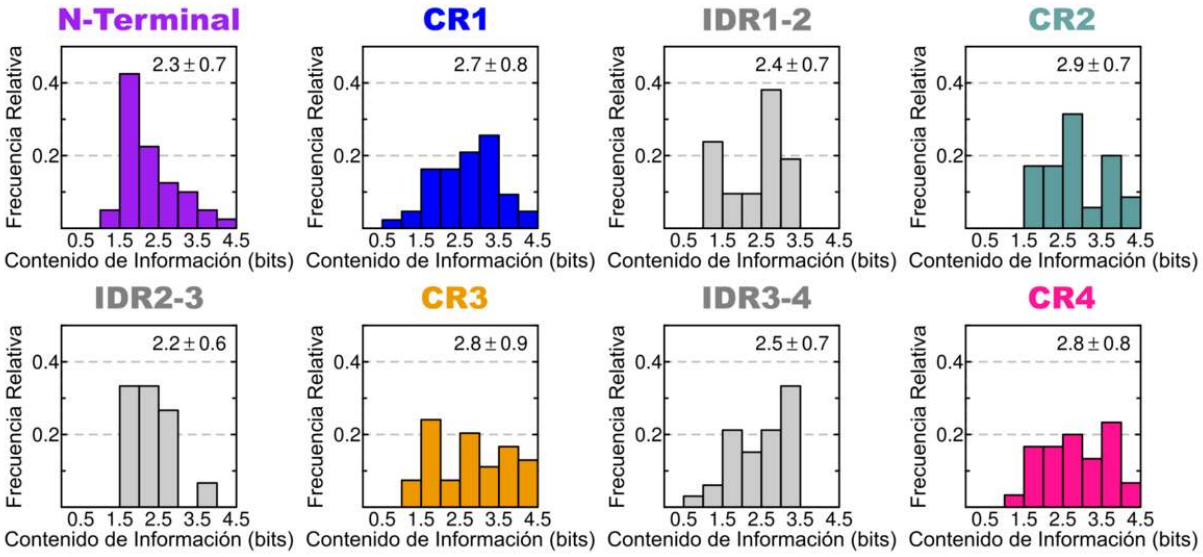


Figura 4.10: Conservación de secuencia de los dominios y regiones de la proteína E1A. Se muestra en forma de histogramas la distribución del contenido de información de cada dominio de la proteína E1A. En la parte superior del histograma se indica el valor promedio y el desvío estándar. El código de colores es el mismo que el utilizado en la Figura 4.2.

En resumen, el grado de conservación general de la proteína E1A es alto y los dominios CR1

a CR4 de la proteína E1A están tan conservados como las regiones interdominio. En las próximas secciones analizaremos la conservación de secuencia de E1A en términos de sus propiedades estructurales y sitios funcionales conocidos.

4.6. Desorden intrínseco en la proteína E1A

El análisis computacional de seis secuencias prototípicas de E1A sugirió que los dominios CR1, CR2 y CR4 carecen una estructura globular bien definida y que pueden ser consideradas intrínsecamente desordenadas, mientras que el dominio N-terminal es principalmente desordenado y el dominio CR3 es principalmente ordenado (Pelka *et al.*, 2008).

Se evaluó la generalidad de este resultado estudiando la tendencia al desorden para nuestra base de datos de 116 secuencias con el algoritmo IUPred (Mészáros *et al.*, 2018) (Sección 2.2.3). El valor de IUPred toma valores entre 0 y 1. Un valor superior a 0.5 indica desorden y un valor menor a 0.5 indica orden (Daughdrill *et al.*, 2011; Mészáros *et al.*, 2018). Estos resultados están representados como un gráfico de puntos (Figura 4.11) donde el eje *x* indica la posición de secuencia y el eje *y* indica el valor promedio de IUPred y el desvío estándar para las posiciones alineadas confiablemente.

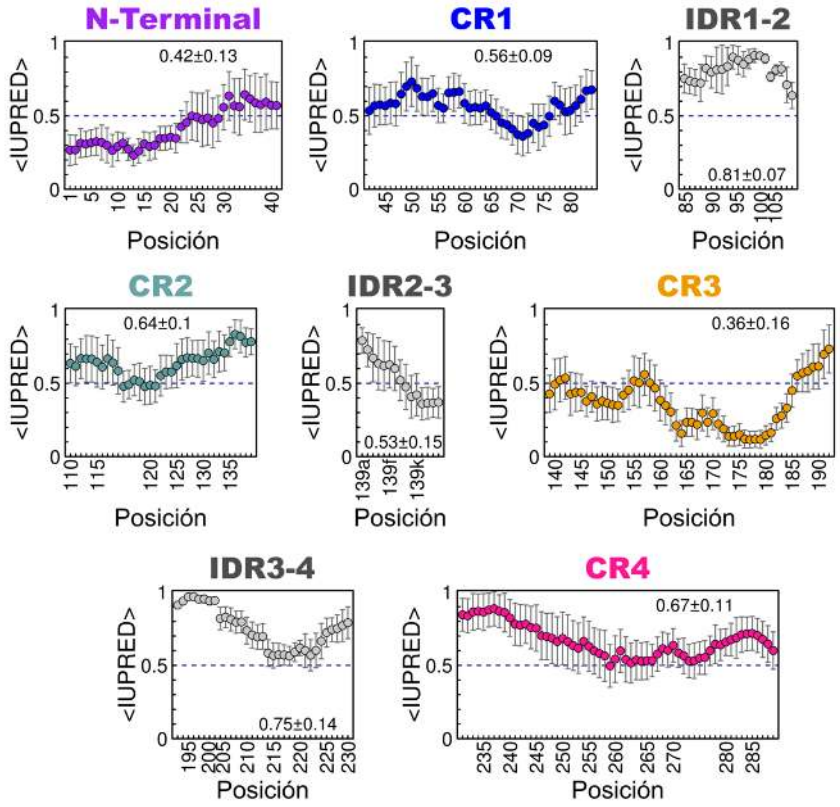


Figura 4.11: Predicción del desorden intrínseco para cada dominio y región de la proteína E1A. Se muestra el índice de IUPred de cada dominio de la proteína E1A para las posiciones alineadas de manera confiable. La numeración de la secuencia se origina en la proteína E1A de HAdV5. El valor promedio y desvío estándar se muestran en la parte superior de cada gráfico. El código de colores es el mismo utilizado en la Figura 4.2.

Los valores promedio de IUPred de los dominios CR1 (0.56 ± 0.09), CR2 (0.64 ± 0.1) y CR4 (0.67 ± 0.11) y las regiones IDR12 (0.81 ± 0.07) e IDR34 (0.75 ± 0.14) sugieren que estos dominios y regiones son desordenadas (valor $p< 0.01$). Estos resultados concuerdan con el hecho de que la mayoría de las posiciones de los dominios CR1 (74 %), CR2 (86 %) y CR4 (98 %) y todas las posiciones de las regiones IDR12 e IDR34 muestran un valor promedio de IUPred superior a 0.5. Los valores promedio de IUPred del dominio N-terminal (0.42 ± 0.13) y del dominio CR3 (0.36 ± 0.16) sugieren que estos dominios son ordenados (valor $p < 0.01$). Al igual que antes, estos resultados concuerdan con el hecho de que la mayoría de las posiciones del dominio N-terminal (73 %) y del CR3 (76 %) muestran un valor promedio de IUPred por debajo de 0.5. El dominio N-terminal parece ser más ordenado que el extremo carboxi terminal. Esto puede deberse a la presencia de una hélice anfipática propuesta en el dominio N-terminal de la proteína E1A de HAdV5 (Pelka *et al.*, 2008). En el caso del dominio CR3, una región globular central parece estar flanqueada por unas regiones más flexibles. De manera similar al dominio N-terminal, el valor promedio de IUPred del IDR23 es cercano a 0.5 (0.53 ± 0.15), con una primera mitad desordenada y una segunda mitad ordenada.

En resumen, confirmamos que los dominios CR1, CR2 y CR4 y las regiones IDR12 e IDR34 son intrínsecamente desordenadas en nuestra base de datos abarcativa, mientras que el dominio N-terminal y la región IDR23 son parcialmente ordenadas y la región CR3 es principalmente globular.

4.6.1. Desorden y conservación de secuencia en la proteína E1A

Comúnmente se espera que los dominios intrínsecamente desordenados estén menos conservados que los dominios globulares (Daughdrill *et al.*, 2011; Toth-Petroczy *et al.*, 2008). Sin embargo, esto no es siempre es así (Chemes *et al.*, 2012a).

Para evaluar esta hipótesis para la proteína E1A se compararon los valores por posición de conservación en *bits* con el valor promedio de desorden por posición y se evaluó la correlación entre ambas mediciones (Figura 4.12). Se realizó un gráfico de puntos (Figura 4.12) donde el eje x indica el valor promedio de IUPred por posición de secuencia y el eje y indica el valor promedio de del contenido de información por posición en *bits*. Para la proteína E1A no se observa correlación entre las mediciones, como tampoco para los dominios desordenados (CR1, CR2 y CR4), o para el dominio N-terminal parcialmente ordenado o para el dominio CR3 predicho globular.

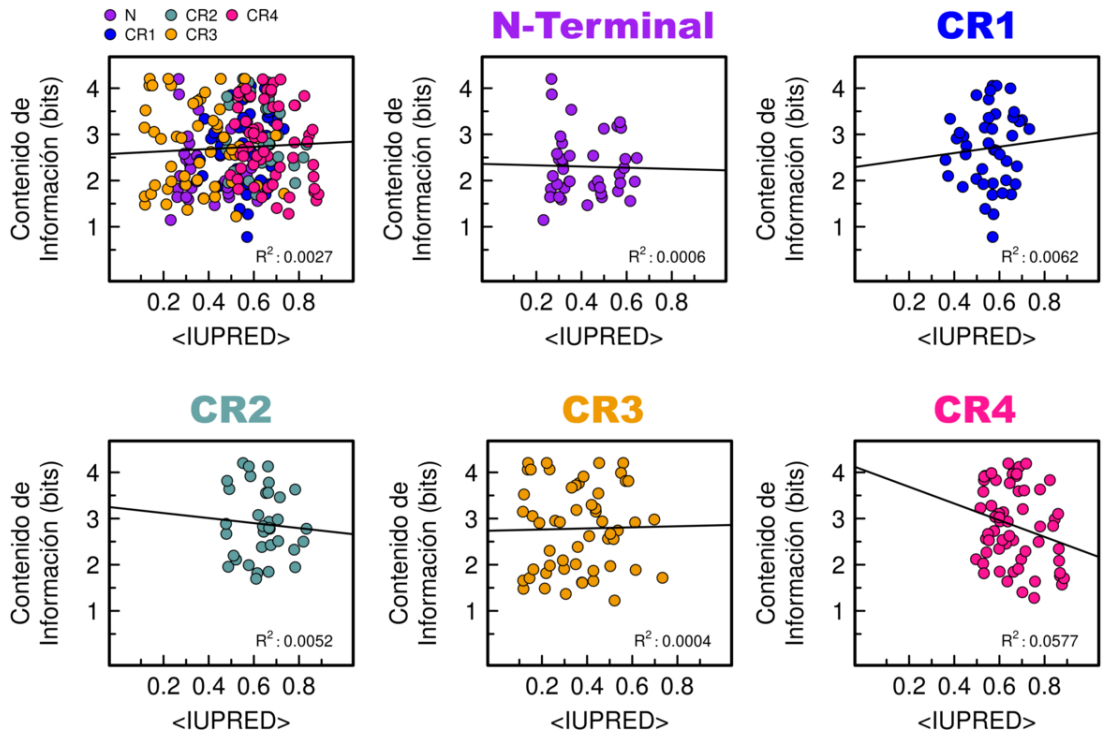


Figura 4.12: Comparación entre conservación de secuencia y desorden de los dominios y regiones de la proteína E1A. En el eje *y* se indica el valor por posición del contenido de información en bits. En el eje *x* se indica el valor promedio de IUPred por posición. En la parte inferior derecha del gráfico se indica el cuadrado del coeficiente de correlación pearson. El código de colores es el mismo que el utilizado en la Figura 4.2.

Por lo tanto, la conservación de secuencia de los dominios de E1A no está únicamente dictada por el grado de desorden intrínseco.

4.7. Motivos lineales y conservación de secuencia

Se analizó la relación entre los doce motivos lineales de la proteína E1A y la conservación de secuencia a nivel de residuo. Los motivos lineales están resaltados en la parte superior e inferior de cada logo de secuencia en la Figura 4.9. En una visión general de la Figura 4.9 se puede decir que la conservación de las posiciones de secuencia que pertenecen a un motivo sigue aproximadamente la prevalencia del motivo.

Existen tres clases distintas de posiciones de secuencia en un motivo lineal (Chemes *et al.*, 2015) (véase Sección 2.1.5). Las posiciones fijas, determinantes del motivo, muestran un pequeño número de aminoácidos permitidos que se muestran en color en la Figura 4.9 y en negro en la Figura 4.13. Las posiciones comodín de un motivo permiten cualquier aminoácido (Figura 4.13, en rojo). Las posiciones adyacentes al motivo incluyen las tres posiciones anteriores al inicio y las tres posiciones siguientes al final del motivo (Figura 4.13, en celeste). Finalmente, las posiciones que no pertenecen a un motivo lineal son distinguidas como “otras” (Figura 4.13, en verde) en este análisis. Las posiciones comodín, adyacentes y “otras” se muestran en gris en la Figura 4.9.

Se analizó el grado de conservación de estas cuatro clases de posiciones de secuencia utilizando el contenido de información, $IC(l)$, como medida de la conservación (véase Sección 2.1.8). Para este análisis las posiciones correspondientes al motivo de fosforilación de CKII, la región acídica y las posiciones ricas en cisteínas fueron incluidas dentro de "otras". Esta decisión se basó en que los tres motivos son variables en número de posiciones que definen su presencia y ubicación en la secuencia (véase Sección 4.4 y Figura 4.8). Los resultados se muestran como histogramas en la Figura 4.13.

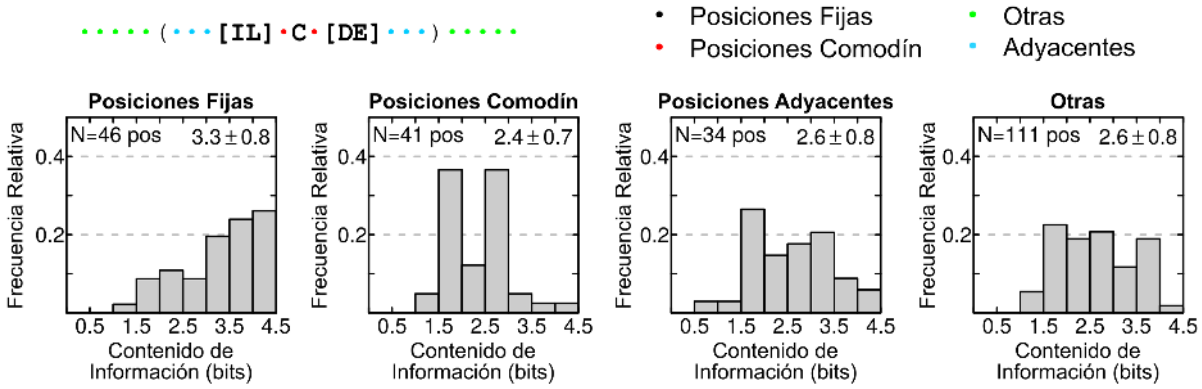


Figura 4.13: Conservación de secuencia de los motivos lineales. Se muestra en forma de histograma la distribución del contenido de información para las posiciones fijas, comodín y adyacentes relativas a cada motivo de E1A y para las posiciones restantes (Otros) de E1A. El valor promedio y el desvío estándar correspondientes se indican en la parte superior de cada histograma así como el número de posiciones consideradas (N). En la parte superior del gráfico se muestra una leyenda que indica las posiciones que se consideran fijas, comodín y adyacentes utilizando como ejemplo la expresión regular del motivo LxCxE.

Para evaluar si existía una diferencia significativa entre los valores de conservación promedio de las cuatro clases de posiciones se realizaron pruebas de permutación para calcular el valor p y utilizamos la corrección de Benjamini-Hochberg para comparaciones múltiples (Benjamini y Hochberg, 1995) (véase Sección 2.5.2). El valor promedio de conservación de las posiciones fijas (3.3±0.8 bits) fue significativamente mayor (valor $p < 0.05$) que el valor promedio de conservación de las posiciones comodín (2.4±0.7 bits), adyacentes (2.6±0.8 bits) y "otras" (2.6±0.8 bits). No se observaron diferencias significativas entre el valor de conservación promedio entre las posiciones comodín, adyacentes y el resto.

En conclusión, los nueve motivos lineales de la proteína E1A analizados en esta sección afectan la conservación de secuencia principalmente en las posiciones fijas, determinantes del motivo.

4.8. Repertorio de motivos lineales

Además de determinar la prevalencia de los motivos individuales (Figura 4.7), se analizó el repertorio de los motivos de nuestras 116 secuencias de E1A (Figura 4.14). Los doce motivos conocidos de E1A pueden aparecer en 2^{12} combinaciones posibles, es decir, existen 4096 potenciales repertorios diferentes de motivos para esta proteína. Sin embargo, solamente 25 combina-

ciones son observadas en las secuencias naturales de E1A y 22 de ellas están presentes en cinco o menos serotipos. Esto sugiere que muchas de las combinaciones de motivos fueron seleccionadas negativamente durante la evolución de adenovirus. La combinación más abundante de los motivos está presente en 42 serotipos e incluye nueve de los motivos lineales más prevalentes: CoRNR Box, pRb_ABGroove, MYND, LxCxE, CKII, la región acídica, el motivo de unión a zinc, CtBP y NLS. Los seis motivos lineales más prevalentes, CoRNR Box, pRb_ABGroove, LxCxE, la región acídica, el motivo de fosforilación de CKII y el motivo de unión a zinc están presentes en las 10 combinaciones más abundantes. Así, existe una correlación entre las prevalencias de motivos y el repertorio de motivos lineales de E1A.

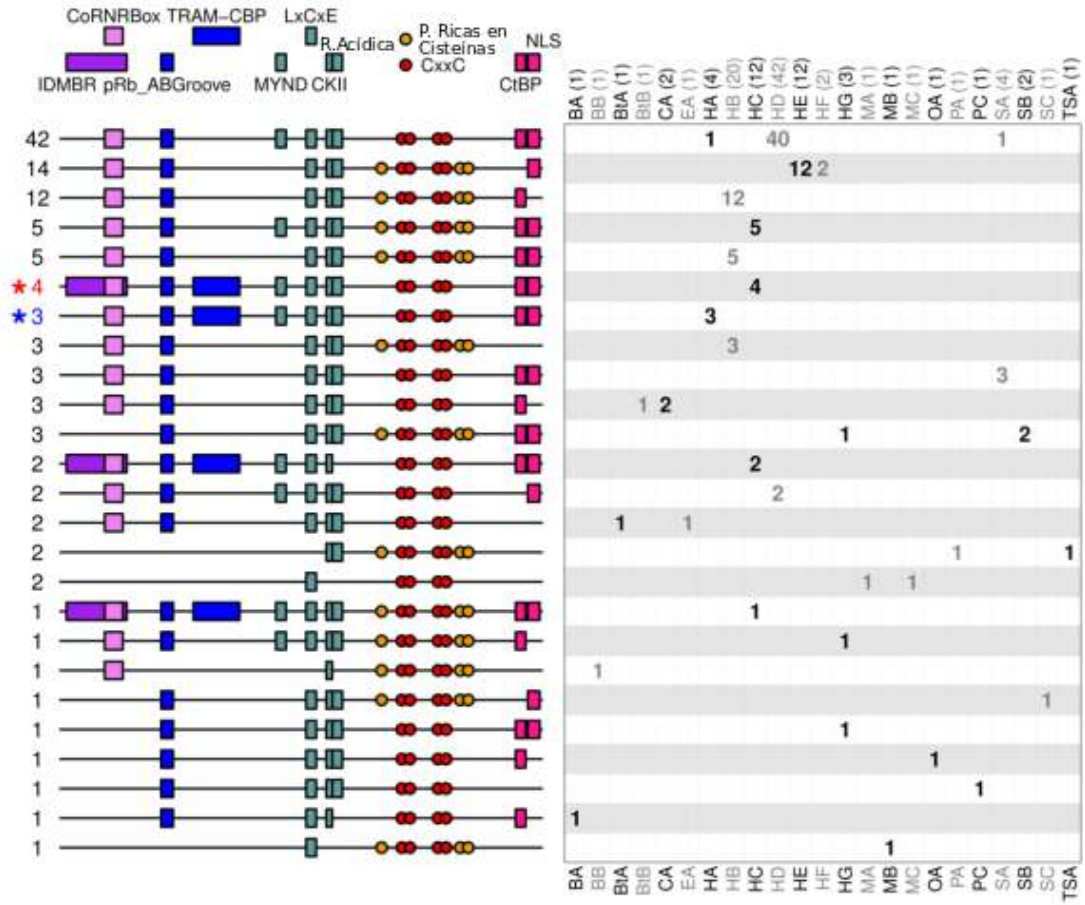


Figura 4.14: Repertorio de las combinaciones de motivos lineales. Representación esquemática de las 25 combinaciones lineales de motivos que halladas en las 116 secuencias. Todos los motivos y la representación gráfica correspondiente se indican en la parte superior del gráfico. El número total de secuencias que posee cada combinación de motivos lineales se indican a la izquierda. Las combinaciones lineales que se encuentran en los serotipos prototípicos HAdV5 y HAdV12 se resaltan con asteriscos rojo y azul respectivamente. El número de secuencias y distribución dentro de las especies de *Mastadenovirus* se muestran a la derecha. El numero total de secuencias de cada especie está indicado entre paréntesis al lado del nombre de la especie en la parte superior. Los motivos lineales están representados como rectángulos utilizando el código de colores de dominios como en la Figura 4.2, con excepción del CoRNR Box (rectángulo rosa claro) y el motivo de unión a zinc (círculos rojos). El tamaño de cada rectángulo, círculo o línea no está relacionado con el tamaño de la proteína o motivo lineal.

Además, se investigó la distribución filogenética de los repertorios de los motivos de E1A. Los resultados se muestran en la Figura 4.14. En el panel de la izquierda se muestran las combinaciones observadas en las secuencias de E1A y en el panel de la derecha el número de serotipos por especie que tienen esa combinación.

Las combinaciones de motivos lineales presentes en múltiples serotipos están frecuentemente presentes en múltiples especies virales, y muchas especies virales con más de un único serotipo en la base de datos poseen múltiples combinaciones de motivos lineales. Esto muestra que la filogenia de adenovirus no determina completamente el repertorio de motivos lineales de la proteína E1A. Por otro lado, se observa una fuerte asociación entre algunos repertorios de motivos y serotipos, por ejemplo, el repertorio de motivos más abundante y la especie *Human adenovirus D* (HD). Es importante resaltar que las proteínas prototípicas HAdV5 E1A y HAdV12 E1A (combinaciones marcadas con un asterisco rojo y azul respectivamente en la Figura 4.14), no pertenecen a ninguna de las cinco combinaciones más abundantes.

En resumen, el repertorio de motivos de E1A correlaciona tanto con la prevalencia de motivos como la filogenia de adenovirus, pero no está totalmente determinado por estos dos factores.

4.9. Coevolución de secuencia en la proteína E1A

Existe poca información estructural de la proteína E1A de adenovirus (véase Sección 1.4.5). En las siguientes secciones se utilizan distintos métodos computacionales para obtener información sobre las relaciones estructura-función presentes en E1A. En primer lugar, se analiza la coevolución de pares de residuos en el alineamiento de secuencia de E1A. Las señales de coevolución son indicador confiable de contacto físico conservado entre pares de residuos, tanto dentro de una proteína (Morcos *et al.*, 2011) como entre cadenas de proteínas (Cheng *et al.*, 2014). Esto también aplica para las proteínas intrínsecamente desordenadas (Toth-Petroczy *et al.*, 2016). Se realizó un análisis de información directa para deducir contactos entre residuos existentes en la proteína E1A (véase Sección 2.2.5). La implementación de este método permite diferenciar señales de coevolución directas e indirectas y tiene en cuenta la redundancia en las secuencias alineadas (Espada *et al.*, 2015).

Se identificaron un total de 95 pares de residuos coevolucionando dentro de E1A. Estos pares están representados en la Figura 4.15 como un mapa de contactos predichos, es decir, una representación bidimensional de los pares de aminoácidos hallados por información directa. De los 95 pares de residuos identificados, 41 pares se encuentran a una distancia de cinco o menos residuos, 31 pares están separados por seis a 30 residuos mientras que los 23 pares restantes se encuentran separados por 31 o más residuos. Los pares de residuos que coevolucionan dentro de un mismo dominio (69 pares, 73 %) están coloreados por dominio en la Figura 4.15. Los cinco dominios de E1A presentan pares intra dominio coevolucionando, aunque con distinta abundancia. El dominio CR3 y la región IDR23 tienen la mayor proporción de pares coevolucionando ($\sim$ 0.5 pares por residuo). Los dominios N-terminal, CR2 y CR4 presentan proporciones intermedias entre 0.2 y 0.3 pares por residuos y el dominio CR1 presenta las proporción más baja con $\sim$ 0.12 pares por re-

siduo. Los restantes 26 pares de residuos que coevolucionan ocurren entre dominios (Figura 4.15, círculos negros). Los pares que coevolucionan están presentes tanto en dominios desordenados como ordenados. En los dominios desordenados, cinco pares coevolucionan dentro del dominio CR1, diez dentro del CR2 y 15 dentro del CR4. En las regiones parcialmente ordenadas, coevolucionan ocho pares dentro del dominio N-terminal y siete pares dentro de la región IDR23. En el dominio CR3 coevolucionan un total de 24 pares.

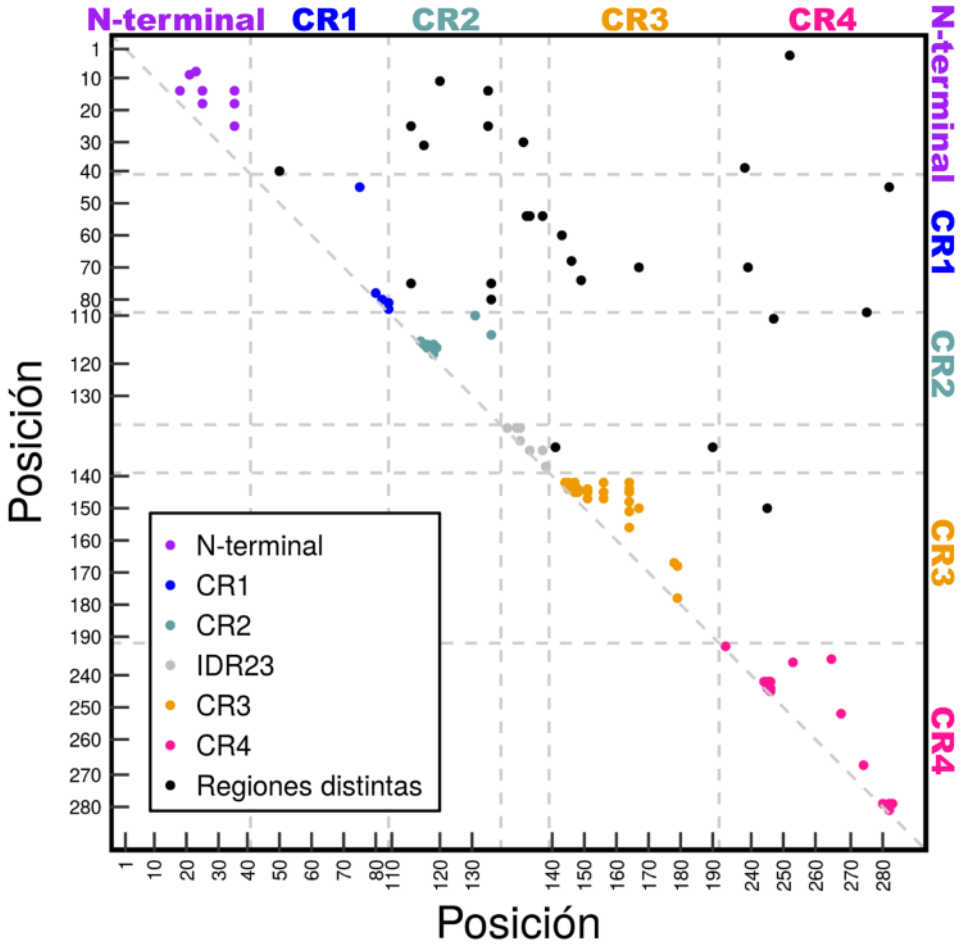


Figura 4.15: Predicción del mapa de contactos de la proteína E1A utilizando información directa. Los pares de residuos con alto valor de información directa están representados en forma de puntos. Los contactos dentro del mismo dominio o misma región están señalados utilizando el mismo código de colores que en la Figura 4.2. Los contactos entre distintos dominios o regiones están indicados en negro. Los distintos dominios están separados por líneas punteadas. La numeración de secuencia utilizada se basa en la proteína E1A de HAdV5. Para una mejor visualización, las predicciones de contacto se muestran únicamente en la parte superior de la diagonal.

En resumen, las señales de coevolución encontradas sugieren una cantidad importante de contactos residuo-residuo tanto intra- como interdominio en la proteína E1A. Este resultado es contrario a lo esperado de la descripción de E1A como una proteína mayormente desordenada.

4.10. E1A no se comporta como un polímero entrópico.

Por un lado, la predicción de IUPred sugiere que la proteína E1A presenta extensas regiones y dominios intrínsecamente desordenados y pequeñas regiones de orden local restringidas al CR3, el dominio N-terminal y la región IDR23. Por otro lado, los resultados obtenidos por el análisis de información directa sugieren que existen múltiples contactos conservados entre los residuos de E1A que están separados por grandes distancias en secuencia. Estos dos descubrimientos parecen ser contradictorios, en cierta medida. A continuación, se estudia si los contactos predichos utilizando el método de información directa se desvían de lo esperado para un polipéptido completamente desestructurado.

Se utilizó como modelo una cadena entrópica (Zhou, 2003, 2004) para representar una proteína E1A completamente desordenada. En este modelo, las únicas restricciones conformacionales son evitar los choques de la proteína consigo misma y la longitud de persistencia del polímero. La longitud de persistencia (l_p) es una propiedad mecánica básica que cuantifica la rigidez de un polímero. Informalmente, para los segmentos del polímero que son más cortos que la longitud de persistencia, la molécula se comporta más bien como un bastón flexible y elástico, mientras que para segmentos del polímero que son mucho más largos que la longitud de persistencia, las propiedades pueden ser solamente descriptas de manera estadística. Utilizando la teoría de polimeros (véase Sección 2.2.6) se puede modelar la formación de un contacto físico entre dos aminoácidos que están separados por L residuos en la cadena de la siguiente manera. Consideramos una constante de equilibrio genérica para la asociación de dos aminoácidos que están libres en solución, $K_{intermolecular}$. Si los dos aminoácidos están embebidos dentro de una proteína, la constante de asociación correspondiente $K_{intramolecular}$ será la constante de asociación intermolecular penalizada entrópicamente ($K_{intermolecular}$) por restringir la conformación de la proteína al atraer los dos aminoácidos dentro de una distancia de contacto. Esta penalidad entrópica es llamada concentración efectiva o C_{ef} (Zhou, 2003). Así,

$$\log K_{intramolecular} = \log K_{intermolecular} + \log C_{ef} \tag{4.2}$$

El valor de C_{ef} depende del número de residuos L entre dos aminoácidos y de la distancia de contacto b y puede ser fácilmente calculada utilizando la ecuación empírica en (Zhou, 2003, 2004). Este enfoque se utilizó de manera exitosa para modelar el comportamiento de conectores proteicos desordenados en diferentes sistemas no relacionados (Zhou, 2003).

Se examinó si los contactos intracatenarios predichos a partir del análisis de pares de residuos en E1A son compatibles con la descripción de E1A como una cadena entrópica. En primer lugar, para cada valor L de distancia de secuencia, se calculó la $K_{intramolecular}$ utilizando el número de todos los contactos posibles y el número de los contactos predichos (véase Sección 2.2.6). Dado que el número de contactos predichos es bajo, se realizó un promedio de la $K_{intramolecular}$ en una ventana de 14 residuos. El tamaño de ventana no alteró de manera significativa los resultados (véase Sección F.6.1). Estos resultados se muestran como puntos en la Figura 4.16, donde en el eje

x se indica la separación entre los residuos L y el eje y se indica la $K_{intramolecular}$ utilizando escala logarítima. $K_{intramolecular}$ presenta un mínimo alrededor de 50 residuos de separación y aumenta tanto hacia valores menores y mayores de L. Luego, se evaluó el poder del modelo de cadena entrópica para describir nuestros resultados ajustando la Ecuación 4.2 a los datos. Se obtuvo un valor de $9.5 \cdot 10^{-5} \pm 0.8 \cdot 10^{-5}$ mM^{-1} para $K_{intermolecular}$ y de 6.1 ± 1.4 Å para la distancia de contacto $C_\alpha - C_\alpha$. Este último valor es aceptado para una distancia de contacto en proteínas globulares. La teoría describe bien los datos para una separación de hasta 100 residuos en secuencia, pero subestima la $K_{intramolecular}$ para distancias de secuencia mayores.

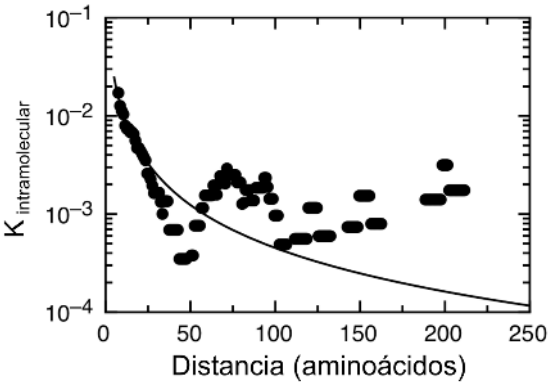


Figura 4.16: Probabilidad de formación de contacto en E1A es función de la separación de secuencia. La constante de equilibrio para la formación de contactos entre aminoácidos $K_{intramolecular}$ fue estimada como función de la separación de secuencia L utilizando el mapa de contactos predicho (puntos) y el modelo de cadena entrópica (Zhou, 2004). Los parámetros que mejor ajustan la ecuación que describe la formación de contactos intramoleculares de una cadena entrópica (línea) son: una distancia de contacto $C_\alpha - C_\alpha$, r, de 6.1 ± 1.4 Å y una constante intermolecular, $K_{intermolecular}$, de $9.5 \cdot 10^{-5} \pm 0.8 \cdot 10^{-5} \mathrm{mM}^{-1}$.

En conclusión, el modelo de cadena entrópica no explica el alto número de contactos de largo rango predichos entre los residuos de E1A. Este resultado sugiere que E1A no es un polipéptido completamente desestructurado.

4.11. Predicción estructural del dominio CR3 de E1A

En conjunto, los estudios de RMN (Ferreon *et al.*, 2009), la predicción de desorden realizada en Pelka *et al.* (2008) y en este trabajo y los estudios funcionales con E1A totalmente desnaturalizada por calor (Krippl *et al.*, 1984) indican que el dominio CR3 de E1A es el único que adopta una conformación globular. La información disponible y métodos computacionales (véase Sección 2.2.7) se utilizaron para elaborar un modelo estructural del dominio CR3 de E1A (Figura 4.17), que puede ser útil para entender la función del CR3.

Construcción del modelo. En total se utilizaron tres tipos de información para construir el modelo: los contactos entre residuos, la coordinación del zinc por las cuatro cisteínas y la predicción de estructura secundaria. El análisis de información directa proporcionó un grupo de 14 pares de

aminoácidos dentro del dominio CR3 que están predichos con alta confianza que están en contacto (Figura 4.17B, arriba de la diagonal). El dominio CR3 de E1A contiene cuatro cisteínas que están totalmente conservadas y unen un átomo de zinc (Culp *et al.*, 1988; Webster *et al.*, 1991). Liu *et al.* (2006) muestra que la estructura del dominio de unión a zinc de la proteína E7 de papilomavirus contiene un sitio de unión a zinc Cys4 de estructura planar con una distancia de 10 Å entre cisteínas opuestas y 7 Å entre las cisteínas enfrentadas y cercanas en secuencia. Estos datos se incluyeron en nuestro análisis como información adicional en la predicción de contactos del modelo (Figura 4.17A) y están indicados como triángulos en la Figura 4.17B. También se consideró un arreglo alternativo de las cuatro cisteínas con la Cys157 enfrentando a la Cys174 y la Cys154 enfrentando a la Cys171. La predicción de estructura secundaria utilizando el algoritmo JPred sugiere que los residuos 160 a 162 y 169 a 178 adquieren una conformación de hélice α (Figura 4.17B).

Se utilizó una estructura desplegada inicial del dominio CR3 de la proteína E1A de HAdV8 para simulaciones de dinámica molecular, utilizando una representación de C_α reducida (Sułkowska *et al.*, 2012) y las restricciones que se muestran en la Figura 4.17A y la Figura 4.17B. La estructura inicial se sometió a ciclos de calentamiento/enfriamiento para desplegar y colapsar el dominio. Como resultado se obtuvo un conjunto recocido (en inglés, *annealed ensemble*) homogéneo de estructuras que satisfacían la mayoría de las restricciones del modelo con el arreglo de unión a zinc mostrado en la Figura 4.17A.

En particular, el 80 % de la predicciones de contacto se encontraban en las estructuras del recocido (Figura 4.17B, debajo de la diagonal). Este valor es típico para las simulaciones de dinámicas del estado nativo para dominios de estructuras conocidas. La distancia entre las cisteínas enfrentadas (Cys154-Cys174: 7.07 ± 0.3 Å, Cys157-Cys171: 7.07 ± 0.25 Å, Figura 4.17C parte inferior) y entre las cisteínas opuestas (Cys154-Cys171: 9.91 ± 0.23 Å, Cys157-Cys174: 9.92 ± 0.27 Å, Figura 4.17C parte superior) fue consistente con las restricciones elegidas (Figura 4.17A). El 93 % de la estructura secundaria predicha se mantuvo al menos el 40 % del tiempo (Figura 4.17B). La Figura 4.17D muestra el RMSD para el conjunto recocido calculado utilizando como referencia las coordenadas promedio. Se observa en el modelo una estructura globular bien definida en la región N-terminal del dominio (Figura 4.17D) seguida de una región C-terminal más flexible (Figura 4.17D).

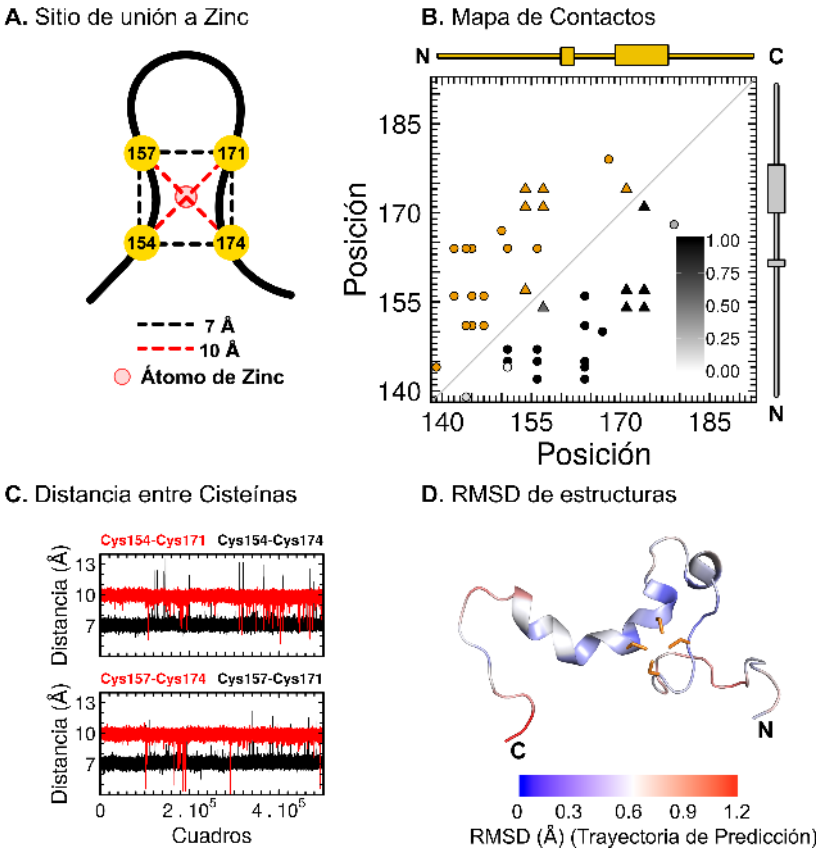


Figura 4.17: Construcción del modelo estructural del dominio CR3 de E1A. (A) Configuración del sitio de unión a zinc. Los cuatro C_α de las cisteínas que coordinan el átomo de zinc (círculo rojo claro) están representados como círculos amarillos y se muestran las interacciones utilizadas como restricciones en la simulación (líneas punteadas). **(B)** Predicción del mapa de contactos y estructura secundaria. Las restricciones de contactos utilizadas para la predicción de la estructura están representadas como puntos amarillos (arriba de la diagonal) o coloreados según la probabilidad de aparición en el conjunto recocido (debajo de la diagonal). Los seis contactos de las cisteínas involucradas en la coordinación del zinc están representados como triángulos. En la parte superior de la gráfica se muestra la predicción de estructura secundaria de JPred y a la derecha de la gráfica la estructura secundaria del conjunto recocido. Las hélices α están representadas como rectángulos. La numeración de secuencia se basa en la secuencia de E1A de HAdV5. **(C)** Distancia entre las cisteínas que coordinan el zinc a lo largo de la trayectoria. **(D)** RMSD de las estructuras predichas del dominio CR3. Se calculó la fluctuación de las posiciones de los átomos en la trayectoria utilizando como referencia el promedio de las coordenadas. Los valores de RMSD están mapeados sobre el esqueleto de la proteína de un cuadro de la trayectoria elegido al azar. Los valores de RMSD se calcularon utilizando GROMACS 4.5 (Pronk *et al.*, 2013). Las cisteínas involucradas en la coordinación del zinc están representadas con palillos amarillos. El N-terminal y C-terminal del dominio CR3 están indicados con una N y una C respectivamente en todos los casos.

Por otro lado, el arreglo alternativo de las cuatro cisteínas no convergió en un grupo de estructuras que satisfaga las restricciones de manera simultánea (véase Sección F.7.1). Por lo tanto, se eligió el conjunto recocido de estructuras resultante del arreglo que se muestra en la Figura 4.17A para un análisis en profundidad.

Análisis de compatibilidad del modelo. Se analizó si el modelo obtenido es compatible con el conocimiento actual de estructuras globulares proteicas (Figura 4.18). Para la representación de los

resultados, se eligió por azar un miembro representativo del conjunto obtenido. En primer lugar, se analizó el perfil energético del modelo estructural predicho midiendo la frustración configuracional local. Típicamente, los dominios globulares consisten de un núcleo mínimamente frustrado con aminoácidos en contacto que son energéticamente favorables en ese contexto estructural (Ferreiro *et al.*, 2007). Por otro lado, las regiones desordenadas pequeñas en el dominio (en inglés, *loops*) y la superficie de la proteína pueden mostrar parches de pares de residuos frustrados que están en conflicto energético con la estructura. Frecuentemente estos parches se relacionan con la función de la proteína (Ferreiro *et al.*, 2007). El modelo del dominio CR3 muestra una región N-terminal mínimamente frustrada que incluye el sitio de unión a zinc (Figura 4.18A) y una región C-terminal frustrada (véase Sección 2.2.7) que se sabe que une factores de transcripción (Figura 4.18A).

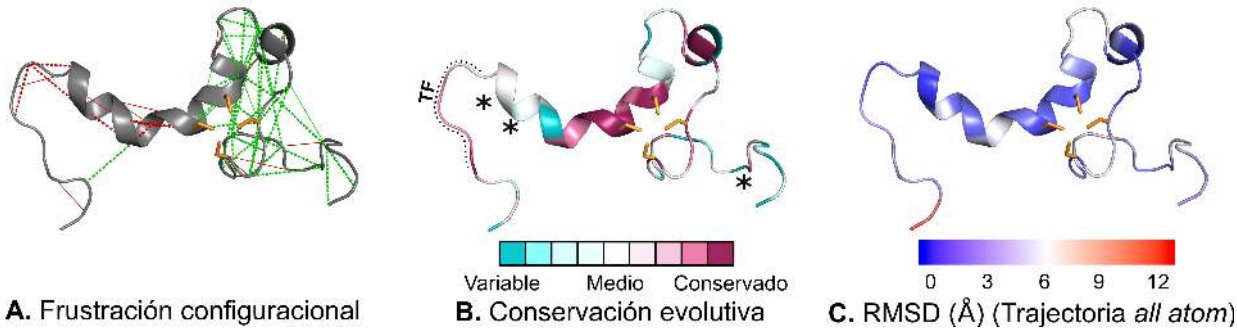


Figura 4.18: Validación del modelo estructural del dominio CR3 de E1A. (A) Frustración configuracional. Se mapearon sobre el esqueleto (en inglés, *backbone*) de la proteína los contactos altamente frustrados (rojo) y mínimamente frustrados (verde). No se muestran los contactos neutros. Los contactos entre residuos están representados con líneas sólidas y los que son mediados por agua están representados como líneas punteadas. El esqueleto de la proteína está representado como un *cartoon* en gris oscuro. **(B)** Conservación evolutiva de los aminoácidos en las posiciones del CR3. Se mapeó sobre el esqueleto de la proteína la conservación evolutiva calcula por ConSurf (Ashkenazy *et al.*, 2016). Los colores varían desde rosa oscuro (muy conservado) a azul claro (poco conservado). Las posiciones ricas en cisteínas están indicadas con asteriscos negros. El sitio de unión de los factores de transcripción (TF) está señalado con una línea negra punteada. **(C)** Estabilidad estructural del modelo predicho para el dominio CR3 en simulaciones de dinámica molecular. Se calcularon los valores de RMSD por residuo luego de una minimización y una simulación corta utilizando como referencia la estructura inicial. Estos valores se mapearon sobre el esqueleto de la proteína. Los valores de RMSD se calcularon utilizando VMD (Humphrey *et al.*, 1996).

En segundo lugar, se utilizó el algoritmo ConSurf (véase Sección 2.2.8) para evaluar la conservación de secuencia de cada posición del dominio (Figura 4.18B) (Ashkenazy *et al.*, 2016). La región N-terminal, mínimamente frustrada y correctamente plegada, muestra un núcleo de residuos altamente conservados, mientras que la región C-terminal, flexible y frustrada, está menos conservada con excepción de un pequeño grupo de residuos involucrados en interacciones proteína-proteína. (Figura 4.18B). En tercer lugar, se realizaron simulaciones cortas de dinámica molecular. En las simulaciones de dinámica molecular de estructuras globulares determinadas experimentalmente se espera que las estructuras muestren cierta estabilidad estructural (Figura 4.18C). Para evaluar la predicción de estructura de grano grueso (en inglés, *coarse grained*), primero se reconstruyeron las cadenas laterales de los aminoácidos para obtener una estructura representati-

va utilizando el programa Pulchra (Rotkiewicz y Skolnick, 2008). Se realizó una simulación de dinámica molecular con solvente explícito para relajar las estructuras durante 2 ns utilizando una dinámica estocástica con un paso de 2 fs y el campo de fuerzas electrostáticas de malla de partículas de Ewald. La Figura 4.18C muestra los valores de RMSD por residuo relativo a la estructura inicial. La mayoría de los valores son menores a 2 Å, valor típico para dominios globulares de este tamaño.

En base a estos resultados, se concluye que el modelo estructural para el dominio CR3 muestra patrones energéticos, de conservación y estabilidad estructural similares a aquellos observados para estructuras experimentales. Por lo tanto, se propone que el modelo del CR3 será útil para entender los resultados de experimentos de mutagénesis y guiar la búsqueda de una estructura experimental.

4.12. Reconstrucción filogenética de *Mastadenovirus*

Hasta la fecha no existe una filogenia disponible en la literatura que abarque a todos los serotipos pertenecientes al género *Mastadenovirus* incluidos en este trabajo de tesis. Con el objetivo de estudiar la evolución de los motivos lineales a lo largo de la historia evolutiva es necesario hacer un árbol que incluya a todos los serotipos utilizados.

Se realizó la construcción del árbol (véase Sección 2.3.2) partiendo de las secuencias de genomas de 139 serotipos de adenovirus (véase Sección A.5). Utilizando HAdV2 como referencia se seleccionaron aquellas regiones del genoma de cada serotipo de adenovirus no sometidas a recombinación homóloga, es decir, no incluye la región codificante de la base del pentón, hexón y fibra (Robinson *et al.*, 2011). Luego se seleccionaron los bloques conservados según Gblocks (Castresana, 2000) obteniendo un alineamiento de 1826 posiciones (disponible en Sección B.3). La reconstrucción filogenética incluyendo los 116 serotipos de interés se muestra en la Figura 4.19. Todas las serotipos de las especies incluidas aparecen agrupados bajo un ancestro común conformando un clado monofilético. 99 nodos de 115 presentan un soporte mayor al 50 % (Figura 4.19, puntos verdes). De los 16 nodos restantes, dos no poseen soporte por como se generó la raíz. Únicamente dos nodos de los 16 son ancestro común de especies diferentes. Uno es el ancestro común a las especies *Murine adenovirus A*, *Murine adenovirus C*, *Tree shrew adenovirus A*, *Ovine adenovirus A*, *Bovine adenovirus A* y *Porcine adenovirus C*. El otro es el ancestro común a las especies *Human adenovirus A*, *F* y *G*. Ocho de los 16 se encuentran distribuidos dentro del clado que contiene la especie *Human adenovirus D*, uno de los 16 se encuentra dentro del clado que contiene la especie *Human adenovirus E*, tres dentro del clado de la especie *Human adenovirus B* y uno dentro del clado de la especie *Human adenovirus C*.

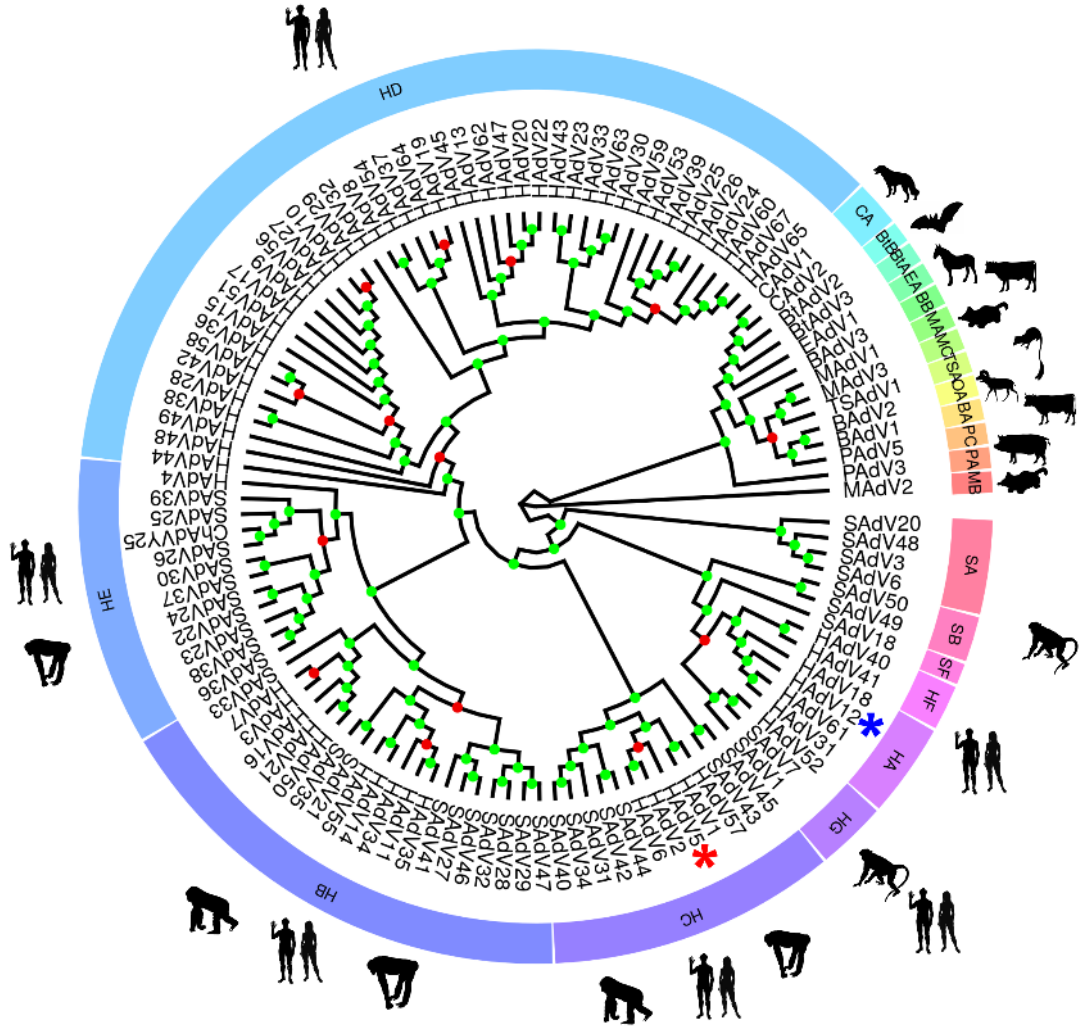


Figura 4.19: Árbol filogenético de *Mastadenovirus*. En el círculo externo se indica la especie a la que pertenecen los serotipos. Los nodos con un soporte mayor al 50 % están señalados con un círculo verde. Las siluetas de los animales indican los hospedadores determinados para cada especie. Las imágenes fueron obtenidas en `phylopic.org` y no están sometidas a derechos de autor. La ubicación de los serotipos HAdV5 y HAdV12 están señaladas con un asterisco rojo y azul respectivamente.

En resumen, obtuvimos un árbol que tiene un soporte adecuado para analizar la evolución de los caracteres.

4.13. Evolución de motivos lineales de E1A en la filogenia de *Mastadenovirus*

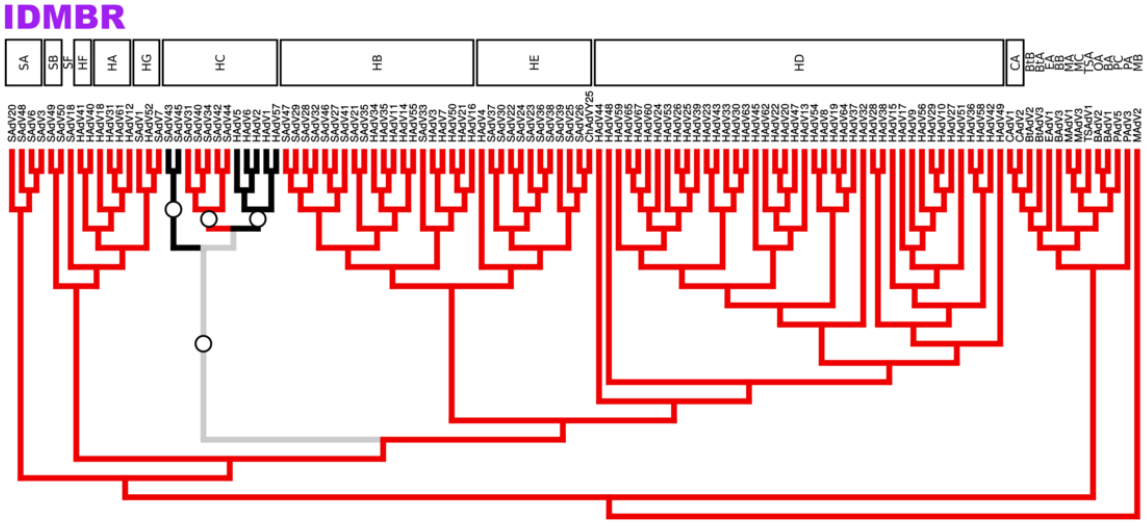
El estudio de conservación mostró que ocho de los doce motivos están altamente conservados en las secuencias de E1A (véase Sección 4.2), mientras que los motivos TRAM-CBP y el motivo IDMBR están poco conservados. Estas observaciones sugieren que los motivos menos conservados aparecieron recientemente en la evolución de *Mastadenovirus* en comparación a los más conservados. Para evaluar esta hipótesis de manera cuantitativa, se realizó la reconstrucción de la evolución

de motivos en E1A utilizando dos métodos, *Máxima Parsimonia* y *Reconstrucción de secuencias ancestrales* (Sección 2.3.4). Es importante destacar que hay tres tipos de información obtenida en ambos métodos. Por un lado se obtiene el dato sobre el estado de presencia o ausencia del motivo en un nodo. Por otro lado, podemos definir estados de presencia o ausencia de un motivo en una rama o eventos de aparición o desaparición del motivo en la rama (el nodo antecesor tiene un estado diferente del nodo descendiente). Un evento de aparición de un motivo implica la ganancia funcional del mismo y un evento de desaparición la pérdida funcional.

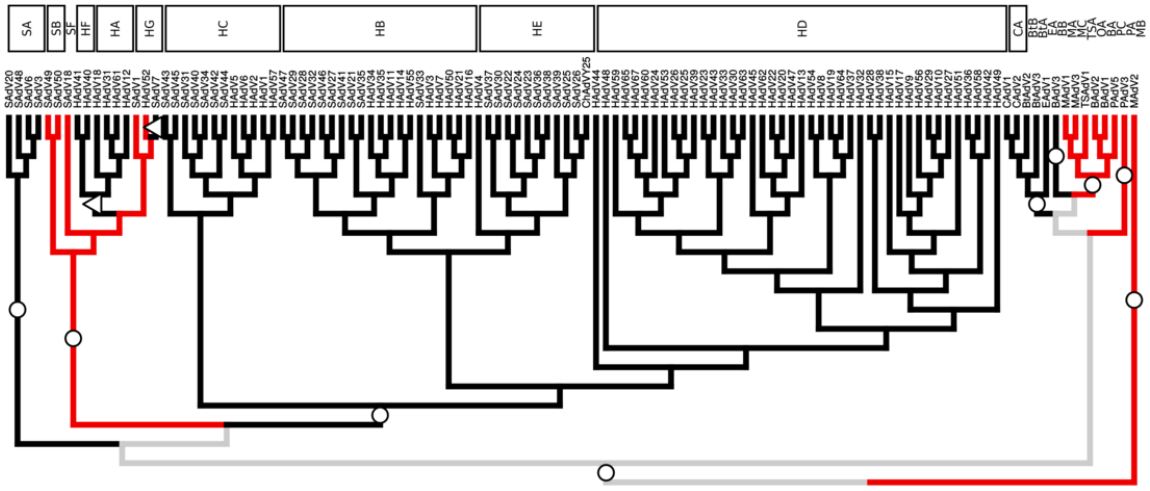
4.13.1. Reconstrucción por máxima parsimonia en la filogenia de *Mastadenovirus*

Utilizando la construcción del árbol filogenético de *Mastadenovirus* (véase Sección 4.12) y la información de presencia o ausencia de cada motivo en los serotipos actuales (véase Sección 4.2 y Sección F.2), se reconstruyeron los estados ancestrales de presencia y ausencia considerando a cada motivo como un caracter utilizando el método de máxima parsimonia (véase Sección 2.3.4) del software Mesquite Versión 2.75 (`http://mesquiteproject.org`). La información de presencia y ausencia de los motivos obtenida es mapeada sobre el árbol filogenético para facilitar su visualización. En la Figura 4.20, las ramas negras indican la presencia del motivo, las ramas rojas indican la ausencia del motivo y las ramas grises indican que no se pudo definir la presencia o ausencia de ese motivo en esa rama. Al mismo tiempo, se indican con círculos los eventos de aparición o desaparición del motivo a lo largo de la rama. Los eventos de aparición están indicados con triángulos, los eventos de desaparición con cuadrados y cuando no se puede definir si ocurrió una aparición o desaparición se utiliza un círculo.

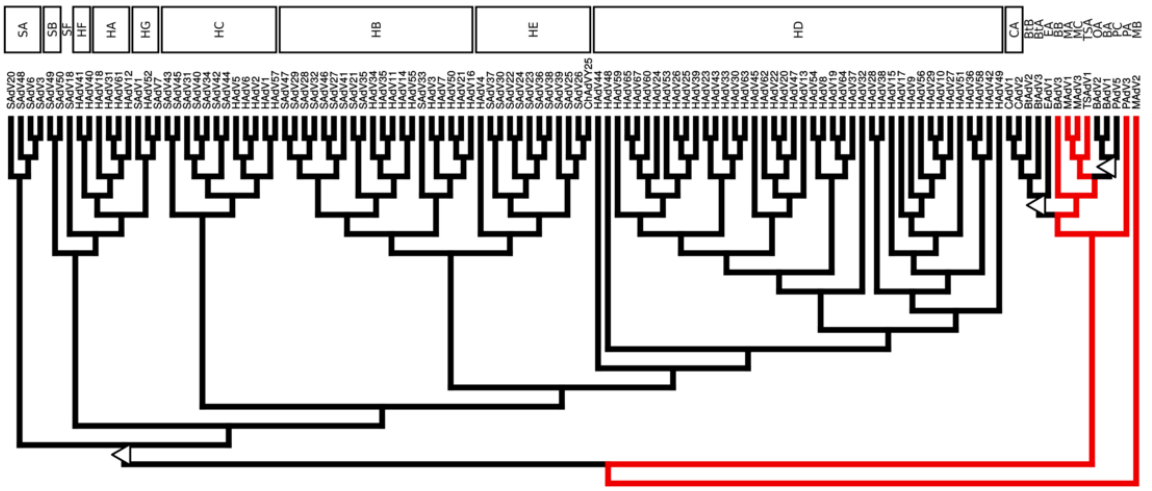
En primer lugar, se puede observar en la Figura 4.20 que los motivos IDMBR y TRAM-CBP aparecen en las ramas menos profundas del árbol de *Mastadenovirus*, los motivos CoRNR Box, pRb_ABGroove, MYND, la región acídica, el motivo CKII, las posiciones ricas en cisteínas, NLS y CtBP aparecen en ramas más profundas y el motivo LxCxE es el único presente en el ancestro.



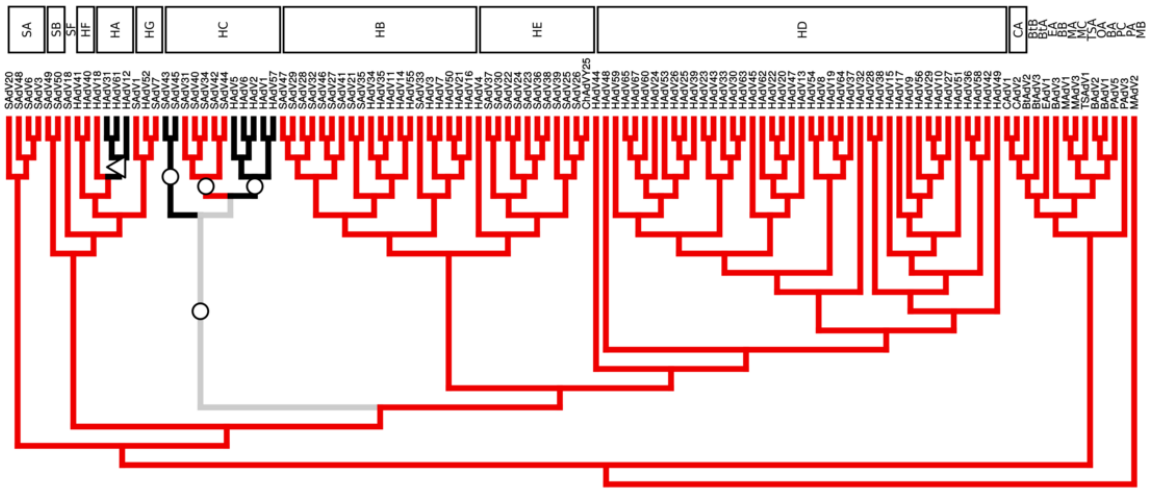
CoRNR Box



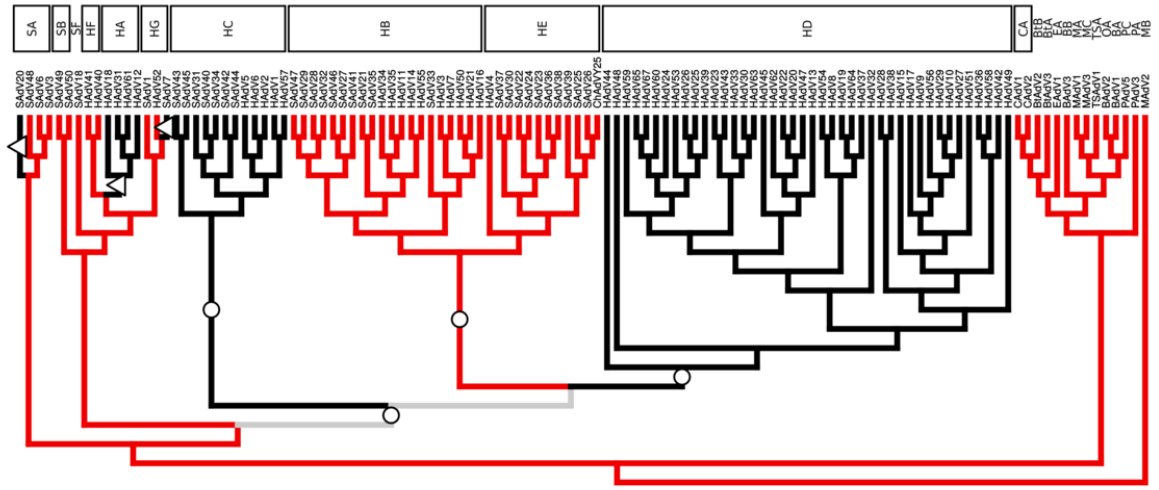
pRb_ABGroove



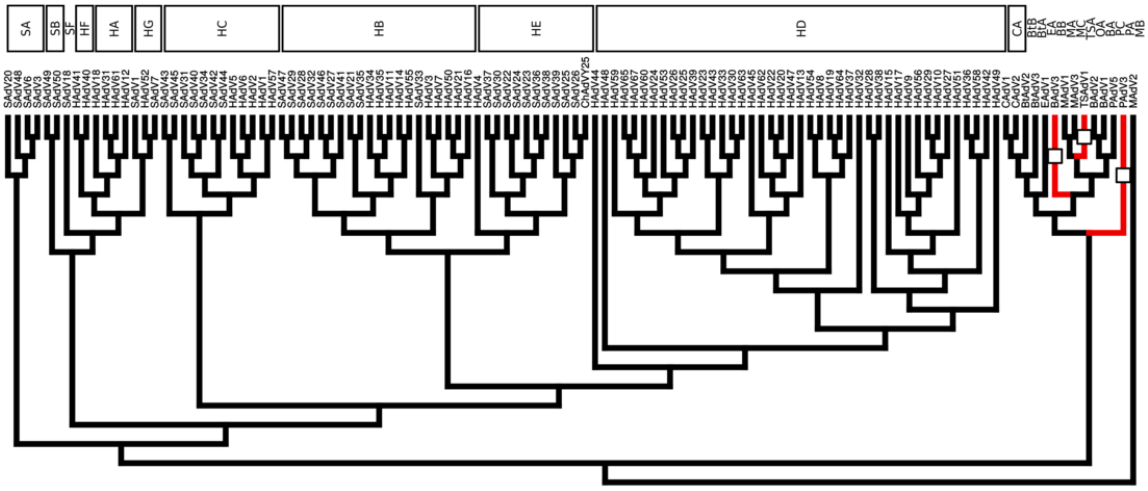
TRAM-CBP



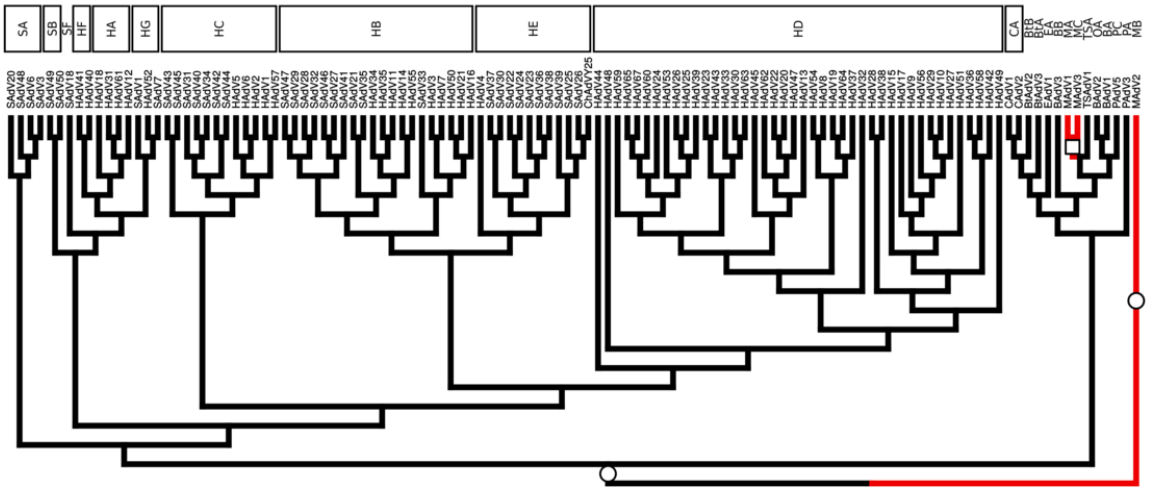
MYND



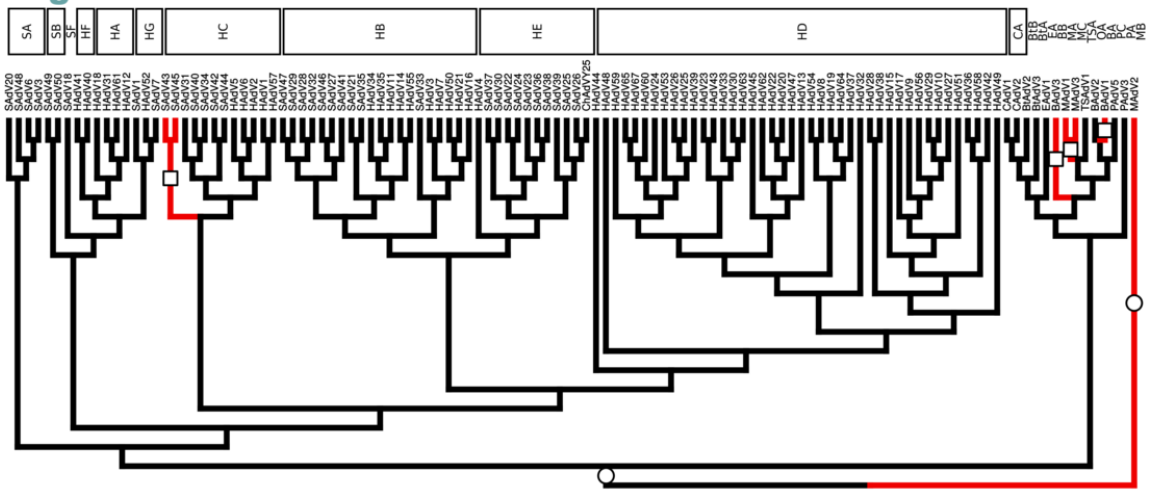
LxCxE



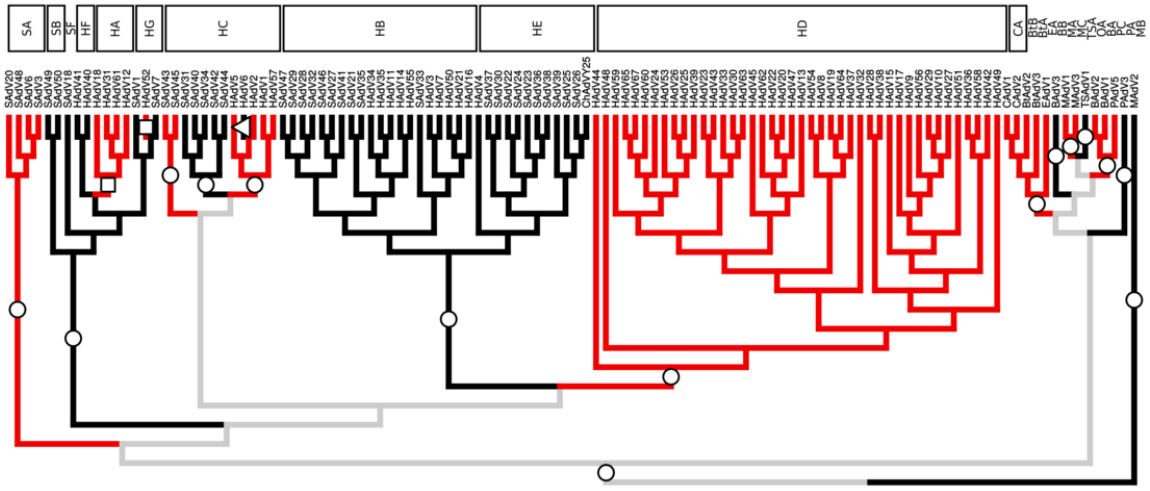
CKII



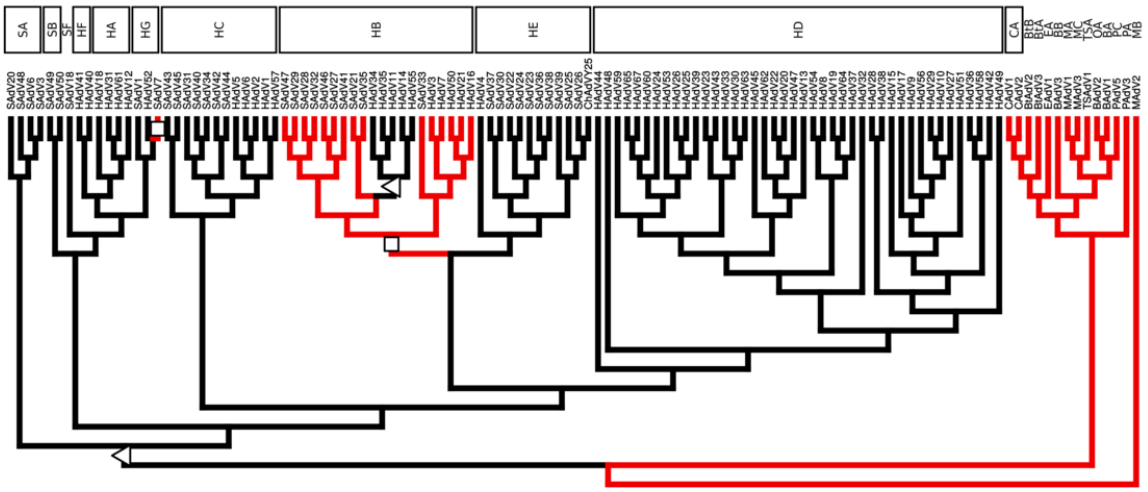
Región Acídica



Pos. Ricas en Cys



NLS



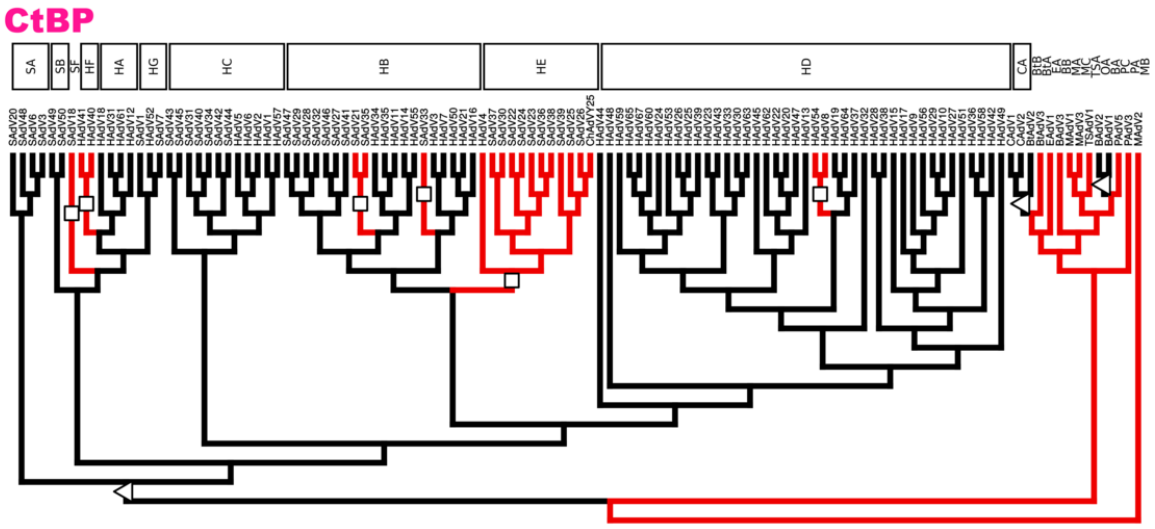


Figura 4.20: Estados ancestrales reconstruidos por parsimonia de los motivos lineales de E1A. Se muestra la reconstrucción obtenida mapeada en la historia evolutiva de *Mastadenovirus* de los motivos de los dominios N-terminal (IDMBR y CoRNR Box), CR1 (pRb_ABGroove y TRAM-CBP), CR2 (MYND, LxCxE, Región Acídica y CKII), CR3 (Posiciones ricas en cisteínas) y CR4 (NLS y CtBP). El estado de presencia, ausencia o indefinido del motivo se muestra sobre la rama en color negro, rojo o gris respectivamente. El evento de aparición, desaparición o indefinido del motivo se muestra con un triángulo, cuadrado o círculo respectivamente. Las hojas de cada árbol corresponden a los serotipos actuales a los que pertenece la secuencia de E1A a partir de la cual se realizó la reconstrucción. En la parte superior se indica la especie a la que pertenecen los serotipos.

En segundo lugar, en la Figura 4.21 se resume el número de eventos de cambio (aparición, desaparición o evento indefinido) para cada motivo. Cuatro de los once motivos analizados (CoRNR Box, pRb_ABGroove, TRAM-CBP y MYND) presentan únicamente eventos de aparición - 2, 3, 1 y 3 respectivamente. Tres de los once motivos analizados (LxCxE, la región acídica y el CKII) presentan únicamente eventos de desaparición - 3, 4 y 1 respectivamente. Tres de los once motivos analizados (las posiciones ricas en cisteínas, los motivos NLS y CtBP) presentan eventos de aparición - 1, 2 y 3 respectivamente, y eventos de desaparición - 2, 2 y 6 respectivamente. El motivo IDMBR no presenta eventos de aparición o desaparición definidos. Aunque hay un gran número de eventos indefinidos, se puede observar que el número de eventos de apariciones y desapariciones es propio de cada motivo.

En tercer lugar, se cuantificó el número total de eventos definidos (Figura 4.21, barras negras). Este número no parece estar definido por la prevalencia del motivo. Los motivos altamente prevalentes pRb_ABGroove, LxCxE, el sitio de fosforilación de CKII y la región acídica varían en el número total de eventos entre 1 y 4. Los de prevalencia media, CoRNR Box, MYND, CtBP, la NLS y las posiciones ricas en cisteínas, varían en el número total de eventos entre 2 y 9.

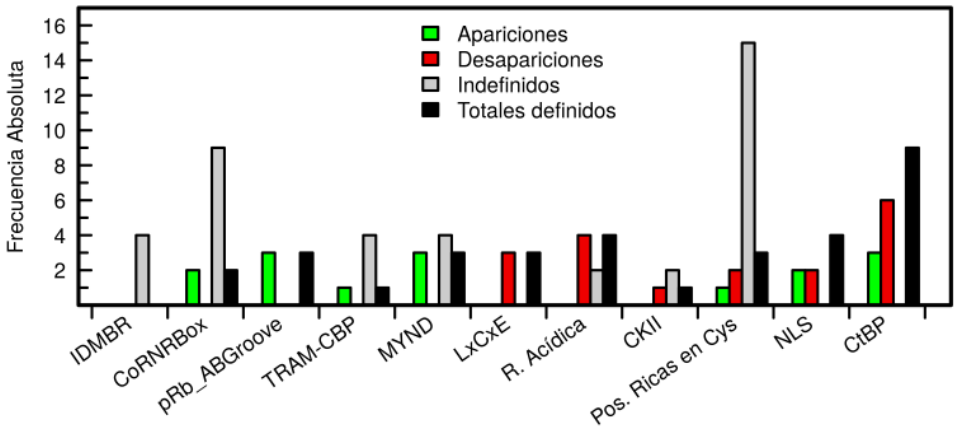


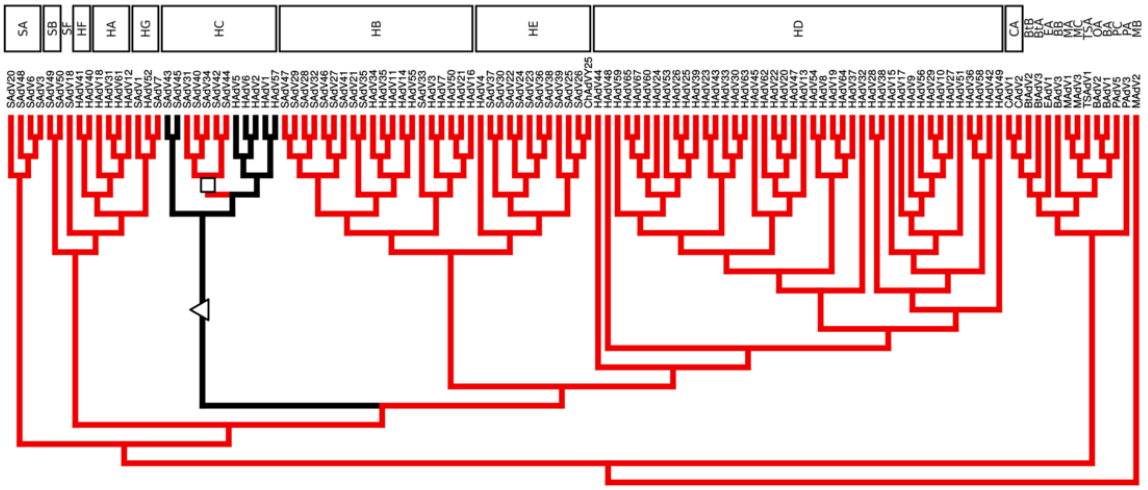
Figura 4.21: Eventos de aparición, desaparición o indefinidos de los motivos lineales de E1A determinados por parsimonia. El número de eventos de cambio de estado de los motivos lineales a lo largo de la filogenia de *Mastadenovirus* está representado en forma de barras. Se muestran en verde las apariciones, en rojo las desapariciones, en gris los eventos indefinidos y en negro los eventos totales definidos.

En conclusión, la reconstrucción de la historia evolutiva de los motivos en la filogenia de *Mastadenovirus* muestra que los motivos de E1A aparecen tanto en las ramas más profundas del árbol como en las menos profundas y aparecen o desaparecen múltiples veces a lo largo de la filogenia de *Mastadenovirus*. En ambos casos siguen un patrón específico de cada motivo. Estos resultados sugieren que existen presiones de selección motivo-específicas.

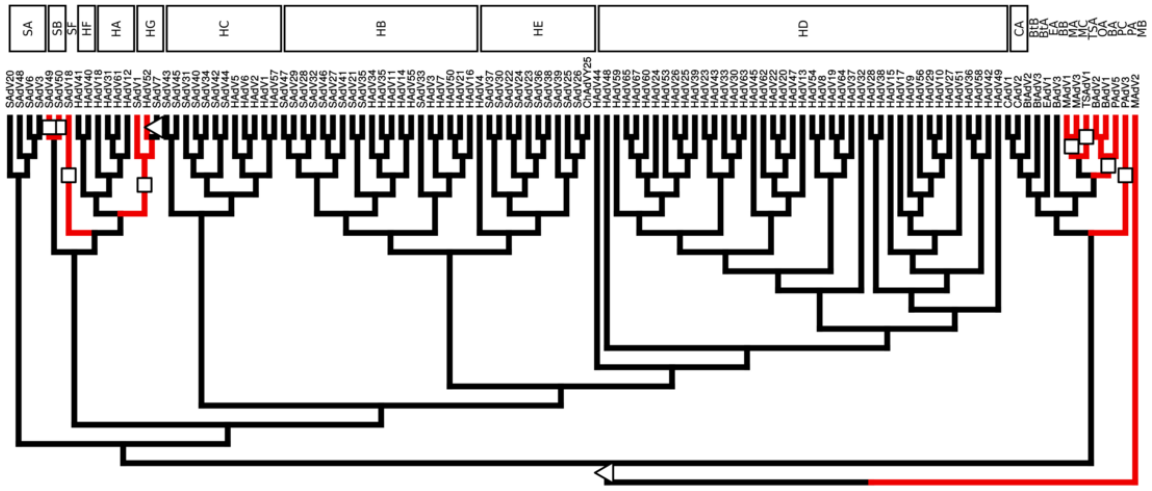
4.13.2. Reconstrucción por el método empírico de Bayes en la filogenia de *Mastadenovirus*

Utilizando la construcción del árbol filogenético de *Mastadenovirus* (véase Sección 4.12) y las secuencias actuales de la proteína E1A, se reconstruyeron las secuencias ancestrales correspondientes a cada uno de los nodos internos del árbol mediante el método empírico de Bayes (véase Sección 2.3.4). Luego, en colaboración con el estudiante César Leonetti, se realizó una búsqueda basada en texto de los motivos lineales de E1A utilizando las expresiones regulares de la Tabla 4.1 para determinar la presencia o ausencia en las secuencias ancestrales. Esta información, junto con la presencia o ausencia de cada motivo en los serotipos actuales, permite determinar la presencia o ausencia en las ramas y los eventos de aparición o desaparición de cada motivo a lo largo de la filogenia de *Mastadenovirus*. Esta información es mapeada sobre el árbol filogenético para facilitar su visualización. En la Figura 4.22, las ramas negras indican la presencia del motivo y las ramas rojas indican la ausencia del motivo. Al mismo tiempo, se indica con círculos los eventos de aparición o desaparición del motivo a lo largo de la rama. Los eventos de aparición están indicados con triángulos y los eventos de desaparición con cuadrados.

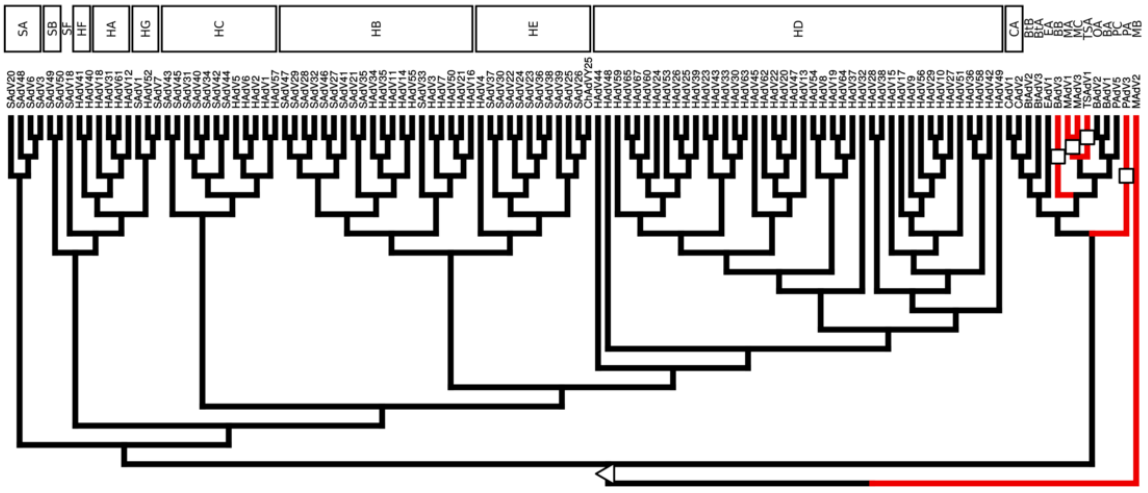
IDMBR



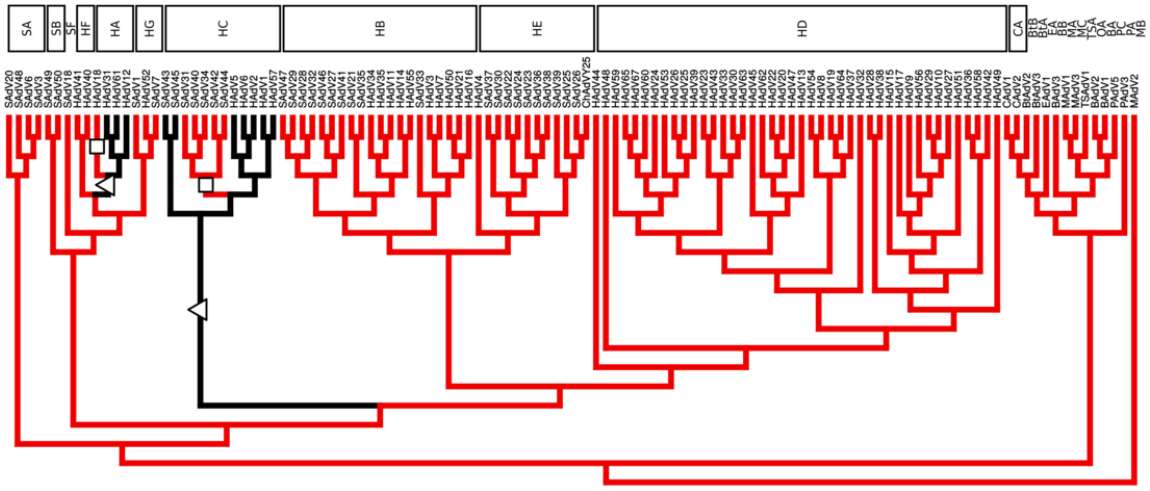
CoRNR Box



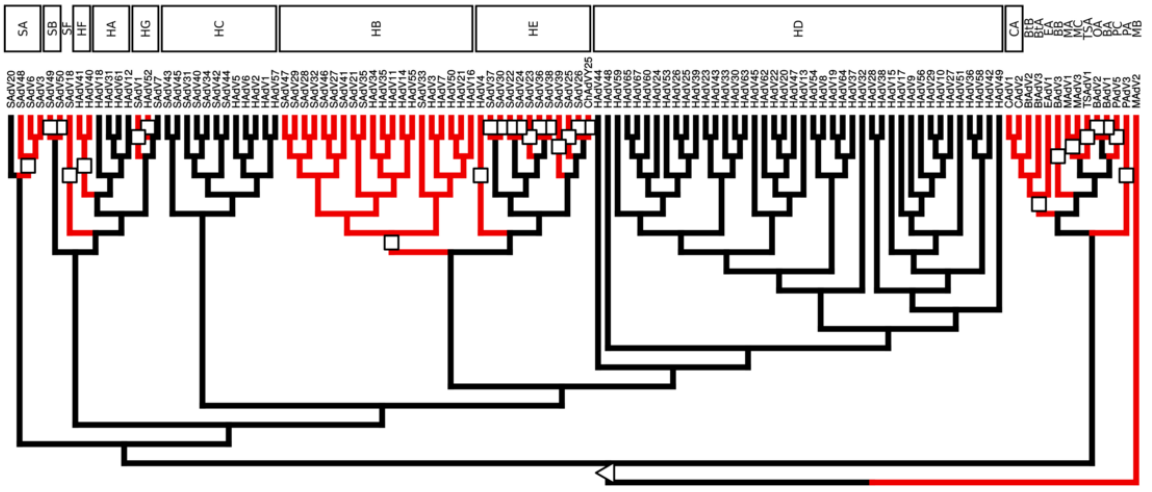
pRb_ABGroove



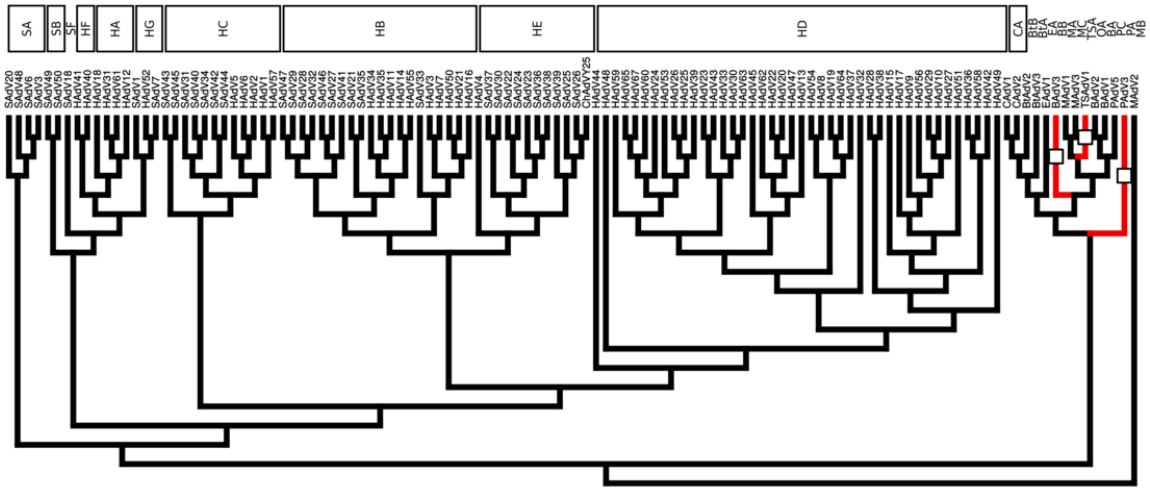
TRAM-CBP



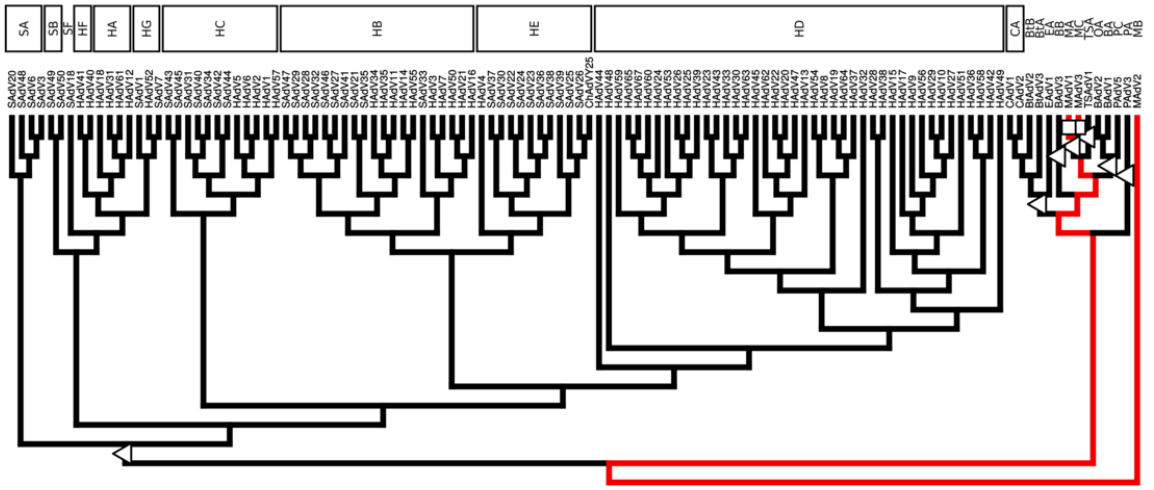
MYND



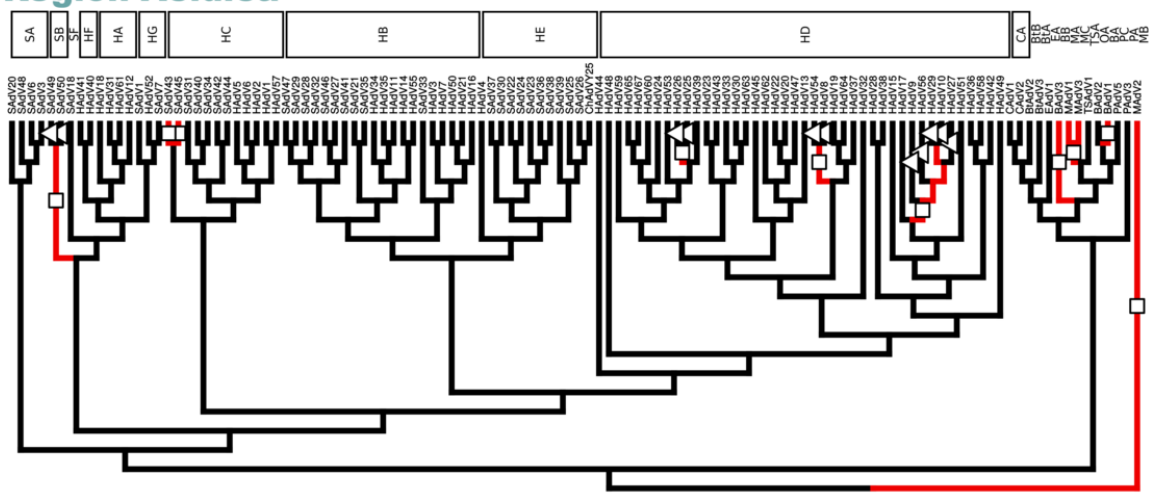
LxCxE



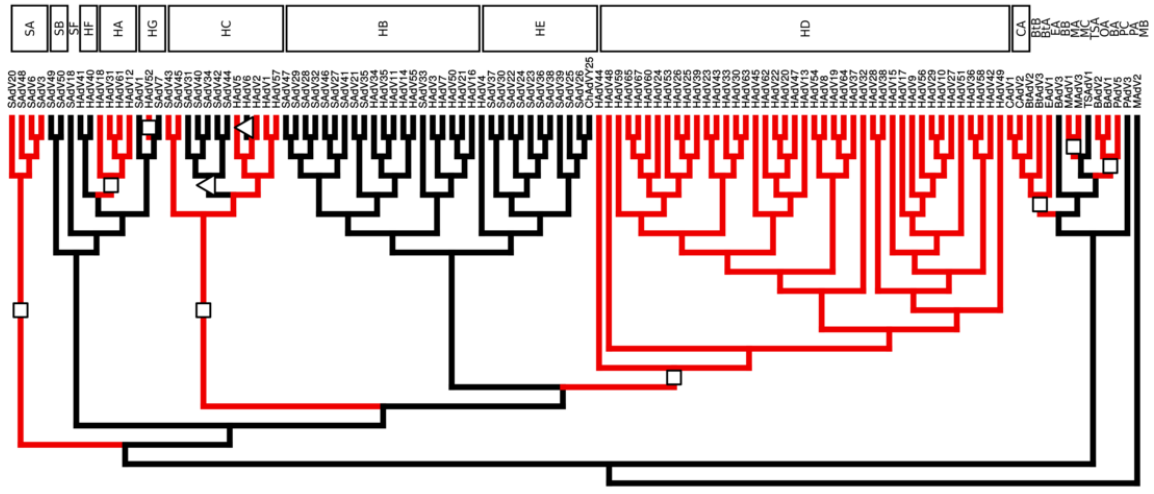
CKII



Región Acídica



Pos. Ricas en Cys



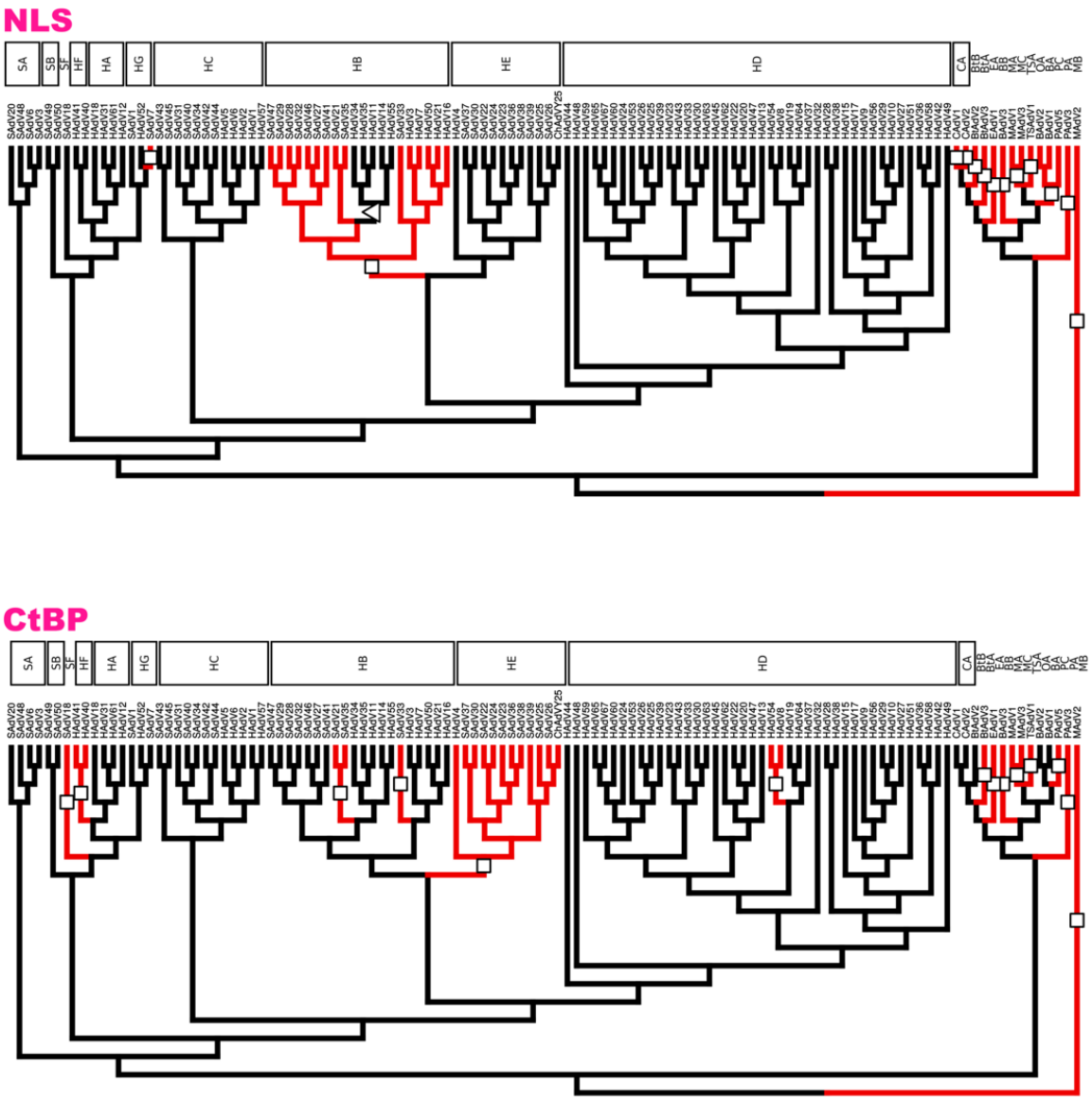


Figura 4.22: Estados ancestrales reconstruidos por el método empírico de Bayes de los motivos lineales de E1A. Se muestra la reconstrucción obtenida mapeada en la historia evolutiva de *Mastadenovirus* de los motivos de los dominios N-terminal (IDMBR y CoRNR Box), CR1 (pRb_ABGroove y TRAM-CBP), CR2 (MYND, LxCxE, Región acídica y CKII), CR3 (Posiciones ricas en cisteínas) y CR4 (NLS y CtBP). El estado de presencia o ausencia del motivo se muestra sobre la rama en color negro o rojo respectivamente. El evento de aparición o desaparición del motivo se muestra con un triángulo o cuadrado respectivamente. Las hojas de cada árbol corresponden a los serotipos actuales a los que pertenece la secuencia de E1A a partir de la cual se realizó la reconstrucción. En la parte superior se indica la especie a la que pertenecen los serotipos.

En primer lugar, se puede observar en la Figura 4.22 que los motivos IDMBR y el motivo TRAM-CBP aparecen por primera vez en las ramas menos profundas del árbol de *Mastadenovirus*. El motivo CKII aparece en una rama profunda del árbol. Los motivos CoRNR Box, pRb_ABGroove y MYND están presentes en el ancestro que da origen a todos los serotipos del género *Mastadenovirus* con excepción del serotipo MAdV2. Por último, los motivos LxCxE, la región acídica, las posiciones ricas en cisteínas, NLS y CtBP están presentes en el ancestro que da

origen al género *Mastadenovirus*.

En segundo lugar, en la Figura 4.23 se resume el número de eventos de cambio aparición o desaparición. Dos de los once motivos analizados, LxCxE y el CtBP, únicamente presentan eventos de desaparición - 3 y 14 respectivamente. Cuatro motivos presentan mayor número de desapariciones que apariciones (CoRNR Box, 2 apariciones y 8 desapariciones, MYND, 1 y 28, las posiciones ricas en Cisteínas, 2 y 8, y CtBP, 1 y 13). Dos de los once motivos analizados, IDMBR y TRAM-CBP presentan igual número de eventos de aparición y desaparición - 1 y 2 respectivamente. Los tres motivos restantes presentan mayor número de apariciones que desapariciones (pRb_ABGroove, 4 apariciones y 1 desaparición, la región acídica, 11 y 10, y CKII, 7 y 2).

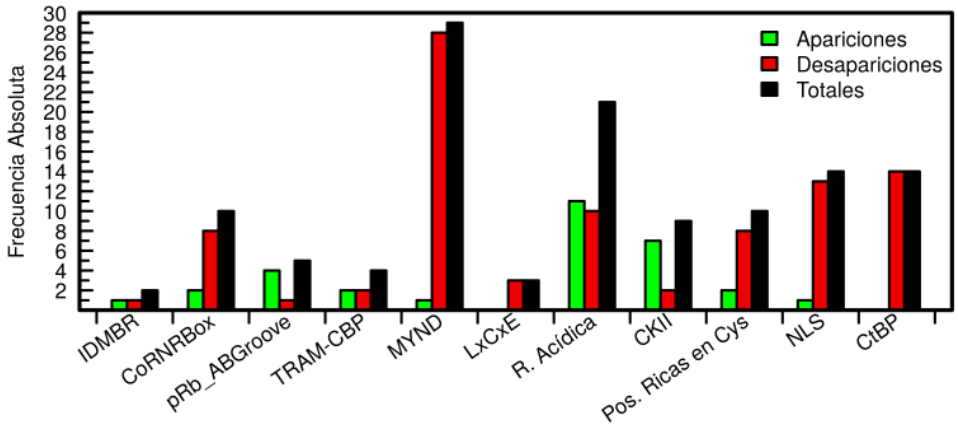


Figura 4.23: Eventos de aparición y desaparición de los motivos lineales de E1A determinados por el método empírico de Bayes. El número de eventos de cambio de estado de los motivos lineales a lo largo de la filogenia de *Mastadenovirus* está representado en forma de barras. Se muestran en verde las apariciones, en rojo las desapariciones y en negro el total.

En tercer lugar, se cuantificó el número total de eventos definidos (Figura 4.23, barras negras). Este número no parece estar definido por la prevalencia del motivo. Los motivos altamente prevalentes pRb_ABGroove, LxCxE, el sitio de fosforilación de CKII y la región acídica, varían en el número total de eventos entre 3 y 21, siendo la región acídica el que más presenta. Los de prevalencia media, CoRNR Box, MYND, CtBP, la NLS y las posiciones ricas en cisteínas, varían en el número total de eventos entre 10 y 29. Los motivos de baja prevalencia, la IDMBR y TRAM-CBP, presenta únicamente 2 y 4 eventos totales respectivamente.

Para resumir, al igual que en la reconstrucción por parsimonia, la reconstrucción de las secuencias ancestrales muestra que los motivos de E1A aparecen tanto en las ramas más profundas del árbol como en las menos profundas. También observamos de nuevo que los motivos aparecen o desaparecen múltiples veces a lo largo de la filogenia de *Mastadenovirus*, siguiendo un patrón motivo específico. Estos resultados sugieren que existen presiones de selección motivo-específicas.

Comparación entre los resultados obtenidos por ambos métodos. A diferencia del método de máxima parsimonia, el método empírico de Bayes permite obtener una secuencia para cada uno de

los nodos internos del árbol y por lo tanto no hay estados indefinidos.

En primer lugar, comparamos el número de eventos totales determinados por ambos métodos. En la Figura 4.24 se puede observar que el número de eventos totales es menor para el método de parsimonia. Esta diferencia se debe posiblemente a que el método de parsimonia se basa en minimizar el número de cambios.

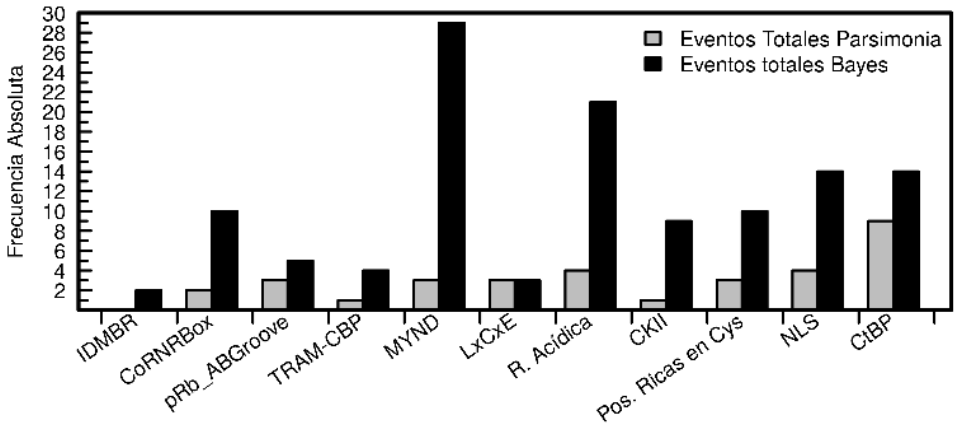


Figura 4.24: Eventos totales de los motivos lineales de E1A determinados por el método de parsimonia y el método empírico de Bayes. El número de eventos totales de cambio de estado de los motivos lineales a lo largo de la filogenia de *Mastadenovirus* está representado en forma de barras. Se muestran en gris lo determinado por parsimonia y en negro los eventos totales determinados por el método empírico de bayes.

En segundo lugar, comparamos para cada nodo el estado de presencia o ausencia de los motivos determinados por ambos métodos y para cada rama los eventos de aparición y desaparición.

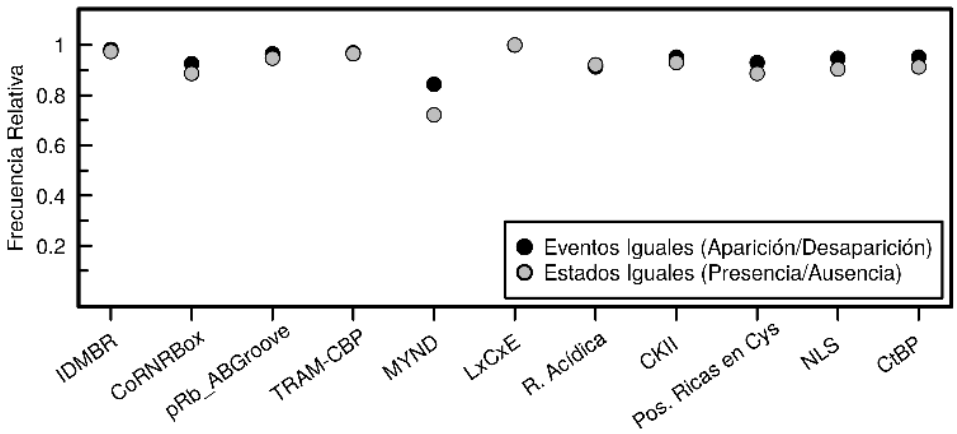


Figura 4.25: Eventos de apariciones/desapariciones y estados de ausencia/presencia en común en la determinación por el método de parsimonia y el método empírico de Bayes. Se muestra frecuencia de ramas en la filogenia de *Mastadenovirus* que muestran el mismo tipo de evento de aparición o desaparición (negros) de motivo en ambas reconstrucciones o el mismo estado de ausencia o presencia (blancos). Sin tener en cuenta las ramas o nodos que en parsimonia se consideran indefinidas.

Las frecuencia de ocurrencia o no ocurrencia de eventos de aparición/desaparición equiva-

lentes en ambos métodos son altas. Lo mismo se observa al analizar la frecuencia de estados de presencia/ausencia de los nodos equivalentes en ambos métodos.

En resumen, podemos decir que los métodos de reconstrucción utilizados arrojan resultados similares, siendo el método empírico de Bayes menos sesgado a diminuir el número de cambios que ocurren a comparación al método de parsimonia.

4.14. Tasa de cambio en el número de interacciones de la proteína E1A a lo largo de la filogenia de *Mastadenovirus*

Como se explicó antes, en la literatura se postula que los motivos lineales pueden ganar o perder funcionalidad por una mutación al azar debido al bajo número de posiciones determinantes de su actividad. Si esto fuera así, la tasa de cambio en el número de interacciones proteína-proteína para una proteína rica en motivos lineales sería mayor que para una proteína con interacciones de otro tipo.

Con el objetivo de evaluar esta hipótesis para la proteína E1A de *Mastadenovirus* se analizó la tasa de cambio en el número de interacciones proteína-proteína extrapolando el método utilizado por Beltrao y Serrano (2007).

Utilizando los proteomas que tienen el mayor número de interacciones reportadas (*Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caernohabditis elegans* y *Homo sapiens*) Beltrao y Serrano (2007) estiman la tasa en el cambio de número de interacciones. En primer lugar, establecen la edad aproximada de la proteína. En segundo lugar, comparan los blancos proteicos de cada proteína con los blancos proteicos de las proteínas parálogas, es decir, proteínas homólogas que surgen por un proceso de duplicación génica. Considerando que las proteínas parálogas tienen un ancestro común, es muy probable que los blancos proteicos compartidos por ambas proteínas sean adquiridos por herencia. Por otro lado, un blanco proteico identificado para una de las proteínas parálogas identifican una ganancia o pérdida de una interacción luego de la aparición del homólogo. Esto último, se corresponde de manera directa con cambios en el conjunto de interacciones durante la evolución. Por lo tanto, cuantifican la tasa de cambio en la redes de interacción promediando el número total de ganancias o pérdidas de interacciones entre todos los pares de proteínas parálogas. En otras palabras, determinan el cambio de interacciones en un proteoma entre proteínas parálogas, cuantificando el cambio en el número de interacciones observadas como la diferencia de blancos proteicos entre dos proteínas parálogas. Dadas dos proteínas parálogas, la presencia de un nuevo blanco proteico en una de ellas implica la ganancia de una nueva interacción y la ausencia de un determinado blanco proteico implica la pérdida de una interacción. Finalmente, definen la tasa de cambio como:

$$\text{tasa} = \frac{\text{Cambio de interacionnes}}{\text{Pares de proteínas parálogas posibles} \cdot \text{tiempo de divergencia}} \tag{4.3}$$

De manera análoga, una ganancia (aparición) o pérdida (desaparición) de un motivo lineal en

una proteína puede considerarse una ganancia o pérdida de una interacción. En base a esto se define la tasa de cambio en el número de interacciones (τ) de la proteína E1A en función del número de cambios, aparición o desaparición, de los motivos (ΔMLs), un número estimado de interacciones que pueden ocurrir por motivo (I_{ML}), el tiempo de divergencia entre las distintas proteínas ($t_{\text{divergencia}}$) y el número de proteínas distintas analizadas (P_{prot}) que en este caso es 1, ya que únicamente se incluye a la familia de proteínas E1A. En base a esto, la tasa de cambio en el número de interacciones está dada por:

$$\tau\left(\frac{\text{Interacciones}}{\text{Prot}\cdot\text{tiempo}}\right)=\frac{\Delta\text{MLs}\cdot I_{ML}}{P_{\text{prot}}\cdot t_{\text{divergencia}}} \tag{4.4}$$

En este trabajo de tesis se utilizaron dos métodos diferentes para determinar el número de apariciones y desapariciones de motivos, el método de parsimonia (véase Sección 4.13.1) y el método bayesiano (véase Sección 4.13.2). El método bayesiano reconstruye las secuencias ancestrales considerando un modelo evolutivo y una probabilidad de cambio por posición. Por otro lado, el método de parsimonia minimiza el número de cambios en el estado de los motivos. Por lo tanto, se utilizó la información obtenida por ambos métodos para calcular el valor mínimo, τ_{min} y el valor máximo de la tasa de interacción, τ_{max}. τ_{min} está determinado por el mínimo número de apariciones y desapariciones, ΔMLs_{min}, de los motivos obtenido por el método de parsimonia y τ_{max} por el número máximo de cambios obtenido por el método bayesiano ΔMLs_{max}. Los valores, τ_{min} y τ_{max} también dependen del número de interacciones por motivo lineal (I_{ML}). Existen interacciones diferentes mediadas por un mismo motivo lineal, por ejemplo, el motivo IDMBR media la interacción al menos con tres proteínas, S8, CBP y TBP. Existen también numerosos blancos proteicos para los que se desconoce si su interacción es mediada o no por un motivo lineal (Figura 4.2), se utilizará un valor estimado máximo de I_{ML}, $I_{ML}{}_{max}$ y un valor estimado mínimo, $I_{ML}{}_{min}$. A continuación se describe la obtención de cada uno de los parámetros utilizados para calcular ambos valores de τ, según:

$$\tau_{max}\left(\frac{\text{Interacciones}}{\text{Prot}\cdot\text{tiempo}}\right)=\frac{\Delta\text{MLs}_{max}\cdot I_{ML}{}_{max}}{P_{\text{prot}}\cdot t_{\text{divergencia}}} \tag{4.5a}$$

$$\tau_{min}\left(\frac{\text{Interacciones}}{\text{Prot}\cdot\text{tiempo}}\right)=\frac{\Delta\text{MLs}_{min}\cdot I_{ML}{}_{min}}{P_{\text{prot}}\cdot t_{\text{divergencia}}} \tag{4.5b}$$

El número de cambios de los motivos, ΔMLs, se calcula como la suma de cambios totales de los motivos lineales a lo largo de la filogenia de *Mastadenovirus*. El valor mínimo de cambios determinado por el método de parsimonia, ΔMLs_{min}, es 26. El valor máximo determinado por el método bayesiano, ΔMLs_{max}, es 121.

El número estimado de interacciones que pueden ocurrir por motivo, I_{ML} está dado por:

$$I_{ML}=\frac{\text{Número de blancos proteicos}}{\text{Número de motivos lineales}} \tag{4.6}$$

El número total de blancos proteicos identificados para la proteína E1A es 68 y el número de motivos es once sin considerar el motivo de unión a zinc. Los valores máximos y mínimos considerados por lo tanto son:

$$\mathrm{I}_{\mathrm{ML}max} = \frac{68 \text{ Interacciones}}{11 \text{ ML}} \sim 6 \text{ Interacciones/ML} \tag{4.7a}$$

$$\mathrm{I}_{\mathrm{ML}min} = 1 \text{ Interacción/ML} \tag{4.7b}$$

Por último, el tiempo de divergencia, $t_{\text{divergencia}}$, está dado por el número de sustituciones por sitio, dN/dS, a lo largo de la filogenia de *Mastadenovirus* por la inversa del reloj molecular de adenovirus, t_{reloj} ($2.8\ 10^{-8}$ *sust/sitio/año*) (Duffy *et al.*, 2008; Hoppe *et al.*, 2015). dN/dS está dado por el largo de las ramas del árbol filogenético de *Mastadenovirus*, y posee un valor de 13.7*sust/sitio*. Es decir,

$$t_{\text{divergencia}} = \frac{dN}{dS} \cdot \frac{1}{t_{\text{reloj}}} = \frac{13.7 \text{ } sust/sitio}{2.8\ 10^{-8} \text{ } sust/sitio/año} \tag{4.8}$$

Reemplazando los valores correspondientes en la Ecuación 4.5a y la Ecuación 4.5b, el valor máximo de la tasa de cambio de interacciones es $\tau_{max} = 1.48$ Interacciones Prot^{-1} Años^{-1} y el valor mínimo de la tasa de cambio de interacciones es $\tau_{min} = 0.05$ Interacciones Prot^{-1} Años^{-1}.

Para analizar la tendencia central de la tasa de cambio de interacciones se utilizó la media geométrica:

$$< \tau_{\text{E1A}} >= \sqrt{\tau_{max} * \tau_{min}} = 0.27 \text{ Interacciones } \text{Prot}^{-1} \text{ } \text{Años}^{-1} \tag{4.9}$$

es decir, que la proteína E1A cambia 0.27 interacciones por cada millón de años.

Beltrao y Serrano (2007) comparan también de esta forma las tasas de cambio de interacciones en dos grupos de proteínas que contienen dominios globulares. El primer grupo está compuesto por proteínas con interacciones específicas e incluye proteínas cuyos dominios poseen pocas interacciones, menos de cinco interacciones, y pocas superficies de interacción. El segundo grupo está compuesto por proteínas no específicas o "promiscuas" e incluye proteínas cuyos dominios interactúan con 15 o más dominios proteicos distintos, y/o poseen muchas superficies de interacción diferente. Sus resultados muestran que la tasa de interacción de las proteínas promiscuas ($1.81\ 10^{-5}$Interacciones Prot^{-1} Años^{-1}) es significativamente mayor que para las específicas ($6.35\ 10^{-6}$ Interacciones Prot^{-1} Años^{-1}).

El valor de la tasa de cambio en el número de interacciones de la proteína E1A es superior en cuatro órdenes de magnitud a las tasas encontradas para las proteínas promiscuas por Beltrao y Serrano (2007)

En resumen, considerando que las interacciones de la proteína E1A son mediadas por motivos lineales, se puede decir que los motivos de la proteína E1A cambian rápidamente a lo largo de la filogenia de *Mastadenovirus* en comparación a las interacciones mediadas por dominios globulares de proteínas específicas o promiscuas.

4.15. Asociaciones entre rasgos fenotípicos

4.15.1. Asociaciones entre motivos de la proteína E1A

La distribución de los motivos lineales entre los distintos serotipos de la proteína E1A de adenovirus o dentro de cada especie no es homogénea y no parece ser azarosa (Sección 4.4). Para determinar si la combinación de motivos difiere del azar se realizó una prueba hipergeométrica (véase Sección 2.5.1) para evaluar cuán significativa es la asociación existente entre motivos dentro de todas las combinaciones posibles para once de los doce motivos de la proteína E1A (se excluyó al motivo de unión a zinc, conservado en el 100 % de los serotipos). El valor p obtenido es finalmente corregido aplicando la corrección de Benjamini-Hochberg (valor p*) para comparaciones múltiples. (Benjamini y Hochberg, 1995). Aquellas combinaciones en las que el valor p* es menor a 0.05 son representadas con un punto en la Figura 4.26.

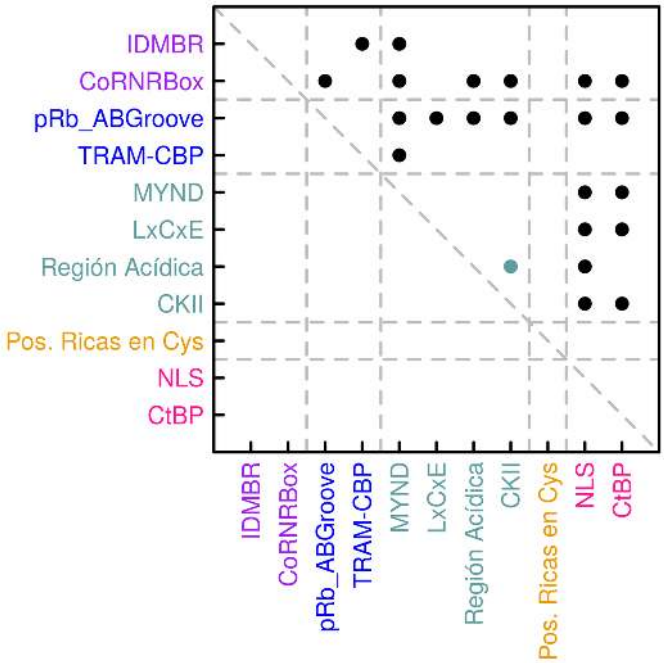


Figura 4.26: Asociaciones entre Motivos-Motivos en secuencias actuales. Las asociaciones positivas (valor $p^* < 0.05$) están marcadas con un punto negro cuando los motivos pertenecen a distintos dominios y coloreado cuando pertenecen al mismo dominio. La línea punteada indica la separación de los dominios. El código de colores para las etiquetas y para los puntos es el utilizado en la Figura 4.2.

Para los once motivos lineales existen un total de 55 asociaciones posibles. Sin embargo, sólo se observaron 23 asociaciones significativas (Figura 4.26). En particular, se observó una asociación significativa entre el sitio de fosforilación de CKII y la región acídica del CR2, y entre el motivo

pRb_ABGroove que se encuentra en el dominio CR1 y los motivos LxCxE, sitio de fosforilación CKII y la región acídica, que se encuentran en el dominio CR2. Estos cuatro motivos están relacionados a la unión con pRb. Con excepción de la asociación entre el motivo CKII y la región acídica, las restantes asociaciones significativas ocurren entre motivos que se encuentran en dominios diferentes. Ninguna asociación significativa ocurrió entre las posiciones ricas en cisteínas y los diez motivos restantes.

Estos resultados sugieren que existe una asociación no azarosa entre los motivos, lo que podría estar indicando una cooperación funcional entre ellos.

4.15.2. Asociación de motivos lineales con rasgos fenotípicos de la infección de adenovirus

La presencia o ausencia de un motivo particular en un serotipo particular de *Mastadenovirus* podría estar en principio asociada a un rasgo fenotípico como ser una preferencia de hospedador o tropismo tisular. Los estudios realizados en la proteína E7 de papilomavirus (Chemes *et al.*, 2012b) mostraron una asociación significativa entre la deleción del motivo LxCxE y la infección de los hospedadores del orden *Perissodactyla*, entre la ausencia del motivo LxCxE y la ausencia de la región acídica y entre la ausencia de la región acídica y la presencia de fibropapillomas en el hospedador. Estos resultados sugieren una conexión entre la plasticidad evolutiva de los motivos de la proteína E7 y eventos adaptativos que dieron forma a la diversificación de papilomavirus. Para determinar si esto ocurría también para la proteína E1A de adenovirus, se evaluó si las asociaciones motivo-hospedador o motivo-tropismo tisular era significativas utilizando la prueba hipergeométrica (véase Sección 2.5.1), aplicando la corrección por comparaciones múltiples de Benjamini-Hochberg (Benjamini y Hochberg, 1995) (Sección 2.5.2).

La información de hospedador para la proteína E1A es definida y se asignó un hospedador a cada serotipo. Sin embargo, la información de tropismo para cada serotipo no es tan simple. Los adenovirus pueden infectar células epiteliales de un gran número de órganos, produciendo diversas sintomatologías que son compartidas por un gran número de serotipos distintos. Además, los adenovirus pueden persistir de manera subclínica en los organismos durante largos períodos de tiempo y dificultando estudiar la asociación entre una sintomatología y la presencia de un determinado serotipo.

Asociaciones entre motivos lineales y hospedador. Los serotipos del género *Mastadenovirus* infectan hospedadores de siete órdenes distintos de mamíferos: Chiroptera, Artiodactyla, Carnívora, Perissodactyla, Primates, Rodentia y Scandentia (véase Sección A.6). Se realizó una prueba hipergeométrica para evaluar si existía alguna asociación entre alguno de los siete órdenes del hospedador y alguno de los once motivos lineales. Aquellas combinaciones en las que el valor $p*$ es menor a 0.05 se representan con un punto en la Figura 4.27.

De las 77 asociaciones posibles, solamente se identificaron ocho asociaciones significativas Figura 4.27). El orden Primates mostró una asociación significativa con ocho de los once moti-

vos. Cuatro de los ocho motivos participan en la unión de la proteína retinoblastoma: el motivo pRb_ABGroove, el motivo LxCxE, el sitio de fosforilación CKII y la región acídica.

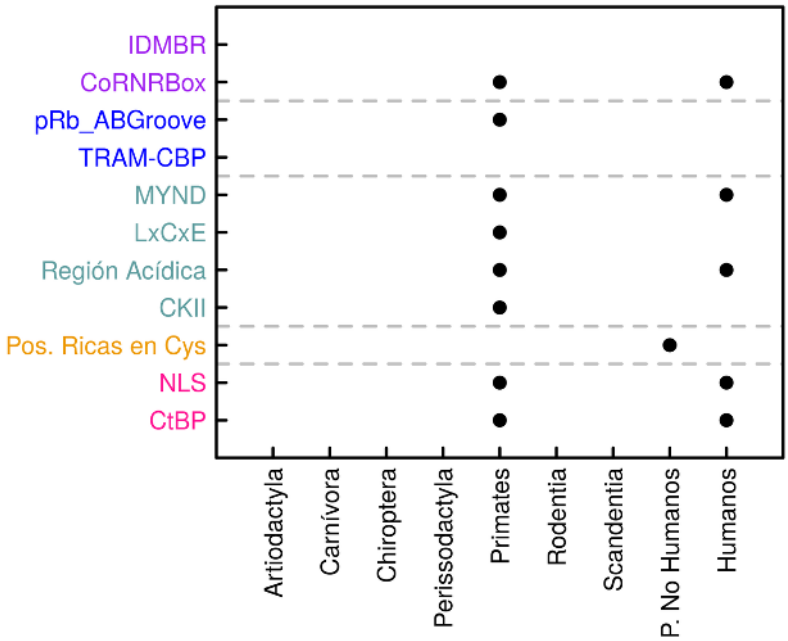


Figura 4.27: Asociaciones entre Motivos-Hospedadores en secuencias actuales. Las asociaciones positivas (valor $p^* < 0.05$) están marcadas con un punto negro. Las líneas punteadas indican la separación de los dominios. El código de colores para las etiquetas y para los puntos es el utilizado en la Figura 4.2.

Para determinar si las asociaciones determinadas entre primates y los motivos estaban asociadas principalmente a humanos, se realizó una segunda prueba de asociación separando el orden Primates en Humanos y Primates No Humanos. De las 88 asociaciones posibles, sólo se identificaron seis asociaciones significativas. El hospedador primates no-humanos mostró una asociación significativa con la presencia de posiciones ricas en cisteínas. De los cuatro motivos que participan en la unión a pRb, sólo la región acídica mostró una asociación positiva con el hospedador humano.

Estos resultados sugieren que existe una asociación no azarosa entre algunos de los motivos y los primates humanos y no-humanos, lo que podría estar indicando una selección evolutiva para determinados motivos.

Asociaciones entre motivos lineales y tropismo tisular. Los serotipos de *Mastadenovirus* son causantes de múltiples enfermedades asociadas a nueve tropismos diferentes: el tracto respiratorio superior e inferior, el tracto gastrointestinal, los órganos restantes del sistema digestivo, la mucosa ocular, y los sistemas linfático, cardiovascular, urinario y nervioso central (véase Sección 1.4.4). Se realizó una prueba hipergeométrica para evaluar si existía alguna asociación entre alguno de los 9 tropismos y alguno de los once motivos lineales. Aquellas combinaciones en las que el valor p^* es menor a 0.05 son representadas con un punto en la Figura 4.28.

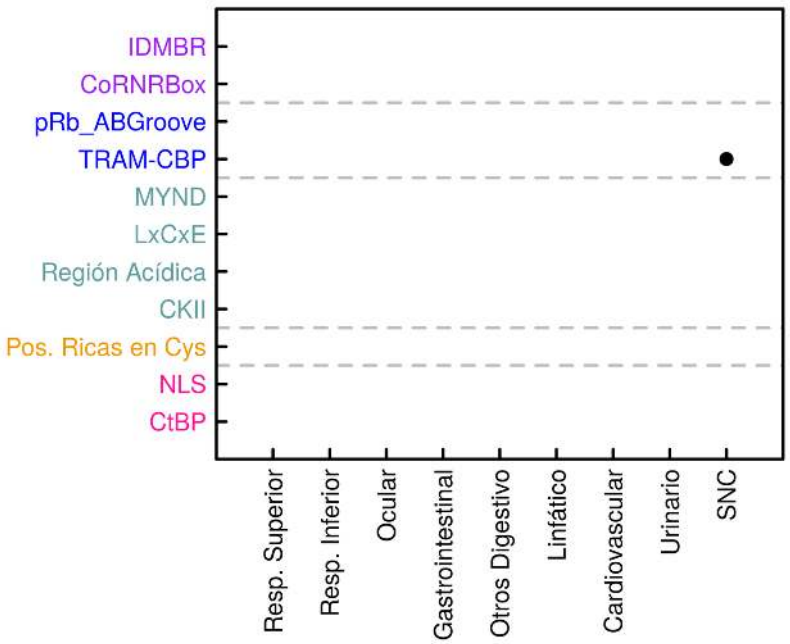


Figura 4.28: Asociaciones entre Motivos-Tropismos en secuencias actuales. Las asociaciones positivas (valor $p^* < 0.05$) están marcadas con un punto negro. El código de colores para las etiquetas y para los puntos es el utilizado en la Figura 4.2.

De las 99 asociaciones posibles entre motivos y tropismo, únicamente se encontró una asociación significativa entre el motivo TRAM-CBP y el tropismo por el Sistema Nervioso Central (Figura 4.28).

En resumen, los estudios de asociación realizados muestran que la presencia y ausencia de motivos está asociada a la especificidad de hospedador, sugiriendo un rol importante de estos motivos en la diversificación de *Mastadenovirus*. Sin embargo, los motivos estudiados no parecen haber tenido un rol significativo en la especificidad de tropismo. Esto puede deberse a la calidad de la información existente en relación al tropismo de adenovirus.

4.15.3. Asociaciones entre eventos de aparición y desaparición de motivos de la proteína E1A

Los eventos de aparición y desaparición de motivos lineales a lo largo de la filogenia de *Mastadenovirus* podrían darse de manera independiente o de manera asociada debido a que, por ejemplo, exista un acople funcional entre ellos. Para determinar si la combinación de motivos difiere del azar se realizó una prueba hipergeométrica (véase Sección 2.5.1) para evaluar cuán significativa es la asociación existente entre los eventos de aparición y desaparición de los once motivos en las distintas ramas del árbol de *Mastadenovirus* por el método empírico de Bayes. El valor p obtenido es finalmente corregido aplicando la corrección de Benjamini-Hochberg (valor p^*) para comparaciones múltiples. (Benjamini y Hochberg, 1995). Aquellas combinaciones en las que el valor p^* es menor a 0.05 son representadas con un punto en la Figura 4.29.

Para los once motivos lineales, considerando que existen dos tipos de eventos, aparición y desaparición, existen un total de 220 asociaciones posibles. Sin embargo, solamente se observaron 30 asociaciones significativas (Figura 4.29). 21 de 30 ocurrían entre eventos del mismo tipo, cuatro entre eventos de aparición y 17 entre eventos de desaparición. Las nueve asociaciones restantes ocurrían entre eventos de distinto tipo. De las 30 asociaciones, solamente cuatro asociaciones ocurren dentro del mismo dominio.

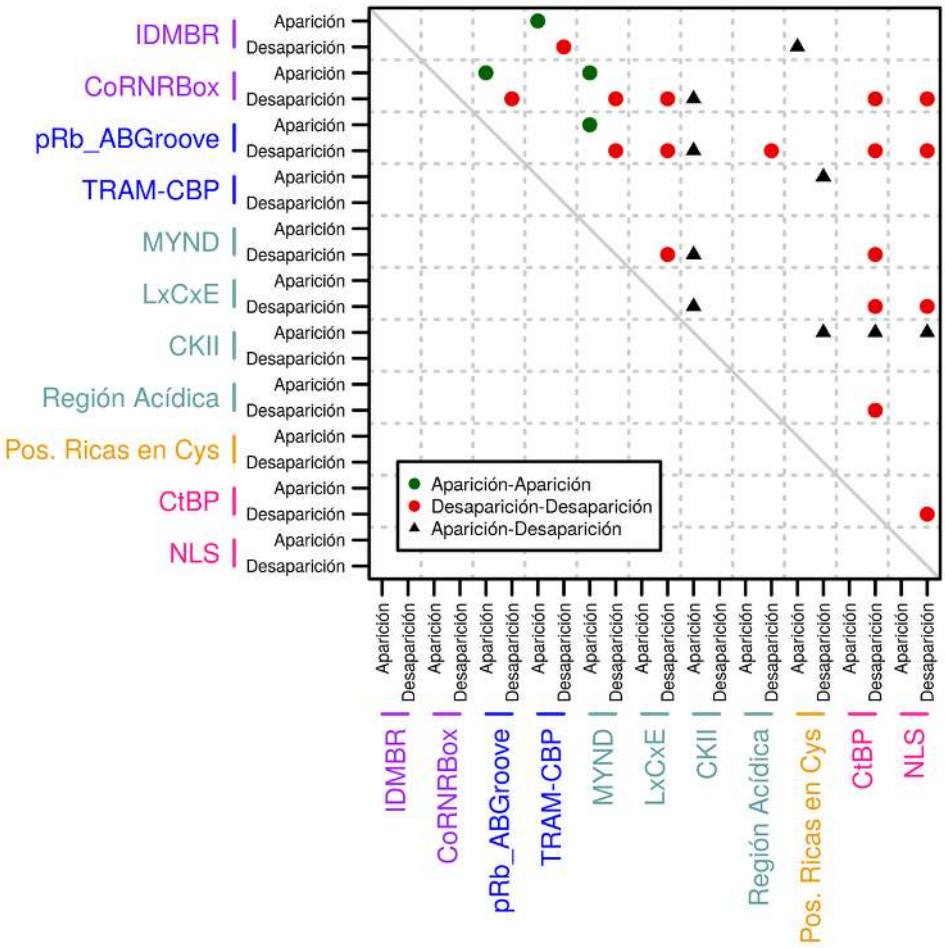


Figura 4.29: Asociaciones entre eventos de aparición y desaparición de motivos a lo largo de la filogenia de *Mastadenovirus*. Las asociaciones positivas (valor $p^* < 0.05$) están marcadas con un punto rojo cuando la asociación ocurre entre eventos de desaparición. Un punto verde cuando la asociación ocurre entre eventos de aparición. Un triángulo negro cuando la asociación ocurre entre un evento de aparición y un evento de desaparición. El código de colores para las etiquetas es el utilizado en la Figura 4.2.

Estos resultados sugieren que existe una asociación no azarosa entre los eventos de aparición y desaparición de motivos, que podría estar indicando una cooperación entre ellos.

4.16. Análisis cofilogenético de los miembros de *Mastadenovirus* y sus hospedadores

Es comúnmente aceptado que los parásitos codivergen con sus hospedadores (Huyse y Volckaert, 2005; Wang *et al.*, 2017) y son numerosos los ejemplos donde se observa coespeciación entre parásito viral y sus hospedadores. Este es el caso de las familias *Polyomaviridae* (Pérez-Losada *et al.*, 2006), *Papillomaviridae* (Gottschling *et al.*, 2011b) y *Reoviridae* (Göker *et al.*, 2011).

Para evaluar la hipótesis de que *Mastadenovirus* coevoluciona con su hospedador en primer lugar se realizó un estudio de cofilogenia, también conocido como coevolución o coespeciación, utilizando un método global. En segundo lugar se realizó un estudio cofilogenético utilizando un método basado en eventos para determinar cuáles podrían ser los eventos evolutivos que participan en la diversificación de hospedadores de *Mastadenovirus*. Inicialmente se intentó utilizar la filogenia completa de *Mastadenovirus*. Esto significaba utilizar las 116 asociaciones observadas (Figura 4.30, arriba). Sin embargo o bien los tiempos de cómputo eran muy grandes (mayor a 2 semanas), o bien el software utilizado colapsaba (véase Sección 2.3.5). Por lo tanto, fue necesario disminuir el tamaño de la filogenia. Para esto se colapsaron aquellos clados monofiléticos que infectaban a un mismo hospedador, disminuyendo el número a 39 asociaciones (Figura 4.30, abajo).

En la Figura 4.30 (arriba) se muestran las asociaciones a agrupar: De 4 serotipos de la especie *Simian adenovirus A*, los serotipos SAdV8 y SAdV3 que infectan a macaco Rhesus (*Macaca mulatta*) tienen un origen monofilético y las ramas fueron colapsadas en un único representante SA_Mm (en lila). La especie *Simian adenovirus B* infecta a macaco cangrejero (*Macaca fascicularis*) y sus ramas fueron colapsadas en un único representante SB (en marrón). Las especies *Human adenovirus A* y *F* infectan a humanos (*Homo sapiens*) y sus ramas fueron colapsadas en HA_HF (azul oscuro). Los serotipos de la especie *Human adenovirus C* infectan a 4 hospedadores diferentes: gorilas (*Gorilla gorilla*), chimpancé (*Pan troglodytes*), bonobo (*Pan paniscus*) y humanos. Los serotipos correspondientes son de origen monofilético y por lo tanto fueron agrupados en: HC_Gg, HC_Pt, HC_Pp y HC_Hs (en violeta, verde oscuro, naranja y fucsia respectivamente). Dentro de la especie *Human adenovirus B*, los serotipos que infectan a los hospedadores chimpancé, humanos y gorila, no poseen un origen monofilético. Se pudieron crear un total de 4 grupos: HB_Pt1, HB_Pt2, HB_Hs1 y HB_Hs2 (amarillo, rosa, celeste y azul claro). El resto de los serotipos no fueron incluidos en ningún grupo. Los serotipos de la especie *Human adenovirus E* que infectan a chimpancé fueron colapsados (verde claro). Los serotipos que pertenecen a la especie *Human adenovirus D* (en rojo en Figura 4.30, arriba) infectan a humanos. Por lo tanto, ese clado fue colapsado en un único representante denominado HD. Por último, fueron colapsadas las ramas de los serotipos que infectan a perros (*Canis lupus familiaris*) en CA (violeta). Este proceso además elimina once de los 14 nodos con bajo soporte.

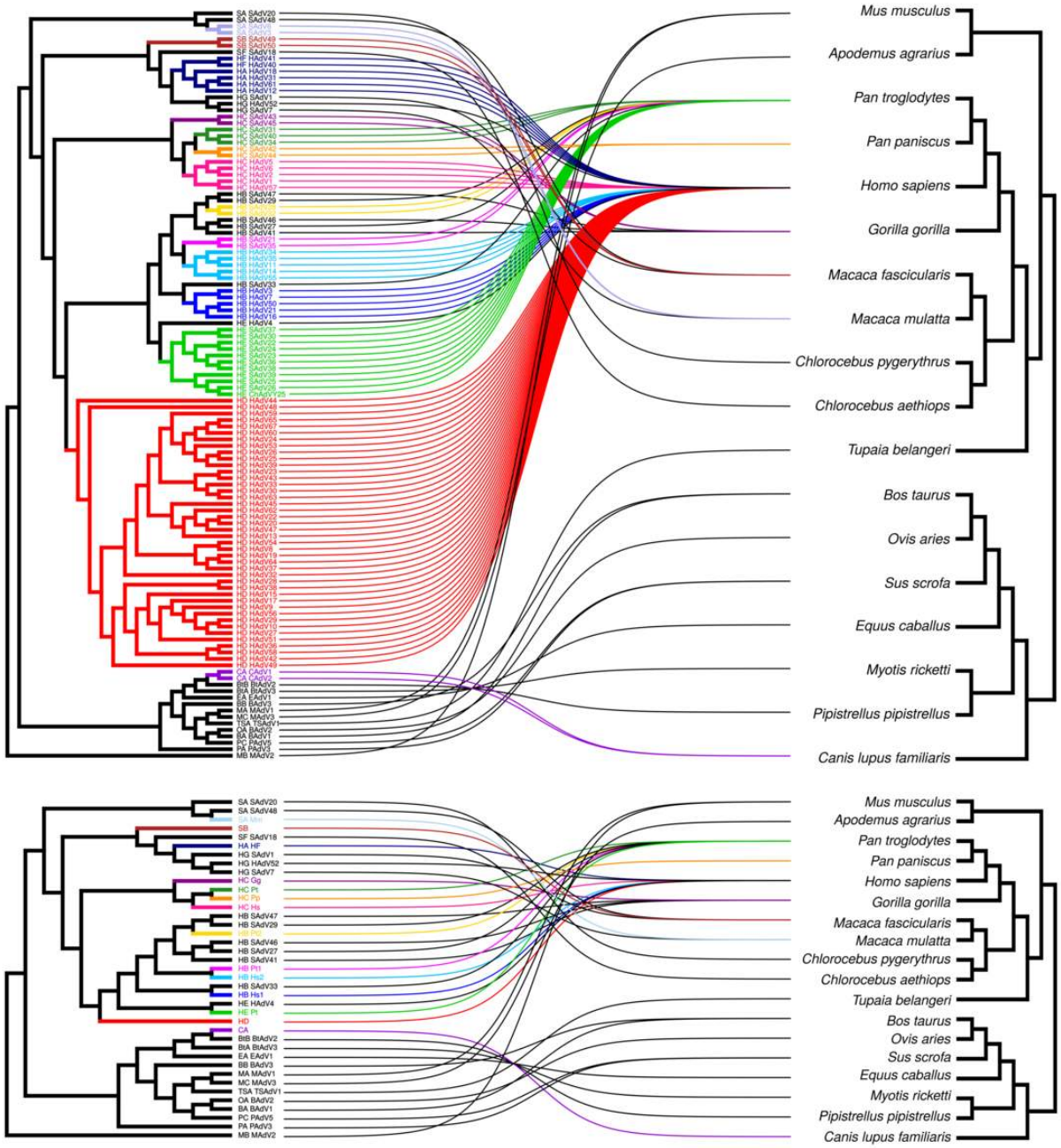


Figura 4.30: Asociaciones entre virus (izquierda) y hospedador (derecha). Las asociaciones entre los clados monofiléticos del parásito (izquierda) que infectan a un mismo hospedador (derecha) están resaltadas con colores en el árbol sin colapsar (*arriba*) y colapsado (*abajo*). Los colores de las asociaciones son los mismos en ambas figuras.

4.16.1. Análisis de cofilogenia global

Con el árbol del hospedador (O'Leary *et al.*, 2013), el árbol de *Mastadenovirus* colapsado (véase Sección 2.3.2) y las asociaciones colapsadas entre los serotipos y sus hospedadores, se realizó un análisis de cofilogenia global utilizando el algoritmo Parafit (véase Sección 2.3.5).

La hipótesis nula global (H_0 *Global*) planteada sería:

> H_0 *Global* = *Dadas la filogenia del hospedador y Mastadenovirus, y las asociaciones observadas, la evolución de los dos grupos fue independiente*

La hipótesis nula Individual (H_0 *Individual*) planteada para cada asociación sería:

> H_0 *Individual* = *Cualquier contribución de cada asociación hospedador-parásito individual al ajuste global no es distinta de la obtenida al azar, y por ende podría ser omitida.*

La hipótesis nula global de evolución independiente es rechazada (valor $p < 0.005$). 25 de 39 hipótesis individuales fueron rechazadas (valor $p < 0.005$ para ambos estadísticos). Estas asociaciones se muestran en color rojo en la Figura 4.31. Diez de las 25 asociaciones significativas corresponden a los diez grupos dentro de la especie *Human adenovirus B* en asociación con los distintos hospedadores (chimpancés, humanos y gorilas). Cuatro de las 25 asociaciones significativas corresponden a los cuatro grupos dentro de la especie *Human adenovirus C* asociados con los hospedadores humanos, bonobos, chimpancés y gorilas. Una de las 25 corresponde al clado de la especie *Human adenovirus D* que tiene como hospedador a humanos. Dos de las 25 corresponden a los dos grupos de la especie *Human adenovirus E* con los hospedadores chimpancés y humanos. Las siete asociaciones restantes significativas corresponden a los serotipos que comparten el hospedador chancho (*Sus scrofa*) (PAdV3 y PAdV5 de las especies *Porcine adenovirus A* y *Porcine adenovirus C* respectivamente), las que comparten el hospedador vaca (*Bos taurus*) (BAdV1 y BAdV3 de las especies *Bovine adenovirus A* y *Bovine adenovirus B* respectivamente), y los serotipos BAdV2 de la especie *Ovine adenovirus A* que tiene como hospedador a la oveja (*Ovis aries*), EAdV1 de la especie *Equine adenovirus A* que tiene como hospedador al caballo (*Equus caballus*), y los serotipos BtAdV3 y BtAdV2 de las especies *Bat adenovirus A* y *B* que tienen como hospedador al murciélago ratonero (*Myotis ricketti*) y al murciélago común (*Pipistrellus pipistrellus*) respectivamente.

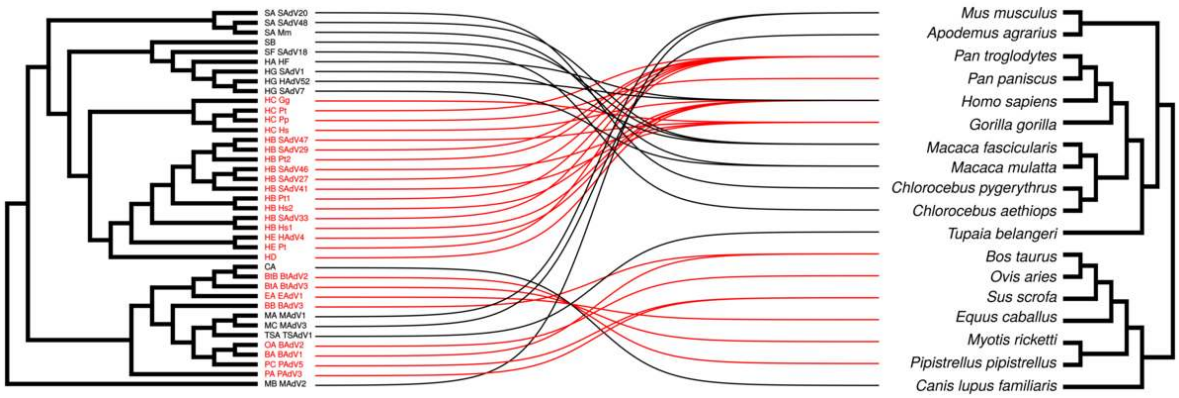


Figura 4.31: Análsis global de coespeciación entre parásito (izquierda) y hospedador (derecha). Se muestran en rojo las asociaciones significativas (valor $p < 0.005$) entre parásito-hospedador.

Las 14 asociaciones restantes (Figura 4.31, líneas negras) no se tiene información suficiente para decir que aportan o no a la coespeciación global (valor $p > 0.005$). En el árbol del parásito

estás asociaciones ocurren con cuatro grandes grupos del virus. El primer grupo contiene dos clados, el clado que da origen a la especie *Simian adenovirus A* y al clado que da origen a las especies *Simian adenovirus B* y *F*, y *Human adenovirus A*, *F* y *G*. El segundo grupo incluye únicamente a *Canine adenovirus A*. El tercer grupo incluye el clado que contiene a las especies *Murine adenovirus A* y *C*, y *Tree shrew adenovirus A*. El último grupo incluye a la especie *Murine adenovirus B*. Por el contrario, las asociaciones significativas (Figura 4.31, líneas rojas) no están agrupadas excepto por el clado que contiene a las especies *Human adenovirus B*, *C*, *D* y *E*.

En el árbol del hospedador, con excepción de *Homo sapiens*, que tiene asociaciones significativas y no significativas, el resto de los hospedadores poseen únicamente un tipo de asociación significativa o no significativa. Las asociaciones significativas (Figura 4.31, líneas rojas) ocurren con dos grandes grupos de hospedadores. El primer grupo incluye a la familia de homínidos (humanos, gorilas, chimpancé y bonobo). El segundo grupo abarca el clado que agrupa hospedadores de los órdenes artiodáctila, perisodáctila y chiroptera. Por otro lado, las asociaciones no significativas (Figura 4.31, líneas negras) ocurren con 5 grupos de hospedadores. El primer grupo incluye los hospedadores del orden rodentia. El segundo grupo incluye únicamente al humano. El tercer grupo incluye al clado que da origen a los cercopitécidos o monos del viejo mundo (*Macaca sp.* y *Chlorocebus sp.*). El cuarto y quinto grupo incluye a tupaya y a los perros.

En resumen, los resultados obtenidos sugieren que la coespeciación tiene un papel evolutivo importante en la diversificación del género *Mastadenovirus*, pero no sería el único evento evolutivo involucrado.

4.16.2. Análisis de cofilogenia basado en eventos

Dado que se demostró que la evolución de *Mastadenovirus* y sus hospedadores no fue independiente y que 25 de las 39 asociaciones aportan de manera significativa a la cofilogenia global, es interesante estudiar cuáles son los otros eventos evolutivos que participaron en la diversificación de hospedadores a lo largo de la historia evolutiva de *Mastadenovirus*. Con este objetivo, se utilizó la filogenia de *Mastadenovirus* colapsada, la filogenia del hospedador, las asociaciones observadas y el software Jane (Conow *et al.*, 2010). Jane utiliza un método basado en eventos para reconciliar la superposición de ambas filogenias (véase Sección 2.3.5).

Dados los cuatro eventos evolutivos coespeciación, duplicación, duplicación con cambio de hospedador (cambio de hospedador de ahora en adelante) y extinción parcial (véase Sección 2.3.5) y los costos asignados Jane devuelve un conjunto de soluciones de igual costo. Ese conjunto de soluciones de igual costo abarca una gran cantidad de soluciones que son isomórficas. Las soluciones isomórficas comparten el mismo costo y el mismo número de eventos para cada evento evolutivo pero la ubicación relativa entre los distintos nodos es diferente. Por lo tanto, las soluciones isomórficas se consideran técnicamente iguales y pueden agruparse.

A cada evento evolutivo es necesario asignarle un costo. En base a ese costo y el número total de eventos que Jane utilice para reconciliar las filogenias se obtiene el costo total de la solución. No está definido en la literatura un criterio para la elección de costos de los distintos eventos

evolutivos. En general, se elige un conjunto de costos y variaciones, se describen los rangos de los números de eventos y se muestra una solución a modo de ejemplo. En este trabajo de tesis se intenta elegir en base a un criterio las soluciones a utilizar ya que serán utilizadas en sucesivos análisis. Esta selección tiene dos etapas. En primer lugar, se realiza una elección de costos y, en segundo lugar, dentro de los conjuntos de soluciones elegidos en la primera etapa el grupo de soluciones isomórficas adecuado.

Selección de costos de eventos. Para expresar los costos utilizaremos un vector de costos, c = (Coespeciación, Duplicación, Cambio de Hospedador, Extinción Parcial). En primer lugar, se evaluaron primero los tres costos comúnmente utilizados en la literatura, $v1 = (0, 1, 2, 1)$, $v2 = (0, 1, 1, 2)$ y $v3 = (0, 1, 1, 1)$ (Charleston y Robertson, 2002; Conow *et al.*, 2010; Cuthill y Charleston, 2012; Huyse y Volckaert, 2005; Pérez-Losada *et al.*, 2006; Wang *et al.*, 2017). Luego, se amplió al grupo de costos fijando coespeciación en cero y duplicación en uno, variando los eventos evolutivos cambio de hospedador y extinción parcial de uno a cinco en pasos de 1. Para los costos comúnmente usados en la literatura, $v1 = (0, 1, 2, 1)$, $v2 = (0, 1, 1, 2)$ y $v3 = (0, 1, 1, 1)$, se obtuvieron los costos totales de 29, 30 y 48. Para el resto de las combinaciones de costos de los eventos evolutivos se obtuvo un rango de costos totales que variaban de 53 a 125. Por lo tanto, nos quedamos con las soluciones que poseen los costos de la literatura por ser las más parsimoniosas.

En la Tabla 4.4 se resumen los costos y número de eventos para los tres grupos de soluciones elegidas. Además, se indica para cada grupo el número total de soluciones obtenido y el número de soluciones en las que se pueden agrupar las soluciones isomórficas. Por ejemplo, para la combinación de costos $v1 = (0, 1, 2, 1)$, se obtuvieron 8244 soluciones que poseían un costo total de 48 y las soluciones isomórficas se podían agrupar en 30 grupos de soluciones óptimas.

	Número de eventos (Costo del evento)					
Costo Total	**Coespeciación**	**Duplicación**	**Cambio de hospedador**	**Extinción Parcial**	**Soluciones agrupadas**	**Soluciones totales**
48	15,16,17 (0)	5,6 (1)	15,16,17,18 (2)	7,8,9,10,11,12 (1)	30	8244
30	8,10 (0)	2,3 (1)	26,27,28 (1)	0,1 (2)	6	6100
29	10,11,12 (0)	2 (1)	24,25,26 (1)	3 (1)	5	6783

Tabla 4.4: Costos asignados a cada evento evolutivo. Para cada costo total (primera columna) se indica el rango de número de eventos obtenidos para cada uno de los cuatros eventos evolutivos y el costo asignado al evento entre paréntesis. En las dos últimas columnas se indican el número de grupos de soluciones isofórmicas y el número total de soluciones isomórficas.

Se puede observar en la Tabla 4.4 que para el conjunto de soluciones de costo total 48 el rango del número de eventos de coespeciación es entre 15 y 17, para duplicaciones es entre cinco y seis, para cambio de hospedador es de 15 a 18 y para las extinciones parciales es de siete a doce. Para el conjunto de soluciones de costo total 30 el rango del número de eventos de coespeciación es entre ocho y diez, para duplicaciones es entre dos y tres, para cambio de hospedador es de 26 a 28 y

para las extinciones parciales es de cero a uno. Para el conjunto de soluciones de costo total 29 el rango del número de eventos de coespeciación es entre diez y doce, para cambio de hospedador es de 24 a 26 y todas las soluciones presentan dos eventos de duplicación y tres extinciones parciales. En comparación al conjunto de soluciones de costo total 48, se puede observar que al aumentar el costo de extinción parcial y disminuir el costo de cambio de hospedador en el conjunto de soluciones de costo total 30, disminuye el número de eventos de coespeciación. Lo mismo ocurre en el conjunto de soluciones de costo total 29 al disminuir únicamente el costo de cambio de hospedador. Dado que se estableció que existe una coespeciación global evaluada por Parafit (véaase Sección 4.16.1), se considera que el conjunto de soluciones de costo total 48 es la que maximiza el número de eventos de coespeciación y se continuará trabajando de ahora en más con este conjunto de soluciones.

Selección de grupos de soluciones isomórficas de costo total 48. Dado que las soluciones isomórficas son técnicamente equivalentes, se utiliza una una solución por cada conjunto de soluciones isomórficas. Es decir, para el conjunto de soluciones de costo total 48, se elige para cada uno de los 30 grupos una única solución. Al estudiar los eventos evolutivos que ocurrían en cada una de las ramas de las 30 soluciones, se observó que compartían al menos el 67 % de eventos en las distintas ramas. Para poder elegir un conjunto de soluciones se utilizaron los valores de soporte de cada evento evolutivo. Dado un conjunto de soluciones de igual costo, Jane asigna un valor de soporte a cada evento evolutivo que aparece en una determinada posición en el árbol del parásito. Este valor de soporte se calcula como el porcentaje de ver ese mismo evento en esa misma posición entre todas las soluciones encontradas. Se realizó un diagrama de cajas (en inglés, *boxplot*) para evaluar la distribución de los valores de soporte para cada una de las soluciones (Figura 4.32). En el eje *x* se indica los identificadores de cada una de las soluciones utilizadas. En el eje *y* los valores de soporte de cada una de los eventos calculado como el porcentaje de veces que aparece en una misma ubicación en el árbol.

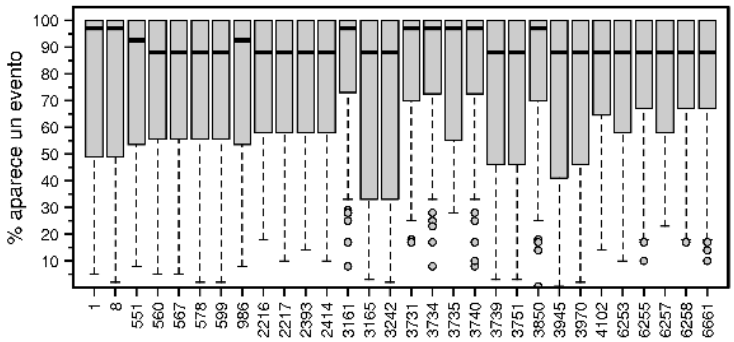


Figura 4.32: Distribución de soporte de los eventos evolutivos para cada solución de costo total 48.

Un diagrama de cajas permite representar gráficamente una serie de datos numéricos a través de sus cuartiles. De esta manera, el diagrama de caja muestra a simple vista la mediana (línea negra en cada caja) y los cuartiles de los datos, pudiendo también representar los valores atípicos de estos

en forma de puntos. El extremo inferior de la caja representa el cuartil uno (Q1). Este es el valor a partir del cual el 25 % de los datos están por debajo y el 75 % de los datos están por arriba. La línea negra representa la media o el cuartil dos (Q2). Este es el valor a partir del cual el 50 % de los datos están por debajo y el 50 % de los datos están por arriba. El extremo superior de la caja representa el cuartil tres (Q3). Este es el valor a partir del cual el 75 % de los datos están por debajo y el 25 % de los datos están por arriba. La distancia entre el extremo superior e inferior, se denomina rango intercuartílico (IQR). Los bigotes de cada caja (líneas punteadas) se calcularon como 1.5 IQR. Por debajo de este valor están los valores atípicos.

Siete soluciones presentan el 50 % de los valores de soporte mayores al 95 %, soluciones 1, 8, 3161, 3731, 3734, 3735, 3740 y 3850. Se analizaron para estas soluciones cuales eran las que tenían menos valores de soporte bajos y cuales eran las que tenían más valores de soporte altos. De las siete soluciones, las soluciones 1 y 8 poseían más del 10 % de los valores de soporte menores al 20 %, y las soluciones 3161, 3731, 3735 y 3850 poseían más del 20 % de valores de soporte menores a 40 %. Las soluciones 3734 y 3740 poseían menos del 20 % de valores de soporte menores a 40 %. De las siete soluciones, las soluciones 1 y 8 poseían menos del 75 % de los valores de soporte mayores al 50 %, y las soluciones 3731 y 3850 poseían menos del 75 % de valores de soporte menores a 70 %. Las soluciones 3161, 3734 y 3740 poseían más del 75 % de valores de soporte mayores al 70 %. En base a estos resultados se eligieron las soluciones 3734 y 3740 para ser utilizadas en los siguientes análisis ya que son las que poseen el mayor número de eventos con un soporte elevado y el menor número de eventos con un soporte bajo. A fines prácticos se muestra únicamente la solución 3734. La solución 3740 está disponible en el apéndice Sección F.10 representada como proyección.

Este experimento puede ampliarse de diversas maneras. En Jane se pueden incluir espacios temporales que no se están considerando, por ejemplo, largos de ramas. También se pueden analizar todas las soluciones alternativas que no se realizó por demanda temporal. Sería interesante además probar hipótesis alternativas de la filogenia de los hospedadores y del parásito, así como también, incluir en la filogenia de *Mastadenovirus* las más de 10 especies nuevas que son de especial interés y enriquecerían considerablemente el análisis ya que no todas son de Primates. Incluye, por ejemplo, delfines, ardillas, león marino, zorrinos y ciervos.

La Figura 4.33 muestra la reconciliación de árboles obtenida por Jane (Conow *et al.*, 2010). La filogenia del hospedador se muestra como una línea negra y sobre ella se superpone la filogenia de *Mastadenovirus* representada en color azul. Todos los eventos evolutivos están marcados sobre la filogenia de *Mastadenovirus*. Los eventos de coespeciación y duplicación están indicados sobre los nodos. Así, los círculos blancos con línea roja indican los eventos de coespeciación. Se puede observar que las ramas que surgen de estos nodos superponen bien. Los eventos de duplicación están indicados con círculos rojos. Los eventos de cambio de hospedador ocurren luego de un evento de duplicación indicado en el nodo, y están marcados con un triángulo azul sobre la rama. Por último, el evento de extinción parcial está señalado con una línea punteada que sigue una de las dos ramas del hospedador.

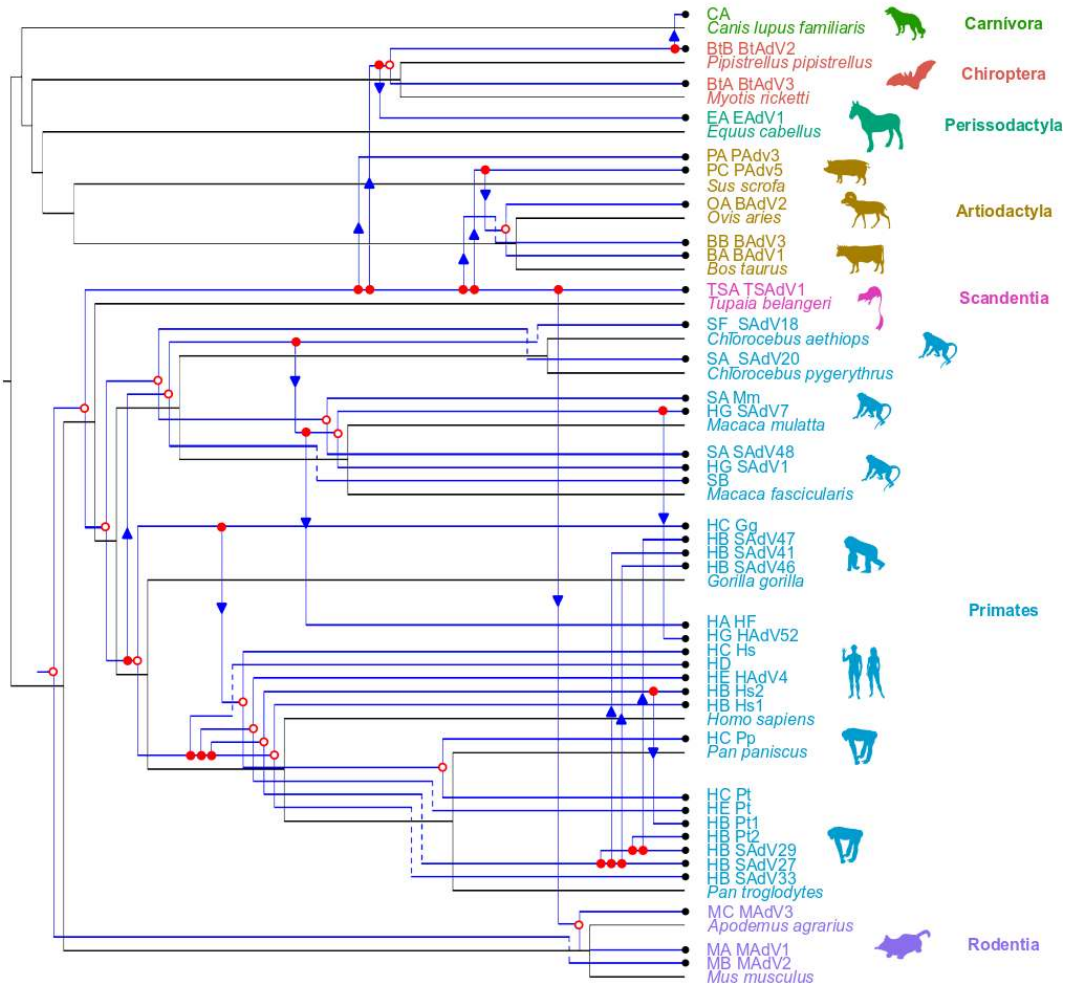


Figura 4.33: Procesos evolutivos en la evolución de *Mastadenovirus*. Se muestra la reconciliación creada por Jane (Conow *et al.*, 2010) del árbol de *Mastadenovirus* (azul) y sus hospedadores (negro). Los símbolos utilizados para los eventos son: círculo blanco con borde rojo para coespeciación, círculos rojos para duplicación, triángulos azules para cambio de hospedador y línea punteada para extinción parcial. En las hojas del árbol se muestran los serotipos con su especie y los grupos monofiléticos que compartían hospedador colapsados, por debajo el hospedador al que infectan. A la derecha se indica el orden al que pertenece el hospedador y la silueta del mismo.

Para poder visualizar más fácilmente lo que ocurre a lo largo de la evolución de *Mastadenovirus* es necesario mapear o proyectar los eventos evolutivos que la dirigen sobre la filogenia de *Mastadenovirus*. El mapeo resultante se muestra en la Figura 4.34 (véase Sección 2.3.5).

Cada uno de los cuatro eventos evolutivos, coespeciación, duplicación, cambio de hospedador y extinción parcial, se representan como proyección sobre el árbol de *Mastadenovirus*. En primer lugar, el símbolo en el nodo y el color de las ramas marca la ocurrencia de uno de los cuatro eventos evolutivos. Así, por ejemplo, el evento de coespeciación está indicado con un nodo blanco con borde rojo y seguido de una rama roja. El evento de duplicación está indicado con un nodo negro y seguido de una rama negra. El evento de cambio de hospedador está indicado con una rama azul y un triángulo azul. Una rama punteada indica que además ocurrió un evento de extinción parcial luego del evento evolutivo correspondiente. En los extremos de las ramas se indican los serotipos

correspondientes, los clados que fueron colapsados, el hospedador y el orden del hospedador.

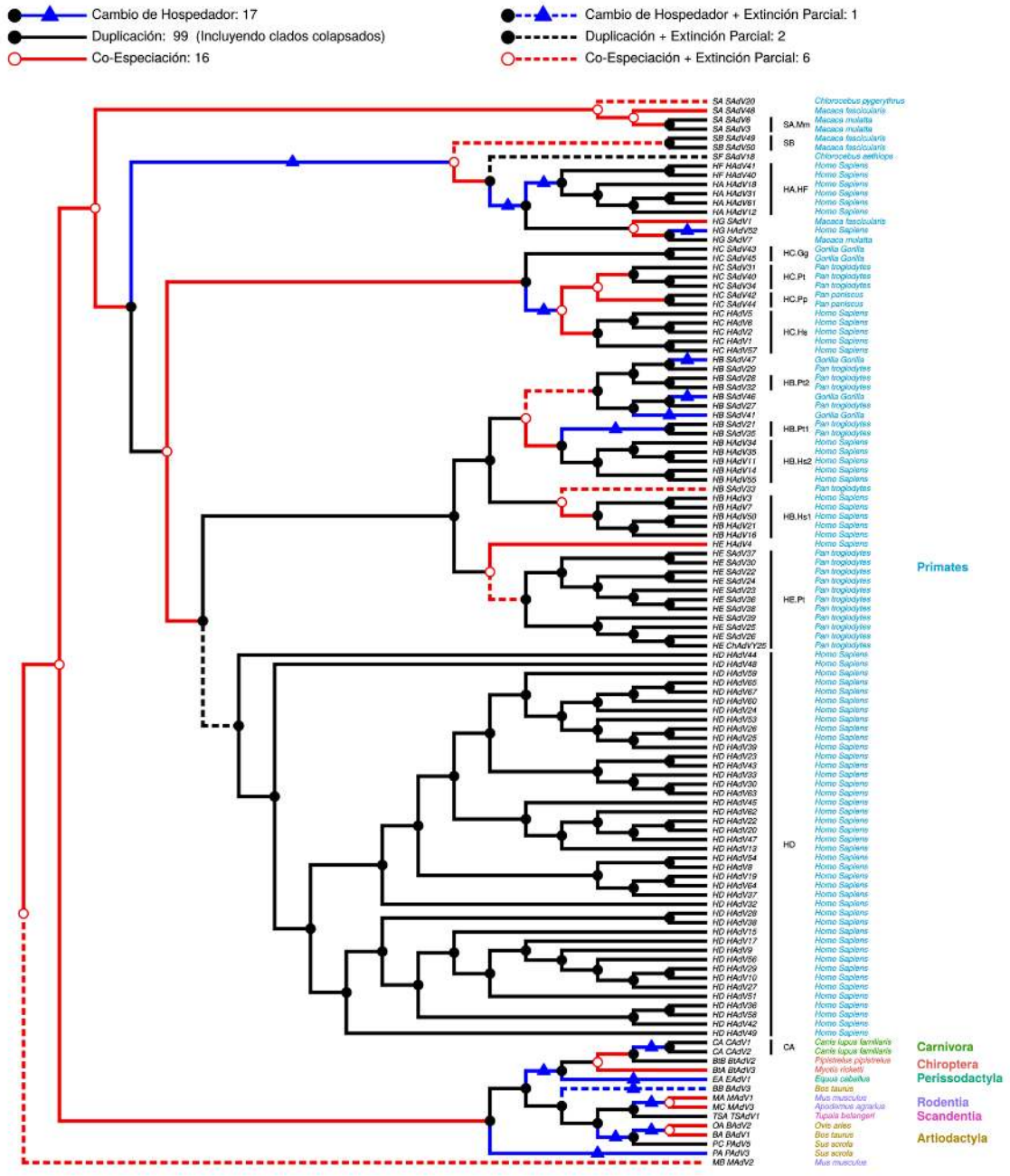


Figura 4.34: Mapeo de los procesos evolutivos en la evolución de *Mastadenovirus* sobre su filogenia. Se muestra el mapeo de la reconciliación creada por Jane (Conow *et al.*, 2010) del árbol de *Mastadenovirus* y el árbol de su hospedador de la Figura 4.33. Los símbolos utilizados para los eventos son: círculo blanco con borde rojo para coespeciación, círculos negros para duplicación, triángulos azules para cambio de hospedador y línea punteada para extinción parcial. En las hojas del árbol se muestran los serotipos con su especie. Se señalan los grupos monofiléticos que compartían hospedador y que fueron colapsados para el análisis. Al lado de los serotipos se indica el hospedador al que infectan. Por último, se indica el orden al que pertenece el hospedador.

En la solución se identifican 16 eventos de coespeciación (Figura 4.33 y Figura 4.34), tanto en la base del árbol como en los extremos. Se observa un evento de coespeciación que da origen a los

ancestros de adenovirus que infectan a los hospedadores pertenecientes a los ordenes Scandentia y Primates. Se observan cinco eventos de duplicación y cambio de hospedador que dan origen a los ancestros que infectan los hospedadores pertenecientes a los órdenes Artiodactyla, Perissodactyla, Chiroptera, Carnívora y Rodentia (con excepción de MAdV2). También se observa diversidad de eventos evolutivos que dan origen a las especies de *Mastadenovirus*. Un evento de coespeciación da origen a la especie *Simian adenovirus B*, otro da origen al ancestro de las especies de adenovirus que infectan a los hospedadores pertenecientes a Chiroptera, *Bat adenovirus A* y *B* y otro da origen a las especies *Murine adenovirus A* y *C*. Un cambio de hospedador da origen a las especies *Human adenovirus A* y *F*. Una duplicación y extinción parcial da origen a la especie *Human adenovirus D*. La diversidad de hospedadores dentro de algunas especies también es explicada por distintos eventos evolutivos: dos eventos de coespeciación y un evento de cambio de hospedador explican la diversidad dentro de la especie *Human adenovirus C*, dos eventos de coespeciación la diversidad de hospedadores dentro de la especie *Simian adenovirus A*, un evento de coespeciación la de la especie *Human adenovirus E*, y por último un evento de coespeciación, un evento de duplicación y tres eventos de cambio de hospedador la de *Human adenovirus B*.

En conclusión, los resultados obtenidos sugieren que la coespeciación tiene un papel evolutivo importante en la diversificación del género *Mastadenovirus* pero no sería el único evento evolutivo que estaría involucrado. También ocurren numerosas duplicaciones, cambios de hospedador y extinciones parciales.

4.17. Asociación entre eventos evolutivos y eventos de aparición y desaparición de motivos.

Es generalmente aceptado en la literatura que los motivos lineales pueden aparecer y desaparecer de manera azarosa (Neduva y Russell, 2005). Este pensamiento se basa en que los motivos lineales consisten mayoritariamente en unos pocos residuos funcionales y una mutación en uno de ellos tiene como consecuencia la pérdida de funcionalidad - o desaparición - del motivo. De la misma manera, una mutación en una determinada secuencia puede llevar a la ganancia de funcionalidad y aparición del motivo por convergencia. Esto les brinda una cierta plasticidad evolutiva que los hace buenos candidatos para cumplir un rol adaptativo en la evolución.

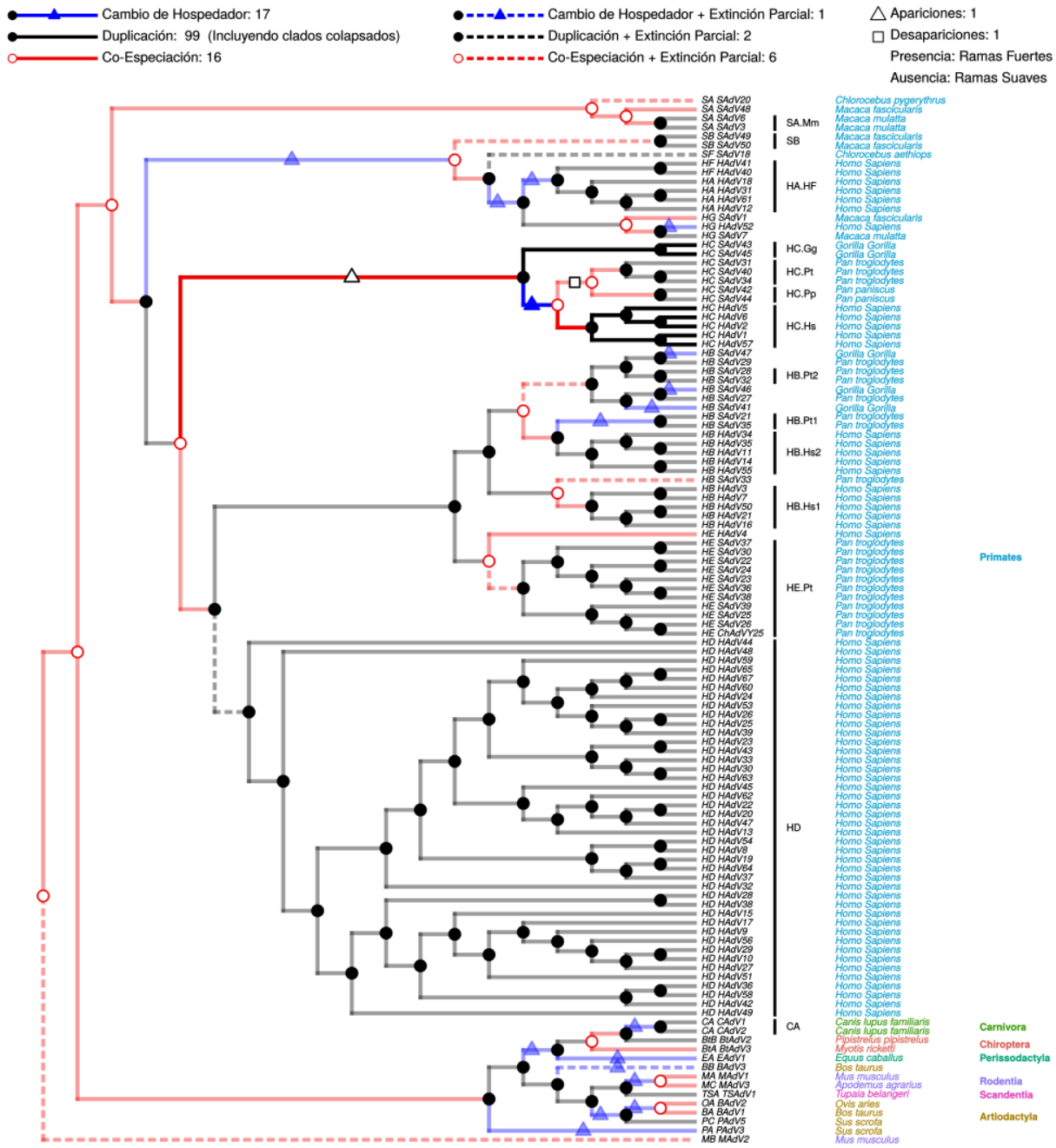
Utilizando los resultados de coespeciación obtenidos mediante el análisis basado en eventos (Sección 4.16.2) y el análisis evolutivo de motivos lineales de la proteína E1A a lo largo de la filogenia de *Mastadenovirus* (Sección 4.13) estudiamos si existía una relación entre los cuatro eventos evolutivos, coespeciación, duplicación, cambio de hospedador y extinción parcial, y la aparición y desaparición de motivos a lo largo de la filogenia de *Mastadenovirus*. Para esto realizamos una superposición de los árboles con el mapeo de motivos (Figura 4.22) y eventos evolutivos (Figura 4.34).

Este análisis se realizó utilizando las reconstrucciones evolutivas de los motivos mediante el método de parsimonia y el método empírico de Bayes. A modo de ejemplo, se muestran los resultados obtenidos en el análisis de coespeciación con la solución mostrada en la Figura 4.33 y la

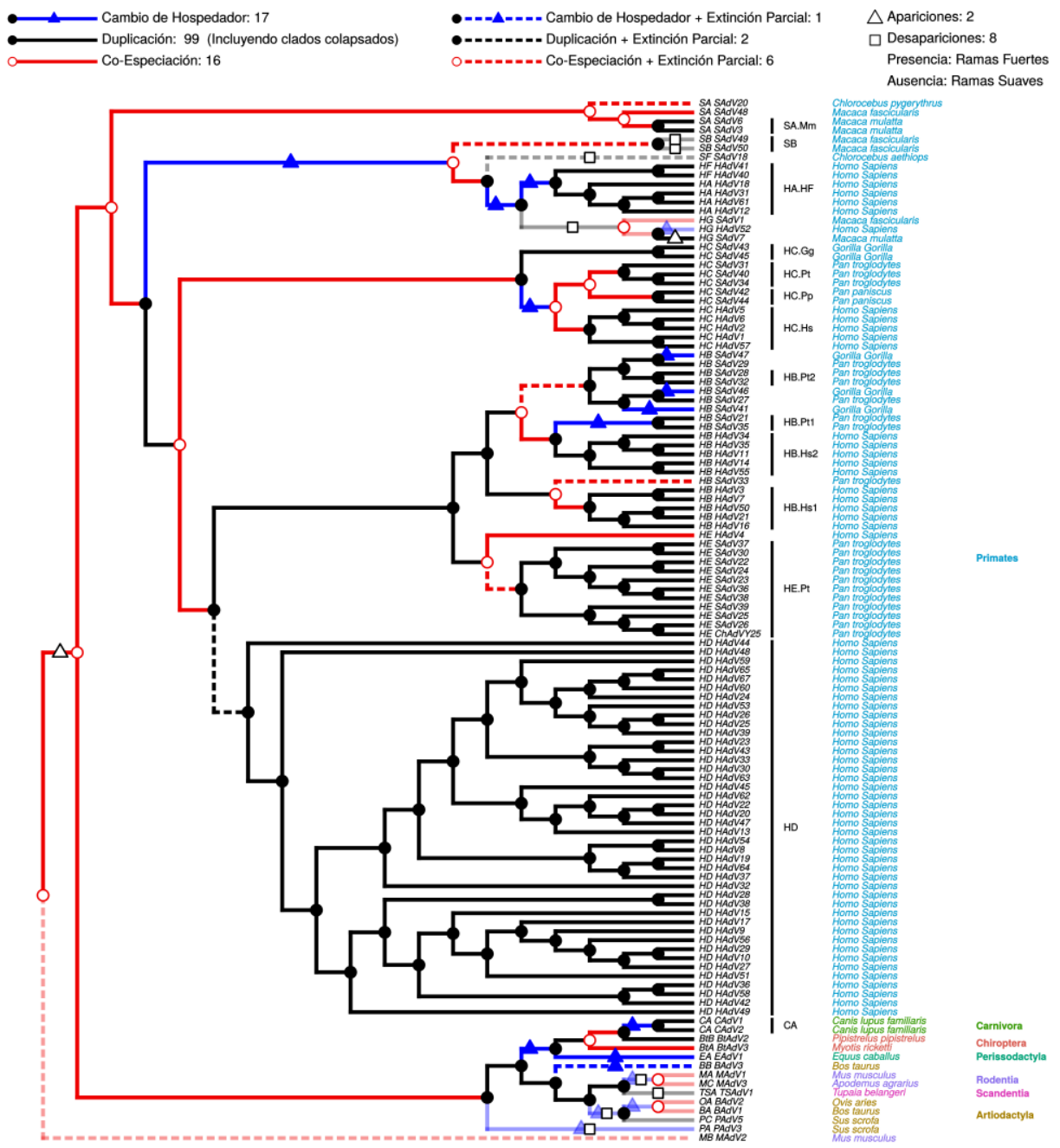
Figura 4.34 (véase Sección 4.16.2) y la reconstrucción evolutiva de motivos mediante el método empírico de Bayes mostrada en la Figura 4.22 (véase Sección 4.13.2).

Los resultados se muestran en la Figura 4.35. Utilizando la proyección de eventos evolutivos sobre el árbol de *Mastadenovirus* se representan en el mismo árbol los eventos de aparición y desaparición de motivos. En primer lugar, el símbolo en el nodo y el color de las ramas marca la ocurrencia de uno de los cuatro eventos evolutivos. Así, por ejemplo, el evento de coespeciación está indicado con un nodo blanco con borde rojo y seguido de una rama roja. Una rama punteada indica que ocurrió un evento de extinción parcial luego del evento de coespeciación. Los eventos de aparición y desaparición de motivos están indicados sobre la rama como triángulos y cuadrados blancos con borde negro. Las ramas más suaves indican la ausencia del motivo y las ramas más fuertes indican la presencia del motivo.

Superposición de eventos evolutivos con el mapeo de motivos lineales del dominio N-terminal

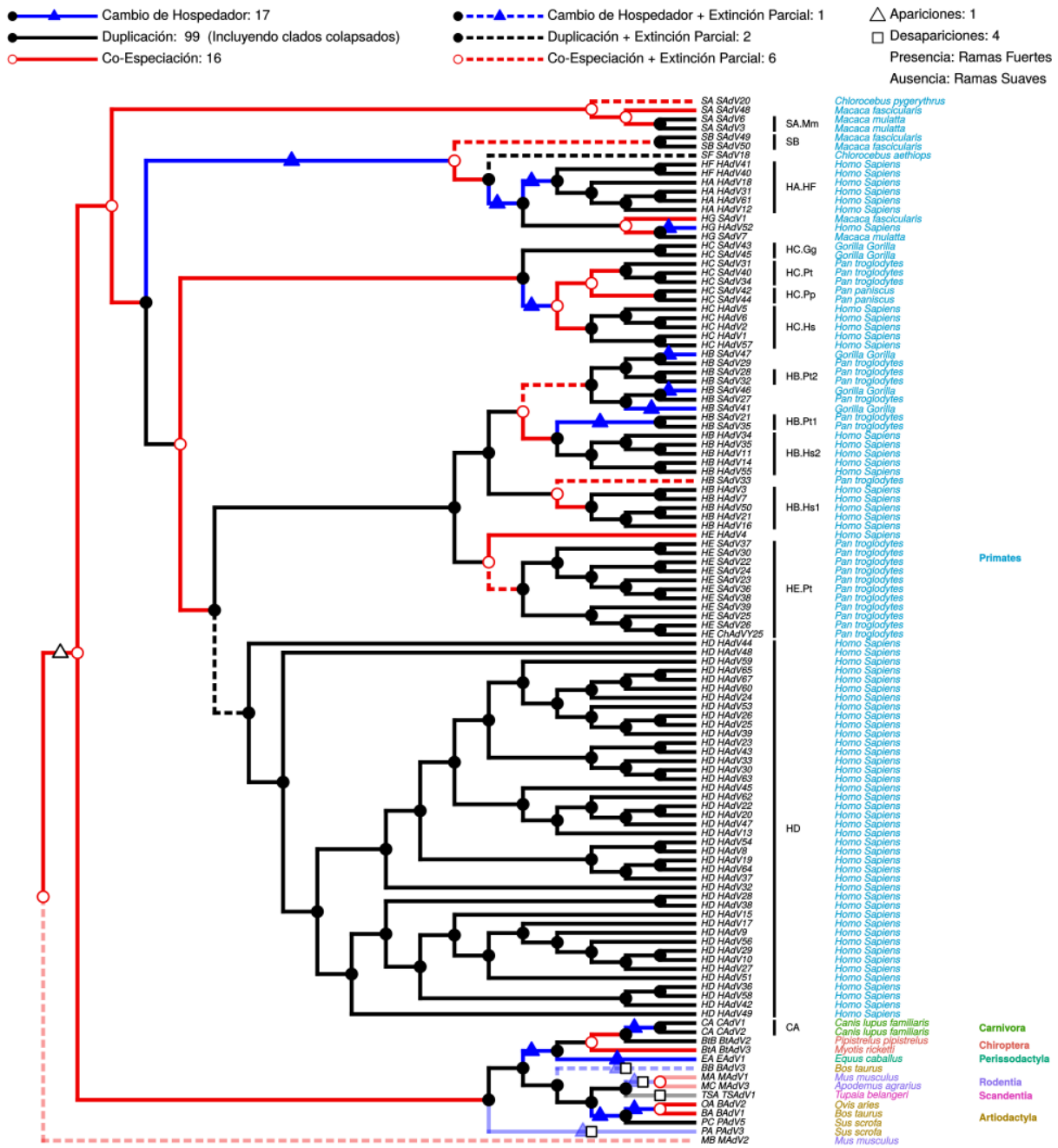


A. IDMBR

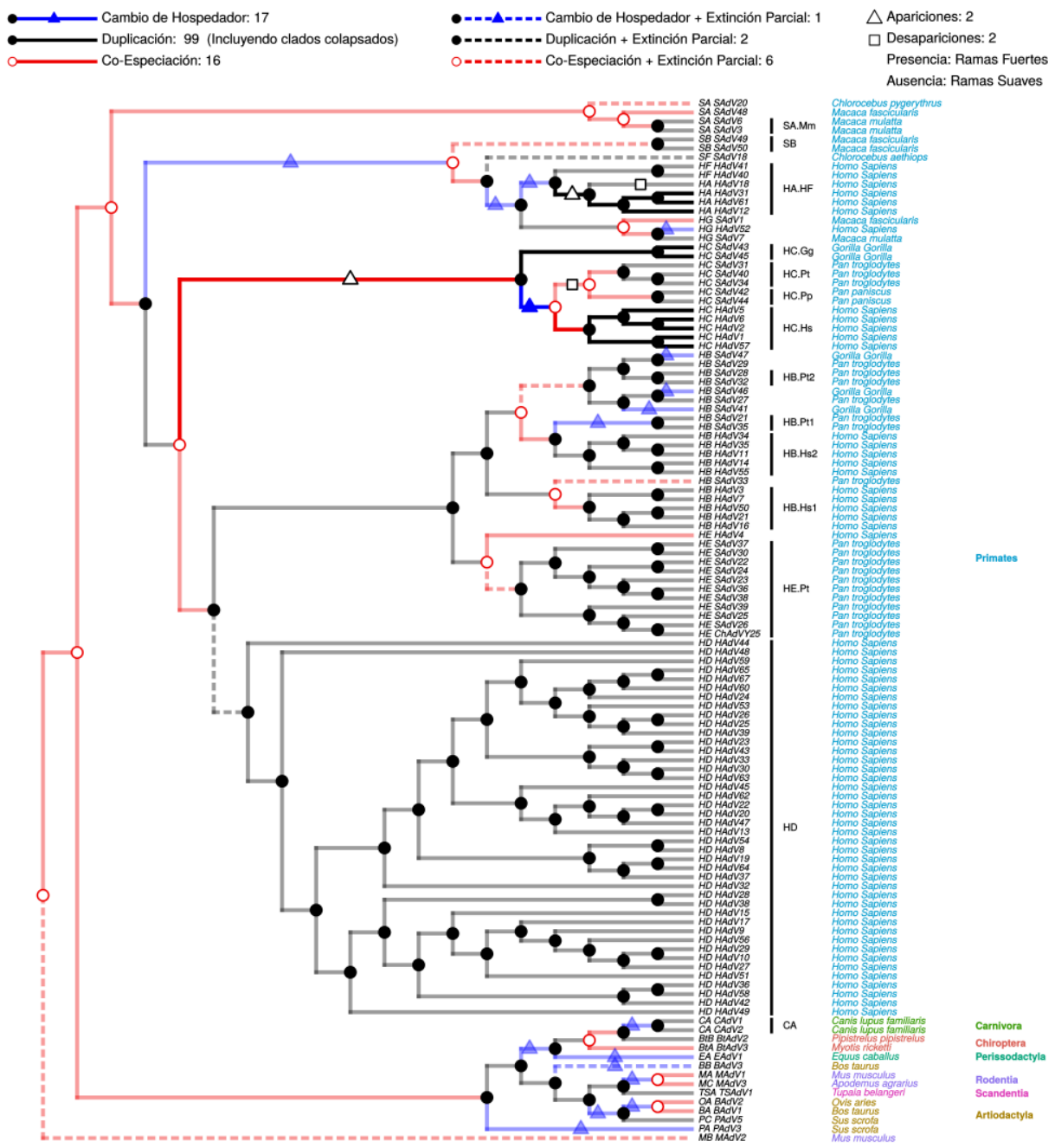

Cambio de Hospedador: 17
Duplicación: 99 (Incluyendo clados colapsados)
Co-Especiación: 16
Cambio de Hospedador + Extinción Parcial: 1
Duplicación + Extinción Parcial: 2
Co-Especiación + Extinción Parcial: 6
Apariciones: 2
Desapariciones: 8
Presencia: Ramas Fuertes
Ausencia: Ramas Suaves
SA.Mm
SB
HA.HF
HC.Gg
HC.Pt
HC.Pp
HC.Hs
HB.Pt2
HB.Pt1
HB.Hs2
HB.Hs1
HE.Pt
HD
CA
Primates
Carnivora
Chiroptera
Perissodactyla
Rodentia
Scandentia
Artiodactyla

B. CoRNRBox

Superposición de eventos evolutivos con el mapeo de motivos lineales del dominio CR1

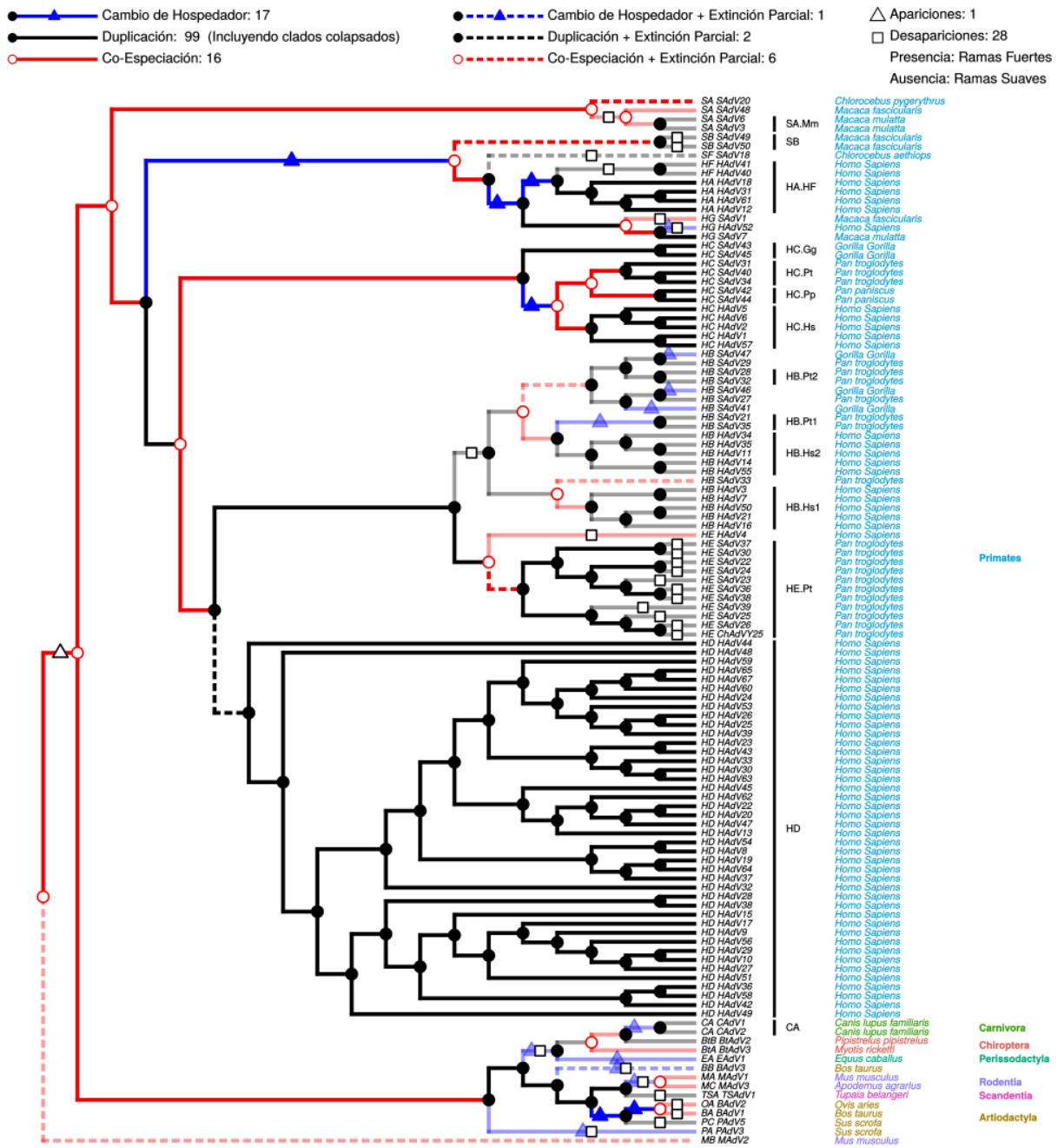


C. pRb_ABGroove

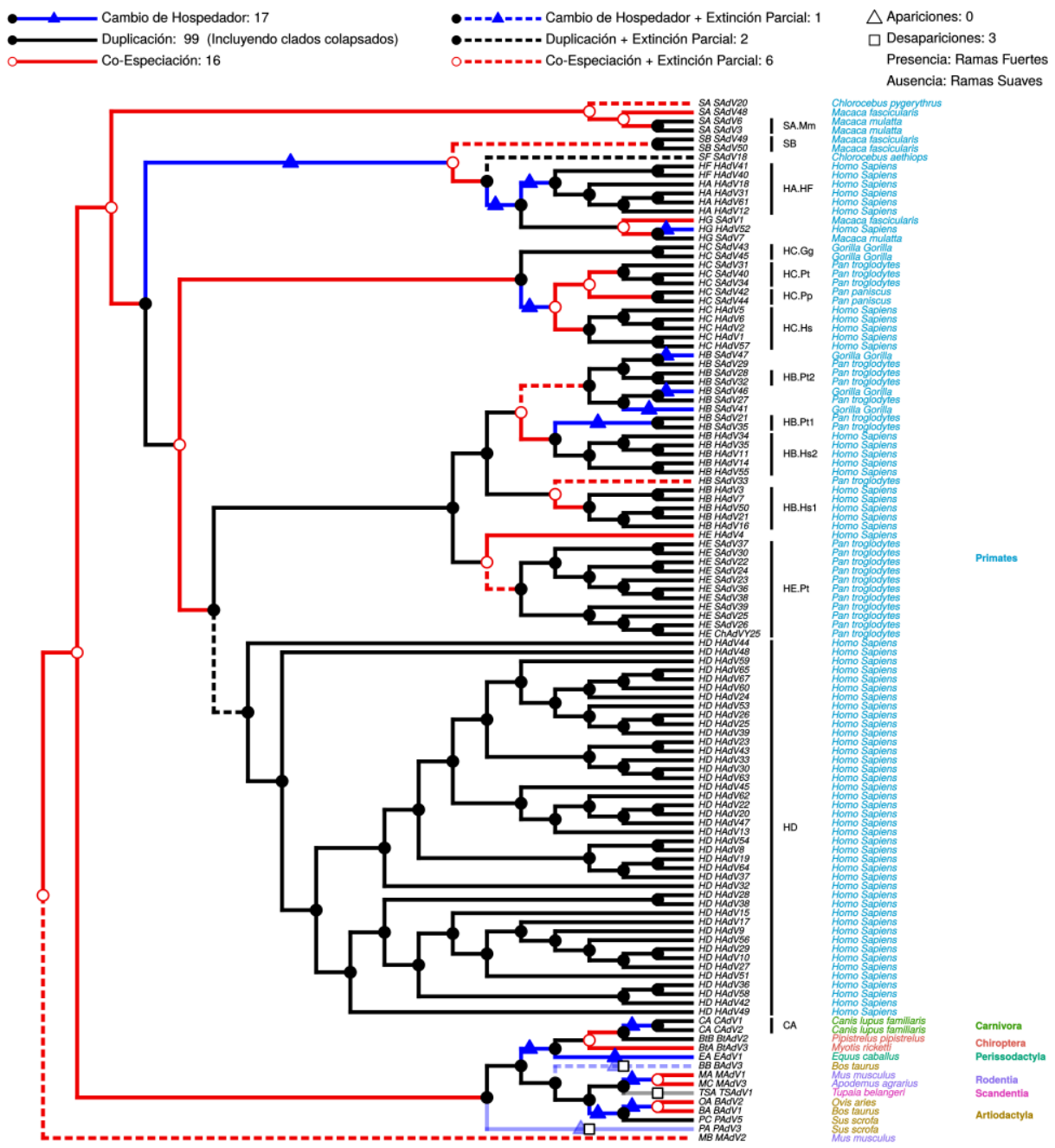

Cambio de Hospedador: 17
Duplicación: 99 (Incluyendo clados colapsados)
Co-Especiación: 16
Cambio de Hospedador + Extinción Parcial: 1
Duplicación + Extinción Parcial: 2
Co-Especiación + Extinción Parcial: 6
Apariciones: 2
Desapariciones: 2
Presencia: Ramas Fuertes
Ausencia: Ramas Suaves
SA.Mm
SB
HA.HF
HC.Gg
HC.Pt
HC.Pp
HC.Hs
HB.Pt2
HB.Pt1
HB.Hs2
HB.Hs1
HE.Pt
HD
CA
Primates
Carnivora
Chiroptera
Perissodactyla
Rodentia
Scandentia
Artiodactyla

D. TRAM-CBP

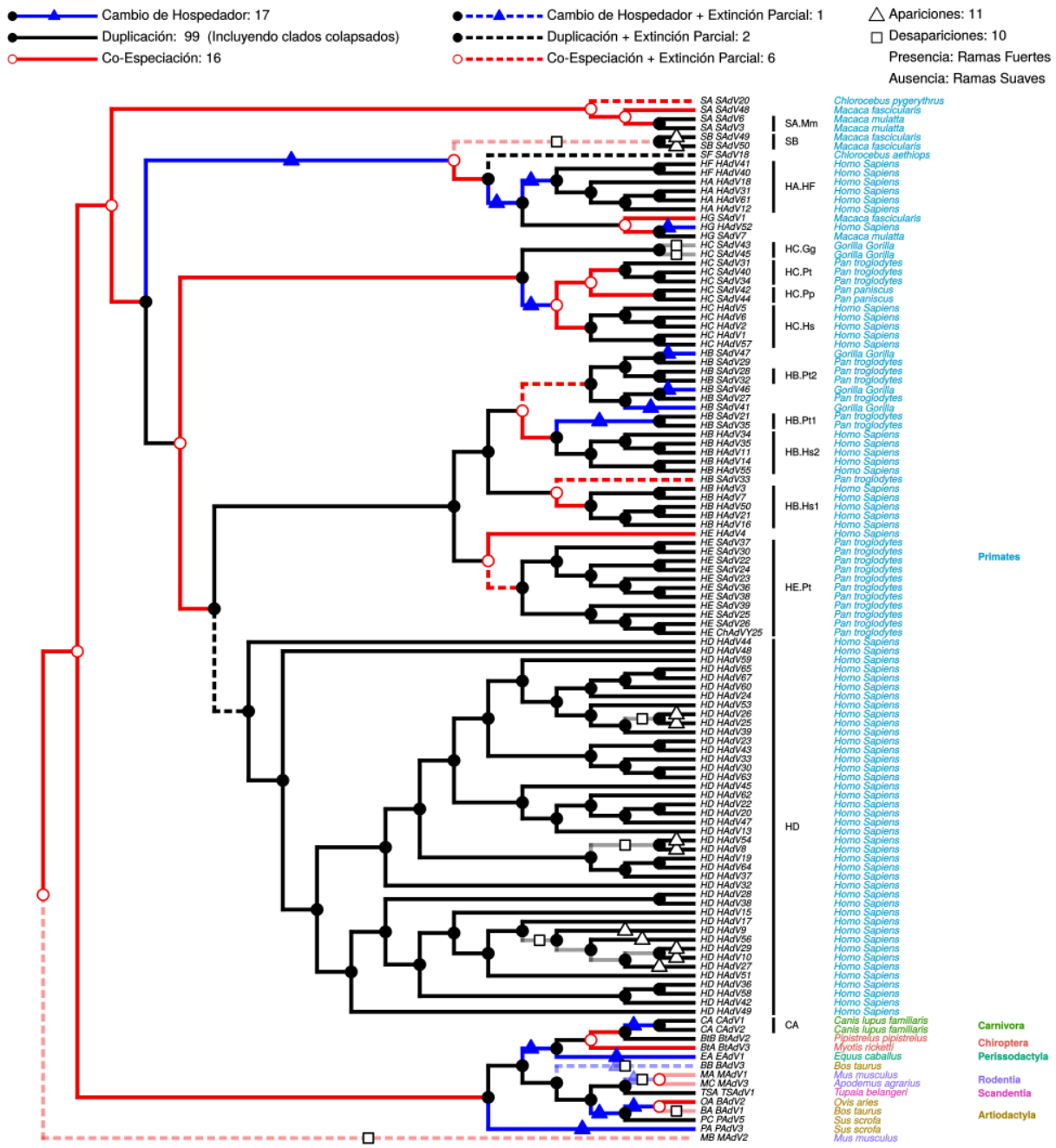
Superposición de eventos evolutivos con el mapeo de motivos lineales del dominio CR2



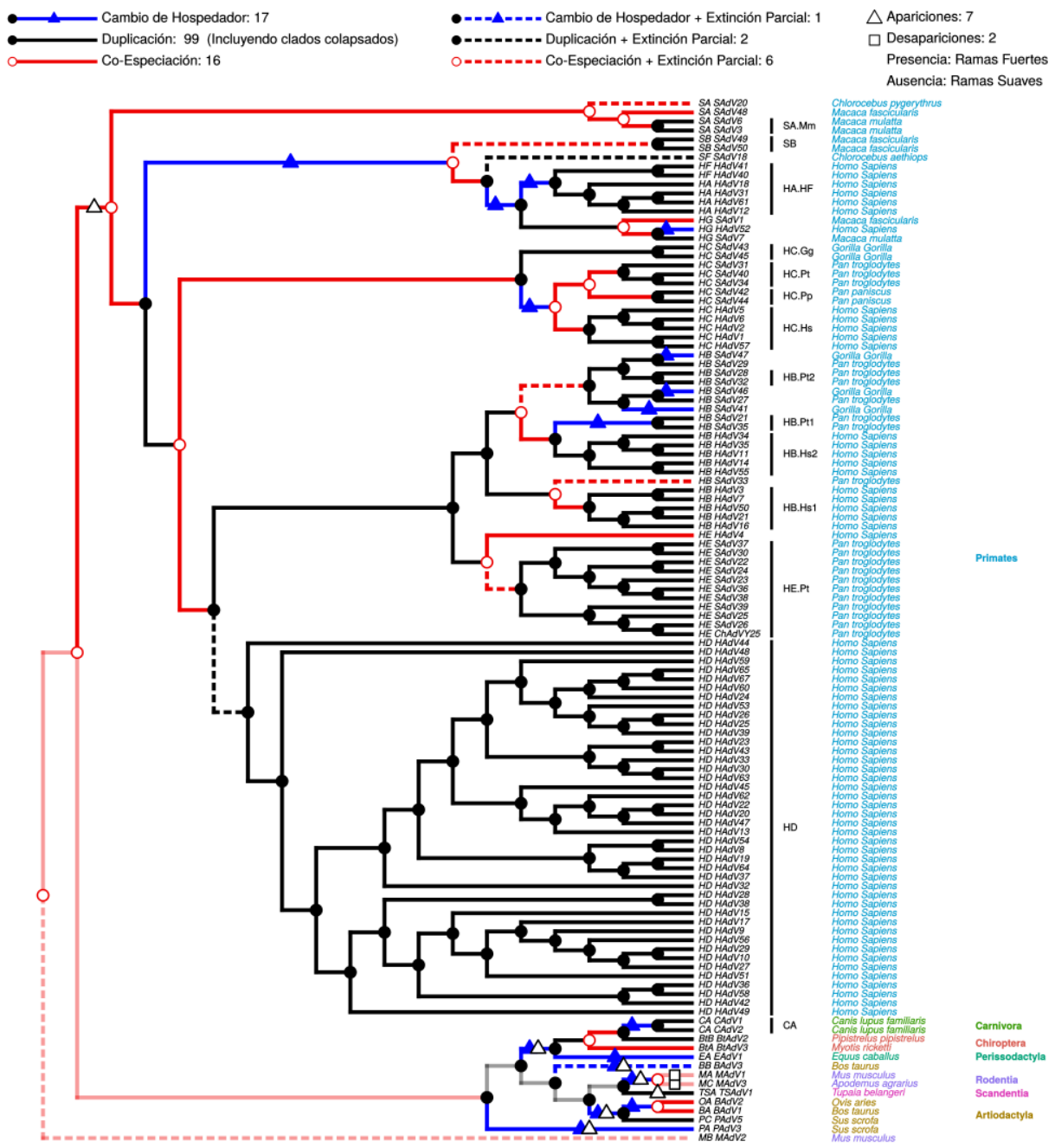
E. MYND


Cambio de Hospedador: 17
Duplicación: 99 (Incluyendo clados colapsados)
Co-Especiación: 16
Cambio de Hospedador + Extinción Parcial: 1
Duplicación + Extinción Parcial: 2
Co-Especiación + Extinción Parcial: 6
Apariciones: 0
Desapariciones: 3
Presencia: Ramas Fuertes
Ausencia: Ramas Suaves
SA.Mm
SB
HA.HF
HC.Gg
HC.Pt
HC.Pp
HC.Hs
HB.Pt2
HB.Pt1
HB.Hs2
HB.Hs1
HE.Pt
HD
CA
Primates
Carnivora
Chiroptera
Perissodactyla
Rodentia
Scandentia
Artiodactyla

F. LxCxE

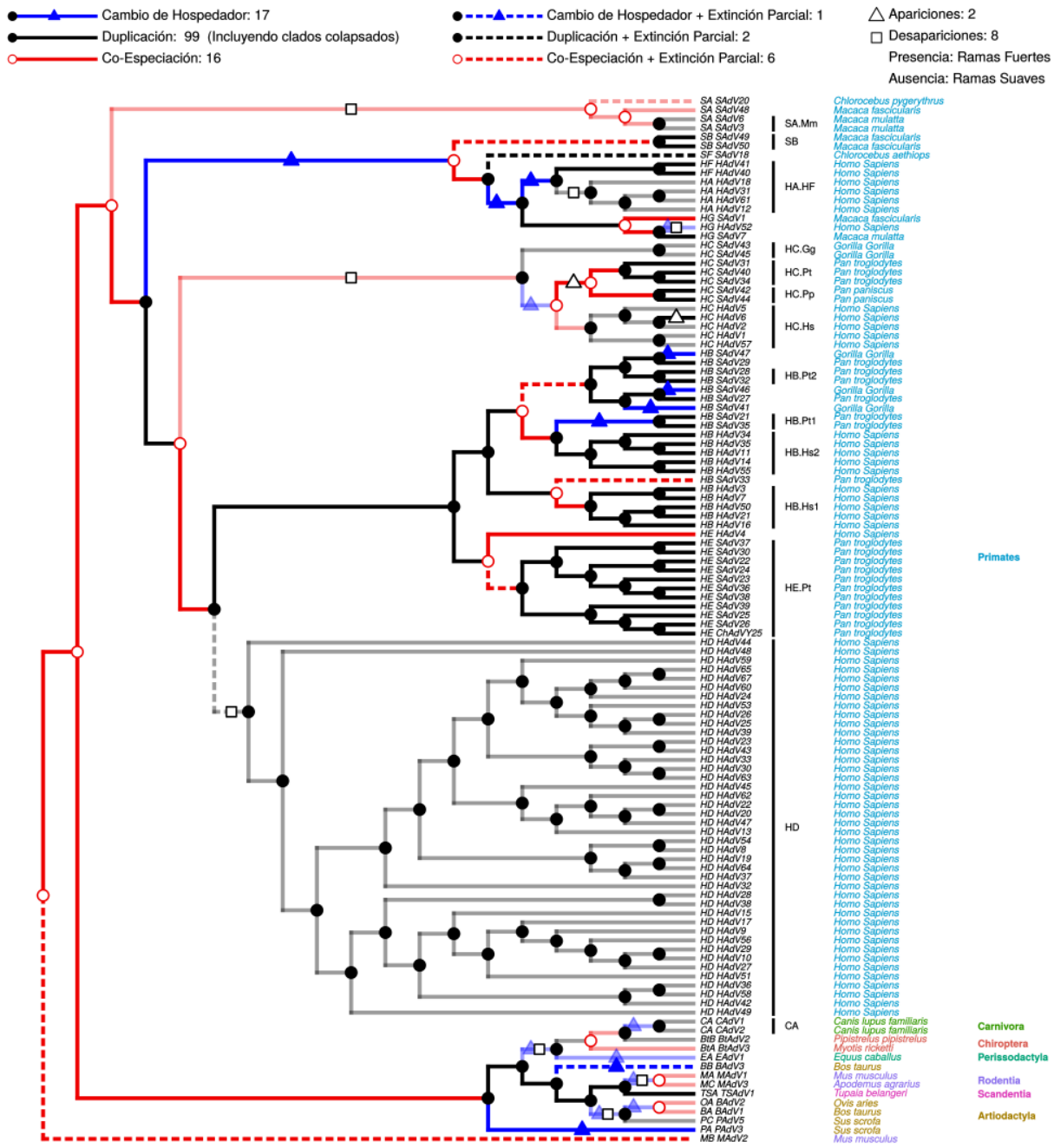


G. Región acídica


Cambio de Hospedador: 17
Duplicación: 99 (Incluyendo clados colapsados)
Co-Especiación: 16
Cambio de Hospedador + Extinción Parcial: 1
Duplicación + Extinción Parcial: 2
Co-Especiación + Extinción Parcial: 6
Apariciones: 7
Desapariciones: 2
Presencia: Ramas Fuertes
Ausencia: Ramas Suaves
SA.Mm
SB
HA.HF
HC.Gg
HC.Pt
HC.Pp
HC.Hs
HB.Pt2
HB.Pt1
HB.Hs2
HB.Hs1
HE.Pt
HD
CA
Primates
Carnivora
Chiroptera
Perissodactyla
Rodentia
Scandentia
Artiodactyla

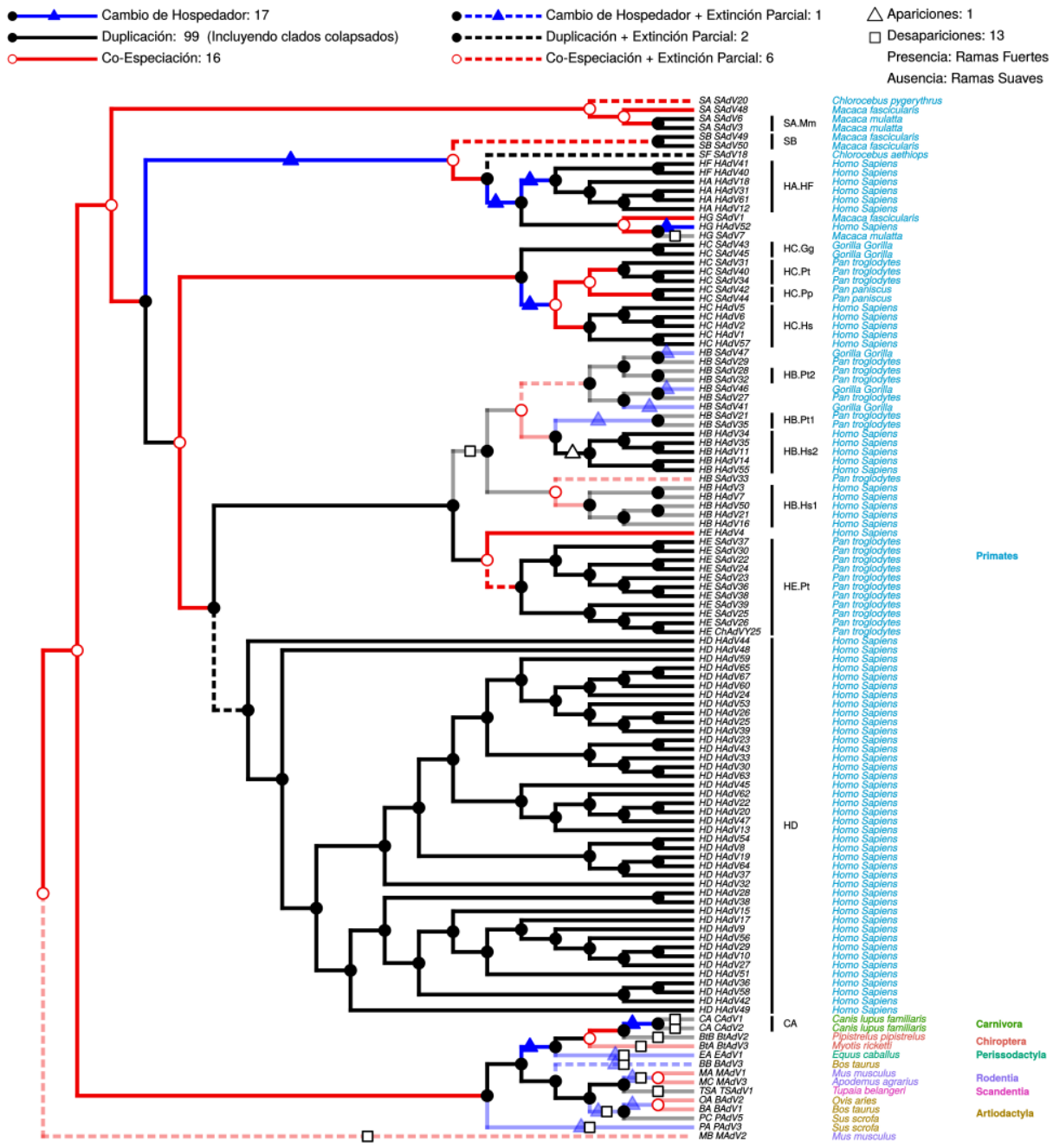
H. CKII

Superposición de eventos evolutivos con el mapeo de motivos lineales del dominio CR3

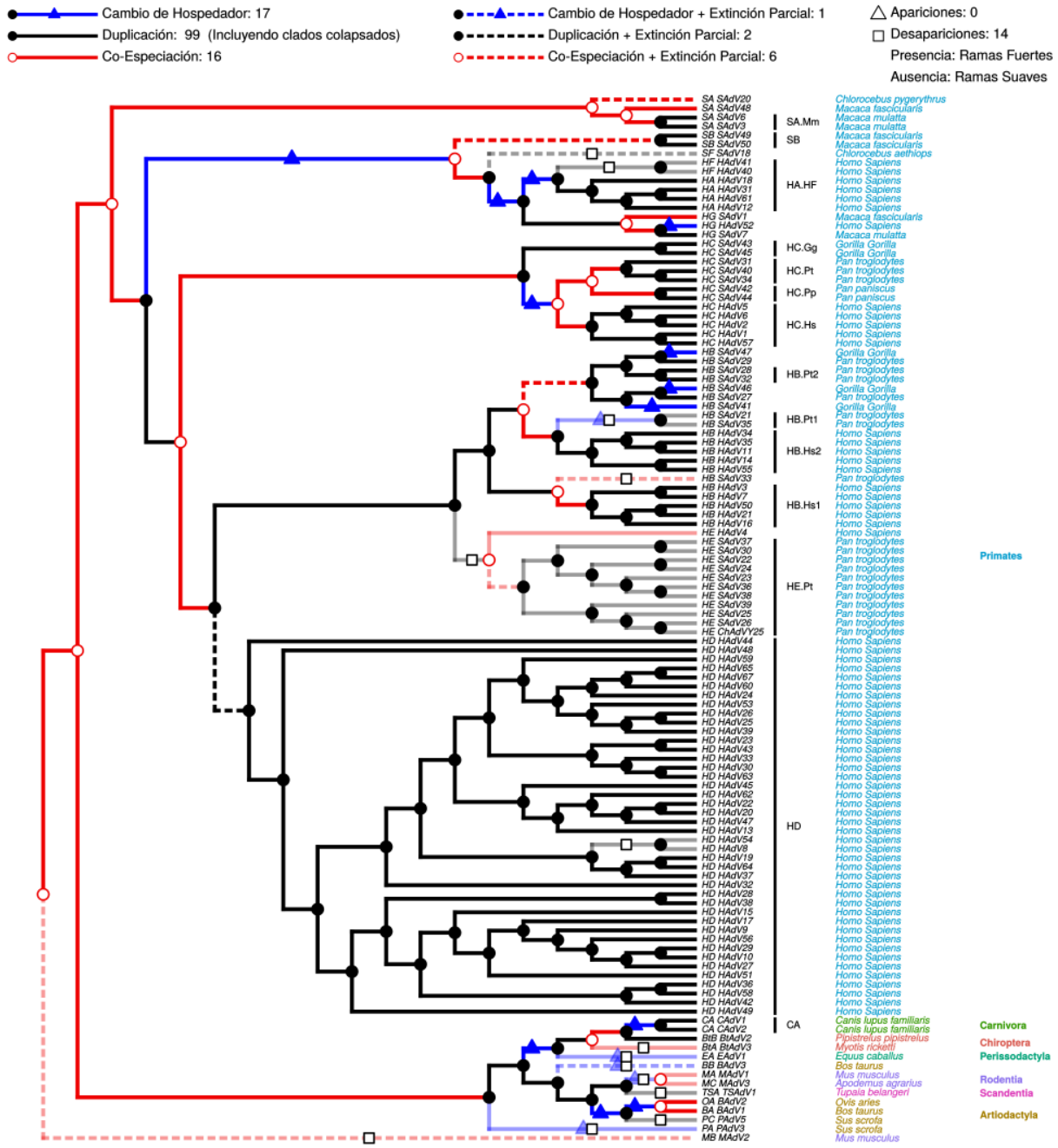


I. Posiciones ricas en cisteínas

Superposición de eventos evolutivos con el mapeo de motivos lineales del dominio CR4



J. NLS



K. CtBP

Figura 4.35: Superposición de eventos evolutivos y estados ancestrales de los motivos lineales de E1A. Se muestra la superposición de los eventos evolutivos y las apariciones y desapariciones de motivos en la historia evolutiva de *Mastadenovirus* a partir de la reconstrucción bayesiana de los estados ancestrales de los motivos de E1A. Para el dominio N-terminal se muestran los dos motivos: **(A)** IDMBR y **(B)** CoRNR Box. Para el dominio CR1 se muestran los motivos **(C)** pRb_ABGroove y **(D)** TRAM-CBP. Para el dominio CR2 se muestran los motivos **(E)** MYND, **(F)** LxCxE, **(G)** Región acídica y **(H)** CKII. Para el dominio CR3 se muestra el motivo **(I)** Posiciones ricas en cisteínas. Para el dominio CR4 se muestran los motivos **(J)** NLS y **(K)** CtBP. En las ramas se indica el estado de presencia (ramas fuertes) o ausencia (ramas suaves) del motivo, el evento de aparición (triángulo blanco) o desaparición (cuadrado blanco), y los cuatro eventos evolutivos: coespeciación (ramas rojas, nodo blanco con borde rojo), duplicación (ramas negras, nodo negro), cambio de hospedador (rama azul, triángulo azul) y extinción parcial (rama punteada). Las hojas del árbol corresponden a los serotipos actuales a los que pertenece la secuencia de E1A a partir de la cual se realizó la reconstrucción se secuencia. Luego se muestran la agrupación de clados, los hospedadores que infecta cada serotipo y los órdenes a los que pertenecen los hospedadores.

De esta manera se puede unificar la información de eventos evolutivos y apariciones o desapariciones de motivos por cada rama del árbol de *Mastadenovirus*.

Para definir si existe una asociación significativa entre los dos tipos de eventos que ocurren en una misma rama se realizó una prueba hipergeométrica (véase Sección 2.5.1). El valor p obtenido es finalmente corregido aplicando la corrección de Benjamini-Hochberg (valor p*) para comparaciones múltiples (Benjamini y Hochberg, 1995) (véase Sección 2.5.1). Aquellas combinaciones en las que el valor p* es menor a 0.05 son representadas con un punto en la Figura 4.36. En el eje x de la Figura 4.36 se indican los eventos de aparición y desaparición de cada motivo y en el eje y los eventos evolutivos.

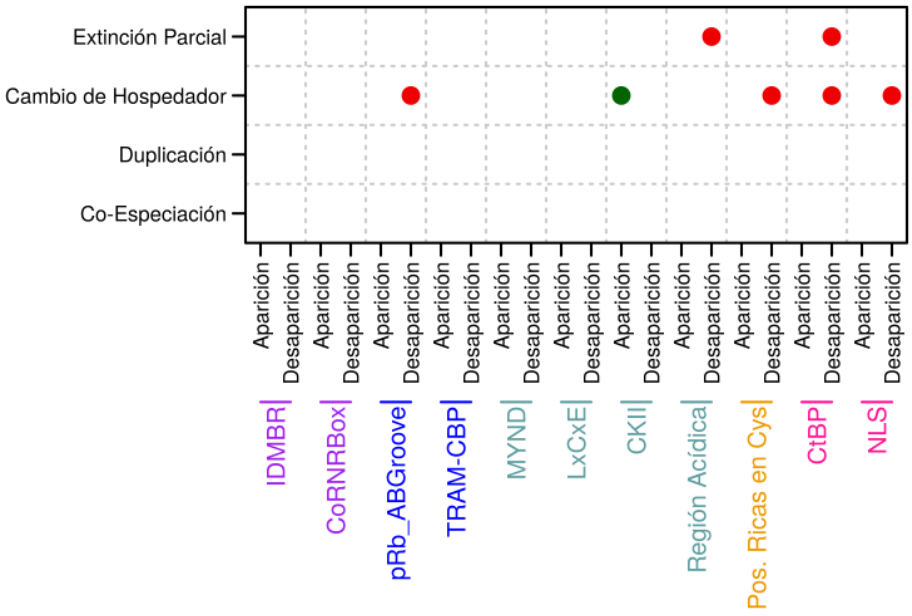


Figura 4.36: Asociaciones entre eventos de aparición y desaparición de motivos a lo largo de la filogenia de *Mastadenovirus* y eventos evolutivos. Las asociaciones positivas (valor $p^* < 0.05$) están marcadas con un punto rojo cuando ocurre entre un evento evolutivo y un evento de desaparición. Un punto verde cuando la asociación ocurre entre un evento evolutivo y un evento de aparición. El código de colores para las etiquetas es el utilizado en la Figura 4.2

El número de asociaciones posibles es 44. Sin embargo, sólo siete muestran una asociación significativa (valor $p^* < 0.05$). Cuatro desapariciones de motivos y dos apariciones muestran asociación con el evento evolutivo cambio de hospedador, y dos desapariciones con el evento evolutivo extinción parcial. Resultados similares se observan para la otra solución y la reconstrucción bayesiana, sin embargo, al utilizar la reconstrucción por parsimonia no se observaron asociaciones (véase Sección F.11). Esta diferencia en los resultados puede deberse a que para hacer el análisis fue necesario eliminar todas aquellas ramas que tenían eventos indefinidos.

Estos resultados sugieren que el cambio de hospedador representa una presión de selección frente a la funcionalidad del motivo favoreciendo la desaparición, o pérdida de funcionalidad, del mismo. Por lo tanto, los motivos lineales tuvieron un rol adaptativo en la evolución de *Mastadenovirus*.

Capítulo 5

Discusión general

En este trabajo de tesis se realizó un estudio abarcativo y sistemático de las proteínas E7 de papilomavirus y E1A de adenovirus. Considerando características de secuencias prototípicas se evaluó cuánto pueden ser extrapoladas a las proteínas aún no descriptas. Utilizando los datos obtenidos a nivel de secuencia y utilizando como modelo la proteína E1A se estudió la relación entre motivos y fenotipos y el rol adaptativo de los motivos en la historia evolutiva viral. Este capítulo está dedicado a la discusión de los resultados obtenidos.

5.1. Implicancias del trabajo con E7 para las predicciones en bioinformática

El avance en las técnicas de secuenciación de los últimos años llevó a la producción de una gran cantidad de información. El costo monetario y los tiempos de las técnicas de biología molecular no permiten el análisis sistemático de las variaciones observadas en el conjunto de secuencias. La bioinformática es un área interdisciplinaria que permite el análisis de grandes sistemas biológicos. Las herramientas informáticas permiten realizar un análisis de la información almacenada en grandes bases de datos de manera sistemática, a nivel secuencia, estructura, fenotipo y evolutivo. Esto permite estudiar aspectos particulares del sistema a analizar que permitan facilitar y orientar la investigación experimental. Por el contrario, en la virología molecular, en general los estudios se enfocan en aquellas proteínas que corresponden a serotipos prevalentes y/o clínicamente relevantes. El estudio en determinadas proteínas, es necesario y permite desarrollar conocimiento sobre el funcionamiento de las mismas, pero incorporar nuevas secuencias al análisis enriquece profundamente el entendimiento de la funcionalidad proteica, evolución viral y patogénesis. Tres de las observaciones realizadas mediante análisis bioinformáticos en esta tesis presentan hoy evidencia experimental.

A partir de una observación visual del alineamiento de la proteína E7 de papilomavirus realizada por el Dr. Alonso, se estudió en esta tesis y se determinó la presencia de posiciones ricas en cisteínas en el dominio globular E7C. Las cisteínas, además de estar normalmente involucradas en reacciones de catálisis y coordinación de metales, son un blanco importante para la regulación redox debido a la reactividad del grupo tiol. Dado que las células transformadas por HPV se encuentran en estrés oxidativo, el grupo de trabajo del Dr. Prat-Gay se propuso estudiar si las cisteínas no canónicas influían en el comportamiento redox de la proteína E7 bajo diferentes condiciones celulares (Camporeale *et al.*, 2017; Chemes *et al.*, 2014). Mediante experimentos de mutagénesis demostraron que el dominio E7C ejerce un efecto de protección en el dominio E7N regulando el estado oxidativo de la cisteína del motivo LxCxE y, por lo tanto, modula de manera indirecta la interacción con pRb. También observaron que la regulación redox por parte de la cisteínas no canónicas estaba involucrada en la formación del dímero de E7 (Chemes *et al.*, 2014). Por último, observaron que las posiciones ricas en cisteínas le permiten a E7 sensar especies reactivas del oxígeno (ROS) (en inglés, *reactive oxygen species*) en la célula infectada y regular la ubicación subcelular de la misma (Camporeale *et al.*, 2017).

Las posiciones ricas en cisteínas fueron también determinadas en el dominio CR3 de la proteína E1A. La abundancia relativa y el número total de posiciones ricas en cisteínas es menor en la proteína E1A, sugiriendo un rol menos importante en E1A. Sin embargo, algunos estudios muestran que E1A impide la respuesta al estrés oxidativo en células que la expresan (Orino *et al.*, 1999), sugiriendo que E1A podría sensar ROS al igual que E7. Experimentos son necesarios para probar la funcionalidad de las posiciones ricas de cisteínas en E1A.

Para la proteína E7 se reportaron cuatro posiciones altamente conservadas en el dominio E7C que se encuentran expuestas en la superficie (52, 66, 77 y 80 Figura 3.9, S naranja). Estudios

experimentales recientes demostraron que estos residuos participan en la interacción con diversas proteínas. La mutación del residuo 66 en la proteína E7 de HPV16 inhibe la asociación con la proteína p190 (Todorovic *et al.*, 2014) perteneciente a la familia de proteínas Rho activadoras de GTPasa. Esta interacción está conservada por lo menos en 13 secuencias distintas de E7 de tres especies diferentes de papilomavirus, sugiriendo un rol importante para la supervivencia viral. El residuo 80 en la proteína E7 de HPV16 participa en la interacción con la fosfatasa de tirosina citoplasmática (PTPN14) potencial supresora de tumores (Szalmás *et al.*, 2017). Análisis *in vitro* e *in vivo* demostraron que las cuatro posiciones contribuían a la interacción del dominio globular E7C de HPV16 con pRb en el bolsillo formado por los dominios A, B y C (Todorovic *et al.*, 2012). Estos experimentos son posteriores al análisis de conservación realizado en E7.

Por último, el grado de conservación en algunas de las posiciones del dominio desordenado E7N pueden estar relacionadas con la necesidad de impedir la formación de una estructura globular. Con el objetivo de evaluar esta hipótesis. el grupo de trabajo del Dr. Prat-Gay realizó experimentos de mutagénesis en el dominio E7N en el contexto de la proteína E7 de HPV16 y analizó la variabilidad conformacional (Borkosky *et al.*, 2017). Estos experimentos permitieron demostrar que los residuos más conservados en el dominio intrínsecamente desordenado estabilizan estructuras locales que se oponen a la formación de hélices α o agregación lenta de láminas β, ya que las mutaciones aumentaban el contenido de hélices α y aceleraban el agregado. El análisis teórico presentado en esta tesis y la evidencia experimental aportada por el grupo de trabajo del Dr. Prat-Gay se suman a la evidencia existente que sugiere que el conjunto de confórmeros observados para las proteínas intrínsecamente desordenadas es un subconjunto específico del espacio posible (Varadi *et al.*, 2014) y por lo tanto no ocurren al azar. Esto implica que la ausencia de estructura de una región intrínsecamente desordenada es tan sensible a las mutaciones como un dominio globular y sugiere que ambas estructuras están sujetas a presiones de selección similares (Borkosky *et al.*, 2017).

Estas tres comprobaciones experimentales son un claro ejemplo de la importancia del trabajo interdisciplinario y del rol de la bioinformática en la investigación. La biología molecular es necesaria para poder realizar estudios bioinformáticos y la bioinformática es un herramienta necesaria para la biología molecular.

5.2. Secuencias

En esta sección se discuten los resultados obtenidos a nivel de secuencia a la luz de las observaciones realizadas sobre los alineamientos, abundancia de motivos, la conservación a nivel posición y blancos proteicos.

5.2.1. Dominios y regiones de las proteínas E7 y E1A

La proteína E7 de papilomavirus está formada por un dominio intrínsecamente desordenado E7N, que contiene las regiones CR1 y CR2, y un dominio globular E7C (véase Sección 1.3.5 y Fi-

gura 3.12). La proteína E1A está formada por tres dominios intrínsecamente desordenados (CR1, CR2 y CR4) y dos dominios parcialmente ordenados (dominio N-terminal y CR3) (véase Sección 1.4.5 y Figura 4.11). Las regiones CR1 y CR2 presentan regiones con similitud de secuencia entre ambas proteínas, mientras que en el caso del CR3 no hay similitud de secuencia más allá de las cuatro cisteínas que coordinan el zinc.

Los estudios realizados en ambas proteínas incluyen unos pocos serotipos. Las bases de datos utilizadas en esta tesis son más abarcativas, actualizadas y representativas de la familia *Papillomaviridae* y el género *Mastadenovirus*. En relación a dominios y regiones, esto permitió realizar tres grandes observaciones. En la proteína E7 se identificaron cuatro regiones conectoras entre regiones conservadas. En la proteína E1A se redifinieron los límites de los dominios y se identificaron regiones entre dominios.

Los alineamientos múltiples de secuencia y los logos permitieron visualizar los dominios y regiones conservadas en cada proteína. La presencia de los dos dominios definidos para E7 (Figura 3.9) y los cinco dominios definidos para E1A (Figura 4.9) en la mayoría de las secuencias de nuestra base de datos demuestra la importancia funcional de los mismos.

Para la proteína E7 además se identificaron cuatro regiones conectoras entre regiones conservadas (Región conectora 1, 2, 3 y 4, Figura 3.6) que presentaron composición de aminoácidos y longitudes propias de cada una.

Para la proteína E1A además se identificaron tres regiones entre dominios (IDR12, IDR23 e IDR34, Figura 4.9). La región IDR23 parcialmente ordenada (Figura 4.9 y 4.11) que está involucrada en la actividad transformante de E1A (Doerfler y Böhm, 2004; Larsen y Tibbetts, 1987; Telling y Williams, 1994) se encuentra presente en un subgrupo específico de secuencias de E1A. Esto sugiere que regiones no conservadas en todas las secuencias podrían proveer evidencia clave para estudiar la oncogenicidad de adenovirus. Por otro lado, la región desordenada IDR12 que interactúa con algunas proteínas del hospedador y la región desordenada IDR34 que participa en la regulación transcripcional (Figura 4.9 y 4.11) de E1A sólo están presentes en un pequeño número de secuencias. Esto sugiere que algunas funciones moleculares de E1A podrían haber surgido como resultado de la ganancia o pérdida de ciertos dominios.

En la proteína E7, se observó que la región conectora entre los motivos LxCxE y el motivo CKII embebido en la región acídica presentaba una longitud restringida. Por otro lado, en la proteína E1A no se observó una región conectora entre estos motivos, confirmando que los tres motivos conforman un módulo funcional (Palopoli *et al.*, 2018).

5.2.2. Motivos lineales comunes a las proteínas E7 y E1A

Ambas proteínas compiten por la interacción con pRb, desplazando a E2F y permitiendo la entrada a la fase S del ciclo celular, induciendo así un ambiente propicio para la replicación viral. Esta interacción está mediada por cuatro motivos lineales presentes en ambas proteínas: el motivo pRb_ABGroove en la región CR1, el motivo LxCxE en la región CR2 y los motivos que lo siguen en secuencia, la región acídica y los sitios de fosforilación de CKII (Figura 3.9 y 4.9). Los tres

últimos motivos, LxCxE-CKII-Región Acídica, constituyen un módulo funcional (Palopoli *et al.*, 2018).

La interacción con pRb está modulada por diversas características de secuencia de E7 y E1A, como ser los residuos adyacentes al motivo LxCxE y sus motivos adyacentes. Experimentos de mutagénesis dirigida sugieren que la mutación del residuo 21 en un péptido derivado de E7 modula la afinidad por pRb (Singh *et al.*, 2005). La inspección de la estructura del complejo E7-LxCxE-pRb revela que el cuarto residuo hidrofóbico en la posición +2 del motivo, se une a un bolsillo formado por los residuos V725, F739, I752 e I753 de pRb (Kim *et al.*, 2001; Lee *et al.*, 1998; Palopoli *et al.*, 2018). Tanto en el motivo LxCxE de E7 como de E1A, existe un residuo aspártico precedente al motivo altamente conservado (posición 21 en E7 Figura 3.9 y posición 121 en E1A Figura 4.9). En ambas proteínas también se puede observar el residuo hidrofóbico conservado en la posición +2 del motivo en el caso de la proteína E7 (residuo 28 en E7, Figura 3.9) y en la proteína E1A, en la posición +3 del motivo (residuo 129, Figura 4.9). La configuración geométrica del bolsillo de pRb sería lo suficientemente amplia para permitir estas variaciones (Palopoli *et al.*, 2018). La región acídica y los sitios de fosforilación CKII también modulan la afinidad por pRb (Chemes y de Prat-Gay, 2010; Chemes *et al.*, 2011). Ambos motivos se encuentran conservados en ambas proteínas (Figura 3.9 y 4.9) y poseen características en común. La región acídica en ambas proteínas posee una carga neta que varía entre -4 y -6 (Figura 3.5 y 4.8B). Casi la mitad de las proteínas E7 y E1A poseen un único sitio CKII (Figura 3.7 y 4.8A), mientras que el resto poseen dos. Tanto la presencia de un aspártico y un cuarto residuo hidrofóbico en los residuos adyacentes al motivo como los motivos CKII y región acídica aumentan la afinidad por pRb. La presencia de estas características en dos proteínas no homólogas sugiere que la propiedad de mimetizar la unión a pRb con alta afinidad fue adquirida por evolución convergente en estas proteínas virales.

Para ambas proteínas se identificaron posiciones ricas en cisteínas, constituyendo un nuevo motivo en común. En el dominio globular de la proteína E7 existen dos grupos de cisteínas, cuatro cisteínas que coordinan el zinc que posee un rol estructural y diez posiciones ricas en cisteínas (Chemes *et al.*, 2012b) (véase Sección 3.2), que se encuentran en la superficie del homodímero y cercanas en estructura a las cisteínas que coordinan el zinc (Figura 3.8). En la región CR3 de la proteína E1A de adenovirus se observan dos grupos de cisteínas. Sin embargo, existen tres diferencias importantes respecto de la proteína E7. En primer lugar, el número de posiciones ricas en cisteínas identificado es mucho menor en E1A que en E7 (Figura 3.9 y 4.9). En segundo lugar, el número de cisteínas no canónicas por secuencia es también menor en E1A que en E7 (Figura 3.8 y 4.8C). Por último, la abundancia del motivo es menor en el caso de E1A que en E7 (Tabla 3.2 y 4.1). El grado de conservación de estas cisteínas en ambas proteínas sugieren que poseen un rol funcional importante. Sin embargo, las tres diferencias observadas entre la proteína E7 y E1A sugieren que el rol de las posiciones ricas en cisteínas es menos importante en E1A.

5.2.3. Abundancia de blancos proteicos en las proteínas E7 y E1A

Los virus son parásitos intracelulares obligados y sus genomas codifican para unas pocas proteínas. Por lo tanto, los virus dependen de la maquinaria celular del hospedador para su supervivencia. Los procesos celulares forman redes complejas de interacción proteína-proteína, y muchas de ellas comparten numerosas proteínas y están altamente interconectadas. Las proteínas centrales de estas redes son los blancos proteicos ideales para los virus ya que les permitiría ganar control sobre la mayoría de los procesos celulares. Sin embargo, la red de interacción entre proteínas virus-hospedador es compleja y aún poco comprendida. La proteína E7 posee un alto número de blancos proteicos reportados en la literatura (Figura 3.2) (Chemes *et al.*, 2012a) que pueden explicarse en parte a partir de los motivos lineales reportados. A su vez, algunos de los blancos proteicos, como pRb, interactúan con múltiples proteínas celulares, formando una compleja red de interacción. La actividad biológica de la proteína E1A involucra la formación de interacciones proteína-proteína (Pelka *et al.*, 2008). En esta tesis, se recolectaron más de 50 blancos proteicos para la proteína E1A de *Mastadenovirus* (Figura 4.2) y algunos de ellos también pueden ser explicados por la presencia de motivos lineales. Sin embargo, al igual que para la proteína E7, se desconoce si la interacción para muchos de los blancos proteicos está mediada o no por un motivo lineal.

5.2.4. Motivos lineales por descubrir

A nivel secuencia muchas posiciones conservadas pueden entenderse en términos de los motivos conocidos en ambas proteínas. En particular, para la proteína E1A se determinó que la conservación estaba relacionada con las posiciones fijas de los motivos lineales predichos (Figura 4.13). Esta observación se puede visualizar considerando las posiciones L, C y E en el motivo LxCxE, que corresponden a las posiciones 122, 124 y 126 en la proteína E1A (Figura 4.9). Estos residuos también se encuentran altamente conservados en la proteína E7 (posiciones 22, 24 y 26, Figura 3.9). Otras posiciones conservadas tienen algún rol biológico asignado, como por ejemplo las posiciones adyacentes al motivo LxCxE, que modulan la afinidad por pRb.

Otras posiciones poseen un alto nivel de conservación, pero se desconoce si poseen alguna actividad biológica, sugiriendo la existencia de motivos lineales aún no descubiertos. Por ejemplo, la posición conservada G3 en la proteína E7 no ha sido asignada a ningún motivo pero podría poseer algún rol en la poco entendida ubiquitinación del extremo N-terminal (Figura 3.9) (véase Sección 1.3.5). En la proteína E1A existe un alto grado de conservación para los residuos 231-278 del dominio CR4. Recientemente se reportó para esa región una posible señal de localización nuclear bipartita (Cohen *et al.*, 2014) que podría involucrar a las argininas 262 y 263 (Figura 4.9) y un sitio de unión para el factor de transcripción DREF (Radko *et al.*, 2014) apoyando la idea de que la alta conservación de secuencia puede estar relacionada con la presencia de motivos lineales aún no descubiertos.

5.3. Desorden y estructuras

En esta sección se discuten los resultados obtenidos a nivel de estructura a la luz de las predicciones de orden y desorden, coevolución y conservación de secuencia.

5.3.1. Conservación de secuencia y estructura

En general, está bien establecido en la literatura que dadas las restricciones estructurales que presentan los dominios globulares es esperable que exista un alto grado de conservación de secuencia. Por el contrario, dada la ausencia de una conformación definida en las proteínas o regiones intrínsecamente desordenadas la lógica sugiere que el grado de conservación de secuencia sea menor.

El grado de desorden intrínseco predicho para los distintos dominios de las proteínas varía poco dentro de las distintas secuencias de ambas proteínas (Figura 3.12 y 4.11) sugiriendo que el desorden (y el orden) es una propiedad estructural conservada de ambas proteínas. Sin embargo, contrario a lo esperado, el análisis de conservación de secuencia de ambas proteínas reveló que para ambas proteínas las regiones desordenadas, predichas o determinadas experimentalmente, estaban tan conservadas como la región globular (véase Sección 3.5 y 4.5).

El alto grado de conservación de secuencia observado en las regiones desordenadas puede explicarse debido a la presencia de motivos lineales y la conservación por fuera de los motivos lineales conocidos a motivos lineales aún no descubiertos, pero también podría explicarse considerando que la estructura (o la no estructura) de una proteína puede regular la funcionalidad o interacciones de la misma. En primer lugar, los estudios de co-evolución de secuencia revelaron para ambas proteínas contactos de largo alcance (entre dominios diferentes) y de corto alcance (dentro del mismo dominio) (Figura 3.16 y 4.15). Los resultados obtenidos en E1A muestran que los contactos predichos se desvían de la distribución esperada para una cadena completamente desordenada (Figura 4.16). Estos resultados, junto con la elevada conservación de secuencia, sugieren que algunos residuos podrían poseer un rol en la modulación de la asamblea conformacional y ser los responsables del desvío de una distribución al azar. Dado que en la proteína E1A muchas de las posiciones altamente conservadas están enriquecidas en aminoácidos que favorecen el desorden, como prolina (Figura 4.9), el alto grado de conservación observado podría, por ejemplo, estar relacionado con impedir la formación de una estructura globular.

Los estudios experimentales sobre E1A muestran que es necesaria la formación de un complejo ternario que incluye a la proteína CBP para inducir la proliferación celular a través de la interacción con pRb y el desplazamiento de E2F. La interacción entre CBP y pRb estimula la acetilación en la región C-terminal de pRb (Ferreon *et al.*, 2013; Wang *et al.*, 1995), promoviendo la degradación de pRb. Al igual que E1A, E7 también interactúa con CBP y pRb. En la proteína E1A los sitios de interacción están separados por aproximadamente 40 residuos y ambas proteínas pueden unirse de manera simultánea (Ferreon *et al.*, 2013), pero en E7 los sitios de interacción están superpuestos y ambas proteínas compiten por la interacción con E7 cuando se utiliza un péptido monomérico que

incluye la región CR1 y CR2 (Jansma *et al.*, 2014). A diferencia de E1A, la proteína E7 dimeriza a través del dominio E7C y posee dos dominios E7N, es decir, dos regiones CR1 y dos regiones CR2 permitiendo la formación de un complejo ternario entre E7, CBP y pRb (Jansma *et al.*, 2014). Esta evidencia, junto con la co-evolución de residuos observada entre los distintos dominios y regiones de las proteínas (Figura 3.16, 3.17 y 4.15) sugiere que la conservación de secuencia puede estar relacionada también con la formación de complejos proteicos que involucran distintos sitios de interacción (Figura 3.2 y 4.2) (Dyson y Wright, 2016; Ferreon *et al.*, 2013; Wang *et al.*, 1995).

En resumen, la conservación y coevolución de secuencia en las proteínas E1A y E7 refleja la presión evolutiva para mantener una conformación desordenada no azarosa y múltiples motivos lineales que en conjunto permiten la interacción con un elevado número de proteínas del hospedador (Figura 3.2 y 4.2).

5.3.2. Coevolución de secuencia y dominios

Aún no existe mucha evidencia que explique cómo la secuencia determina el grado de desorden en las IDPs/IDRs. Residuos conservados y que coevolucionan dentro de dominio globular se considera que cumplen una función estructural (Buslje *et al.*, 2009). Sin embargo, el rol estructural de los residuos en proteínas desordenadas aún está en proceso de comprensión (véase Sección 1.1.2).

Los estudios de coevolución en E1A y en E7 identificaron contactos de corto (dentro de un mismo dominio) y largo (entre dominios distintos) alcance (Figura 3.16, 3.17 y 4.15). Dentro de los contactos de corto alcance, en ambas proteínas se observó un alto porcentaje ocurriendo dentro de los dominios desordenados. Para la proteína E7, 67 % de los contactos predichos ocurren dentro del dominio desordenado, E7N, y 33 % en el dominio globular, E7C (véase Sección 3.8.2). Para la proteína E1A, 43 % de los contactos ocurren dentro de los dominios desordenados, CR1, CR2 y CR4, 22 % en las regiones parcialmente ordenadas, Dominio N e IDR23, y el 35 % restante ocurre dentro del dominio globular CR3 (véase Sección 4.9). Estos resultados se suman a la evidencia de que la distribución de conformaciones de una proteína desordenada dentro de las asamblea conformacional no es al azar (Varadi *et al.*, 2014). Dentro de los contactos de largo alcance, se observó para la proteína E1A un 27 % y para la proteína E7 un 19 %, sugiriendo que los dominios de ambas proteínas no son estructuras modulares independientes, sino que pueden presentar un acoplamiento funcional y evolución coordinada.

5.3.3. Regiones ordenadas

El análisis de las regiones ordenadas en términos de las propiedades de conservación de secuencia, estructural y perfil energético aportaron información sobre las funcionalidad de los dominios ordenados de las proteínas E1A y E7.

El análisis de co-evolución dentro del dominio globular de la proteína E7 permitió identificar pares de residuos que co-evolucionan y se encuentran a una distancia compatible con la formación de contactos en la estructura cristalográfica de E7C, dentro de la misma cadena y entre cadenas

del dímero (véase Sección 3.8 y Figura 3.16). Esta observación impulsó el uso de los datos de coevolución de secuencia para predecir un modelo estructural *de novo* para el dominio globular de E1A (Figura 4.17). El análisis de contactos (Figura 4.17), perfil energético y conservación de secuencia del modelo estructural predicho (Figura 4.18A y 4.18B) resultaron compatibles con el conocimiento actual estructural de proteínas globulares. En el modelo estructural propuesto se puede observar que el sitio de unión de factores de transcripción reportado en el CR3 (Figura 4.9) se encuentra altamente conservado pero a la vez pobremente plegado y frustrado energéticamente (Figura 4.17 y 4.18). Estos resultados sugieren que el sitio de unión para los factores de transcripción de la familia ATF, USF y Sp1 y los factores asociados a TBP, $TAF_{II}250$ y $TAF_{II}135/130$ es un módulo independiente dentro del dominio CR3 y probablemente un motivo lineal aún no reportado. Dada la alta heterogeneidad del dominio CR3 en solución (Hošek *et al.*, 2016) nuestro modelo estructural facilitará una futura búsqueda de una estructura experimental.

Por último, la representación de la conservación de secuencia en la estructura de E7C reveló un posible sitio de interacción para blancos celulares como p21 (Figura 3.14). Existen 35 proteínas cuya interacción con el dominio E7C está reportada en la literatura (Figura 3.2). Nuestro grupo de trabajo identificó utilizando dos algoritmos de predicción distintos un posible motivo lineal rico en serinas y prolinas que podría mediar la interacción (Chemes *et al.*, 2012a).

5.4. Evolución de motivos lineales

En esta sección se discuten los resultados obtenidos en relación a la evolución de los motivos lineales a la luz del estudio de la abundancia de los motivos lineales en las proteínas E1A y E7, las reconstrucciones de la historia evolutiva de los motivos lineales de la proteína E1A a lo largo de la filogenia de *Mastadenovirus* y las asociaciones observadas entre ellos a lo largo de la historia evolutiva y las asocaciones observadas con rasgos fenotípicos.

5.4.1. Secuencias actuales

La información recolectada para la proteína E7 permitió identificar siete motivos lineales, ampliar una definición y crear una nueva definición, mientras que para la proteína E1A se identificaron un total de diez motivos lineales conocidos y se definieron dos motivos lineales *de novo*. Ambas proteínas poseen cinco motivos lineales en común, los motivos pRb_ABGroove, LxCxE, CKII, la región acídica y las posiciones ricas en cisteínas. La presencia de estos motivos en estas proteínas poco relacionadas evolutivamente apoya la hipótesis de la evolución por convergencia de los motivos lineales.

Los dos motivos de la proteína E1A definidos *de novo*, la IDMBR y TRAM-CBP, mostraron una baja prevalencia. Esto podría deberse a que la evidencia experimental existente para ambos motivos está muy restringida debido a que hasta la fecha no se reportaron instancias del motivo en otras proteínas. Futuros experimentos permitirán definir si estos motivos propuestos están definidos correctamente.

Para ambas proteínas se observó que los motivos lineales conocidos para las secuencias prototípicas no son extrapolables a las secuencias no caracterizadas (Tabla 3.2 y 4.1). Dado que los motivos lineales de E7 y E1A participan en la unión directa a las proteínas del hospedador, el cambio en el repertorio de motivos podría llevar a cambios en las interacciones y por lo tanto a las funciones descriptas en la literatura. Para E1A, el repertorio de motivos predichos más común abarca solamente el 36 % de las secuencias (Figura 4.14) y no incluye a los serotipos más estudiados, indicando dos aspectos importantes. Por un lado, no se debería asumir que las características funcionales descubiertas en las secuencias más estudiadas son extrapolables a la mayoría de las proteínas homólogas. Por otro lado, es probable que existan diferencias funcionales en las proteínas esperando a ser descubiertas.

5.4.2. Historia evolutiva de los motivos lineales

La variabilidad observada en el repertorio de los motivos lineales (Figura 4.14) en las distintas secuencias de E1A demostró que la distribución de motivos no es igual para todas las especies de *Mastadenovirus*. Estas diferencias se pueden explicar a partir de que los motivos lineales de la proteína E1A cambian de manera aproximadamente independiente a lo largo de la evolución de *Mastadenovirus* (véase Sección 4.13).

El mapeo de la historia evolutiva en la filogenia de *Mastadenovirus* reveló que la aparición y desaparición de motivos ocurre tanto en las ramas profundas como en las ramas terminales del árbol. En particular, el motivo CKII posee un evento de aparición en una de las ramas profundas del árbol, y distintos eventos de aparición en las ramas menos profundas (Figura 4.22). La NLS tiene un evento de aparición luego de un evento de desaparición en una rama más profunda (Figura 4.22). Estos resultados apoyan la hipótesis de que los motivos lineales evolucionan por convergencia.

El número de eventos de apariciones y desapariciones no correlaciona con la abundancia actual del motivo o la aparición del motivo en las ramas más profundas del árbol (Figura 4.22 y 4.23). Por ejemplo, el motivo LxCxE presenta tres eventos de aparición/desaparición mientras que la región acídica presenta 21 eventos de aparición/desaparición, el motivo CKII presenta nueve eventos de aparición/desaparición y el pRb_ABGroove presenta cinco eventos de aparición/desaparición. Los cuatro motivos son altamente prevalentes y aparecen en las ramas más profundas del árbol, pero el número de eventos de aparición/desaparición es claramente diferente. Estos resultados sugieren que el repertorio de motivos lineales en la proteína E1A está dictado por presiones de selección específicas para cada motivo o pares de motivos.

El patrón evolutivo de cada motivo no depende del grado de desorden observado para cada dominio. Por ejemplo, de los motivos lineales que se encuentran en los dominios ordenados de E1A (N-terminal y CR3) el motivo CoRNR Box y las cisteínas poseen diez y nueve eventos de aparición/desaparición respectivamente. Por otro lado, los motivos LxCxE y CKII que se encuentran en el dominio CR2 desordenado presentan tres y nueve eventos de aparición/desaparición respectivamente. Estos resultados sugieren que el contexto estructural no es el principal determinante de la variación de motivos a lo largo de la evolución. Además, al igual que la prevalencia de los

motivos, el patrón evolutivo no depende de si los motivos están dentro del mismo dominio o no. Por ejemplo, el dominio CR2 contiene al motivo LxCxE que posee tres eventos de desaparición, y al motivo MYND que posee 29 eventos de aparición/desaparición. Por último, el estudio de la asociación entre motivos lineales de la proteína E1A reveló un alto número de asociaciones (Figura 4.26) entre motivos que pertenecen a distintos dominios y una asociación entre el motivo CKII y la región acídica que pertenecen ambos al dominio. Estos resultados sugieren que la presión de selección es ejercida sobre el motivo, o pares de motivos, y no sobre el dominio.

Este análisis permitió también reconstruir el repertorio de motivos en la proteína E1A del ancestro común a *Mastadenovirus*. La teórica proteína E1A ancestral poseía un motivo LxCxE, la región acídica, las posiciones ricas en cisteínas y los motivos CtBP y NLS. La teórica proteína E7 ancestral también presentaba un motivo LxCxE (Chemes *et al.*, 2012b) sugiriendo que las actividades ancestrales de ambas proteínas incluyen la interacción con pRb.

5.4.3. Asociación entre motivos lineales

Estudios previos realizados por la Dra. Chemes en la proteína E7 de papilomavirus (Chemes *et al.*, 2012b) mostraron una asociación entre el motivo LxCxE y el sitio de fosforilación CKII, y de este con la región acídica. Los estudios de asociación entre motivos revelaron en la proteína E1A una asociación entre el motivo pRb_ABGroove y los motivos LxCxE, CKII y la región acídica. Estos cuatros motivos median la interacción con pRb. Además, en la proteína E7 la distancia en secuencia entre el motivo LxCxE y el sitio CKII está altamente restringida (Figura 3.6) mientras que en E1A no se observa separación de secuencia entre ellos (Figura 4.9). Estos resultados en su conjunto se suman a la evidencia experimental que indica que los tres motivos LxCxE, CKII y la región acídica son un módulo funcional (Palopoli *et al.*, 2018).

El estudio de la asociación entre los eventos de aparición y desaparición de motivos a lo largo de la filogenia de *Mastadenovirus* reveló un alto número de asociaciones (Figura 4.29) entre las apariciones/desapariciones de los distintos motivos. De las 23 asociaciones identificadas entre los distintos motivos en las secuencias actuales, 17 también presentan asociaciones entre los eventos de aparición y desaparición. Estos resultados, junto con los resultados de coevolución de secuencia, pueden estar también relacionados con la formación de complejos proteicos que involucran distintos sitios de interacción (Figura 4.2) y sugieren que existe un acoplamiento funcional entre los motivos lineales. Además, indican que los distintos dominios no son independientes entre sí sino que funcionan y evolucionan de manera coordinada, resaltando la importancia de considerar a la proteína como un todo.

Actualmente, se cree que los motivos lineales evolucionan de manera rápida en comparación a la evolución de un dominio globular. Esta creencia se basa en que los motivos lineales consisten en unos pocos residuos funcionales que pueden ser modificados por mutaciones puntuales (véase Sección 1.2.3). Considerando que la aparición y desaparición de motivos lineales puede reflejarse como el cambio en el número de interacciones, los resultados obtenidos al analizar la tasa de cambio en el número de interacciones (véase Sección 4.14) concuerdan con esta observación.

5.4.4. Motivos lineales y rasgos fenotípicos

Estudios previos en la proteína E7 de papilomavirus (Chemes *et al.*, 2012b) mostraron una asociación significativa entre la presencia/ausencia de ciertos motivos y la especificidad del hospedador o el tipo de lesión, sugiriendo que algunos de los motivos lineales contribuyeron a la evolución adaptativa y cambios en el fenotipo de papilomavirus. Sin embargo, los motivos de la proteína E1A no parecen haber tenido un rol principal en cambios de tropismo (Figura 4.27). Esta diferencia puede deberse en primer lugar a que adenovirus puede infectar células epiteliales de un gran número de órganos produciendo diversas sintomatologías. En segundo lugar, la información clínica relacionada a la infección de adenovirus es mucho más compleja en comparación a la papilomavirus. Mientras que para papilomavirus una especie puede asociarse con determinada lesión, las distintas especies de adenovirus se asocian con un mayor número de tropismos y sintomatología. Por último, determinar la asociación entre una sintomatología y la presencia de un determinado serotipo de adenovirus presenta mayor dificultad ya que muchos adenovirus pueden persistir de manera subclínica en diversos órganos durante largos períodos de tiempo. Esto contrasta con las lesiones locales producidas por papilomavirus, que facilitan la asociación de un tejido con un serotipo viral.

El estudio de asociación entre la presencia/ausencia de un motivo lineal y la especificidad de hospedador reveló que en la proteína E1A se observa la asociación de numerosos motivos lineales con primates, humanos y no humanos (Figura 4.27), sugiriendo un rol importante de estos motivos en la evolución adaptativa que determinó la diversificación de *Mastadenovirus*.

5.5. Rol adaptativo de los motivos lineales en la evolución de *Mastadenovirus*

Los estudios de co-evolución entre *Mastadenovirus* y sus hospedadores indican que existe una co-divergencia entre ambos. Sin embargo, menos de la mitad de los eventos evolutivos determinados corresponden a eventos de co-especiación (véase Sección 4.16) y la diversidad de hospedadores de *Mastadenovirus* se explica incluyendo otros eventos como la duplicación viral dentro del mismo hospedador, la extinción parcial o la adaptación viral luego de un cambio de hospedador. Esta observación concuerda con estudios de co-evolución realizados recientemente entre secuencias de *Mastadenovirus* recolectadas de monos africanos y poblaciones humanas simpátricas (Hoppe *et al.*, 2015). Este estudio reveló que si bien la diversidad de hospedadores de la especie *Human adenovirus C* ocurre mayoritariamente por coespeciación con su hospedador, otros eventos evolutivos como cambio de hospedador y extinción parcial, contribuyeron a la diversidad de hospedadores de las especies *Human adenovirus B*, *D* y *E*.

La co-ocurrencia entre los eventos de desaparición de motivos y el evento evolutivo de cambio de hospedador, junto con la co-ocurrencia de presencia de motivos con hospedadores específicos sugiere que los motivos lineales juegan un rol adaptativo en la diversidad de *Mastadenovirus*. Estudios de co-ocurrencia realizados en la proteína E7 de papilomavirus revelaron que la ausencia

de determinados motivos estaba asociada con la infección de determinados órdenes de hospedador, por ejemplo, la ausencia del motivo LxCxE está asociada a la infección de miembros del orden Artiodactyla (Chemes *et al.*, 2012b). Es interesante que para la proteína E1A la asociación ocurra entre la pérdida de un motivo y el evento evolutivo que implica un cambio de hospedador. Esto puede deberse, en primer lugar, a la pérdida de las funciones conocidas de la proteína prototípica o a la ganancia de funciones aún no descubiertas. La mayoría de los estudios funcionales de la proteína E1A se realizan en los serotipos de mayor relevancia clínica que infectan a humanos. Por lo tanto, se desconoce a ciencia cierta si las funciones que cumple la proteína E1A en un nuevo hospedador son equivalentes a las identificadas en el humano. En segundo lugar, los experimentos en base a los cuales se crean las definiciones de motivos ocurren en un rango acotado de variabilidad de secuencias y más aún acotado respecto a la variabilidad de organismos. En la base de datos ELMdb el número total de instancias de motivos lineales es 3078, de las cuales 1852 corresponden a secuencias de proteínas humanas. El siguiente taxón más representado en la base de datos es *Mus musculus* con 290 secuencias. *Bos taurus*, que se encuentra dentro de los diez más representados, cuenta sólo con 25 instancias (véase `http://elm.eu.org/`, sección estadísticas). Por lo tanto, las definiciones generadas para los motivos lineales que existen están sesgadas. Es decir, que de existir diferencias en la definición de los residuos funcionales de los motivos lineales para cada organismo, la expresión regular utilizada para su búsqueda no es suficiente para su identificación. En tercer lugar, podría ser que el motivo esté presente, pero en otra de las proteínas del proteoma viral. Por ejemplo, en el caso de la proteína E7 el motivo CKII está ausente en la proteína E7 de Bovine papillomavirus 1 pero presente en la proteína E2 (Noval *et al.*, 2013). Ninguna de las tres observaciones realizadas invalidan el posible rol adaptativo de los motivos lineales en la diversificación de *Mastadenovirus*. En el primer caso, la pérdida de una funcionalidad conocida, y el correspondiente motivo lineal en la proteína E1A implica una adaptación a un nuevo ambiente. Por otro lado, la ganancia de una nueva funcionalidad aún no identificada en la proteína E1A sugiere a la vez la adquisición de un motivo lineal aún no conocido. En el segundo caso, una diferencia en la definición de los residuos funcionales del motivo lineal implica que el nuevo ambiente representa una presión de selección. La misma interacción mediada por el motivo lineal conocido y la posible nueva configuración son lo suficientemente diferentes como para no ser identificados. En el último caso, la presión de selección favorece la aparición del motivo lineal en una proteína diferente para que lleve a cabo su función en el nuevo hospedador. Sería interesante en el futuro estudiar la evolución de los motivos lineales utilizando en conjunto la filogenia y el proteoma viral para evaluar esta hipótesis.

5.6. Conclusiones generales

Los estudios sistemáticos y el uso de una base de datos abarcativa permitieron obtener resultados que parten de un nivel de secuencia, mediante la creación de alineamientos, logos y búsqueda de patrones, a un nivel de estructura, mediante la predicción de desorden y estudios de coevolución de secuencia, a un segundo nivel de secuencia mediante los estudios de asociaciones entre motivos.

Estos resultados pudieron luego conectarse con rasgos fenotípicos y patrones evolutivos.

El primer resultado es que las observaciones realizadas muestran que los estudios realizados en las proteínas virales prototípicas no son extrapolables a las proteínas homólogas.

En cuanto a la estructura general de las proteínas, los resultados obtenidos sugieren, en primer lugar, que las proteínas desordenadas presentan un grado de conservación similar al de las proteínas globulares. Estos resultados son opuesto a la creencia establecida en el área de estudio. Este grado de conservación puede explicarse principalmente desde dos aspectos principales. En primer lugar, la conservación de secuencia está fuertemente asociada a la presencia de módulos funcionales. En segundo lugar, la conservación de secuencia se debe a la necesidad de mantener una estructura desordenada pero no azarosa. Los estudios realizados también sugieren que los dominios desordenados y globulares no son módulos funcionales independientes entre sí, sino que presentan un acoplamiento funcional y evolutivo dado por la interacción entre ellos o la interacción por blancos proteicos comunes.

En cuanto a los motivos lineales, en primer lugar, los resultados obtenidos evidencian que no son módulos funcionales independientes, sino que muestran un acoplamiento funcional y evolutivo. En segundo lugar, las diferencias observadas entre la variabilidad de motivos y características de los mismos apoyan la visión de que los motivos lineales poseen características específicas asociadas a la proteína y organismo en estudio. En tercer lugar, los resultados obtenidos concuerdan con la idea establecida de que los motivos lineales son elementos que cambian con facilidad y también muestran que estos cambios están asociados a un rol adaptativo en la historia evolutiva.

Abreviaturas

ADN Ácido desoxirribonucleico.

AdV Adenovirus.

aLRT Tasa de probabilidad aproximada (en inglés, *Approximate Likelihood-Ratio Test*).

ARN Ácido ribonucleico.

ARNm ARN mensajero.

ARNr ARN ribosomal.

ARNt ARN de transferencia.

CAR Receptor de adenovirus y virus Coksackie B (en inglés, *Cocksackie B and Adenovirus Receptor*).

CBP Proteína de unión a la proteína CREB (en inglés, *CREB Binding Protein*).

CKII Quinasa de caseína II.

CoRNR Box Caja del receptor del corepresor nuclear (en inglés, *CoRepressor Nuclear Receptor box*).

CtBP Proteína de unión al extremo C-terminal (en inglés, C-terminal Binding Protein).

ELMdb Base de datos de motivos lineales eucariotas (en inglés, *Eukaryotic Linear Motif database*).
`http://elm.eu.org`.

IC Contenido de información.

ICTV Comité internacional de taxonomía de virus (*International Committe on Taxonomy of Viruses*)
`https://talk.ictvonline.org`.

IDMBR Región intrínsecamente desordenada de unión múltiple (*Intrinsically Disordered Multiple Binding Region*, en inglés).

IDP Proteína intrínsecamente desordenada (en inglés, *Intrinsically Disordered Protein*).

IDR Región intrínsecamente desordenada (en inglés, *Intrinsically Disordered Region*).

ITC Calorimetría de titulación isotérmica (en inglés, *Isothermal Titration Calorimetry*).

NCBI National Center for Biotechnology Information
`http://www.ncbi.nlm.nih.gov`.

NES Señal de exportación del núcleo (en inglés, *Nuclear Export Signal*).

NJ Unión de vecinos (en inglés, *Neighbor-Joining*).

NLS Señal de localización nuclear (en inglés, *Nuclear Localization Signal*).

NNI Intercambio de vecinos cercanos (NNIs) (en inglés, *Nearest Neighbor Interchanges*).

PDB Banco de datos de proteínas (en inglés, *Protein Data Bank*).
`https://www.rcsb.org/`.

pRb Proteína retinoblastoma.

PV Papilomavirus.

RMN Resonancia magnética nuclear.

RMSD Desviación de la raíz media cuadrática (RMSD) (en inglés, *Root Mean Square Deviation*).

SPR Movimientos topológicos de subárboles (SPR) (en inglés, *Subtree Pruning and Regrafting*).

TBP Proteína de unión a la región TATA (en inglés, *TATA Binding Protein*).

Bibliografía

Aksoy, P., Gottschalk, E. Y., y Meneses, P. I. (2017). *HPV entry into cells*. Mutation Research - Reviews in Mutation Research, 772:13–22.

Allende, J. E. y Allende, C. C. (1995). *Protein kinases. 4. Protein kinase CK2: an enzyme with multiple substrates and a puzzling regulation.* FASEB journal : official publication of the Federation of American Societies for Experimental Biology, 9(5):313–23.

Alonso, L. G., García-Alai, M. M., Nadra, A. D., Lapeña, A. N., Almeida, F. L., Gualfetti, P., y de Prat-Gay, G. (2002). *High-risk (HPV16) human papillomavirus E7 oncoprotein is highly stable and extended, with conformational transitions that could explain its multiple cellular binding partners.* Biochemistry, 41(33):10510–8.

Anfinsen, C. B. (1973). *Principles that govern the folding of protein chains.* Science, 181(4096):223–30.

Anisimova, M. y Gascuel, O. (2006). *Approximate likelihood ratio test for branchs: a fast, accurate and powerful alternative*. Systematic Biology, 55(4):539–552.

Ansieau, S. y Leutz, A. (2002). *The conserved Mynd domain of BS69 binds cellular and oncoviral proteins through a common PXLXP motif.* The Journal of Biological Chemistry, 277(7):4906–10.

Arnone, A., Bier, C., Cotton, F., y Day, V. (1971). *A high resolution structure of an inhibitor complex of the extracellular nuclease of Staphylococcus aureus*. The Journal of Biological Chemistry, 246(7):2302–2316.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., y Ben-Tal, N. (2016). *ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules*. Nucleic acids research, 44(W1):W344–W350.

Avvakumov, N., Kajon, a. E., Hoeben, R. C., y Mymryk, J. S. (2004). *Comprehensive sequence analysis of the E1A proteins of human and simian adenoviruses.* Virology, 329(2):477–92.

Avvakumov, N., Wheeler, R., D'Halluin, J. C., y Mymryk, J. S. (2002). *Comparative Sequence Analysis of the Largest E1A Proteins of Human and Simian Adenoviruses*. Journal of Virology, 76(16):7968–7975.

Balla, S., Thapar, V., Verma, S., Luong, T., Faghri, T., Huang, C. H., Rajasekaran, S., del Campo, J. J., Shinn, J. H., Mohler, W. A., Maciejewski, M. W., Gryk, M. R., Piccirillo, B., Schiller, S. R., y Schiller, M. R. (2006). *Minimotif Miner: a tool for investigating protein function.* Nature Methods, 3(3):175–177.

Barbosa, M. S., Edmonds, C., Fisher, C., Schiller, J. T., Lowy, D. R., y Vousden, K. H. (1990). *The region of the HPV E7 oncoprotein homologous to adenovirus E1a and Sv40 large T antigen contains separate domains for Rb binding and casein kinase II phosphorylation.* The EMBO journal, 9(1):153–60.

Bayley, S. T. y Mymryk, J. S. (1994). *Adenovirus e1a proteins and transformation (review).* International Journal of Oncology, 5(3):425–44.

Beltrao, P. y Serrano, L. (2007). *Specificity and evolvability in eukaryotic protein interaction networks.* PLoS Computational Biology, 3(2):e25.

Ben-Saadon, R., Fajerman, I., Ziv, T., Hellman, U., Schwartz, A. L., y Ciechanover, A. (2004). *The tumor suppressor protein p16INK4aand the human papillomavirus oncoprotein-58 E7 are naturally occurring lysine-less proteins that are degraded by the ubiquitin system: Direct evidence for ubiquitination at the N-terminal residue*. The Journal of Biological Chemistry, 279(40):41414–41421.

Benjamini, Y. y Hochberg, Y. (1995). *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing*. Journal of the Royal Statistical Society. Series B. Methodological, 57(1):289–300.

Benkö, M., Élo, P., Ursu, K., Ahne, W., Lapatra, S. E., Thomson, D., y Harrach, B. (2002). *First Molecular Evidence for the Existence of Distinct Fish and Snake Adenoviruses First Molecular Evidence for the Existence of Distinct Fish and Snake Adenoviruses*. Journal of Virology, 76(19):10056–10059.

Benkö, M. y Harrach, B. (2003). *Molecular evolution of adenoviruses.* Current Topics in Microbiology and Immunology, 272:3–35.

Bergelson, J. M., Cunningham, J. A., Droguett, G., Kurt-Jones, E. A., Krithivas, A., Hong, J. S., Horwitz, M. S., Crowell, R. L., y Finberg, R. W. (1997). *Isolation of a common receptor for Coxsackie B viruses and adenoviruses 2 and 5.* Science, 275(5304):1320–3.

Bergvall, M., Melendy, T., y Archambault, J. (2013). *The E1 proteins*. Virology, 445(1-2):35–56.

Berk, A. J. (2005). *Recent lessons in gene expression, cell cycle control, and cell biology from adenovirus.* Oncogene, 24(52):7673–85.

Berk, A. J. (2013). *Adenoviridae*. En Knipe, D. M. y Howley, P. M., editores, *Fields Virology*, capítulo Vol 2. 55, pages 1704–1731. Lippincott Williams & Wilkins, Philadelphia, PA, 6 edition.

Berk, A. J., Lee, F., Harrison, T., Williams, J., y Sharp, P. A. (1979). *Pre-early adenovirus 5 gene product regulates synthesis of early viral messenger RNAs*. Cell, 17(4):935–944.

Bernard, H.-U., Burk, R. D., Chen, Z., van Doorslaer, K., zur Hausen, H., y de Villiers, E.-M. (2010). *Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments.* Virology, 401(1):70–9.

Beuming, T., Skrabanek, L., Niv, M. Y., Mukherjee, P., y Weinstein, H. (2005). *PDZBase: A protein-protein interaction database for PDZ-domains*. Bioinformatics, 21(6):827–828.

Bode, W., Fehlhammer, H., y Huber, R. (1976). *Crystal structure of bovine trypsinogen at 1.8 Å resolution. I. Data collection, application of Patterson search techniques and preliminary structural interpretation*. Journal of Molecular Biology, 106(2):325–335.

Bondesson, M., Svensson, C., Linder, S., y Akusjärvi, G. (1992). *The carboxy-terminal exon of the adenovirus E1A protein is required for E4F-dependent transcription activation.* The EMBO journal, 11(9):3347–54.

Borcherds, W., Becker, A., Chen, L., Chen, J., Chemes, L. B., y Daughdrill, G. W. (2017). *Optimal Affinity Enhancement by a Conserved Flexible Linker Controls p53 Mimicry in MdmX*. Biophysical Journal, 112(10):2038–2042.

Borkosky, S. S., Camporeale, G., Chemes, L. B., Risso, M., Noval, M. G., Sánchez, I. E., Alonso, L. G., y de Prat-Gay, G. (2017). *Hidden Structural Codes in Protein Intrinsic Disorder*. Biochemistry, 56(41):5560–5569.

Bowles, N. E., Ni, J., Kearney, D. L., Pauschinger, M., Schultheiss, H. P., McCarthy, R., Hare, J., Bricker, J. T., Bowles, K. R., y Towbin, J. A. (2003). *Detection of viruses in myocardial tissues by polymerase chain reaction: Evidence of adenovirus as a common cause of myocarditis in children and adults*. Journal of the American College of Cardiology, 42(3):466–472.

Boyd, J. M., Loewenstein, P. M., Tang, Q.-q., Yu, L., y Green, M. (2002). *Adenovirus E1A N-terminal amino acid sequence requirements for repression of transcription in vitro and in vivo correlate with those required for E1A interference with TBP-TATA complex formation.* Journal of Virology, 76(3):1461–74.

Boyd, J. M., Subramanian, T., Schaeper, U., La Regina, M., Bayley, S. T., y Chinnadurai, G. (1993). *A region in the C-terminus of adenovirus 2/5 E1a protein is required for association with a cellular phosphoprotein and important for the negative modulation of T24-ras mediated transformation, tumorigenesis and metastasis.* The EMBO journal, 12(2):469–78.

Boyer, S. N., Wazer, D. E., y Band, V. (1996). *E7 protein of human papilloma virus-16 induces degradation of retinoblastoma protein through the ubiquitin-proteasome pathway.* Cancer Research, 56(20):4620–4.

Bravo, I. G., de Sanjosé, S., y Gottschling, M. (2010). *The clinical importance of understanding the evolution of papillomaviruses.* Trends in Microbiology, 18(10):432–8.

Brehm, A., Nielsen, S. J., Miska, E. A., McCance, D. J., Reid, J. L., Bannister, A. J., y Kouzarides, T. (1999). *The E7 oncoprotein associates with Mi2 and histone deacetylase activity to promote cell growth.* The EMBO journal, 18(9):2449–58.

Brown, C. J., Johnson, A. K., y Daughdrill, G. W. (2010). *Comparing Models of Evolution for Ordered and Disordered Proteins*. Molecular Biology and Evolution, 27(3):609–621.

Brown, C. J., Johnson, A. K., Dunker, A. K., y Daughdrill, G. W. (2011). *Evolution and disorder.* Current Opinion in Structural Biology, 21(3):441–6.

Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Green, E. D., Sidow, A., y Batzoglou, S. (2003). *LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA*. Genome Research, 13(4):721–31.

Bryngelson, J. D., Onuchic, J. N., Socci, N. D., y Wolynes, P. G. (1995). *Funnels, pathways, and the energy landscape of protein folding: a synthesis.* Proteins, 21(3):167–95.

Buck, C. B., Day, P. M., y Trus, B. L. (2013). *The papillomavirus major capsid protein L1*. Virology, 445(1-2):169–174.

Buneman, P. (1974). *A note on the metric properties of trees*. Journal of Combinatorial Theory, 17(1):48–50.

Burke, J. R., Deshong, A. J., Pelton, J. G., y Rubin, S. M. (2010). *Phosphorylation-induced conformational changes in the retinoblastoma protein inhibit E2F transactivation domain binding.* The Journal of Biological Chemistry, 285(21):16286–93.

Burke, J. R., Hura, G. L., y Rubin, S. M. (2012). *Structures of inactive retinoblastoma protein reveal multiple mechanisms for cell cycle control.* Genes & Development, 26(11):1156–66.

Burkhart, D. L. y Sage, J. (2008). *Cellular mechanisms of tumour suppression by the retinoblastoma gene*. Nature Reviews. Cancer, 8(9):671–682.

Buslje, C. M., Santos, J., Delfino, J. M., y Nielsen, M. (2009). *Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information.* Bioinformatics, 25(9):1125–31.

Camporeale, G., Lorenzo, J. R., Thomas, M. G., Salvatierra, E., Borkosky, S. S., Risso, M. G., Sánchez, I. E., de Prat-Gay, G., y Alonso, L. G. (2017). *Degenerate cysteine patterns mediate two redox sensing mechanisms in the papillomavirus E7 oncoprotein*. Redox Biology, 11(October 2016):38–50.

Carr, S. M., Munro, S., Zalmas, L.-P., Fedorov, O., Johansson, C., Krojer, T., Sagum, C. A., Bedford, M. T., Oppermann, U., y La Thangue, N. B. (2014). *Lysine methylation-dependent binding of 53BP1 to the pRb tumor suppressor*. Proceedings of the National Academy of Sciences of the United States of America, 111(31):11341–11346.

Castresana, J. (2000). *Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.* Molecular Biology and Evolution, 17(4):540–52.

Charleston, M. A. y Robertson, D. L. (2002). *Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny.* Systematic Biology, 51(3):528–535.

Chatton, B., Bocco, J. L., Gaire, M., Hauss, C., Reimund, B., Goetz, J., y Kedinger, C. (1993). *Transcriptional activation by the adenovirus larger E1a product is mediated by members of the cellular transcription factor ATF family which can directly associate with E1a.* Molecular and Cellular Biology, 13(1):561–70.

Chellappan, S., Kraus, V. B., Kroger, B., Munger, K., Howley, P. M., Phelps, W. C., y Nevins, J. R. (1992). *Adenovirus E1A, simian virus 40 tumor antigen, and human papillomavirus E7 protein share the capacity to disrupt the interaction between transcription factor E2F and the retinoblastoma gene product*. Biochemistry, 89(May):4549–4553.

Chemes, L. B., Camporeale, G., Sánchez, I. E., de Prat-Gay, G., y Alonso, L. G. (2014). *Cysteine-rich positions outside the structural zinc motif of human papillomavirus E7 provide conformational modulation and suggest functional redox roles.* Biochemistry, 53(10):1680–96.

Chemes, L. B. y de Prat-Gay, G. (2010). *La proteína supresora de tumores Retinoblastoma : caracterización de su dominio AB y mecanismo de interacción con la oncoproteína E7 del papilomavirus humano*. PhD thesis.

Chemes, L. B., de Prat-Gay, G., y Sánchez, I. E. (2015). *Convergent evolution and mimicry of protein linear motifs in host-pathogen interactions.* Current Opinion in Structural Biology, 32:91–101.

Chemes, L. B., Glavina, J., Alonso, L. G., Marino-Buslje, C., de Prat-Gay, G., y Sánchez, I. E. (2012a). *Sequence evolution of the intrinsically disordered and globular domains of a model viral oncoprotein.* PLoS One, 7(10):e47661.

Chemes, L. B., Glavina, J., Faivovich, J., de Prat-Gay, G., y Sánchez, I. E. (2012b). *Evolution of linear motifs within the papillomavirus E7 oncoprotein.* Journal of molecular biology, 422(3):336–46.

Chemes, L. B., Sánchez, I. E., y de Prat-Gay, G. (2011). *Kinetic recognition of the retinoblastoma tumor suppressor by a specific protein target.* Journal of molecular biology, 412(2):267–84.

Chemes, L. B., Sánchez, I. E., Smal, C., y de Prat-Gay, G. (2010). *Targeting mechanism of the retinoblastoma tumor suppressor by a prototypical viral oncoprotein. Structural modularity, intrinsic disorder and phosphorylation of human papillomavirus E7.* The FEBS Journal, 277(4):973–88.

Cheng, R. R., Morcos, F., Levine, H., y Onuchic, J. N. (2014). *Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information.* Proceedings of the National Academy of Sciences of the United States of America, 111:E563–71.

Classon, M. y Dyson, N. J. (2001). *p107 and p130: Versatile proteins with interesting pockets*. Experimental Cell Research, 264(1):135–147.

Classon, M. y Harlow, E. (2002). *The retinoblastoma tumour suppressor in development and cancer*. Nature Reviews. Cancer, 2(12):910–917.

Clementi, C., Nymeyer, H., y Onuchic, J. N. (2000). *Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins.* Journal of molecular biology, 298(5):937–53.

Cobrinik, D. (2005). *Pocket proteins and cell cycle control*. Oncogene, 24(17):2796–2809.

Cohen, M. J., King, C. R., Dikeakos, J. D., y Mymryk, J. S. (2014). *Functional analysis of the C-terminal region of human adenovirus E1A reveals a misidentified nuclear localization signal.* Virology, 468-470C:238–243.

Cohen, M. J., Yousef, A. F., Massimi, P., Fonseca, G. J., Todorovic, B., Pelka, P., Turnell, A. S., Banks, L., y Mymryk, J. S. (2013). *Dissection of the C-terminal region of E1A redefines the roles of CtBP and other cellular targets in oncogenic transformation.* Journal of Virology, 87(18):10348–55.

Conow, C., Fielder, D., Ovadia, Y., y Libeskind-Hadas, R. (2010). *Jane: a new tool for the cophylogeny reconstruction problem.* Algorithms for Molecular Biology, 5:16.

Consortium, T. U. (2017). *UniProt: The universal protein knowledgebase*. Nucleic Acids Research, 45(D1):D158–D169.

Corbeil, H. B. y Branton, P. E. (1994). *Functional importance of complex formation between the retinoblastoma tumor suppressor family and adenovirus E1A proteins as determined by mutational analysis of E1A conserved region 2.* Journal of Virology, 68(10):6697–709.

Crooks, G. E., Hon, G., Chandonia, J.-m., y Brenner, S. E. (2004). *WebLogo: a sequence logo generator.* Genome Research, 14(6):1188–90.

Crosbie, E. J., Einstein, M. H., Franceschi, S., y Kitchener, H. C. (2013). *Human papillomavirus and cervical cancer*. The Lancet, 382(9895):889–899.

Cubie, H. (2013). *Diseases associated with human papillomavirus infection.* Virology, 445(1-2):21–34.

Culp, J. S., Webster, L. C., Friedman, D. J., Smith, C. L., Huang, W.-j., Wu, F. Y., Rosenberg, M., y Ricciardi, R. P. (1988). *The 289-amino acid E1A protein of adenovirus binds zinc in a region that is important for trans-activation.* Proceedings of the National Academy of Sciences of the United States of America, 85(17):6450–4.

Cuthill, J. H. y Charleston, M. A. (2012). *Phylogenetic codivergence supports coevolution of mimetic Heliconius butterflies.* PLoS One, 7(5):e36464.

Dai, X., Wu, L., Sun, R., y Zhou, Z. H. (2017). *Atomic Structures of Minor Proteins VI and VII in the Human Adenovirus.* Journal of Virology, 91(24):1–15.

Daughdrill, G. W., Borcherds, W. M., y Wu, H. (2011). *Disorder predictors also predict backbone dynamics for a family of disordered proteins*. PLoS One, 6(12):e29207.

Davey, N. E., Cyert, M. S., y Moses, A. M. (2015). *Short linear motifs – ex nihilo evolution of protein regulation*. Cell Communication and Signaling, 13(1):43.

Davey, N. E., Haslam, N. J., Shields, D. C., y Edwards, R. J. (2011a). *SLiMSearch 2.0: Biological context for short linear motifs in proteins*. Nucleic Acids Research, 39(SUPPL. 2):56–60.

Davey, N. E., Travé, G., y Gibson, T. J. (2011b). *How viruses hijack cell regulation.* Trends in Biochemical Sciences, 36(3):159–69.

Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., y Gibson, T. J. (2012). *Attributes of short linear motifs.* Molecular BioSystems, 8(1):268–81.

Davison, A. J., Benkö, M., y Harrach, B. (2003). *Genetic content and evolution of adenoviruses*. Journal of General Virology, 84(11):2895–2908.

de Castro, E., Sigrist, C. J. A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., y Hulo, N. (2006). *ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins*. Nucleic Acids Research, 34(Web Server issue):362–365.

de Villiers, E.-M., Fauquet, C., Broker, T. R., Bernard, H.-U., y zur Hausen, H. (2004). *Classification of papillomaviruses.* Virology, 324(1):17–27.

DeCaprio, J. A., Ludlow, J. W., Figge, J., Shew, J.-Y., Huang, C.-M., Lee, W.-H., Marsilio, E., Paucha, E., y Livingston, D. M. (1988). *SV40 large tumor antigen forms a specific complex with the product of the retinoblastoma susceptibility gene.* Cell, 54(2):275–83.

DeGregori, J., Kowalik, T., y Nevins, J. R. (1995). *Cellular targets for activation by the E2F1 transcription factor include DNA synthesis- and G1/S-regulatory genes.* Molecular and Cellular Biology, 15(8):4215–24.

Delston, R. B., Matatall, K. A., Sun, Y., Onken, M. D., y Harbour, J. W. (2011). *p38 phosphorylates Rb on Ser567 by a novel, cell cycle-independent mechanism that triggers Rb-Hdm2 interaction and apoptosis.* Oncogene, 30(5):588–99.

Desdevises, Y. (2007). *Cophylogeny: insights from fish-parasite systems.* Parassitologia, 49(3):125–8.

Dick, F. A. (2007). *Structure-function analysis of the retinoblastoma tumor suppressor protein - Is the whole a sum of its parts?* Cell Division, 2:1–15.

Dick, F. A., Goodrich, D. W., Sage, J., y Dyson, N. J. (2018). *Non-canonical functions of the RB protein in cancer.* Nature Reviews. Cancer, 18(7):442–451.

Dick, F. A. y Rubin, S. M. (2013). *Molecular mechanisms underlying RB protein function*. Nature Reviews. Molecular Cell Biology, 14(5):297–306.

Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N. P., Trave, G., y Gibson, T. J. (2008). *Understanding eukaryotic linear motifs and their role in cell signaling and regulation.* Frontiers in Bioscience, 13(1):6580–603.

DiMaio, D. y Petti, L. M. (2013). *The E5 proteins*. Virology, 445(1-2):99–114.

Dingle, J. H. y Langmuir, A. D. (1968). *Epidemiology of acute, respiratory disease in military recruits.* The American Review of Respiratory Disease, 97(6):Suppl:1–65.

Dingwall, C. y Laskey, R. A. (1991). *Nuclear targeting sequences–a consensus?* Trends in biochemical sciences, 16(12):478–81.

Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., y Diella, F. (2011). *Phospho.ELM: A database of phosphorylation sites-update 2011*. Nucleic Acids Research, 39(SUPPL. 1):261–267.

Dinkel, H. y Sticht, H. (2007). *A computational strategy for the prediction of functional linear peptide motifs in proteins.* Bioinformatics, 23(24):3297–303.

Dinkel, H., Van Roey, K., Michael, S., Davey, N. E., Weatheritt, R. J., Born, D., Speck, T., Krüger, D., Grebnev, G., Kuban, M., Strumillo, M., Uyar, B., Budd, A., Altenberg, B., Seiler, M., Chemes, L. B., Glavina, J., Sánchez, I. E., Diella, F., y Gibson, T. J. (2014). *The eukaryotic linear motif resource ELM: 10 years and counting.* Nucleic acids research, 42(Database issue):D259–66.

Doerfler, W. y Böhm, P. (2004). *Adenoviruses: Model and Vectors in Virus-Host Interactions*, volume 273 of *Current Topics in Microbiology and Immunology*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1 edition.

Doolittle, R. F. (1973). *Structural aspects of the fibrinogen to fibrin conversion.* Advances in Protein Chemistry, 27:1–109.

Doorbar, J. (2013). *The E4 protein; structure, function and patterns of expression.* Virology, 445(1-2):80–98.

Dosztányi, Z. (2018). *Prediction of protein disorder based on IUPred.* Protein Science, 27(1):331–340.

Dosztányi, Z., Csizmók, V., Tompa, P., y Simon, I. (2005a). *IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.* Bioinformatics, 21(16):3433–3434.

Dosztányi, Z., Csizmók, V., Tompa, P., y Simon, I. (2005b). *The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins.* Journal of Molecular Biology, 347(4):827–839.

Dryden, D. T., Thomson, A. R., y White, J. H. (2008). *How much of protein sequence space has been explored by life on Earth?* Journal of the Royal Society Interface, 5(25):953–956.

Duffy, S., Shackelton, L. A., y Holmes, E. C. (2008). *Rates of evolutionary change in viruses: patterns and determinants.* Nature Reviews. Genetics, 9(4):267–276.

Dunker, A. K., Babu, M. M., Barbar, E., Blackledge, M., Bondos, S. E., Dosztányi, Z., Dyson, H. J., Forman-Kay, J. D., Fuxreiter, M., Gsponer, J., Han, K.-H., Jones, D. T., Longhi, S., Metallo, S. J., Nishikawa, K., Nussinov, R., Obradovic, Z., Pappu, R. V., Rost, B., Selenko, P., Subramaniam, V., Sussman, J. L., Tompa, P., y Uversky, V. N. (2013). *What's in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered.* Intrinsically Disordered Proteins, 1(1):e24157.

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., y Obradovic, Z. (2002). *Intrinsic Disorder and Protein Function.* Proteins, 41(21):6573–6582.

Dunker, A. K., Oldfield, C. J., Meng, J., Romero, P., Yang, J. Y., Chen, J. W., Vacic, V., Obradovic, Z., y Uversky, V. N. (2008). *The unfoldomics decade: an update on intrinsically disordered proteins.* BMC Genomics, 9 Suppl 2:S1.

Dunn, S. D., Wahl, L. M., y Gloor, G. B. (2008). *Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.* Bioinformatics, 24(3):333–340.

Dyer, M. D., Murali, T. M., y Sobral, B. W. (2008). *The landscape of human proteins interacting with viruses and other pathogens.* PLoS Pathogens, 4(2):e32.

Dyson, H. J. y Wright, P. E. (2005). *Intrinsically unstructured proteins and their functions.* Nature Reviews. Molecular Cell Biology, 6(3):197–208.

Dyson, H. J. y Wright, P. E. (2016). *Role of intrinsic protein disorder in the function and interactions of the transcriptional coactivators CREB-binding Protein (CBP) and p300.* The Journal of Biological Chemistry, 291(13):6714–6722.

Dyson, N. J., Guida, P., McCall, C., y Harlow, E. (1992a). *Adenovirus E1A makes two distinct contacts with the retinoblastoma protein.* Journal of Virology, 66(7):4606–11.

Dyson, N. J., Guida, P., Münger, K., y Harlow, E. (1992b). *Homologous sequences in adenovirus E1A and human papillomavirus E7 proteins mediate interaction with the same set of cellular proteins.* Journal of Virology, 66(12):6893–902.

Dyson, N. J., Howley, P. M., Münger, K., y Harlow, E. (1989). *The Human Papilloma Virus-16E7 Oncoprotein is Able to Bind to the Retinoblastoma Gene Product*. Science, 243:934–936.

Edgar, R. C. (2004). *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic acids research, 32(5):1792–7.

Edwards, R. J., Davey, N. E., y Shields, D. C. (2007). *SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins.* PLoS One, 2(10):e967.

Edwards, R. J. y Palopoli, N. (2015). *Computational prediction of short linear motifs from protein sequences.* Methods in molecular biology (Clifton, N.J.), 1268:89–141.

Eliezer, D. (2009). *Biophysical characterization of intrinsically disordered proteins*. Current Opinion in Structural Biology, 19(1):23–30.

Espada, R., Parra, R. G., Mora, T., Walczak, A. M., y Ferreiro, D. U. (2015). *Capturing coevolutionary signals inrepeat proteins.* BMC Bioinformatics, 16(1):207.

Felsani, A., Mileo, A. M., y Paggi, M. G. (2006). *Retinoblastoma family proteins as key targets of the small DNA virus oncoproteins.* Oncogene, 25(38):5277–85.

Felsenstein, J. (1981). *Evolutionary trees from DNA sequences: a maximum likelihood approach.* Journal of Molecular Evolution, 17(6):368–76.

Ferreiro, D. U., Hegler, J. A., Komives, E. A., y Wolynes, P. G. (2007). *Localizing frustration in native proteins and protein assemblies.* Proceedings of the National Academy of Sciences of the United States of America, 104(50):19819–24.

Ferreon, A. C. M., Ferreon, J. C., Wright, P. E., y Deniz, A. a. (2013). *Modulation of allostery by protein intrinsic disorder.* Nature, 498(7454):390–4.

Ferreon, J. C., Martinez-Yamout, M. A., Dyson, H. J., y Wright, P. E. (2009). *Structural basis for subversion of cellular control mechanisms by the adenoviral E1A oncoprotein.* Proceedings of the National Academy of Sciences of the United States of America, 106(32):13260–5.

Firzlaff, J. M., Galloway, D. A., Eisenman, R. N., y Lüscher, B. (1989). *The E7 protein of human papillomavirus type 16 is phosphorylated by casein kinase II.* The New Biologist, 1(1):44–53.

Fitch, W. M. (1971). *Toward defining the course of evolution: Minimum change for a specific tree topology*. Systematic Biology, 20:406–416.

Flint, J. y Shenk, T. (1997). *Viral transactivating proteins.* Annual Review of Genetics, 31:177–212.

Fontes, M. R. M., Teh, T., Jan, D., Brinkworth, R. I., y Kobe, B. (2003). *Structural basis for the specificity of bipartite nuclear localization sequence binding by importin-α*. The Journal of Biological Chemistry, 278(30):27981–27987.

Forman-Kay, J. D. y Mittag, T. (2013). *From sequence and forces to structure, function, and evolution of intrinsically disordered proteins*. Structure, 21(9):1492–1499.

Fox, J. P., Brandt, C. D., Wassermann, F. E., Hall, C. E., Spigland, I., Kogon, A., y Elveback, L. R. (1969). *The virus watch program: a continuing surveillance of viral infections in metropolitan New York families. VI. Observations of adenovirus infections: virus excretion patterns, antibody response, efficiency of surveillance, patterns of infections, and relation to illness*. American Journal of Epidemiology, 89(1):25–50.

Fred Dice, J. (1990). *Peptide sequences that target cytosolic proteins for lysosomal proteolysis*. Trends in Biochemical Sciences, 15(8):305–309.

Friend, S. H., Bernards, R., Rogelj, S., Weinberg, R. A., Rapaport, J. M., Albert, D. M., y Dryja, T. P. (1986). *A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma.* Nature, 323(6089):643–6.

Fukuchi, S., Amemiya, T., Sakamoto, S., Nobe, Y., Hosoda, K., Kado, Y., Murakami, S. D., Koike, R., Hiroaki, H., y Ota, M. (2014). *IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners*. Nucleic Acids Research, 42(D1):320–325.

Gaggar, A., Shayakhmetov, D. M., y Lieber, A. (2003). *CD46 is a cellular receptor for group B adenoviruses*. Nature Medicine, 9(11):1408–1412.

Gallimore, P. H. y Turnell, A. S. (2001). *Adenovirus E1A: remodelling the host cell, a life or death experience.* Oncogene, 20(54):7824–35.

García-Alai, M. M., Alonso, L. G., y de Prat-Gay, G. (2007). *The N-terminal module of HPV16 E7 is an intrinsically disordered domain that confers conformational and recognition plasticity to the oncoprotein.* Biochemistry, 46(37):10405–12.

García-Alai, M. M., Gallo, M., Salame, M., Wetzler, D. E., McBride, A. A., Paci, M., Cicero, D. O., y De Prat-Gay, G. (2006). *Molecular basis for phosphorylation-dependent, PEST-mediated protein turnover*. Structure, 14(2):309–319.

Gascuel, O. (1996). *BIONJ: An Improved Version of the NJ Algorithm Based on a Simple Model of Sequence Data*. Molecular Biology and Evolution, 14:685–695.

Geisberg, J. V., Chen, J. L., y Ricciardi, R. P. (1995). *Subregions of the adenovirus E1A transactivation domain target multiple components of the TFIID complex.* Molecular and Cellular Biology, 15(11):6283–90.

Ginsberg, H. S., Lundholm-Beauchamp, U., Horswood, R. L., Pernis, B., Wold, W. S. M., Chanock, R. M., y Prince, G. A. (1989). *Role of early region 3 (E3) in pathogenesis of adenovirus disease.* Proceedings of the National Academy of Sciences of the United States of America, 86(10):3823–7.

Glavina, J., Román, E. A., Espada, R., de Prat-Gay, G., Chemes, L. B., y Sánchez, I. E. (2018). *Interplay between sequence, structure and linear motifs in the adenovirus E1A hub protein*. Virology, 525(May):117–131.

Gloor, G. B., Martin, L. C., Wahl, L. M., y Dunn, S. D. (2005). *Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions*. Biochemistry, 44(19):7156–7165.

Göker, M., Scheuner, C., Klenk, H.-p., Stielow, J. B., y Menzel, W. (2011). *Codivergence of mycoviruses with their hosts.* PLoS One, 6(7):e22252.

Good, P. I. (2006). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer Series in Statistics. Springer-Verlag New York, 3 edition.

Gottschling, M., Bravo, I. G., Schulz, E., Bracho, M. A., Deaville, R., Jepson, P. D., Van Bressem, M.-F., Stockfleth, E., y Nindl, I. (2011a). *Modular organizations of novel cetacean papillomaviruses.* Molecular Phylogenetics and Evolution, 59(1):34–42.

Gottschling, M., Göker, M., Stamatakis, A., Bininda-Emonds, O. R. P., Nindl, I., y Bravo, I. G. (2011b). *Quantifying the phylodynamic forces driving papillomavirus evolution.* Molecular Biology and Evolution, 28(7):2101–13.

Gould, C. M., Diella, F., Via, A., Puntervoll, P., Gemünd, C., Chabanis-Davidson, S., Michael, S., Sayadi, A., Bryne, J. C., Chica, C., Seiler, M., Davey, N. E., Haslam, N., Weatheritt, R. J., Budd, A., Hughes, T., Pas, J., Rychlewski, L., Travé, G., Aasland, R., Helmer-Citterich, M., Linding, R., y Gibson, T. J. (2010). *ELM: the status of the 2010 eukaryotic linear motif resource.* Nucleic acids research, 38(Database issue):D167–80.

Gouw, M., Michael, S., Sámano-Sánchez, H., Kumar, M., Zeke, A., Lang, B., Bely, B., Chemes, L. B., Davey, N. E., Deng, Z., Diella, F., Gürth, C. M., Huber, A. K., Kleinsorg, S., Schlegel, L. S., Palopoli, N., Roey, K. V., Altenberg, B., Reményi, A., Dinkel, H., y Gibson, T. J. (2018). *The eukaryotic linear motif resource - 2018 update*. Nucleic Acids Research, 46(D1):D428–D434.

Graham, F. L., Smiley, J., Russell, W. C., y Nairn, R. (1977). *Characteristics of a human cell line transformed by DNA from human adenovirus type 5.* The Journal of General Virology, 36(1):59–74.

Graham, S. V. (2017). *Keratinocyte differentiation-dependent human papillomavirus gene regulation*. Viruses, 9(9):1–18.

Guan, J., Bywaters, S. M., Brendle, S. A., Ashley, R. E., Makhov, A. M., Conway, J. F., Christensen, N. D., y Hafenstein, S. (2017). *Cryoelectron Microscopy Maps of Human Papillomavirus 16 Reveal L2 Densities and Heparin Binding Site*. Structure, 25(2):253–263.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., y Gascuel, O. (2010). *New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.* Systematic Biology, 59(3):307–21.

Guindon, S. y Gascuel, O. (2003). *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.* Systematic Biology, 52(5):696–704.

Guo, Z., Yikang, S., Zacksenhaus, E., Yoshida, H., y Mak, T. W. (2001). *Inactivation of the retinoblastoma tumor suppressor induces apoptosis protease-activating factor-1 dependent and independent apoptotic pathways during embryogenesis*. Cancer Research, 61(23):8395–8400.

Gutman, R., Berezin, C., Wollman, R., Rosenberg, Y., y Ben-Tal, N. (2005). *QuasiMotiFinder: Protein annotation by searching for evolutionarily conserved motif-like patterns*. Nucleic Acids Research, 33(SUPPL. 2):255–261.

Güttler, T., Madl, T., Neumann, P., Deichsel, D., Corsini, L., Monecke, T., Ficner, R., Sattler, M., y Görlich, D. (2010). *NES consensus redefined by structures of PKI-type and Rev-type nuclear export signals bound to CRM1*. Nature Structural and Molecular Biology, 17(11):1367–76.

Haberz, P., Arai, M., Martinez-Yamout, M. A., Dyson, H. J., y Wright, P. E. (2016). *Mapping the interactions of adenoviral E1A proteins with the p160 nuclear receptor coactivator binding domain of CBP*. Protein Science, 25(12):2256–2267.

Hamada, N., Gotoh, K., Hara, K., Iwahashi, J., Imamura, Y., Nakamura, S., Taguchi, C., Sugita, M., Yamakawa, R., Etoh, Y., Sera, N., Ishibashi, T., Chijiwa, K., y Watanabe, H. (2008). *Nosocomial outbreak of epidemic keratoconjunctivitis accompanying environmental contamination with adenoviruses*. Journal of Hospital Infection, 68(3):262–268.

Hanson, J., Yang, Y., Paliwal, K., y Zhou, Y. (2017). *Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks*. Bioinformatics, 33(5):685–692.

Harrach, B., Benkö, M., Both, G. W., Brown, M., Davison, A. J., Echavarría, M., Hess, M., Jones, M. S., Kajon, A., Lehmkuhl, H. D., Mautner, V., Mittal, S. K., y Wadell, G. (2012). *Family Adenoviridae*. En King, A. M., Adams, M. J., Carstens, E. B., y Lefkowitz, E. J., editores, *Virus Taxonomy*, pages 125–141. Elsevier, San Diego, 9 edition.

Hasegawa, M., Kishino, H., y aki Yano, T. (1985). *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA*. Journal of Molecular Evolution, 22:160–174.

Hateboer, G., Gennissen, A., Ramos, Y. F. M., Kerkhoven, R. M., Sonntag-Buck, V., Stunnenberg, H. G., y Bernards, R. (1995). *BS69, a novel adenovirus E1A-associated protein that inhibits E1A transactivation.* The EMBO journal, 14(13):3159–69.

Henikoff, S. y Henikoff, J. G. (1994). *Position-based sequence weights.* Journal of molecular biology, 243(4):574–8.

Himpel, S., Tegge, W., Frank, R., Leder, S., Joost, H. G., y Becker, W. (2000). *Specificity determinants of substrate recognition by the protein kinase DYRK1A*. The Journal of Biological Chemistry, 275(4):2431–2438.

Hirschi, A., Cecchini, M., Steinhardt, R. C., Schamber, M. R., Dick, F. A., y Rubin, S. M. (2010). *An overlapping kinase and phosphatase docking site regulates activity of the retinoblastoma protein*. Nature Structural and Molecular Biology, 17(9):1051–1057.

Ho, C. M., Lee, B. H., Chang, S. F., Chien, T. Y., Huang, S. H., Yan, C. C., y Cheng, W. F. (2011). *Integration of human papillomavirus correlates with high levels of viral oncogene transcripts in cervical carcinogenesis*. Virus Research, 161(2):124–130.

Holder, M. y Lewis, P. O. (2003). *Phylogeny estimation: traditional and Bayesian approaches.* Nature reviews. Genetics, 4(4):275–84.

Hoppe, E., Pauly, M., Gillespie, T. R., Akoua-Koffi, C., Hohmann, G., Fruth, B., Karhemere, S., Madinda, N. F., Mugisha, L., Muyembe, J. J., Todd, A., Petrzelkova, K. J., Gray, M., Robbins, M., Bergl, R. A., Wittig, R. M., Zuberbühler, K., Boesch, C., Schubert, G., Leendertz, F. H., Ehlers, B., y Calvignac-Spencer, S. (2015). *Multiple cross-species transmission events of human adenoviruses (HAdV) during hominine evolution*. Molecular Biology and Evolution, 32(8):2072–2084.

Hošek, T., Calçada, E. O., Nogueira, M. O., Salvi, M., Pagani, T. D., Felli, I. C., y Pierattelli, R. (2016). *Structural and Dynamic Characterization of the Molecular Hub Early Region 1A (E1A) from Human Adenovirus*. Chemistry - A European Journal, 22(37):13010–13013.

Huang, A. y Stultz, C. M. (2009). *Finding order within disorder: Elucidating the structure of proteins associated with neurodegenerative disease*. Future Medicinal Chemistry, 1(3):467–482.

Huang, P. S., Patrick, D. R., Edwards, G., Goodhart, P. J., Huber, H. E., Miles, L., Garsky, V. M., Oliff, A., y Heimbrook, D. C. (1993). *Protein domains governing interactions between E2F, the retinoblastoma gene product, and human papillomavirus type 16 E7 protein*. Molecular and Cellular Biology, 13(2):953–960.

Huh, K.-W., Zhou, X., Hayakawa, H., Cho, J.-Y., Libermann, T. A., Jin, J., Wade Harper, J., y Munger, K. (2007). *Human Papillomavirus Type 16 E7 Oncoprotein Associates with the Cullin 2 Ubiquitin Ligase Complex, Which Contributes to Degradation of the Retinoblastoma Tumor Suppressor*. Journal of Virology, 81(18):9737–9747.

Hume, A. J., Finkel, J. S., Kamil, J. P., Coen, D. M., Culbertson, M. R., y Kalejta, R. F. (2008). *Phosphorylation of retinoblastoma protein by viral protein with cyclin-dependent kinase function.* Science, 320(5877):797–9.

Humphrey, W., Dalke, A., y Schulten, K. (1996). *VMD: visual molecular dynamics.* Journal of Molecular Graphics, 14(1):33–38.

Huyse, T. y Volckaert, F. A. M. (2005). *Comparing host and parasite phylogenies: gyrodactylus flatworms jumping from goby to goby.* Systematic Biology, 54(5):710–8.

Ikeda, M. A. y Nevins, J. R. (1993). *Identification of distinct roles for separate E1A domains in disruption of E2F complexes.* Molecular and Cellular Biology, 13(11):7029–35.

Inoue, Y., Kitagawa, M., y Taya, Y. (2007). *Phosphorylation of pRB at Ser612 by Chk1/2 leads to a complex between pRB and E2F-1 after DNA damage*. The EMBO Journal, 26(8):2083–2093.

Ishida, T. y Kinoshita, K. (2008). *Prediction of disordered regions in proteins based on the meta approach*. Bioinformatics, 24(11):1344–1348.

Isobe, T., Uchida, C., Hattori, T., Kitagawa, K., Oda, T., y Kitagawa, M. (2006). *Ubiquitin-dependent degradation of adenovirus E1A protein is inhibited by BS69.* Biochemical and Biophysical Research Communications, 339(1):367–74.

Iwanaga, R., Ohtani, K., Hayashi, T., y Nakamura, M. (2001). *Molecular mechanism of cell cycle progression induced by the oncogene product Tax of human T-cell leukemia virus type I.* Oncogene, 20(17):2055–2067.

Jansma, A. L., Martinez-Yamout, M. A., Liao, R., Sun, P., Dyson, H. J., y Wright, P. E. (2014). *The high-risk HPV16 E7 oncoprotein mediates interaction between the transcriptional coactivator CBP and the retinoblastoma protein pRb.* Journal of molecular biology, 426(24):4030–4048.

Jirgensons, B. (1958). *Optical rotation and viscosity of native and denatured proteins. X. Further studies on optical rotatory dispersion.* Archives of Biochemistry and Biophysics, 74(1):57–69.

Jogler, C., Hoffmann, D., Theegarten, D., Grunwald, T., Uberla, K., y Wildner, O. (2006). *Replication Properties of Human Adenovirus In Vivo and in Cultures of Primary Cells from Different Animal Species*. Journal of Virology, 80(7):3549–3558.

Jones, D. T. y Ward, J. J. (2003). *Prediction of Disordered Regions in Proteins From Position Specific Score Matrices*. Proteins: Structure, Function and Genetics, 53(SUPPL. 6):573–578.

Jones, N. y Shenk, T. (1979). *An adenovirus type 5 early gene function regulates expression of other early viral genes.* Proceedings of the National Academy of Sciences of the United States of America, 76(8):3665–9.

Kadaveru, K., Vyas, J., y Schiller, M. R. (2008). *Viral infection and human disease–insights from minimotifs.* Frontiers in Bioscience : a journal and virtual library, 13:6455–71.

Kalderon, D., Roberts, B. L., Richardson, W. D., y Smith, A. E. (1984). *A short amino acid sequence able to specify nuclear location*. Cell, 39(3 PART 2):499–509.

Kalejta, R. F. y Shenk, T. (2003). *Proteasome-dependent, ubiquitin-independent degradation of the Rb family of tumor suppressors by the human cytomegalovirus pp71 protein*. Proceedings of the National Academy of Sciences of the United States of America, 100(6):3263–3268.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., y Phillips, D. C. (1958). *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis*. Nature, 181(4610):662–666.

Kim, H. Y., Ahn, B. Y., y Cho, Y. (2001). *Structural basis for the inactivation of retinoblastoma tumor suppressor by SV40 large T antigen*. The EMBO Journal, 20(1-2):295–304.

Kimelman, D., Miller, J. S., Porter, D., y Roberts, B. E. (1985). *E1a regions of the human adenoviruses and of the highly oncogenic simian adenovirus 7 are closely related.* Journal of Virology, 53(2):399–409.

King, A. M. Q., Lefkowitz, E. J., Mushegian, A. R., Adams, M. J., Dutilh, B. E., Gorbalenya, A. E., Harrach, B., Harrison, R. L., Junglen, S., Knowles, N. J., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Nibert, M. L., Rubino, L., Sabanadzovic, S., Sanfaçon, H., Siddell, S. G., Simmonds, P., Varsani, A., Zerbini, F. M., y Davison, A. J. (2018). *Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018)*. Springer Vienna.

King, C. R., Zhang, A., y Mymryk, J. S. (2016). *The persistent mystery of adenovirus persistence*. Trends in Microbiology, 24(5):323–324.

Knapp, A. a., McManus, P. M., Bockstall, K., y Moroianu, J. (2009). *Identification of the nuclear localization and export signals of high risk HPV16 E7 oncoprotein.* Virology, 383(1):60–8.

Knight, J. S., Sharma, N., y Robertson, E. S. (2005). *Epstein-Barr virus latent antigen 3C can mediate the degradation of the retinoblastoma protein through an SCF cellular ubiquitin ligase.* Proceedings of the National Academy of Sciences of the United States of America, 102(51):18562–6.

Kohl, C., Vidovszky, M. Z., Mühldorfer, K., Dabrowski, P. W., Radonić, A., Nitsche, A., Wibbelt, G., Kurth, A., y Harrach, B. (2012). *Genome analysis of Bat adenovirus 2: indications of interspecies transmission.* Journal of Virology, 86(3):1888–92.

Köhler, M., Görlich, D., Hartmann, E., y Franke, J. (2001). *Adenoviral E1A protein nuclear import is preferentially mediated by importin alpha 3 in vitro.* Virology, 289(2):186–91.

Kovács, G. M., LaPatra, S. E., D'Halluin, J. C., y Benkö, M. (2003). *Phylogenetic analysis of the hexon and protease genes of a fish adenovirus isolated from white sturgeon (Acipenser transmontanus) supports the proposal for a new adenovirus genus*. Virus Research, 98(1):27–34.

Krippl, B., Ferguson, B. Q., Rosenberg, M., y Westphal, H. (1984). *Functions of purified E1A protein microinjected into mammalian cells.* Proceedings of the National Academy of Sciences of the United States of America, 81(22):6988–92.

Krystkowiak, I. y Davey, N. E. (2017). *SLiMSearch: A framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions*. Nucleic Acids Research, 45(W1):W464–W469.

Larsen, P. L. y Tibbetts, C. (1987). *Adenovirus E1A gene autorepression: revertants of an E1A promoter mutation encode altered E1A proteins.* Proceedings of the National Academy of Sciences of the United States of America, 84(December):8185–8189.

Lee, C., Chang, J. H., Lee, H. S., y Cho, Y. (2002). *Structural basis for the recognition of the E2F transactivation domain by the retinoblastoma tumor suppressor*. Genes & Development, 16(24):3199–3212.

Lee, J.-O., Russo, A. A., y Pavletich, N. P. (1998). *Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7.* Nature, 391(6670):859–65.

Lees, J. A., Buchkovich, K. J., Marshak, D. R., Anderson, C. W., y Harlow, E. (1991). *The retinoblastoma protein is phosphorylated on multiple sites by human cdc2.* The EMBO journal, 10(13):4279–90.

Legendre, P., Desdevises, Y., y Bazin, E. (2002). *A statistical test for host-parasite coevolution.* Systematic Biology, 51(2):217–234.

Lemey, P., Salemi, M., y Vandamme, A.-M. (2009). *The phylogenetic handbook: a practical approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press, Cambridge, 2nd edition.

Lenman, A., Liaci, A. M., Liu, Y., Årdahl, C., Rajan, A., Nilsson, E., Bradford, W., Kaeshammer, L., Jones, M. S., Frängsmyr, L., Feizi, T., Stehle, T., y Arnberg, N. (2015). *Human Adenovirus 52 Uses Sialic Acid-containing Glycoproteins and the Coxsackie and Adenovirus Receptor for Binding to Target Cells*. PLoS Pathogens, 11(2):1–23.

Liang, Y.-J., Chang, H.-S., Wang, C.-Y., y Yu, W. C. Y. (2008). *DYRK1A stabilizes HPV16E7 oncoprotein through phosphorylation of the threonine 5 and threonine 7 residues.* The International Journal of Biochemistry & Cell Biology, 40(11):2431–41.

Liban, T. J., Thwaites, M. J., Dick, F. A., y Rubin, S. M. (2016). *Structural Conservation and E2F Binding Specificity within the Retinoblastoma Pocket Protein Family.* Journal of molecular biology, 428(20):3960–3971.

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., y Russell, R. B. (2003a). *Protein disorder prediction: Implications for structural proteomics*. Structure, 11(11):1453–1459.

Linding, R., Russell, R. B., Neduva, V., y Gibson, T. J. (2003b). *GlobPlot: Exploring protein sequences for globularity and disorder*. Nucleic Acids Research, 31(13):3701–3708.

Lion, T. (2014). *Adenovirus infections in immunocompetent and immunocompromised patients*. Clinical Microbiology Reviews, 27(3):441–462.

Liu, F. y Green, M. R. (1994). *Promoter targeting by adenovirus E1a through interaction with different cellular DNA-binding domains.* Nature, 368(6471):520–5.

Liu, J., Boehme, P., Zhang, W., Fu, J., Yumul, R., Mese, K., Tsoukas, R., Solanki, M., Kaufmann, M., Lu, R., Schmidtko, A., Stewart, A. F., Lieber, A., y Ehrhardt, A. (2018). *Human adenovirus type 17 from species D transduces endothelial cells and human CD46 is involved in cell entry*. Scientific Reports, 8(1):1–14.

Liu, X., Clements, A., Zhao, K., y Marmorstein, R. (2006). *Structure of the human Papillomavirus E7 oncoprotein and its mechanism for inactivation of the retinoblastoma tumor suppressor.* The Journal of Biological Chemistry, 281(1):578–86.

Liu, X. y Marmorstein, R. (2007). *Structure of the retinoblastoma protein bound to adenovirus E1A reveals the molecular basis for viral oncoprotein inactivation of a tumor suppressor.* Genes & Development, 21(21):2711–6.

Liu, Y., Chen, W., Gaudet, J., Cheney, M. D., Roudaia, L., Cierpicki, T., Klet, R. C., Hartman, K., Laue, T. M., Speck, N. a., y Bushweller, J. H. (2007). *Structural basis for recognition of SMRT/N-CoR by the MYND domain and its contribution to AML1/ETO's activity.* Cancer Cell, 11(6):483–97.

Liu, Z. y Huang, Y. (2014). *Advantages of proteins being disordered*. Protein Science, 23(5):539–550.

Liu, Z., Song, Y., Bia, B., y Cowell, J. K. (1995). *Germline mutations in the RB1 gene in patients with hereditary retinoblastoma*. Genes, Chromosomes and Cancer, 14(4):277–284.

Lohmann, D. R. (1999). *RB1 gene mutations in retinoblastoma*. Human Mutation, 14(4):283–288.

López-Bueno, A., Mavian, C., Labella, A. M., Castro, D., Borrego, J. J., Alcami, A., y Alejo, A. (2016). *Concurrence of Iridovirus, Polyomavirus, and a Unique Member of a New Group of Fish Papillomaviruses in Lymphocystis Disease-Affected Gilthead Sea Bream.* Journal of Virology, 90(19):8768–79.

Lynch, J. P. y Kajon, A. E. (2016). *Adenovirus: Epidemiology, Global Spread of Novel Serotypes, and Advances in Treatment and Prevention*. Seminars in Respiratory and Critical Care Medicine, 37(4):586–602.

Lyon, K. F., Cai, X., Young, R. J., Mamun, A. A., Rajasekaran, S., y Schiller, M. R. (2018). *Minimotif Miner 4: A million peptide minimotifs and counting*. Nucleic Acids Research, 46(D1):D465–D470.

Lyons, R. H., Ferguson, B. Q., y Rosenberg, M. (1987). *Pentapeptide nuclear localization signal in adenovirus E1a.* Molecular and Cellular Biology, 7(7):2451–6.

Maddison, W. P. y Maddison, D. R. (2015). *Mesquite: a modular system for evolutionary analysis. Version 3.11 http://www.mesquiteproject.org*.

Madison, D. L., Yaciuk, P., Kwok, R. P. S., y Lundblad, J. R. (2002). *Acetylation of the adenovirus-transforming protein E1A determines nuclear localization by disrupting association with importin-alpha.* The Journal of Biological Chemistry, 277(41):38755–38763.

Magnaghi-Jaulin, L., Groisman, R., Naguibneva, I., Robin, P., Lorain, S., Le Villain, J. P., Troalen, F., Trouche, D., y Harel-Bellan, A. (1998). *Retinoblastoma protein represses transcription by recruiting a histone deacetylase.* Nature, 391(6667):601–5.

Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L., y Pappu, R. V. (2010). *Net charge per residue modulates conformational ensembles of intrinsically disordered proteins*. Proceedings of the National Academy of Sciences of the United States of America, 107(18):8183–8188.

Mao, A. H., Lyle, N., y Pappu, R. V. (2013). *Describing sequence-ensemble relationships for intrinsically disordered proteins.* The Biochemical Journal, 449(2):307–18.

Marsh, J. A. y Forman-Kay, J. D. (2012). *Ensemble modeling of protein disordered states: experimental restraint contributions and validation.* Proteins, 80(2):556–72.

Martin, A. J. M., Walsh, I., y Tosatto, S. C. E. (2010). *MOBI: a web server to define and visualize structural mobility in NMR protein ensembles.* Bioinformatics, 26(22):2916–7.

Marttila, M., Persson, D., Gustafsson, D., Liszewski, M. K., Atkinson, J. P., Wadell, G., y Arnberg, N. (2005). *CD46 Is a Cellular Receptor for All Species B Adenoviruses except Types 3 and 7*. Journal of Virology, 79(22):14429–14436.

Massimi, P., Pim, D., Kühne, C., y Banks, L. (2001). *Regulation of the human papillomavirus oncoproteins by differential phosphorylation.* Molecular and Cellular Biochemistry, 227(1-2):137–44.

Mazzarelli, J. M., Mengus, G., Davidson, I., y Ricciardi, R. P. (1997). *The transactivation domain of adenovirus E1A interacts with the C terminus of human TAF(II)135.* Journal of Virology, 71(10):7978–83.

McBride, A. A. (2013). *The Papillomavirus E2 proteins*. Virology, 445(1-2):57–79.

McBride, A. A. (2017). *Mechanisms and strategies of papillomavirus replication*. Biological Chemistry, 398(8):919–927.

McIntyre, M. C., Ruesch, M. N., y Laimins, L. A. (1996). *Human papillomavirus E7 oncoproteins bind a single form of cyclin E in a complex with cdk2 and p107*. Virology, 215(1):73–82.

Meier-Kolthoff, J. P., Auch, A. F., Huson, D. H., y Göker, M. (2007). *CopyCat: Cophylogenetic analysis tool*. Bioinformatics, 23(7):898–900.

Meng, X., Webb, P., Yang, Y.-f., Shuen, M., Yousef, A. F., Baxter, J. D., Mymryk, J. S., y Walfish, P. G. (2005). *E1A and a nuclear receptor corepressor splice variant (N-CoRI) are thyroid hormone receptor coactivators that bind in the corepressor mode.* Proceedings of the National Academy of Sciences of the United States of America, 102(18):6267–72.

Mészáros, B., Erdös, G., y Dosztányi, Z. (2018). *IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding*. Nucleic Acids Research, 46(W1):W329–W337.

Mi, T., Merlin, J. C., Deverasetty, S., Gryk, M. R., Bill, T. J., Brooks, A. W., Lee, L. Y., Rathnayake, V., Ross, C. A., Sargeant, D. P., Strong, C. L., Watts, P., Rajasekaran, S., y Schiller, M. R. (2012). *Minimotif Miner 3.0: Database expansion and significantly improved reduction of false-positive predictions from consensus sequences*. Nucleic Acids Research, 40(D1):252–260.

Miller, M. S., Pelka, P., Fonseca, G. J., Cohen, M. J., Kelly, J. N., Barr, S. D., Grand, R. J. A., Turnell, A. S., Whyte, P., y Mymryk, J. S. (2012). *Characterization of the 55-Residue Protein Encoded by the 9S E1A mRNA of Species C Adenovirus*. Journal of Virology, 86(8):4222–4233.

Miskei, M., Antal, C., y Fuxreiter, M. (2017). *FuzDB: Database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies*. Nucleic Acids Research, 45(D1):D228–D235.

Mizianty, M. J., Stach, W., Chen, K., Kedarisetti, K. D., Disfani, F. M., y Kurgan, L. (2011). *Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources*. Bioinformatics, 27(13):i489–i496.

Molloy, D. P., Barral, P. M., Bremner, K. H., Gallimore, P. H., y Grand, R. J. A. (2000). *Structural determinants in adenovirus 12 E1A involved in the interaction with C-terminal binding protein 1*. Virology, 277(1):156–66.

Molloy, D. P., Barral, P. M., Gallimore, P. H., y Grand, R. J. a. (2007). *The effect of CtBP1 binding on the structure of the C-terminal region of adenovirus 12 early region 1A*. Virology, 363(2):342–56.

Molloy, D. P., Mapp, K. L., Webster, R., Gallimore, P. H., y Grand, R. J. a. (2006). *Acetylation at a lysine residue adjacent to the CtBP binding motif within adenovirus 12 E1A causes structural disruption and limited reduction of CtBP binding.* Virology, 355(2):115–26.

Molloy, D. P., Milner, A. E., Yakub, I. K., Chinnadurai, G., Gallimore, P. H., y Grand, R. J. A. (1998). *Structural Determinants Present in the C-terminal Binding Protein Binding Site of Adenovirus Early Region 1A Proteins*. The Journal of Biological Chemistry, 273(33):20867–20876.

Molloy, D. P., Smith, K. J., Milner, A. E., Gallimore, P. H., y Grand, R. J. A. (1999). *The Structure of the Site on Adenovirus Early Region 1A Responsible for Binding to TATA-binding Protein Determined by NMR Spectroscopy*. The Journal of Biological Chemistry, 274(6):3503–3512.

Montell, C., Courtois, G., Eng, C., y Berk, A. J. (1984). *Complete transformation by adenovirus 2 requires both E1A proteins*. Cell, 36(4):951–961.

Montell, C., Fisher, E. F., Caruthers, M. H., y Berk, A. J. (1982). *Resolving the functions of overlapping viral genes by site-specific mutagenesis at a mRNA splice site*. Nature, 295(5848):380–4.

Moody, C. A. (2017). *Mechanisms by which HPV induces a replication competent environment in differentiating keratinocytes*. Viruses, 9(9):1–21.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., y Weigt, M. (2011). *Direct-coupling analysis of residue coevolution captures native contacts across many protein families.* Proceedings of the National Academy of Sciences of the United States of America, 108(49):E1293–301.

Morgunova, E., Yin, Y., Jolma, A., Dave, K., Schmierer, B., Popov, A., Eremina, N., Nilsson, L., y Taipale, J. (2015). *Structural insights into the DNA-binding specificity of E2F family transcription factors*. Nature Communications, 6:4–11.

Morris, E. J. y Dyson, N. J. (2001). *Retinoblastoma protein partners.* Advances in Cancer Research, 82(617):1–54.

Munakata, T., Liang, Y., Kim, S., McGivern, D. R., Huibregtse, J., Nomoto, A., y Lemon, S. M. (2007). *Hepatitis C virus induces E6AP-dependent degradation of the retinoblastoma protein*. PLoS Pathogens, 3(9):e139.

Munakata, T., Nakamura, M., Liang, Y., Li, K., y Lemon, S. M. (2005). *Down-regulation of the retinoblastoma tumor suppressor by the hepatitis C virus NS5B RNA-dependent RNA polymerase.* Proceedings of the National Academy of Sciences of the United States of America, 102(50):18159–64.

Münger, K., Basile, J. R., Duensing, S., Eichten, A., Gonzalez, S. L., Grace, M., y Zacny, V. L. (2001). *Biological activities and molecular targets of the human papillomavirus E7 oncoprotein*. Oncogene, 20(54 REV. ISS. 7):7888–7898.

Münger, K., Werness, B. A., Dyson, N. J., Phelps, W. C., Harlow, E., y Howley, P. M. (1989). *Complex formation of human papillomavirus E7 proteins with the retinoblastoma tumor suppressor gene product.* The EMBO journal, 8(13):4099–105.

Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W., y Springer, M. S. (2001). *Resolution of the early placental mammal radiation using Bayesian phylogenetics.* Science, 294(5550):2348–51.

Murtagh, P., Cerqueiro, C., Halac, A., Avila, M., y Kajon, A. (1993). *Adenovirus type 7h respiratory infections: a report of 29 cases of acute lower respiratory disease.* Acta paediatrica, 82(6-7):557–61.

Na, J. H., Lee, W. K., y Yu, Y. G. (2018). *How do we study the dynamic structure of unstructured proteins: A case study on nopp140 as an example of a large, intrinsically disordered protein*. International Journal of Molecular Sciences, 19(2).

Neduva, V., Linding, R., Su-Angrand, I., Stark, A., de Masi, F., Gibson, T. J., Lewis, J., Serrano, L., y Russell, R. B. (2005). *Supplemental Data Systematic discovery of new recognition peptides mediating protein interaction networks.* PLoS Biology, 3(12):e405.

Neduva, V. y Russell, R. B. (2005). *Linear motifs: evolutionary interaction switches.* FEBS letters, 579(15):3342–5.

Neduva, V. y Russell, R. B. (2006). *DILIMOT: Discovery of linear motifs in proteins*. Nucleic Acids Research, 34(WEB. SERV. ISS.):350–355.

Notredame, C., Higgins, D. G., y Heringa, J. (2000). *T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton*. Journal of Molecular Biology, 302(1):205–217.

Noval, M. G., Gallo, M., Perrone, S., Salvay, A. G., Chemes, L. B., y de Prat-Gay, G. (2013). *Conformational dissection of a viral intrinsically disordered domain involved in cellular transformation.* PLoS One, 8(9):e72760.

Oates, M. E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M. J., Xue, B., Dosztányi, Z., Uversky, V. N., Obradovic, Z., Kurgan, L., Dunker, A. K., y Gough, J. (2013). *D2P2: Database of disordered protein predictions*. Nucleic Acids Research, 41(D1):508–516.

Obenauer, J. C., Cantley, L. C., y Yaffe, M. B. (2003). *Scansite 2.0: Proteome-wide prediction of cell signalling interactions using short sequence motifs*. Nucleic Acids Research, 31(13):3635–3641.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., y Dunker, A. K. (2003). *Predicting intrinsic disorder from amino acid sequence.* Proteins, 53 Suppl 6(February):566–72.

O'Connor, M. J., Zimmermann, H., Nielsen, S., Bernard, H.-U., y Kouzarides, T. (1999). *Characterization of an E1A-CBP interaction defines a novel transcriptional adapter motif (TRAM) in CBP/p300.* Journal of Virology, 73(5):3574–81.

Oh, K.-J., Kalinina, A., Wang, J., Nakayama, K. K. I., y Bagchi, S. (2004). *The papillomavirus E7 oncoprotein is ubiquitinated by UbcH7 and Cullin 1- and Skp2-containing E3 ligase.* Journal of Virology, 78(10):5338–46.

Ohlenschläger, O., Seiboth, T., Zengerling, H., Briese, L., Marchanka, A., Ramachandran, R., Baum, M., Korbas, M., Meyer-Klaucke, W., Dürst, M., y Görlach, M. (2006). *Solution structure of the partially folded high-risk human papilloma virus 45 oncoprotein E7.* Oncogene, 25(44):5953–9.

Oldfield, C. J. y Dunker, A. K. (2014). *Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions*. Annual Review of Biochemistry, 83(1):553–584.

O'Leary, M. A., Bloch, J. I., Flynn, J. J., Gaudin, T. J., Giallombardo, A., Giannini, N. P., Goldberg, S. L., Kraatz, B. P., Luo, Z.-x., Meng, J., Ni, X., Novacek, M. J., Perini, F. A., Randall, Z. S., Rougier, G. W., Sargis, E. J., Silcox, M. T., Simmons, N. B., Spaulding, M., Velazco, P. M., Weksler, M., Wible, J. R., y Cirranello, A. L. (2013). *The placental mammal ancestor and the post-K-Pg radiation of placentals.* Science, 339(6120):662–7.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., Van Roey, K., Cesareni, G., y Hermjakob, H. (2014). *The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases*. Nucleic Acids Research, 42(D1):358–363.

Orino, K., Tsuji, Y., Torti, F. M., y Torti, S. V. (1999). *Adenovirus E1A blocks oxidant-dependent ferritin induction and sensitizes cells to pro-oxidant cytotoxicity*. FEBS Letters, 461(3):334–338.

Ovadia, Y., Fielder, D., Conow, C., y Libeskind-Hadas, R. (2011). *The Cophylogeny Reconstruction Problem Is NP-Complete*. Journal of Computational Biology, 18(1):59–65.

Ozenne, V., Bauer, F., Salmon, L., Huang, J.-R., Jensen, M. R., Segard, S., Bernadó, P., Charavay, C., y Blackledge, M. (2012). *Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables.* Bioinformatics, 28(11):1463–70.

Page, R. D. (1994). *Parallel phylogenies: reconstructing the history of host-parasite assemblages*. Cladistics, 10(2):155–173.

Palopoli, N., González Foutel, N. S., Gibson, T. J., y Chemes, L. B. (2018). *Short linear motif core and flanking regions modulate retinoblastoma protein binding affinity and specificity.* Protein Engineering, Design & Selection, 31(3):69–77.

Palopoli, N., Lythgow, K. T., y Edwards, R. J. (2015). *QSLiM Finder: Improved short linear motif prediction using specific query protein data*. Bioinformatics, 31(14):2284–2293.

Patel, M. R., Loo, Y. M., Horner, S. M., Gale, M., y Malik, H. S. (2012). *Convergent evolution of escape from hepaciviral antagonism in primates*. PLoS Biology, 10(3):e1001282.

Pauwels, K., Lebrun, P., y Tompa, P. (2017). *To be disordered or not to be disordered: is that still a question for proteins in the cell?* Cellular and Molecular Life Sciences, 74(17):3185–3204.

Pelka, P., Ablack, J. N. G., Fonseca, G. J., Yousef, A. F., y Mymryk, J. S. (2008). *Intrinsic structural disorder in adenovirus E1A: a viral molecular hub linking multiple diverse processes.* Journal of virology, 82(15):7252–63.

Peng, Z., Mizianty, M. J., y Kurgan, L. (2014). *Genome-scale prediction of proteins with long intrinsically disordered regions*. Proteins: Structure, Function and Bioinformatics, 82(1):145–158.

Pérez-Losada, M., Christensen, R. G., McClellan, D. a., Adams, B. J., Viscidi, R. P., Demma, J. C., y Crandall, K. a. (2006). *Comparing phylogenetic codivergence between polyomaviruses and their hosts.* Journal of Virology, 80(12):5663–9.

Perricaudet, M., Akusjärvi, G., Virtanen, A., y Pettersson, U. (1979). *Structure of two spliced mRNAs from the transforming region of human subgroup C adenoviruses.* Nature, 281(5733):694–6.

Petsalaki, E., Stark, A., García-Urdiales, E., y Russell, R. B. (2009). *Accurate prediction of peptide binding sites on protein surfaces*. PLoS Computational Biology, 5(3):e1000335.

Phelan, C. A., Gampe, R. T., Lambert, M. H., Parks, D. J., Montana, V., Bynum, J., Broderick, T. M., Hu, X., Williams, S. P., Nolte, R. T., y Lazar, M. a. (2010). *Structure of Rev-erbalpha bound to N-CoR reveals a unique mechanism of nuclear receptor-co-repressor interaction.* Nature Structural and Molecular Biology, 17(7):808–14.

Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C. J., Aspromonte, M. C., Davey, N. E., Davidović, R., Dosztányi, Z., Elofsson, A., Gasparini, A., Hatos, A., Kajava, A. V., Kalmar, L., Leonardi, E., Lazar, T., Macedo-Ribeiro, S., Macossay-Castillo, M., Meszaros, A., Minervini, G., Murvai, N., Pujols, J., Roche, D. B., Salladini, E., Schad, E., Schramm, A., Szabo, B., Tantos, A., Tonello, F., Tsirigos, K. D., Veljković, N., Ventura, S., Vranken, W. F., Warholm, P., Uversky, V. N., Dunker, A. K., Longhi, S., Tompa, P., y Tosatto, S. C. E. (2017). *DisProt 7.0: A major update of the database of disordered proteins*. Nucleic Acids Research, 45(D1):D219–D227.

Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Mieti, I., Camilloni, C., Davey, N. E., Dosztányi, Z., Mészáros, B., Monzon, A. M., Parisi, G., Schad, E., Sormanni, P., Tompa, P., Vendruscolo, M., Vranken, W. F., y Tosatto, S. C. (2018). *MobiDB 3.0: More annotations for intrinsic disorder, conformational diversity and interactions in proteins*. Nucleic Acids Research, 46(D1):D471–D476.

Prichard, M. N., Sztul, E., Daily, S. L., Perry, A. L., Frederick, S. L., Gill, R. B., Hartline, C. B., Streblow, D. N., Varnum, S. M., Smith, R. D., y Kern, E. R. (2008). *Human Cytomegalovirus UL97 Kinase Activity Is Required for the Hyperphosphorylation of Retinoblastoma Protein and Inhibits the Formation of Nuclear Aggresomes*. Journal of Virology, 82(10):5054–5067.

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., Van Der Spoel, D., Hess, B., y Lindahl, E. (2013). *GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit*. Bioinformatics, 29(7):845–854.

Puntervoll, P., Linding, R., Gemünd, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D. M. A., Ausiello, G., Brannetti, B., Costantini, A., Ferrè, F., Maselli, V., Via, A., Cesareni, G., Diella, F., Superti-Furga, G., Wyrwicz, L. S., Ramu, C., McGuigan, C., Gudavalli, R., Letunic, I., Bork, P., Rychlewski, L., Küster, B., Helmer-Citterich, M., Hunter, W. N., Aasland, R., y Gibson, T. J. (2003). *ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins*. Nucleic acids research, 31(13):3625–30.

Radko, S., Koleva, M., James, K. M., Jung, R., Mymryk, J. S., y Pelka, P. (2014). *Adenovirus E1A Targets the DREF Nuclear Factor To Regulate Virus Gene Expression, DNA Replication, and Growth*. Journal of Virology, 88(22):13469–13481.

Rasti, M., Grand, R. J. a., Mymryk, J. S., Gallimore, P. H., y Turnell, A. S. (2005). *Recruitment of CBP/p300, TATA-binding protein, and S8 to distinct regions at the N terminus of adenovirus E1A*. Journal of Virology, 79(9):5594–605.

Rechsteiner, M. y Rogers, S. W. (1996). *PEST sequences and regulation by proteolysis*. Trends in Biochemical Sciences, 21(7):267–271.

Rehtanz, M., Ghim, S. J., Rector, A., Van Ranst, M., Fair, P. A., Bossart, G. D., y Jenson, A. B. (2006). *Isolation and characterization of the first American bottlenose dolphin papillomavirus: Tursiops truncatus papillomavirus type 2*. Journal of General Virology, 87(12):3559–3565.

Reinstein, E., Scheffner, M., Oren, M., Ciechanover, A., y Schwartz, A. (2000). *Degradation of the E7 human papillomavirus oncoprotein by the ubiquitin-proteasome system: targeting via ubiquitination of the N-terminal residue.* Oncogene, 19(51):5944–50.

Revell, L. J. (2012). *phytools: An R package for phylogenetic comparative biology (and other things)*. Methods in Ecology and Evolution, 3(2):217–223.

Rivals, I., Personnaz, L., Taing, L., y Potier, M.-C. (2007). *Enrichment or depletion of a GO category within a class of genes: which test?* Bioinformatics, 23(4):401–7.

Robinson, C. M., Seto, D., Jones, M. S., Dyer, D. W., y Chodosh, J. (2011). *Molecular evolution of human species D adenoviruses.* Infection, Genetics and Evolution, 11(6):1208–17.

Roelvink, P. W., Lizonova, A., Lee, J. G. M., Li, Y., Bergelson, J. M., Finberg, R. W., Brough, D. E., Kovesdi, I., y Wickham, T. J. (1998). *The Coxsackievirus-Adenovirus receptor protein can function as a cellular attachment protein for adenovirus serotypes from subgroups A,C, D, E, and F*. Journal of Virology, 72(10):7909–7915.

Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J. E., y Dunker, A. K. (1997). *Identifying disordered regions in proteins from amino acid sequence*. Proceedings of International Conference on Neural Networks, 1(June):1–6.

Ronquist, F. (1995). *Reconstructing the history of host-parasite associations using generalised parsimony*. Cladistics, 11(1):73–89.

Rotkiewicz, P. y Skolnick, J. (2008). *Fast procedure for reconstruction of full-atom protein models from reduced representations.* Journal of Computational Chemistry, 29(9):1460–5.

Rowe, D. T., Graham, F. L., y Branton, P. E. (1983). *Intracellular localization of adenovirus type 5 tumor antigens in productively infected cells*. Virology, 129(2):456–468.

Rowe, W. P., Huebner, R. J., Gilmore, L. K., Parrot, R. H., y Ward, T. J. (1953). *Isolation of a cytopathogenic agent from human adenoids undergoing spontaneous degeneration in tissue culture.* Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine (New York, N.Y.), 84(3):570–3.

Saitou, N. y Nei, M. (1987). *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Molecular Biology and Evolution, 4(4):406–425.

Sarkar, D., Jana, T., y Saha, S. (2015). *LMPID: A manually curated database of linear motifs mediating protein-protein interactions*. Database, 2015:1–6.

Schaefer, C., Schlessinger, A., y Rost, B. (2010). *Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be.* Bioinformatics, 26(5):625–31.

Schaeper, U., Boyd, J. M., Verma, S., Uhlmann, E., Subramanian, T., y Chinnadurai, G. (1995). *Molecular cloning and characterization of a cellular phosphoprotein that interacts with a conserved C-terminal domain of adenovirus E1A involved in negative modulation of oncogenic transformation.* Proceedings of the National Academy of Sciences of the United States of America, 92(23):10467–71.

Scheffner, M., Werness, B. A., Huibregtse, J. M., Levine, A. J., y Howley, P. M. (1990). *The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53.* Cell, 63(6):1129–36.

Schneider, T. D. y Stephens, R. M. (1990). *Sequence logos: a new way to display consensus sequences.* Nucleic acids research, 18(20):6097–100.

Schneider, T. D., Stormo, G. D., Gold, L., y Ehrenfeucht, A. (1986). *Information content of binding sites on nucleotide sequences.* Journal of Molecular Biology, 188(3):415–31.

Schulze, A., Zerfass, K., Spitkovsky, D., Middendorp, S., Bergès, J., Helin, K., Jansen-Dürr, P., y Henglein, B. (1995). *Cell cycle regulation of the cyclin A gene promoter is mediated by a variant E2F site.* Proceedings of the National Academy of Sciences of the United States of America, 92(24):11264–11268.

Schutze, D. M., Snijders, P. J. F., Bosch, L., Kramer, D., Meijer, C. J. L. M., y Steenbergen, R. D. M. (2014). *Differential In Vitro Immortalization Capacity of Eleven, Probable High-Risk Human Papillomavirus Types*. Journal of Virology, 88(3):1714–1724.

Schwartz, K. L., Richardson, S. E., MacGregor, D., Mahant, S., Raghuram, K., y Bitnun, A. (2018). *Adenovirus-Associated Central Nervous System Disease in Children*. Journal of Pediatrics.

Schwartz, S. (2013). *Papillomavirus transcripts and posttranscriptional regulation*. Virology, 445(1-2):187–196.

Serrano, B., Brotons, M., Bosch, F. X., y Bruni, L. (2018). *Epidemiology and burden of HPV-related disease*. Best Practice and Research: Clinical Obstetrics and Gynaecology, 47:14–26.

Shaanan, B. (1983). *Structure of human oxyhaemoglobin at 2·1resolution*. Journal of Molecular Biology, 171(1):31–59.

Shan, B., Durfee, T., y Lee, W.-H. (1996). *Disruption of RB/E2F-1 interaction by single point mutations in E2F-1 enhances S-phase entry and apoptosis.* Proceedings of the National Academy of Sciences of the United States of America, 93(2):679–84.

Shannon, C. E. (1948). *A Mathematical Theory of Communication*. The Bell System Technical Journal, 27(July 1928):379–423.

Shapiro, S. S. y Wilk, M. B. (1965). *An Analysis of Variance Test for Normality (Complete Samples)*. Biometrika, 52(3):591–611.

Singh, G., Robinson, C. M., Dehghan, S., Schmidt, T., Seto, D., Jones, M. S., Dyer, D. W., y Chodosh, J. (2012). *Overreliance on the hexon gene, leading to misclassification of human adenoviruses.* Journal of Virology, 86(8):4693–5.

Singh, M., Krajewski, M., Mikolajka, A., y Holak, T. A. (2005). *Molecular determinants for the complex formation between the retinoblastoma protein and LXCXE sequences.* The Journal of Biological Chemistry, 280(45):37868–76.

Sippl, M. J. y Wiederstein, M. (2012). *Detection of spatial correlations in protein structures and molecular complexes*. Structure, 20(4):718–728.

Smotkin, D. y Wettstein, F. O. (1987). *The major human papillomavirus protein in cervical cancers is a cytoplasmic phosphoprotein.* Journal of Virology, 61(5):1686–9.

Stephens, C. y Harlow, E. (1987). *Differential splicing yields novel adenovirus 5 E1A mRNAs that encode 30 kd and 35 kd proteins.* The EMBO journal, 6(7):2027–35.

Stevens, H., Rector, A., Bertelsen, M. F., Leifsson, P. S., y Van Ranst, M. (2008a). *Novel papillomavirus isolated from the oral mucosa of a polar bear does not cluster with other papillomaviruses of carnivores.* Veterinary Microbiology, 129(1-2):108–16.

Stevens, H., Rector, A., Van Der Kroght, K., y Van Ranst, M. (2008b). *Isolation and cloning of two variant papillomaviruses from domestic pigs: Sus scrofa papillomaviruses type 1 variants a and b.* The Journal of General Virology, 89(Pt 10):2475–81.

Stokes, P. H., Thompson, L. S., Marianayagam, N. J., y Matthews, J. M. (2007). *Dimerization of CtIP may stabilize in vivo interactions with the Retinoblastoma-pocket domain.* Biochemical and Biophysical Research Communications, 354(1):197–202.

Subramanian, T., Kuppuswamy, M., Nasr, R. J., y Chinnadurai, G. (1988). *An N-terminal region of adenovirus E1a essential for cell transformation and induction of an epithelial cell growth factor.* Oncogene, 2(2):105–12.

Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T., y Onuchic, J. N. (2012). *Genomics-aided structure prediction.* Proceedings of the National Academy of Sciences of the United States of America, 109(26):10340–5.

Szalmás, A., Tomaić, V., Basukala, O., Massimi, P., Mittal, S., Kónya, J., y Banks, L. (2017). *The PTPN14 Tumor Suppressor Is a Degradation Target of Human Papillomavirus E7.* Journal of Virology, 91(7):1–13.

Tan, S. H., Hugo, W., Sung, W. K., y Ng, S. K. (2006). *A correlated motif approach for finding short linear motifs from protein interaction networks.* BMC Bioinformatics, 7.

Tedesco, D., Lukas, J., y Reed, S. I. (2002). *The pRb-related protein p130 is regulated by phosphorylation-dependent proteolysis via the protein-ubiquitin ligase SCFSkp2.* Genes & Development, 16(22):2946–2957.

Telling, G. C. y Williams, J. (1994). *Constructing chimeric type 12/type 5 adenovirus E1A genes and using them to identify an oncogenic determinant of adenovirus type 12.* Journal of Virology, 68(2):877–887.

Thompson, J. D., Higgins, D. G., y Gibson, T. J. (1994). *CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Research, 22(22):4673–4680.

Thompson, J. D., Koehl, P., Ripp, R., y Poch, O. (2005). *BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark.* Proteins: Structure, Function and Genetics, 61(1):127–136.

Todorovic, B., Hung, K., Massimi, P., Avvakumov, N., Dick, F. A., Shaw, G. S., Banks, L., y Mymryk, J. S. (2012). *Conserved Region 3 of Human Papillomavirus 16 E7 Contributes to Deregulation of the Retinoblastoma Tumor Suppressor.* Journal of Virology, 86(24):13313–13323.

Todorovic, B., Nichols, A. C., Chitilian, J. M., Myers, M. P., Shepherd, T. G., Parsons, S. J., Barrett, J. W., Banks, L., y Mymryk, J. S. (2014). *The human papillomavirus E7 proteins associate with p190RhoGAP and alter its function.* Journal of Virology, 88(7):3653–63.

Tokuriki, N., Oldfield, C. J., Uversky, V. N., Berezovsky, I. N., y Tawfik, D. S. (2009). *Do viral proteins possess unique biophysical features?* Trends in Biochemical Sciences, 34(2):53–59.

Tollefson, A. E., Ryerse, J. S., Scaria, A., Hermiston, T. W., y Wold, W. S. (1996). *The E3-11.6-kDa adenovirus death protein (ADP) is required for efficient cell death: Characterization of cells infected with adp mutants.* Virology, 220(1):152–162.

Tomaić, V., Gardiol, D., Massimi, P., Ozbun, M., Myers, M. P., y Banks, L. (2009). *Human and primate tumour viruses use PDZ binding as an evolutionarily conserved mechanism of targeting cell polarity regulators.* Oncogene, 28(1):1–8.

Tomko, R. P., Xu, R., y Philipson, L. (1997). *HCAR and MCAR: The human and mouse cellular receptors for subgroup C adenoviruses and group B coxsackieviruses.* Proceedings of the National Academy of Sciences of the United States of America, 94(7):3352–3356.

Tompa, P. (2005). *The interplay between structure and function in intrinsically unstructured proteins.* FEBS Letters, 579(15):3346–3354.

Tompa, P., Davey, N. E., Gibson, T. J., y Babu, M. M. (2014). *A Million peptide motifs for the molecular biologist.* Molecular Cell, 55(2):161–169.

Toth-Petroczy, A., Meszaros, B., Simon, I., Dunker, A. K., Uversky, V. N., y Fuxreiter, M. (2008). *Assessing Conservation of Disordered Regions in Proteins*. The Open Proteomics Journal, 1(1):46–53.

Toth-Petroczy, A., Palmedo, P., Ingraham, J., Hopf, T. A., Berger, B., Sander, C., y Marks, D. S. (2016). *Structured States of Disordered Proteins from Genomic Sequences.* Cell, 167(1):158–170.e12.

Trentin, J. J., Yabe, Y., y Taylor, G. (1962). *The quest for human cancer viruses.* Science, 137(3533):835–41.

Turnell, A. S., Grand, R. J. A., Gorbea, C., Zhang, X., Wang, W., Mymryk, J. S., y Gallimore, P. H. (2000). *Regulation of the 26S proteasome by adenovirus E1A.* The EMBO journal, 19(17):4759–73.

Uversky, V. N. (2011). *Intrinsically disordered proteins from A to Z.* The International Journal of Biochemistry & Cell Biology, 43(8):1090–1103.

Uversky, V. N. (2013). *Unusual biophysics of intrinsically disordered proteins*. Biochimica et Biophysica Acta, 1834(5):932–951.

Uversky, V. N. (2015). *Biophysical Methods to Investigate Intrinsically Disordered Proteins: Avoiding an "Elephant and Blind Men" Situation.* Advances in Experimental Medicine and Biology, 870(Chapter 159):215–60.

Uversky, V. N. (2017). *How to Predict Disorder in a Protein of Interest.* Methods in molecular biology (Clifton, N.J.), 1484:137–158.

Uversky, V. N., Roman, A., Oldfield, C. J., y Dunker, A. K. (2006). *Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in E6 and E7 oncoproteins from high risk HPVs.* Journal of Proteome Research, 5(8):1829–42.

Valdés, O., Acosta, B., Piñón, A., Savón, C., Goyenechea, A., Gonzalez, G., Gonzalez, G., Palerm, L., Sarmiento, L., Pedro, M. L., Martínez, P. A., Rosario, D., Kourí, V., Guzmán, M. G., Llop, A., Casas, I., y Perez Breña, M. P. (2008). *First report on fatal myocarditis associated with adenovirus infection in Cuba.* Journal of Medical Virology, 80(10):1756–61.

Valverde, J. R., Alonso, J., Palacios, I., y Pestaña, Á. (2005). *RB1 gene mutation up-date, a meta-analysis based on 932 reported mutations available in a searchable database*. BMC Genetics, 6:1–9.

Van Bressem, M.-F., Cassonnet, P., Rector, A., Desaintes, C., Van Waerebeek, K., Alfaro-Shigueto, J., Van Ranst, M., y Orth, G. (2007). *Genital warts in Burmeister's porpoises: characterization of Phocoena spinipinnis papillomavirus type 1 (PsPV-1) and evidence for a second, distantly related PsPV.* The Journal of General Virology, 88:1928–33.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., y Babu, M. M. (2014). *Classification of intrinsically disordered regions and proteins.* Chemical Reviews, 114(13):6589–631.

Van Doorslaer, K. (2013). *Evolution of the Papillomaviridae*. Virology, 445(1-2):11–20.

Van Doorslaer, K., Chen, Z., Bernard, H.-U., Chan, P. K. S., DeSalle, R., Dillner, J., Forslund, O., Haga, T., McBride, A. A., Villa, L. L., Burk, R. D., y Ictv Report Consortium (2018). *ICTV Virus Taxonomy Profile: Papillomaviridae.* The Journal of General Virology, 99(8):989–990.

van Ormondt, H. y Hesper, B. (1983). *Comparison of the nucleotide sequences of early region E1b DNA of human adenovirus types 12, 7 and 5 (subgroups A, B and C).* Gene, 21(3):217–26.

Van Roey, K. y Davey, N. E. (2015). *Motif co-regulation and co-operativity are common mechanisms in transcriptional, post-transcriptional and post-translational regulation.* Cell Communication and Signaling, 13(1):45.

Van Roey, K., Uyar, B., Weatheritt, R. J., Dinkel, H., Seiler, M., Budd, A., Gibson, T. J., y Davey, N. E. (2014). *Short linear motifs: Ubiquitous and functionally diverse protein interaction modules directing cell regulation*. Chemical Reviews, 114(13):6733–6778.

Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A. K., Felli, I. C., Forman-Kay, J. D., Kriwacki, R. W., Pierattelli, R., Sussman, J. L., Svergun, D. I., Uversky, V. N., Vendruscolo, M., Wishart, D., Wright, P. E., y Tompa, P. (2014). *PE-DB: A database of structural ensembles of intrinsically disordered and of unfolded proteins*. Nucleic Acids Research, 42(D1):326–335.

Via, A., Gherardini, P. F., Ferraro, E., Ausiello, G., Tomba, G. S., y Helmer-Citterich, M. (2007). *False occurrences of functional motifs in protein sequences highlight evolutionary constraints*. BMC Bioinformatics, 8:1–13.

Virtanen, A. y Pettersson, U. (1983). *The molecular structure of the 9S mRNA from early region 1A of adenovirus serotype 2.* Journal of molecular biology, 165(3):496–9.

Vullo, A., Bortolamil, O., Pollastri, G., y Tosatto, S. C. E. (2006). *Spritz: A server for the prediction of intrinsically disordered regions in protein sequences using kernel machines*. Nucleic Acids Research, 34(WEB. SERV. ISS.):164–168.

Walsh, I., Martin, A. J., Di Domenico, T., Vullo, A., Pollastri, G., y Tosatto, S. C. (2011). *CSpritz: Accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs*. Nucleic Acids Research, 39(SUPPL. 2):190–196.

Walsh, I., Martin, A. J. M., Di Domenico, T., y Tosatto, S. C. E. (2012). *Espritz: Accurate and fast prediction of protein disorder*. Bioinformatics, 28(4):503–509.

Wang, H.-g. G., Moran, E., y Yaciuk, P. (1995). *E1A promotes association between p300 and pRB in multimeric complexes required for normal biological activity.* Journal of Virology, 69(12):7917–24.

Wang, H.-g. H., Draetta, G., y Moran, E. (1991). *E1A induces phosphorylation of the retinoblastoma protein independently of direct physical association between the E1A and retinoblastoma products*. Molecular and Cellular Biology, 11(8):4253–4265.

Wang, I.-n., Yeh, W.-B., y Lin, N.-S. (2017). *Phylogeography and Coevolution of Bamboo Mosaic Virus and Its Associated Satellite RNA.* Frontiers in Microbiology, 8(May):886.

Wang, J., Zhou, D., Prabhu, A., Schlegel, R., y Yuan, H. (2010). *The canine papillomavirus and gamma HPV E7 proteins use an alternative domain to bind and destabilize the retinoblastoma protein.* PLoS pathogens, 6(9):e1001089.

Wang, J. W. y Roden, R. B. S. (2013). *L2, the minor capsid protein of papillomavirus*. Virology, 445(1-2):175–186.

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., y Jones, D. T. (2004). *Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life*. Journal of Molecular Biology, 337(3):635–645.

Webster, L. C., Zhang, K., Chance, B., Ayene, I., Culp, J. S., Huang, W.-j., Wu, F. Y., y Ricciardi, R. P. (1991). *Conversion of the E1A Cys4 zinc finger to a nonfunctional His2,Cys2 zinc finger by a single point mutation.* Proceedings of the National Academy of Sciences of the United States of America, 88(22):9989–93.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., y Hwa, T. (2009). *Identification of direct residue contacts in protein-protein interaction by message passing.* Proceedings of the National Academy of Sciences of the United States of America, 106(1):67–72.

Whalen, S. G., Marcellus, R. C., Barbeau, D., y Branton, P. E. (1996). *Importance of the Ser-132 phosphorylation site in cell transformation and apoptosis induced by the adenovirus type 5 E1A protein.* Journal of Virology, 70(8):5373–83.

Whelan, S. y Goldman, N. (2001). *A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.* Molecular Biology and Evolution, 18(5):691–9.

Whyte, P., Buchkovich, K. J., Horowitz, J. M., Friend, S. H., Raybuck, M., Weinberg, R. A., y Harlow, E. (1988a). *Association between an oncogene and an anti-oncogene: the adenovirus E1A proteins bind to the retinoblastoma gene product.* Nature, 334(6178):124–9.

Whyte, P., Ruley, H. E., y Harlow, E. (1988b). *Two regions of the adenovirus early region 1A proteins are required for transformation.* Journal of Virology, 62(1):257–265.

Whyte, P., Williamson, N. M., y Harlow, E. (1989). *Cellular targets for transformation by the adenovirus E1A proteins*. Cell, 56(1):67–75.

Williams, J. F., Zhang, Y., Williams, M. A., Hou, S., Kushner, D., y Ricciardi, R. P. (2004). *E1A-Based Determinants of Oncogenicity in Human Adenovirus Groups A and C*. En *Current Topics in Microbiology and Immunology*, volume 273, pages 245–288.

Winberg, G. y Shenk, T. (1984). *Dissection of overlapping functions within the adenovirus type 5 E1A gene*. The EMBO Journal, 3(8):1907–1912.

Wright, P. E. y Dyson, H. J. (1999). *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.* Journal of molecular biology, 293(2):321–31.

Wright, P. E. y Dyson, H. J. (2015). *Intrinsically disordered proteins in cellular signalling and regulation.* Nature Reviews. Molecular Cell Biology, 16(1):18–29.

Wu, F. (1982). *The Potts model*. Reviews of Modern Physics, 54(1):235–268.

Wu, Z., Ren, X., Yang, L., Hu, Y., Yang, J., He, G., Zhang, J., Dong, J., Sun, L., Du, J., Liu, L., Xue, Y., Wang, J., Yang, F., Zhang, S., y Jin, Q. (2012). *Virome Analysis for Identification of Novel Mammalian Viruses in Bat Species from Chinese Provinces*. Journal of Virology, 86(20):10999–11012.

Xiao, B., Spencer, J., Clements, A., Ali-Khan, N., Mittnacht, S., Broceño, C., Burghammer, M., Perrakis, A., Marmorstein, R., y Gamblin, S. J. (2003). *Crystal structure of the retinoblastoma tumor suppressor protein bound to E2F and the molecular basis of its regulation.* Proceedings of the National Academy of Sciences of the United States of America, 100(5):2363–8.

Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., Uversky, V. N., y Obradovic, Z. (2007). *Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions.* Journal of Proteome Research, 6(5):1882–98.

Xie, Q., Arnold, G. E., Romero, P., Obradovic, Z., Garner, E., y Dunker, A. K. (1998). *The Sequence Attribute Method for Determining Relationships Between Sequence and Protein Disorder.* Genome Informatics. Workshop on Genome Informatics, 9:193–200.

Xue, B., Blocquel, D., Habchi, J., Uversky, A. V., Kurgan, L., Uversky, V. N., y Longhi, S. (2014). *Structural disorder in viral proteins.* Chemical Reviews, 114(13):6880–911.

Yaffe, M. B., Leparc, G. G., Lai, J., Obata, T., Volinia, S., y Cantley, L. C. (2001). *A motif-based profile scanning approach for genome-wide prediction of signaling pathways*. Nature Biotechnology, 19(4):348–353.

Yang, Z. (2007). *PAML 4: Phylogenetic analysis by maximum likelihood*. Molecular Biology and Evolution, 24(8):1586–1591.

Yang, Z., Kumar, S., y Nei, M. (1995). *A new method of inference of ancestral nucleotide and amino acid sequences*. Genetics, 141(4):1641–1650.

Yokose, N., Hirakawa, T., y Inokuchi, K. (2009). *Adenovirus-associated hemorrhagic cystitis in a patient with plasma cell myeloma treated with bortezomib*. Leukemia Research, 33(8):2009.

Zheng, Y., Stamminger, T., y Hearing, P. (2016). *E2F/Rb Family Proteins Mediate Interferon Induced Repression of Adenovirus Immediate Early Transcription to Promote Persistent Viral Infection*. PLoS Pathogens, 12(1):1–24.

Zheng, Z.-M. y Baker, C. C. (2006). *Papillomavirus genome structure, expression, and post-transcriptional regulation.* Frontiers in Bioscience : a journal and virtual library, 11:2286–302.

Zhou, H. X. (2003). *Quantitative account of the enhanced affinity of two linked scFvs specific for different epitopes on the same antigen*. Journal of Molecular Biology, 329(03):1–8.

Zhou, H. X. (2004). *Polymer Models of Protein Stability, Folding, and Interactions*. Biochemistry, 43(8):2141–2154.

Zhu, Y., Li, H., Long, C., Hu, L., Xu, H., Liu, L., Chen, S., Wang, D. C., y Shao, F. (2007). *Structural Insights into the Enzymatic Mechanism of the Pathogenic MAPK Phosphothreonine Lyase*. Molecular Cell, 28(5):899–913.

Zielinska, D. F., Gnad, F., Schropp, K., Wiśniewski, J. R., y Mann, M. (2012). *Mapping N-Glycosylation Sites across Seven Evolutionarily Distant Species Reveals a Divergent Substrate Proteome Despite a Common Core Machinery*. Molecular Cell, 46(4):542–548.

Zur Hausen, H. (2009). *The search for infectious causes of human cancers: where and why.* Virology, 392(1):1–10.

Apéndice A

Bases de datos

A.1. Secuencias de la proteína E7: Base de datos 1

En la Tabla A.1 se listan las secuencias de la proteína de E7 de la familia *Papillomaviridae* recolectadas. En las cinco primeras columnas se indica la taxonomía correspondiente y la abreviatura utilizada en el alineamiento para su identificación. La columna GI indica el identificador del gen codificante (en inglés, *Gene id*). Aquellos tipos virales para los cuales la región codificante para la proteína E7 se identificó manualmente a partir del genoma viral están identificados con una "g" al lado del identificador del gen (GI) y en la siguiente columna se indica el número de acceso (en inglés, *Accession Number*). Se indican también la especie y orden del hospedador del tipo viral. Por último, las columnas E7N y E7C indican si las secuencias correspondientes se incluyeron o no en el alineamiento del dominio E7N o E7C.

Tabla A.1: Lista de secuencias de la proteína E7 de la familia *Papillomaviridae* incluidas en la base de datos 1.

Género	Especie	Tipo	Abrv. Esp	Abrv. Tipo	GI	Nro de Acceso	Esp. del Hosp.	Orden del Hosp.	E7N	E7C
Alphapapillomavirus	*Alphapapillomavirus 1*	Human papillomavirus 32	Alpha1	HPV32	549284	P36827	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 1*	Human papillomavirus 42	Alpha1	HPV42	137800	P27231	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 2*	Human papillomavirus 3	Alpha2	HPV3	X74462.1 (g)	NA	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 2*	Human papillomavirus 10	Alpha2	HPV10	549275	P36818	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 2*	Human papillomavirus 28	Alpha2	HPV28	1718144	P50783	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 2*	Human papillomavirus 29	Alpha2	HPV29	1718145	P50784	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 2*	Human papillomavirus 77	Alpha2	HPV77	2911560	CAA75464.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 2*	Human papillomavirus 94	Alpha2	HPV94	57013120	Q705D1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 2*	Human papillomavirus 117	Alpha2	HPV117	256807726	ACV30143.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 2*	Human papillomavirus 125	Alpha2	HPV125	339639366	CBD35695.2	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 3*	Human papillomavirus 61	Alpha3	HPV61	3024825	Q80949	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 3*	Human papillomavirus 62	Alpha3	HPV62	39932601	AAR32247.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 3*	Human papillomavirus 72	Alpha3	HPV72	1491685	CAA63874.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 3*	Human papillomavirus 81	Alpha3	HPV81	40804511	CAF05693.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 3*	Human papillomavirus 83	Alpha3	HPV83	5059326	AAD38969.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 3*	Human papillomavirus 84	Alpha3	HPV84	12958169	AAK09272.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 3*	Human papillomavirus 86	Alpha3	HPV86	15741129	AAL06736.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 3*	Human papillomavirus 87	Alpha3	HPV87	14475580	CAC17713.2	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 3*	Human papillomavirus 89	Alpha3	HPV89	22095	AAM92152.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 3*	Human papillomavirus 102	Alpha3	HPV102	71726720	AAZ39521.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 3*	Human papillomavirus 114	Alpha3	HPV114	289656453	ADD14046.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 4*	Human papillomavirus 2	Alpha4	HPV2	125660688	ABN49465.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 4*	Human papillomavirus 2a	Alpha4	HPV2a	137794	P25485	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 4*	Human papillomavirus 27	Alpha4	HPV27	549282	P36825	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 4*	Human papillomavirus 27b	Alpha4	HPV27b	71061111	BAE16264.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 4*	Human papillomavirus 57	Alpha4	HPV57	137804	P22160	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 4*	Human papillomavirus 57b	Alpha4	HPV57b	U37537.1 (g)	NA	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 4*	Human papillomavirus 57c	Alpha4	HPV57c	157678965	BAF80481.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 5*	Human papillomavirus 26	Alpha5	HPV26	549281	P36824	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 5*	Human papillomavirus 51	Alpha5	HPV51	137803	P26558	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 5*	Human papillomavirus 69	Alpha5	HPV69	76363469	Q9JH50	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 5*	Human papillomavirus 82	Alpha5	HPV82	6970429	BAA90736.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 6*	Human papillomavirus 30	Alpha6	HPV30	549283	P36826	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 6*	Human papillomavirus 53	Alpha6	HPV53	549290	P36832	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 6*	Human papillomavirus 56	Alpha6	HPV56	549291	P36833	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 6*	Human papillomavirus 66	Alpha6	HPV66	3024844	Q80956	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 7*	Human papillomavirus 18	Alpha7	HPV18	137792	P06788	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 7*	Human papillomavirus 39	Alpha7	HPV39	137798	P24837	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 7*	Human papillomavirus 45	Alpha7	HPV45	549287	P21736	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 7*	Human papillomavirus 59	Alpha7	HPV59	557238	CAA54850.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 7*	Human papillomavirus 68	Alpha7	HPV68	1718147	P54668	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 7*	Human papillomavirus 68a	Alpha7	HPV68a	71726687	AAZ39492.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 7*	Human papillomavirus 68b	Alpha7	HPV68b	323669634	CBY85086.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 7*	Human papillomavirus 70	Alpha7	HPV70	1718148	P50785	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 7*	Human papillomavirus 85	Alpha7	HPV85	45747	AAD24182.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 7*	Human papillomavirus 97	Alpha7	HPV97	126738385	ABO27077.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 7*	Human papillomavirus Me180	Alpha7	HPVMe180	137808	P27963.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 8*	Human papillomavirus 7	Alpha8	HPV7	549273	P36816	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 8*	Human papillomavirus 40	Alpha8	HPV40	549286	P36829	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 8*	Human papillomavirus 43	Alpha8	HPV43	40804476	CAF05784.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 8*	Human papillomavirus 91	Alpha8	HPV91	22023570	AAM89131.1	*Homo sapiens*	Primates	SI	SI

continúa en la siguiente hoja

Género	Especie	Tipo	Abrv. Esp	Abrv. Tipo	GI	Nro de Acceso	Esp. del Hosp.	Orden del Hosp.	E7N	E7C
Alphapapillomavirus	*Alphapapillomavirus 9*	Human papillomavirus 16	Alpha9	HPV16	137791	P03129	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 9*	Human papillomavirus 31	Alpha9	HPV31	137795	P17387	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 9*	Human papillomavirus 33	Alpha9	HPV33	137796	P06429	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 9*	Human papillomavirus 35	Alpha9	HPV35	137797	P27230	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 9*	Human papillomavirus 35H	Alpha9	HPV35H	396999	CAA52562.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 9*	Human papillomavirus 52	Alpha9	HPV52	549289	P36831	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 9*	Human papillomavirus 58	Alpha9	HPV58	137805	P26557	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 9*	Human papillomavirus 67	Alpha9	HPV67	3228269	BAA66110.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 10*	Human papillomavirus 6	Alpha10	HPV6	6002619	AAF00065.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 10*	Human papillomavirus 6A	Alpha10	HPV6A	3024848	Q84292	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 10*	Human papillomavirus 6B	Alpha10	HPV6B	137807	P06464	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 10*	Human papillomavirus 11	Alpha10	HPV11	137790	P04020	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 10*	Human papillomavirus 13	Alpha10	HPV13	267300	Q02271	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 10*	Human papillomavirus 44	Alpha10	HPV44	3024819	Q80914	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 10*	Human papillomavirus 55	Alpha10	HPV55	3183193	Q80935	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 10*	Human papillomavirus 74	Alpha10	HPV74	1491799	AAC55128.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 10*	Pan paniscus papillomavirus 1	Alpha10	PpPV1	267301	Q02272	*Pan paniscus*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 10*	Pan troglodytes papillomavirus 1	Alpha10	PtPV1	9629722	NP_045012.1	*Pan troglodytes*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 11*	Human papillomavirus 34	Alpha11	HPV34	549285	P36828	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 11*	Human papillomavirus 73	Alpha11	HPV73	1491694	CAA63883.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Macaca fascicularis papillomavirus 3	Alpha12	MFPV3	161019496	ABX56066.1	*Macaca fascicularis*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Macaca fascicularis papillomavirus 3b	Alpha12	MFPV3b	156633499	ABU90822.1	*Macaca fascicularis*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Macaca fascicularis papillomavirus 4	Alpha12	MFPV4	161019516	ABX56084.1	*Macaca fascicularis*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Macaca fascicularis papillomavirus 5	Alpha12	MFPV5	161019536	ABX56102.1	*Macaca fascicularis*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Macaca fascicularis papillomavirus 6	Alpha12	MFPV6	161019509	ABX56078.1	*Macaca fascicularis*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Macaca fascicularis papillomavirus 7	Alpha12	MFPV7	161019486	ABX56057.1	*Macaca fascicularis*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Macaca fascicularis papillomavirus 8	Alpha12	MFPV8	161019526	ABX56093.1	*Macaca fascicularis*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Macaca fascicularis papillomavirus 9	Alpha12	MFPV9	186972296	ACC99409.1	*Macaca fascicularis*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Macaca fascicularis papillomavirus 10	Alpha12	MFPV10	186972286	ACC99400.1	*Macaca fascicularis*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Macaca fascicularis papillomavirus 11	Alpha12	MFPV11	253558908	ACT32129.1	*Macaca fascicularis*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Macaca mulata papillomavirus 1	Alpha12	MmPV1	137811	P22161	*Macacca mulata*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Rhesus macaque papillomavirus 1	Alpha12	RhPV1	256860449	ACV32159.1	*Macacca mulata*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 13*	Human papillomavirus 54	Alpha13	HPV54	3024830	Q81019	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 14*	Colobus guereza papillomavirus 1	Alpha14	CgPV1	320526312	ADW41702.1	*Colobus guereza*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 14*	Human papillomavirus 71	Alpha14	HPV71	37622203	AAQ95199.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 14*	Human papillomavirus 90	Alpha14	HPV90	22138124	NP_671504.1	*Homo sapiens*	Primates	SI	SI
Alphapapillomavirus	*Alphapapillomavirus 14*	Human papillomavirus 106	Alpha14	HPV106	71726712	AAZ39514.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Colobus guereza papillomavirus 2	Beta1	CgPV2	338209367	YP_004646339.1	*Colobus guereza*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 5	Beta1	HPV5	137788	P06932	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 5B	Beta1	HPV5B	137806	P26559	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 8	Beta1	HPV8	137789	P06430	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 12	Beta1	HPV12	549276	P36819	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 14	Beta1	HPV14	808880	BAA09114.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 19	Beta1	HPV19	549279	P36822	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 20	Beta1	HPV20	1718139	P50778	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 21	Beta1	HPV21	1718140	P50779	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 24	Beta1	HPV24	1718143	P50782.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 25	Beta1	HPV25	549280	P36823	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 36	Beta1	HPV36	1718146	P50811	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 47	Beta1	HPV47	137802	P22423	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 93	Beta1	HPV93	37089394	AAQ88282.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 98	Beta1	HPV98	238623429	CAW42214.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 99	Beta1	HPV99	238623437	CAW42227.1	*Homo sapiens*	Primates	SI	SI

continúa en la siguiente hoja

Género	Especie	Tipo	Abrv. Esp	Abrv. Tipo	GI	Nro de Acceso	Esp. del Hosp.	Orden del Hosp.	E7N	E7C
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 105	Beta1	HPV105	238623460	CAW42261.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 118	Beta1	HPV118	256807733	ACV30149.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 124	Beta1	HPV124	296495870	ADH29827.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus RTRX7	Beta1	HPVRTRX7	2218081	AAB61641.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 1*	Macaca fascicularis papillomavirus 1	Beta1	MFPV1	122891972	ABM67066.1	*Macaca fascicularis*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 9	Beta2	HPV9	549274	P36817	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 15	Beta2	HPV15	549277	P36820	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 17	Beta2	HPV17	549278	P36821	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 22	Beta2	HPV22	1718141	P50780	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 23	Beta2	HPV23	1718142	P50781	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 37	Beta2	HPV37	3024816	Q80901	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 38	Beta2	HPV38	3024818	Q80908	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 38b	Beta2	HPV38b	68304488	AAY89820.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 80	Beta2	HPV80	2911567	CAA75471.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 100	Beta2	HPV100	238623444	CAW42236.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 104	Beta2	HPV104	238623452	CAW42248.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 107	Beta2	HPV107	126131395	ABN79868.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 110	Beta2	HPV110	186461196	ACC78264.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 111	Beta2	HPV111	186461204	ACC78271.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 113	Beta2	HPV113	238623467	CAW42271.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 120	Beta2	HPV120	296495838	ADH29799.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 122	Beta2	HPV122	296495854	ADH29813.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 151	Beta2	HPV151	290964863	CBK38954.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus FA75/KI88	Beta2	HPVFA75/KI88	186461188	ACC78257.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus SIBX-3a	Beta2	HPVSIBX-3a	270119450	CBI67299.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 3*	Human papillomavirus 49	Beta3	HPV49	549288	P36830	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 3*	Human papillomavirus 75	Beta3	HPV75	2911546	CAA75450.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 3*	Human papillomavirus 76	Beta3	HPV76	2911553	CAA75457.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 3*	Human papillomavirus 115	Beta3	HPV115	270048214	ACZ58404.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 4*	Human papillomavirus 92	Beta4	HPV92	27531788	NP_775306.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 5*	Human papillomavirus 96	Beta5	HPV96	37595912	NP_932320.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 5*	Human papillomavirus 150	Beta5	HPV150	290964855	CBK38947.1	*Homo sapiens*	Primates	SI	SI
Betapapillomavirus	*Betapapillomavirus 6*	Macaca fascicularis papillomavirus 2	Beta6	MFPV2	338209359	YP_004646332.1	*Macaca fascicularis*	Primates	SI	SI
Chipapillomavirus	*Chipapillomavirus 1*	Canis familiaris papillomavirus 3	Chi1	CPV3	83314074	ABC02204.1	*Canis familiaris*	Carnivora	SI	SI
Chipapillomavirus	*Chipapillomavirus 1*	Canis familiaris papillomavirus 5	Chi1	CPV5	258611056	YP_003204669.1	*Canis familiaris*	Carnivora	SI	SI
Chipapillomavirus	*Chipapillomavirus 2*	Canis familiaris papillomavirus 4	Chi2	CPV4	156601579	ABU86869.1	*Canis familiaris*	Carnivora	SI	SI
Deltapapillomavirus	*Deltapapillomavirus 1*	Alces alces papillomavirus 1	Delta1	AaPV1	137810	P11332.1	*Alces alces*	Artiodactyla	NO	SI
Deltapapillomavirus	*Deltapapillomavirus 1*	Rangifer tarandus papillomavirus 1	Delta1	RtPV1	22901247	AAN09916.1	*Rangifer tarandus*	Artiodactyla	NO	SI
Deltapapillomavirus	*Deltapapillomavirus 2*	Odocoileus virginianus papillomavirus 1	Delta2	OvPV1	1352844	P03131.2	*Odocoileus virginianus*	Artiodactyla	NO	SI
Deltapapillomavirus	*Deltapapillomavirus 4*	Bos taurus papillomavirus 1	Delta4	BPV1	137784	P06933.1	*Bos taurus*	Artiodactyla	NO	SI
Deltapapillomavirus	*Deltapapillomavirus 4*	Bos taurus papillomavirus 2	Delta4	BPV2	137785	P11303.1	*Bos taurus*	Artiodactyla	NO	SI
Deltapapillomavirus	*Deltapapillomavirus 5*	Capreolus capreolus papillomavirus 1	Delta5	CcaPV1	194268066	YP_002004568.1	*Capreolus capreolus*	Artiodactyla	NO	SI
Deltapapillomavirus	*Deltapapillomavirus UNK*	Bos taurus papillomavirus	DeltaUNK	BPV	238914480	ACR78109.1	*Bos taurus*	Artiodactyla	NO	SI
Deltapapillomavirus	*Deltapapillomavirus UNK*	Camelus dromedarius papillomavirus 1	DeltaUNK	CdPV1	326631678	YP_004306466.1	*Camelus dromedarius*	Artiodactyla	NO	SI
Deltapapillomavirus	*Deltapapillomavirus UNK*	Camelus dromedarius papillomavirus 2	DeltaUNK	CdPV2	326631688	YP_004306473.1	*Camelus dromedarius*	Artiodactyla	NO	SI
DyoEpsilonpapillomavirus	*DyoEpsilonpapillomavirus 1*	Francolinus leucoscepus papillomavirus 1	DyoEpsilon1	FlPV1	256352176	YP_003104798.1	*Francolinus leucoscepus*	Galliformes	NO	NO
DyoEtapapillomavirus	*DyoEtapapillomavirus 1*	Erinaceus europaeus papillomavirus 1	DyoEta1	EePV1	218685639	YP_002427691.1	*Erinaceus europaeus*	Erinaceomorpha	SI	SI
DyoIotapapillomavirus	*DyoIotapapillomavirus 1*	Equus ferus caballus papillomavirus 2	DyoIota1	EcPV2	187938333	ACD38212.1	*Equus ferus caballus*	Perissodactyla	SI	SI
DyoThetapapillomavirus	*DyoThetapapillomavirus 1*	Felis domesticus papillomavirus 2	DyoTheta1	FdPV2	209402851	ACI46049.1	*Felis domesticus*	Carnivora	SI	SI
DyoZetapapillomavirus	*DyoZetapapillomavirus 1*	Caretta caretta papillomavirus 1	DyoZeta1	CcPV1	212286050	YP_002308358.1	*Caretta caretta*	Testudines	NO	NO
DyoZetapapillomavirus	*DyoZetapapillomavirus 1*	Chelonia mydas papillomavirus 1	DyoZeta1	CmPV1	187941628	ACD39806.1	*Chelonia mydas*	Testudines	NO	NO
Epsilonpapillomavirus	*Epsilonpapillomavirus 1*	Bos taurus papillomavirus 5	Epsilon1	BPV5	57013118	Q8BDG6.1	*Bos taurus*	Artiodactyla	NO	SI
Epsilonpapillomavirus	*Epsilonpapillomavirus 1*	Bos taurus papillomavirus 8	Epsilon1	BPV8	68160304	AAY86737.1	*Bos taurus*	Artiodactyla	NO	SI

continúa en la siguiente hoja

Género	Especie	Tipo	Abrv. Esp	Abrv. Tipo	GI	Nro de Acceso	Esp. del Hosp.	Orden del Hosp.	E7N	E7C
Etapapillomavirus	*Etapapillomavirus 1*	Fringilla coelebs papillomavirus	Eta1	FcPV	21844536	NP_663762.1	*Fringilla coelebs*	Passeriformes	NO	NO
Gammapapillomavirus	*Gammapapillomavirus 1*	Human papillomavirus 4	Gamma1	HPV4	586228	Q07857.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 1*	Human papillomavirus 65	Gamma1	HPV65	586230	Q07859.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 1*	Human papillomavirus 95	Gamma1	HPV95	40804522	CAF05703.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 2*	Human papillomavirus 48	Gamma2	HPV48	3024821	Q80921.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 3*	Human papillomavirus 50	Gamma3	HPV50	3183191	Q80928.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 3*	Human papillomavirus 131	Gamma3	HPV131	312451802	ADQ85958.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 4*	Human papillomavirus 60	Gamma4	HPV60	3024823	Q80942.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 5*	Human papillomavirus 88	Gamma5	HPV88	167600367	YP_001672009.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 6*	Human papillomavirus 101	Gamma6	HPV101	109390390	YP_656499.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 6*	Human papillomavirus 103	Gamma6	HPV103	109390383	YP_656493.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 6*	Human papillomavirus 108	Gamma6	HPV108	224983323	YP_002647034.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 6*	Human papillomavirus 128	Gamma6	HPV128	313755947	ADR77926.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 7*	Human papillomavirus 109	Gamma7	HPV109	225927562	YP_002756539.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 7*	Human papillomavirus 123	Gamma7	HPV123	296495862	ADH29820.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 7*	Human papillomavirus 134	Gamma7	HPV134	312451826	ADQ85979.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 7*	Human papillomavirus 149	Gamma7	HPV149	312451786	ADQ85944.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 8*	Human papillomavirus 112	Gamma8	HPV112	225927570	YP_002756546.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 8*	Human papillomavirus 119	Gamma8	HPV119	296495830	ADH29792.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 9*	Human papillomavirus 116	Gamma9	HPV116	254810665	YP_003084347.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 9*	Human papillomavirus 129	Gamma9	HPV129	313755955	ADR77933.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 10*	Human papillomavirus 121	Gamma10	HPV121	297342358	YP_003668026.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 10*	Human papillomavirus 130	Gamma10	HPV130	312451794	ADQ85951.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 10*	Human papillomavirus 133	Gamma10	HPV133	312451818	ADQ85972.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 11*	Human papillomavirus 127	Gamma11	HPV127	293596088	ADE45483.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 11*	Human papillomavirus 132	Gamma11	HPV132	312451810	ADQ85965.1	*Homo sapiens*	Primates	SI	SI
Gammapapillomavirus	*Gammapapillomavirus 11*	Human papillomavirus 148	Gamma11	HPV148	317046149	YP_004111310.1	*Homo sapiens*	Primates	SI	SI
Iotapapillomavirus	*Iotapapillomavirus 1*	Mastomys natalensis papillomavirus 1	Iota1	MnPV1	9627488	NP_042015.1	*Mastomys natalensis*	Rodentia	SI	SI
Kappapapillomavirus	*Kappapapillomavirus 1*	Oryctolagus cuniculus papillomavirus 1	Kappa1	OcPV1	9635135	NP_057842.1	*Oryctolagus cuniculus*	Lagomorpha	SI	SI
Kappapapillomavirus	*Kappapapillomavirus 2*	Sylvilagus floridanus papillomavirus 1k	Kappa2	SfPV1k	137787	P03130.1	*Sylvilagus floridanus*	Lagomorpha	SI	SI
Kappapapillomavirus	*Kappapapillomavirus 2*	Sylvilagus floridanus papillomavirus 1w	Kappa2	SfPV1w	1718138	P51895.1	*Sylvilagus floridanus*	Lagomorpha	SI	SI
Lambdapapillomavirus	*Lambdapapillomavirus 1*	Felis domesticus papillomavirus 1	Lambda1	FdPV1	19224324	AAL86454.1	*Felis domesticus*	Carnivora	SI	SI
Lambdapapillomavirus	*Lambdapapillomavirus 1*	Lynx rufus papillomavirus 1	Lambda1	LrPV1	62547899	AAX86613.1	*Lynx rufus*	Carnivora	SI	SI
Lambdapapillomavirus	*Lambdapapillomavirus 1*	Panthera leo persica papillomavirus 1	Lambda1	PlpPV1	62547915	AAX86627.1	*Panthera leo persica*	Carnivora	SI	SI
Lambdapapillomavirus	*Lambdapapillomavirus 1*	Puma concolor papillomavirus 1	Lambda1	PcPV1	62547907	AAX86620.1	*Puma concolor*	Carnivora	SI	SI
Lambdapapillomavirus	*Lambdapapillomavirus 1*	Uncia uncia papillomavirus 1	Lambda1	UuPV1	77157777	ABA61872.1	*Uncia uncia*	Carnivora	SI	SI
Lambdapapillomavirus	*Lambdapapillomavirus 2*	Canis familiaris papillomavirus 1	Lambda2	CPV1	3915239	Q89759.1	*Canis familiaris*	Carnivora	SI	SI
Lambdapapillomavirus	*Lambdapapillomavirus 3*	Canis familiaris papillomavirus 6	Lambda3	CPV6	258611064	YP_003204676.1	*Canis familiaris*	Carnivora	SI	SI
Lambdapapillomavirus	*Lambdapapillomavirus 4*	Procyon lotor papillomavirus 1	Lambda4	PlPV1	68304290	YP_249599.1	*Procyon lotor*	Carnivora	SI	SI
Mupapillomavirus	*Mupapillomavirus 1*	Human papillomavirus 1a	Mu1	HPV1a	137793	P06465.1	*Homo sapiens*	Primates	SI	SI
Mupapillomavirus	*Mupapillomavirus 2*	Human papillomavirus 63	Mu2	HPV63	586229	Q07858.1	*Homo sapiens*	Primates	SI	SI
No Clasificadas	*No Clasificadas*	Bettongia penicillata papillomavirus 1	UNK	BPP1	296040253	YP_003622564.1	*Bettongia penicillata*	Diprotodontia	SI	SI
No Clasificadas	*No Clasificadas*	Bos taurus papillomavirus 7	UNK	BPV7	82547782	YP_406558.1	*Bos taurus*	Artiodactyla	NO	SI
No Clasificadas	*No Clasificadas*	Equus ferus caballus papillomavirus 3	UNK	EcPV3	317135038	ADV03082.1	*Equus ferus caballus*	Perissodactyla	SI	SI
No Clasificadas	*No Clasificadas*	Mus musculus papillomavirus 1	UNK	MMPV1	295924010	ADG63120.1	*Mus musculus*	Rodentia	SI	SI
No Clasificadas	*No Clasificadas*	Ovis aries papillomavirus 3	UNK	OPV3	226446759	ACO58657.1	*Ovis aries*	Artiodactyla	SI	SI
No Clasificadas	*No Clasificadas*	Zalophus californianus papillomavirus 1	UNK	ZcPV1	332288079	YP_004346963.1	*Zalophus californianus*	Carnivora	SI	SI
Nupapillomavirus	*Nupapillomavirus 1*	Human papillomavirus 41	Nu1	HPV41	137799	P27556.1	*Homo sapiens*	Primates	SI	SI
Phipapillomavirus	*Phipapillomavirus 1*	Capra hircus papillomavirus 1	Phi1	ChPV1	97331428	YP_610954.1	*Capra hircus*	Artiodactyla	SI	SI
Pipapillomavirus	*Pipapillomavirus 1*	Mastomys coucha papillomavirus 2	Pi1	McPV2	116536730	YP_803389.1	*Mastomys coucha*	Rodentia	SI	SI
Pipapillomavirus	*Pipapillomavirus 1*	Micromys minutus papillomavirus 1	Pi1	MmiPV1	118129782	YP_873940.1	*Micromys minutus*	Rodentia	SI	SI
Pipapillomavirus	*Pipapillomavirus 1*	Rattus novergicus papillomavirus 1	Pi1	RnPV1	257136428	YP_003169701.1	*Rattus novergicus*	Rodentia	SI	SI
Psipapillomavirus	*Psipapillomavirus 1*	Rousettus aegyptiacus papillomavirus 1	Psi1	RaPV1	113200765	YP_717908.1	*Rousettus aegyptiacus*	Chiroptera	SI	SI

continúa en la siguiente hoja

Género	Especie	Tipo	Abrv. Esp	Abrv. Tipo	GI	Nro de Acceso	Esp. del Hosp.	Orden del Hosp.	E7N	E7C
Rhopapillomavirus	*Rhopapillomavirus 1*	Trichechus manatus latirostris papillomavirus 1	Rho1	FmPV1	56567503	AAV98686.1	*Trichechus manatus latirostris*	Sirenia	SI	SI
Rhopapillomavirus	*Rhopapillomavirus 1*	Trichechus manatus manatus papillomavirus 1	Rho1	TmPV1	56698750	YP_164622.1	*Trichechus manatus manatus*	Sirenia	SI	SI
Sigmapapillomavirus	*Sigmapapillomavirus 1*	Erethizon dorsatum papillomavirus 1	Sigma1	EdPV1	62362148	YP_224222.1	*Erethizon dorsatum*	Rodentia	SI	SI
Taupapillomavirus	*Taupapillomavirus 1*	Canis familiaris papillomavirus 2	Tau1	CPV2	56693038	YP_164629.1	*Canis familiaris*	Carnivora	SI	SI
Taupapillomavirus	*Taupapillomavirus 1*	Canis familiaris papillomavirus 7	Tau1	CPV7	255683758	ACU27441.1	*Canis familiaris*	Carnivora	SI	SI
Thetapapillomavirus	*Thetapapillomavirus 1*	Psittacus erithacus papillomavirus	Theta1	PePV	21693251	AAM75200.1	*Psittacus erithacus*	Psittaciformes	NO	NO
Xipapillomavirus	*Xipapillomavirus 1*	Bos taurus papillomavirus 3	Xi1	BPV3	57013114	Q8BDD8.1	*Bos taurus*	Artiodactyla	SI	SI
Xipapillomavirus	*Xipapillomavirus 1*	Bos taurus papillomavirus 4	Xi1	BPV4	187608846	P08350.2	*Bos taurus*	Artiodactyla	SI	SI
Xipapillomavirus	*Xipapillomavirus 1*	Bos taurus papillomavirus 6	Xi1	BPV6	57013108	Q705F5.1	*Bos taurus*	Artiodactyla	SI	SI
Xipapillomavirus	*Xipapillomavirus 1*	Bos taurus papillomavirus 9	Xi1	BPV9	163914076	BAF95809.1	*Bos taurus*	Artiodactyla	SI	SI
Xipapillomavirus	*Xipapillomavirus 1*	Bos taurus papillomavirus 10	Xi1	BPV10	163914083	BAF95815.1	*Bos taurus*	Artiodactyla	SI	SI
Xipapillomavirus	*Xipapillomavirus 1*	Bos taurus papillomavirus 11	Xi1	BPV11	301335133	BAJ12073.1	*Bos taurus*	Artiodactyla	SI	SI
Xipapillomavirus	*Xipapillomavirus 1*	Bos taurus papillomavirus AA5	Xi1	BPVAA5	169641578	ACA61500.1	*Bos taurus*	Artiodactyla	SI	SI
Zetapapillomavirus	*Zetapapillomavirus 1*	Equus ferus caballus papillomavirus 1	Zeta1	EcPV1	20428630	NP_620508.1	*Equus ferus caballus*	Perissodactyla	SI	SI

A.2. Secuencias de la proteína E7: Base de datos 2

En la Tabla A.2 se listan las secuencias de la proteína de E7 de la familia *Papillomaviridae* agregadas en la base de datos en abril de 2017 actualizando la base de datos 1 (Sección A.1). En las cinco primeras columnas se indica la taxonomía correspondiente y la abreviatura utilizada en el alineamiento para su identificación. La columna GI indica el identificador del gen codificante (en inglés, *Gene id*). Aquellos tipos virales para los cuales la región codificante para la proteína E7 fue identificada manualmente a partir del genoma viral están identificados con una "g" al lado del identificador del gen (GI) y en la siguiente columna se indica el número de acceso (en inglés, *Accession Number*). Se indica también el hospedador del tipo viral. Por último, la columna alineamiento indica si la secuencia correspondiente se incluyó o no en el alineamiento.

Tabla A.2: Lista de secuencias de la proteína E7 de la familia *Papillomaviridae* incluidas en la base de datos 1.

Género	Especie	Tipo	Abrv. Esp	Abrv. Tipo	GI	Nro de Acceso	Hospedador	Alineamiento
Alphapapillomavirus	*Alphapapillomavirus 2*	Human papillomavirus 78	Alpha2	HPV78	629266011	BAO74123.1	*Homo sapiens*	SI
Alphapapillomavirus	*Alphapapillomavirus 2*	Human papillomavirus XS2	Alpha2	HPV-mXS2	444303711	AGD99638.1	*Homo sapiens*	SI
Alphapapillomavirus	*Alphapapillomavirus UNK*	Macaca fuscata papillomavirus 1	AlphaUNK	MfuPV1	1011764716	AMR98485.1	*Macaca fuscata*	SI
Alphapapillomavirus	*Alphapapillomavirus 12*	Papio hamadryas papillomavirus 1	Alpha12	PhPV1	386576352	YP_006202689.1	*Papio hamadryas*	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 143	Beta1	HPV143	343411579	AEM24650.1	*Homo sapiens*	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 145	Beta2	HPV145	343411595	AEM24664.1	*Homo sapiens*	SI
Betapapillomavirus	*Betapapillomavirus 1*	Human papillomavirus 152	Beta1	HPV152	327195189	AEA35074.1	*Homo sapiens*	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 159	Beta2	HPV159	395627597	CCJ27716.1	*Homo sapiens*	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 174	Beta2	HPV174	469662961	CCV02859.1	*Homo sapiens*	SI
Betapapillomavirus	*Betapapillomavirus 2*	Human papillomavirus 209	Beta2	HPV209	1150189781	AQT33338.1	*Homo sapiens*	SI
Chipapillomavirus	*Chipapillomavirus 3*	Canis familiaris papillomavirus 8	Chi3	CPV8	347750426	YP_004857843.1	*Canis familiaris*	SI
Chipapillomavirus	*Chipapillomavirus 1*	Canis familiaris papillomavirus 9	Chi1	CPV9	363540890	YP_004895374.1	*Canis familiaris*	SI
Chipapillomavirus	*Chipapillomavirus 3*	Canis familiaris papillomavirus 10	Chi3	CPV10	363540898	YP_004895381.1	*Canis familiaris*	SI
Chipapillomavirus	*Chipapillomavirus 1*	Canis familiaris papillomavirus 12	Chi1	CPV12	388542471	AFK65662.1	*Canis familiaris*	SI
Chipapillomavirus	*Chipapillomavirus 3*	Canis familiaris papillomavirus 14	Chi3	CPV14	430025792	YP_007195265.1	*Canis familiaris*	SI
Chipapillomavirus	*Chipapillomavirus 3*	Canis familiaris papillomavirus 15	Chi3	CPV15	429841974	AGA15832.1	*Canis familiaris*	SI
Chipapillomavirus	*Chipapillomavirus 2*	Canis familiaris papillomavirus 16	Chi2	CPV16	765702650	YP_009126907.1	*Canis familiaris*	SI
Chipapillomavirus	*Chipapillomavirus UNK*	Canis familiaris papillomavirus 18	ChiUNK	CPV18	1046841334	ANW12190.1	*Canis familiaris*	SI
Chipapillomavirus	*Chipapillomavirus UNK*	Canis familiaris papillomavirus 20	ChiUNK	CPV20	1008264058	AMQ81154.1	*Canis familiaris*	SI
Deltapapillomavirus	*Deltapapillomavirus 4*	Bos grunniens papillomavirus 1	Delta4	BgPV1	402694783	AFQ90263.1	*Bos grunniens*	NO
Deltapapillomavirus	*Deltapapillomavirus 4*	Bos taurus papillomavirus 13	Delta4	BPV13	1059198210	YP_009272573.1	*Bos taurus*	NO
Deltapapillomavirus	*Deltapapillomavirus 4*	Bovine papillomavirus 13	Delta4	BPV13a	685165390	AIN81125.1	*Bos taurus*	NO
Deltapapillomavirus	*Deltapapillomavirus 4*	Bovine papillomavirus 14	Delta4	BPV14	943351594	ALL29331.1	*Bos taurus*	NO
Deltapapillomavirus	*Deltapapillomavirus UNK*	Giraffa camelopardalis papillomavirus 1	DeltaUNK	GcPV1	1109485633	APG30980.1	*Giraffa camelopardalis*	SI
Deltapapillomavirus	*Deltapapillomavirus 3*	Ovis aries papillomavirus 2	Delta3	OaPV2	U83595.1_2 (g)	U83595.1_2 (g)	*Ovis aries*	NO
Dyochipapillomavirus	*Dyochipapillomavirus 1*	Equus asinus papillomavirus 1	Dyochi1	EaPV1	605039094	YP_009021882.1	*Equus Asinus*	SI
Dyoiotapapillomavirus	*Dyoiotapapillomavirus 2*	Equus caballus papillomavirus 4	Dyoiota2	EcPV4	448261105	YP_007349390.1	*Equus ferus caballus*	SI
Dyoiotapapillomavirus	*Dyoiotapapillomavirus 2*	Equus caballus papillomavirus 5	Dyoiota2	EcPV5	448261097	YP_007349383.1	*Equus ferus caballus*	SI
Dyokappapapillomavirus	*Dyokappapapillomavirus UNK*	Bovine papillomavirus 16	DyokappaUNK	BPV16	1059198222	YP_009272581.1	*Bos taurus*	SI
Dyokappapapillomavirus	*Dyokappapapillomavirus UNK*	Bovine papillomavirus 18	DyokappaUNK	BPV18	1059198233	YP_009272594.1	*Bos taurus*	SI
Dyokappapapillomavirus	*Dyokappapapillomavirus 5*	Pudu puda papillomavirus 1	DyoKappa5	PpuPV1	954539444	ALP46952.1	*Pudu puda*	SI
Dyokappapapillomavirus	*Dyokappapapillomavirus 2*	Rupicapra rupicapra papillomavirus 1	Dyokappa2	RrPV1	607064667	YP_009022079.1	*Rupicapra rupicapra*	SI
Dyomupapillomavirus	*Dyomupapillomavirus 1*	Morelia spilota papillomavirus 1	Dyomu1	MsPV1	347750418	YP_004857836.1	*Morelia spilota*	SI
Dyoomegapapillomavirus	*Dyoomegapapillomavirus 1*	Eptesicus serotinus papillomavirus 2	Dyoomega1	EsPV2	586831313	AHJ81391.1	*Eptesicus serotinus*	SI
Dyoomikronpapillomavirus	*Dyoomikronpapillomavirus UNK*	Alouatta guariba papillomavirus 1	DyoomikronUNK	AgPV1	982957560	AMB19786.1	*Alouatta guariba*	SI
Dyoomikronpapillomavirus	*Dyoomikronpapillomavirus 1*	Saimiri sciureus papillomavirus 1	Dyoomikron1	SscPV1	586946449	YP_009002598.1	*Saimiri sciureus*	SI
Dyoomikronpapillomavirus	*Dyoomikronpapillomavirus 1*	Saimiri sciureus papillomavirus 2	Dyoomikron1	SscPV2	327195175	AEA35062.1	*Saimiri sciureus*	SI
Dyoomikronpapillomavirus	*Dyoomikronpapillomavirus 1*	Saimiri sciureus papillomavirus 3	Dyoomikron1	SscPV3	327195182	AEA35068.1	*Saimiri sciureus*	SI
Dyophipapillomavirus	*Dyophipapillomavirus 1*	Talpa europaea papillomavirus 1	Dyophi1	TePV1	507864203	AGM75113.1	*Talpa europaea*	SI
Dyopsipapillomavirus	*Dyopsipapillomavirus 1*	Eptesicus serotinus papillomavirus 1	Dyopsi1	EsPV1	586831306	AHJ81385.1	*Eptesicus serotinus*	SI
Dyopsipapillomavirus	*Dyopsipapillomavirus 1*	Eptesicus serotinus papillomavirus 3	Dyopsi1	EsPV3	586831320	AHJ81397.1	*Eptesicus serotinus*	SI
Dyorhopapillomavirus	*Dyorhopapillomavirus 1*	Equus caballus papillomavirus 6	Dyorho1	EcPV6	470457105	YP_007518492.1	*Equus ferus caballus*	SI
Dyorhopapillomavirus	*Dyorhopapillomavirus 1*	Equus caballus papillomavirus 7	Dyorho1	EcPV7	470457145	YP_007518499.1	*Equus ferus caballus*	SI
Dyosigmapapillomavirus	*Dyosigmapapillomavirus 1*	Castor canadensis papillomavirus 1	Dyosigma1	CcanPV1	570364814	YP_008992241.1	*Castor canadensis*	SI
Dyotaupapillomavirus	*Dyotaupapillomavirus 1*	Miniopterus schreibersii papillomavirus 1	Dyotau1	MscPV1	389568554	AFK84996.1	*Miniopterus schreibersii*	SI
Dyoupsilonpapillomavirus	*Dyoupsilonpapillomavirus 1*	Eidolon helvum papillomavirus 1	Dyoupsilon1	EhPV1	433335595	AGB34176.1	*Eidolon helvum*	SI
Epsilonpapillomavirus	*Epsilonpapillomavirus UNK*	Cervus elaphus papillomavirus 1	EpsilonUNK	CePV1	388460891	AFK32234.1	*Cervus elaphus*	SI
Epsilonpapillomavirus	*Epsilonpapillomavirus UNK*	Cervus papillomavirus 2	EpsilonUNK	CePV2a	973414555	ALX18687.1	*Cervus elaphus*	NO

continúa en la siguiente hoja

Género	Especie	Tipo	Abrv. Esp	Abrv. Tipo	GI	Nro de Acceso	Hospedador	Alineamiento
Gammapapillomavirus	*Gammapapillomavirus 11*	Human papillomavirus 126	Gamma11	HPV126	358356462	YP_004934013.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 15*	Human papillomavirus 135	Gamma15	HPV135	389656402	YP_006393282.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 11*	Human papillomavirus 136	Gamma11	HPV136	389656410	YP_006393289.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 16*	Human papillomavirus 137	Gamma16	HPV137	389656418	YP_006393296.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 7*	Human papillomavirus 139	Gamma7	HPV139	343411547	AEM24622.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 11*	Human papillomavirus 140	Gamma11	HPV140	389656426	YP_006393303.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 11*	Human papillomavirus 141	Gamma11	HPV141	343411563	AEM24636.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 17*	Human papillomavirus 144	Gamma17	HPV144	389656434	YP_006393310.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 15*	Human papillomavirus 146	Gamma15	HPV146	343411603	AEM24671.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 8*	Human papillomavirus 147	Gamma8	HPV147	343411611	AEM24678.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 13*	Human papillomavirus 153	Gamma13	HPV153	356483526	AET11873.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 11*	Human papillomavirus 154	Gamma11	HPV154	512721868	YP_008083732.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 7*	Human papillomavirus 155	Gamma7	HPV155	353441722	AEQ98806.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 18*	Human papillomavirus 156	Gamma18	HPV156	1149685231	YP_009345873.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 19*	Human papillomavirus 161	Gamma19	HPV161	409183137	AFV27123.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 19*	Human papillomavirus 162	Gamma19	HPV162	409183131	AFV27118.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 20*	Human papillomavirus 163	Gamma20	HPV163	944326231	YP_009175015.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 8*	Human papillomavirus 164	Gamma8	HPV164	409183115	AFV27104.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 12*	Human papillomavirus 165	Gamma12	HPV165	409183151	AFV27135.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 19*	Human papillomavirus 166	Gamma19	HPV166	410493705	YP_006908973.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 21*	Human papillomavirus 167	Gamma21	HPV167	559797728	YP_008828146.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 8*	Human papillomavirus 168	Gamma8	HPV168	557825768	AHA37342.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 11*	Human papillomavirus 169	Gamma11	HPV169	409183107	AFV27097.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 7*	Human papillomavirus 170	Gamma7	HPV170	409183144	AFV27129.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 11*	Human papillomavirus 171	Gamma11	HPV171	564732518	AHC00338.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 22*	Human papillomavirus 172	Gamma22	HPV172	564732526	AHC00345.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 1*	Human papillomavirus 173	Gamma1	HPV173	564732534	AHC00352.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 23*	Human papillomavirus 175	Gamma23	HPV175	443498372	AGC93429.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 24*	Human papillomavirus 178	Gamma24	HPV178	607064614	YP_009022058.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 15*	Human papillomavirus 179	Gamma15	HPV179	530787709	YP_008433327.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 10*	Human papillomavirus 180	Gamma10	HPV180	443498380	AGC93436.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 25*	Human papillomavirus 184	Gamma25	HPV184	551364094	CDI44918.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 24*	Human papillomavirus 197	Gamma24	HPV197	677569935	AIM47229.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 12*	Human papillomavirus 199	Gamma12	HPV199	929996747	ALF36870.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 2*	Human papillomavirus 200	Gamma2	HPV200	870702418	AKP16335.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 27*	Human papillomavirus 201	Gamma27	HPV201	896684340	YP_009158853.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus 11*	Human papillomavirus 202	Gamma11	HPV202	870702436	AKP16348.1	*Homo sapiens*	SI
Gammapapillomavirus	*Gammapapillomavirus UNK*	Human papillomavirus mKC5	GammaUNK	HPV-mKC5	817524634	YP_009134744.1	*Homo sapiens*	SI
Iotapapillomavirus	*Iotapapillomavirus UNK*	Peromyscus papillomavirus 1	IotaUNK	PmPV1	343196967	AEM05817.1	*Peromyscus*	SI
Iotapapillomavirus	*Iotapapillomavirus 1*	Rattus norvegicus papillomavirus 2	Iota1	RnPV2	336092143	AEI00717.1	*Rattus norvegicus*	SI
Iotapapillomavirus	*Iotapapillomavirus 1*	Rattus norvegicus papillomavirus 3	Iota1	RnPV3	959121884	YP_009182325.1	*Rattus norvegicus*	SI
Lambdapapillomavirus	*Lambdapapillomavirus 5*	Crocuta crocuta papillomavirus 1	Lambda5	CcrPV1	404184234	YP_006666515.1	*Crocuta crocuta*	SI
Lambdapapillomavirus	*Lambdapapillomavirus 4*	Enhydra lutris papillomavirus 1	Lambda4	ElPV1	605281648	YP_009021864.1	*Enhydra lutris*	SI
Mupapillomavirus	*Mupapillomavirus 3*	Human papillomavirus 204	Mu3	HPV204	819925268	AKG54925.1	*Homo sapiens*	SI
Pipapillomavirus	*Pipapillomavirus 2*	Apodemus sylvaticus papillomavirus 1	Pi2	AsPV1	685511399	YP_009058913.1	*Apodemus sylvaticus*	SI
Pipapillomavirus	*Pipapillomavirus 2*	Mesocricetus auratus papillomavirus 1	Pi2	MaPV1	556503839	YP_008720072.1	*Mesocricetus auratus*	NO
Pipapillomavirus	*Pipapillomavirus 1*	Phodopus sungorus papillomavirus 1	Pi1	PsuPV1	736455514	AJA71478.1	*Phodopus sungorus*	NO
Psipapillomavirus	*Psipapillomavirus UNK*	Eidolon helvum papillomavirus 2	PsiUNK	EhPV2	1147809816	YP_009345078.1	*Eidolon helvum*	NO
Psipapillomavirus	*Psipapillomavirus UNK*	Eidolon helvum papillomavirus 3	PsiUNK	EhPV3	1147809842	YP_009345099.1	*Eidolon helvum*	SI
Rhopapillomavirus	*Rhopapillomavirus 1*	Trichechus manatus latirostris papillomavirus 2	Rho1	TmPV2	379059603	YP_005271217.1	*Trichechus manatus latirostris*	SI
Rhopapillomavirus	*Rhopapillomavirus 2*	Trichechus manatus latirostris papillomavirus 3	Rho2	TmPV3	815937407	AKE50897.1	*Trichechus manatus latirostris*	SI
Rhopapillomavirus	*Rhopapillomavirus 2*	Trichechus manatus latirostris papillomavirus 4	Rho2	TmPV4	948298198	YP_009177726.1	*Trichechus manatus latirostris*	SI
Sin clasificar	*Sin clasificar*	Bovine papillomavirus 19	Sin clasificar	BPV19	1059198240	YP_009272600.1	*Bos taurus*	SI

continúa en la siguiente hoja

Género	Especie	Tipo	Abrv. Esp	Abrv. Tipo	GI	Nro de Acceso	Hospedador	Alineamiento
Sin clasificar	*Sin clasificar*	Bovine papillomavirus 21	Sin clasificar	BPV21	1059198256	YP_009272614.1	*Bos taurus*	SI
Taupapillomavirus	*Taupapillomavirus 2*	Canis familiaris papillomavirus 13	Tau2	CPV13	601448881	YP_009021235.1	*Canis familiaris*	SI
Taupapillomavirus	*Taupapillomavirus UNK*	Canis familiaris papillomavirus 17	TauUNK	CPV17	974142338	ALY11243.1	*Canis familiaris*	SI
Taupapillomavirus	*Taupapillomavirus UNK*	Canis familiaris papillomavirus 19	TauUNK	CPV19	1064859045	AOP12496.1	*Canis familiaris*	SI
Taupapillomavirus	*Taupapillomavirus 3*	Felis catus papillomavirus 3	Tau3	FcaPV3	511534173	YP_008083105.1	*Felis domesticus*	SI
Taupapillomavirus	*Taupapillomavirus 3*	Felis catus papillomavirus 4	Tau3	FcaPV4	545302082	YP_008574811.1	*Felis domesticus*	SI
Taupapillomavirus	*Taupapillomavirus UNK*	Mustela putorius papillomavirus 1	TauUNK	MpPV1	538654786	YP_008519309.1	*Mustela putorius*	SI
Treisdeltapapillomavirus	*Treisdeltapapillomavirus 1*	Rhinolophus ferrumequinum papillomavirus 1	Treisdelta1	RfPV1	586831327	AHJ81403.1	*Rhinolophus ferrumequinum*	SI
Treisepsilonpapillomavirus	*Treisepsilonpapillomavirus 1*	Pygoscelis adeliae papillomavirus 1	Treisepsilon1	PaPV2	607064651	YP_009022072.1	*Pygoscelis adeliae*	NO
Treisetapapillomavirus	*Treisetapapillomavirus 1*	Vulpes vulpes papillomavirus 1	Treiseta1	VvPV1	595389413	AHM27267.1	*Vulpes vulpes*	SI
Treiskappapapillomavirus	*Treiskappapapillomavirus 1*	Equus caballus papillomavirus 8	Treiskappa1	EcPV8	1098945506	YP_009315921.1	*Equus ferus caballus*	SI
Treisthetapapillomavirus	*Treisthetapapillomavirus 1*	Rusa timorensis papillomavirus 1	TreisTheta1	RtiPV1	973413983	ALX18464.1	*Rusa timorensis*	SI
Treiszetapapillomavirus	*Treiszetapapillomavirus 1*	Fulmarus glacialis papillomavirus 1	Treiszeta1	FgPV1	655454929	YP_009041468.1	*Fulmarus Glacialis*	NO
Xipapillomavirus	*Xipapillomavirus 2*	Bos taurus papillomavirus 12	Xi2	BPV12	944326240	YP_009175021.1	*Bos taurus*	SI
Xipapillomavirus	*Xipapillomavirus 1*	Bovine papillomavirus 04AC14	Xi1	BPV-04AC14	1126547286	APR72338.1	*Bos taurus*	SI
Xipapillomavirus	*Xipapillomavirus 1*	Bovine papillomavirus 15	Xi1	BPV15	751997736	AJG05908.1	*Bos taurus*	SI
Xipapillomavirus	*Xipapillomavirus UNK*	Bovine papillomavirus 17	XiUNK	BPV17	1059198226	YP_009272588.1	*Bos taurus*	SI
Xipapillomavirus	*Xipapillomavirus UNK*	Bovine papillomavirus 20	XiUNK	BPV20	1059198248	YP_009272607.1	*Bos taurus*	SI
Xipapillomavirus	*Xipapillomavirus UNK*	Cervus elaphus papillomavirus 2	XiUNK	CePV2	954539437	ALP46946.1	*Cervus elaphus*	SI
Xipapillomavirus	*Xipapillomavirus 3*	Rangifer tarandus papillomavirus 2	Xi3	RtPV2	529217048	YP_008378667.1	*Rangifer tarandus*	SI

A.3. Hospedadores de los serotipos virales de papilomavirus

En la Tabla A.3 se listan los hospedadores de los serotipos virales para los cuales hay un representante en la base de datos 1 de papilomavirus (Sección A.1) incluyendo el orden al que pertenecen, la especie y el nombre común.

Orden	Especie	Nombre común
Artiodactyla	*Alces alces*	Alce
	Bos taurus	Vaca o toro
	Camelus dromedarius	Camello
	Capra hircus	Cabra
	Capreolus capreolus	Corzo
	Odocoileus virginianus	Ciervo gris
	Ovis aries	Oveja
	Rangifer tarandus	Reno
Carnivora	*Canis familiaris*	Perro doméstico
	Felis domesticus	Gato doméstico
	Lynx rufus	Lince rojo
	Panthera leo persica	León asiático
	Procyon lotor	Mapache boreal
	Puma concolor	Puma
	Uncia uncia	Leopardo de las nieves
	Zalophus californianus	Lobo marino de California
Chiroptera	*Rousettus aegyptiacus*	Murciélago egipcio de la fruta
Diprotodontia	*Bettongia penicillata*	*Bettong woylie* o de cola de cepillo
Erinaceomorpha	*Erinaceus europaeus*	Erizo común
Lagomorpha	*Oryctolagus cuniculus*	Conejo común o Europeo
	Sylvilagus floridanus	Conejo de Florida
Perissodactyla	*Equus caballus*	Caballo
Primates	*Homo sapiens*	Humano
	Pan paniscus	Bonobo o chimpancé pigmeo
	Pan troglodytes	Chimpancé común
	Colobus guereza	Guereza abisinio o colobo oriental negro y blanco
	Macaca fascicularis	Macaco cangrejero
	Macacca mulata	Macaco Rhesus
Rodentia	*Erethizon dorsatum*	Puercoespín norteamericano
	Mastomys coucha	Rata lupita
	Mastomys natalensis	Rata común africana
	Micromys minutus	Ratón espiguero
	Mus musculus	Ratón doméstico
	Rattus novergicus	Rata común
Sirenia	*Trichechus manatus latirostris*	Manatí de Florida
	Trichechus manatus manatus	Manatí del Caribe
Testudines	*Caretta caretta*	Tortuga cabezona
	Chelonia mydas	Tortuga verde
Galliformes	*Francolinus leucoscepus*	Francolín gorgiamarillo
Psittaciformes	*Psittacus erithacus*	Loro gris
Passeriformes	*Fringilla coelebs*	Pinzón vulgar

Tabla A.3: Hospedadores de los serotipos de papilomavirus incluidos en la base de datos 1. En la primera columna se indica el orden de los hospedadores, en la segunda columna el nombre científico y en la tercera columna el nombre comúnmente utilizado.

A.4. Secuencias de la proteína E1A

En la Tabla A.4 se listan las secuencias de la proteína de E1A del género *Mastadenovirus* de la familia *Adenoviridae* recolectadas. En las cinco primeras columnas se indica la taxonomía correspondiente y la abreviatura utilizada en el alineamiento para su identificación. La columna GI indica el identificador del gen codificante (en inglés, *Gene id*). Aquellos tipos virales para los cuales la región codificante para la proteína E1A fue identificada manualmente a partir del genoma viral están identificados con una "g" al lado del identificador del gen (GI) y en la siguiente columna se indica el número de acceso (en inglés, *Accession Number*). Se indica también el hospedador del tipo viral. Por último, las columnas N, CR1, CR2, CR3 y CR4 indican si el serotipo viral se incluyó o no en el alineamiento correspondiente a cada dominio.

Tabla A.4: Lista de secuencias de la proteína E1A del género *Mastadenovirus* de la familia *Adenoviridae*.

Género	Especie	Tipo	Abrv. Esp	Abrv. Tipo	GI	Nro de Acceso	Hospedador	N	CR1	CR2	CR3	CR4
Mastadenovirus	*Bat adenovirus A*	Bat adenovirus 3	BtA	BtAdV3	289719041	ADD17097.1	*Myotis ricketti*	SI	SI	SI	SI	NO
Mastadenovirus	*Bat adenovirus B*	Bat adenovirus 2	BtB	BtAdV2	343201198	AEM06262.1	*Pipistrellus pipistrellus*	SI	SI	SI	SI	SI
Mastadenovirus	*Bovine adenovirus A*	Bovine adenovirus 1	BA	BAdV1	52801680	YP_094027.1	*Bos taurus*	SI	SI	SI	SI	SI
Mastadenovirus	*Bovine adenovirus B*	Bovine adenovirus 3	BB	BAdV3	465386	BAA04816.1	*Bos taurus*	NO	NO	SI	SI	NO
Mastadenovirus	*Canine adenovirus A*	Canine adenovirus 1	CA	CAdV1	56158848	AP_000045.1	*Canis lupus familiaris*	SI	SI	SI	SI	SI
Mastadenovirus	*Canine adenovirus A*	Canine adenovirus 2	CA	CAdV2	56160953	AP_000608.1	*Canis lupus familiaris*	SI	SI	SI	SI	SI
Mastadenovirus	*Equine adenovirus A*	Equine adenovirus 1	EA	EAdV1	347602199	AEP16405.1	*Equus caballus*	SI	SI	SI	SI	NO
Mastadenovirus	*Human adenovirus A*	Human adenovirus 12	HA	HAdV12	119022	P03259.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus A*	Human adenovirus 18	HA	HAdV18	45476726	AAS65971.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus A*	Human adenovirus 31	HA	HAdV31	45476732	AAS65974.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus A*	Human adenovirus 61	HA	HAdV61	341573869	AEK79915.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Human adenovirus 3	HB	HAdV3	78059381	ABB17767.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Human adenovirus 7	HB	HAdV7	119021	P03256.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Human adenovirus 11	HB	HAdV11	24711765	AAN62486.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Human adenovirus 14	HB	HAdV14	45476722	AAS65969.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Human adenovirus 16	HB	HAdV16	45476724	AAS65970.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Human adenovirus 21	HB	HAdV21	21311721	AAM46822.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Human adenovirus 34	HB	HAdV34	57115663	AAW33470.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Human adenovirus 35	HB	HAdV35	56160915	AP_000571.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Human adenovirus 50	HB	HAdV50	57115704	AAW33510.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Human adenovirus 55	HB	HAdV55	256028989	ACU57004.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Simian adenovirus 21	HB	SAdV21	56160596	AP_000261.1	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Simian adenovirus 27	HB	SAdV27	219522356 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Simian adenovirus 28	HB	SAdV28	219522361 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Simian adenovirus 29	HB	SAdV29	219522363 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Simian adenovirus 32	HB	SAdV32	219522358 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Simian adenovirus 33	HB	SAdV33	219522355 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Simian adenovirus 35	HB	SAdV35	219522359 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Simian adenovirus 41	HB	SAdV41	219522360 (g)	-	*Gorilla gorilla*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Simian adenovirus 46	HB	SAdV46	219522377 (g)	-	*Gorilla gorilla*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus B*	Simian adenovirus 47	HB	SAdV47	219522376 (g)	-	*Gorilla gorilla*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus C*	Human adenovirus 1	HC	HAdV1	45476716	AAS65966.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus C*	Human adenovirus 2	HC	HAdV2	119018	P03254.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus C*	Human adenovirus 5	HC	HAdV5	119020	P03255.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus C*	Human adenovirus 6	HC	HAdV6	45476718	AAS65967.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus C*	Human adenovirus 57	HC	HAdV57	304633248	ADM46129.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus C*	Simian adenovirus 31	HC	SAdV31	219522353 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus C*	Simian adenovirus 34	HC	SAdV34	219522352 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus C*	Simian adenovirus 40	HC	SAdV40	219522354 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus C*	Simian adenovirus 42	HC	SAdV42	219522350 (g)	-	*Pan paniscus*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus C*	Simian adenovirus 43	HC	SAdV43	219522347 (g)	-	*Gorilla gorilla*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus C*	Simian adenovirus 44	HC	SAdV44	219522346 (g)	-	*Pan paniscus*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus C*	Simian adenovirus 45	HC	SAdV45	219522348 (g)	-	*Gorilla gorilla*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 8	HD	HAdV8	45476720	AAS65968.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 9	HD	HAdV9	4323354	AAD16301.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 10	HD	HAdV10	389617860	AFK92121.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 13	HD	HAdV13	389617901	AFK92161.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 15	HD	HAdV15	308152911	BAJ22275.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 17	HD	HAdV17	324105304	ADY18413.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 19	HD	HAdV19	222143833	BAH19021.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 20	HD	HAdV20	389617983	AFK92241.1	*Homo sapiens*	SI	SI	SI	SI	SI

continúa en la siguiente hoja

Género	Especie	Tipo	Abrv. Esp	Abrv. Tipo	GI	Nro de Acceso	Hospedador	N	CR1	CR2	CR3	CR4
Mastadenovirus	*Human adenovirus D*	Human adenovirus 22	HD	HAdV22	238915415	ACR78202.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 23	HD	HAdV23	45476728	AAS65972.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 24	HD	HAdV24	389618064	AFK92320.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 25	HD	HAdV25	389618105	AFK92360.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 26	HD	HAdV26	57869778	AAW57770.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 27	HD	HAdV27	389618146	AFK92400.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 28	HD	HAdV28	233770159	ACQ91146.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 29	HD	HAdV29	308152948	BAJ22311.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 30	HD	HAdV30	373938859	AEY79557.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 32	HD	HAdV32	389618228	AFK92480.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 33	HD	HAdV33	389618310	AFK92560.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 36	HD	HAdV36	261875890	ACY04456.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 37	HD	HAdV37	222143981	BAH19165.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 38	HD	HAdV38	389618351	AFK92600.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 39	HD	HAdV39	389618392	AFK92640.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 42	HD	HAdV42	389618433	AFK92680.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 43	HD	HAdV43	389618474	AFK92720.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 44	HD	HAdV44	389618515	AFK92760.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 45	HD	HAdV45	389618556	AFK92800.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 47	HD	HAdV47	389618269	AFK92520.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 48	HD	HAdV48	134141802 (g)	-	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 49	HD	HAdV49	57869780	AAW57771.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 51	HD	HAdV51	57869782	AAW57772.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 53	HD	HAdV53	315259038	BAJ46442.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 54	HD	HAdV54	253683361	BAH84785.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 56	HD	HAdV56	305693904	ADM66102.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 58	HD	HAdV58	322366663	ADW95403.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 59	HD	HAdV59	338235706	AEI91275.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 60	HD	HAdV60	341818645	AEK87011.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 62	HD	HAdV62	342898884	AEL78828.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 63	HD	HAdV63	360042478	AEV92952.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 64	HD	HAdV64	209427723	ACI47087.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 65	HD	HAdV65	363987166	BAL41705.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus D*	Human adenovirus 67	HD	HAdV67	378786192	BAL63171.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus E*	Chimpanzee adenovirus Y25	HE	ChAdVY25	386522684	YP_006272949.1	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus E*	Human adenovirus 4	HE	HAdV4	119019	P10407.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus E*	Simian adenovirus 22	HE	SAdV22	41763966	AAS10355.1	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus E*	Simian adenovirus 23	HE	SAdV23	41764022	AAS10391.1	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus E*	Simian adenovirus 24	HE	SAdV24	41764065	AAS10427.1	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus E*	Simian adenovirus 25	HE	SAdV25	56160635	AP_000299.1	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus E*	Simian adenovirus 26	HE	SAdV26	219522370 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus E*	Simian adenovirus 30	HE	SAdV30	219522367 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus E*	Simian adenovirus 36	HE	SAdV36	219522364 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus E*	Simian adenovirus 37	HE	SAdV37	219522368 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus E*	Simian adenovirus 38	HE	SAdV38	219522369 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus E*	Simian adenovirus 39	HE	SAdV39	219522371 (g)	-	*Pan troglodytes*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus F*	Human adenovirus 40	HF	HAdV40	119023	P10541.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus F*	Human adenovirus 41	HF	HAdV41	119024	P10542.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus G*	Human adenovirus 52	HG	HAdV52	117503036	ABK35030.1	*Homo sapiens*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus G*	Simian adenovirus 1	HG	SAdV1	61602119	YP_213961.1	*Macaca fascicularis*	SI	SI	SI	SI	SI
Mastadenovirus	*Human adenovirus G*	Simian adenovirus 7	HG	SAdV7	119027	P06499.1	*Macaca mulatta*	SI	SI	SI	SI	SI
Mastadenovirus	*Murine adenovirus A*	Murine adenovirus 1	MA	MAdV1	119026	P12534.1	*Mus musculus*	SI	SI	SI	SI	NO
Mastadenovirus	*Murine adenovirus B*	Murine adenovirus 2	MB	MAdV2	318065966	YP_004123733.1	*Mus musculus*	NO	NO	SI	SI	NO

continúa en la siguiente hoja

Género	Especie	Tipo	Abrv. Esp	Abrv. Tipo	GI	Nro de Acceso	Hospedador	N	CR1	CR2	CR3	CR4
Mastadenovirus	*Murine adenovirus C*	Murine adenovirus 3	MC	MAdV3	227122415	YP_002822201.1	*Apodemus agrarius*	SI	NO	SI	SI	NO
Mastadenovirus	*Ovine adenovirus A*	Bovine adenovirus 2	OA	BAdV2	56158827	AP_000001.1	*Ovis aries*	SI	SI	SI	SI	SI
Mastadenovirus	*Porcine adenovirus A*	Porcine adenovirus 3	PA	PAdV3	4678604	CAB41021.1	*Sus scrofa*	SI	SI	NO	SI	NO
Mastadenovirus	*Porcine adenovirus C*	Porcine adenovirus 5	PC	PAdV5	13446710	AAK26478.1	*Sus scrofa*	SI	SI	SI	SI	NO
Mastadenovirus	*Simian adenovirus A*	Simian adenovirus 3	SA	SAdV3	51518816	YP_067903.1	*Macaca mulatta*	SI	SI	SI	SI	SI
Mastadenovirus	*Simian adenovirus A*	Simian adenovirus 6	SA	SAdV6	381283770	AFG19582.1	*Macaca mulatta*	SI	SI	SI	SI	SI
Mastadenovirus	*Simian adenovirus A*	Simian adenovirus 20	SA	SAdV20	333601452	AEF59040.1	*Chlorocebus pygerythrus*	SI	SI	SI	SI	SI
Mastadenovirus	*Simian adenovirus A*	Simian adenovirus 48	SA	SAdV48	325658996	ADZ39799.1	*Macaca fascicularis*	SI	SI	SI	SI	SI
Mastadenovirus	*Simian adenovirus B*	Simian adenovirus 49	SB	SAdV49	325659029	ADZ39831.1	*Macaca fascicularis*	SI	SI	SI	SI	SI
Mastadenovirus	*Simian adenovirus B*	Simian adenovirus 50	SB	SAdV50	325659062	ADZ39863.1	*Macaca fascicularis*	SI	SI	SI	SI	SI
Mastadenovirus	*Simian adenovirus F*	Simian adenovirus 18	SF	SAdV18	379692853	AFD10556.1	*Chlorocebus aethiops*	SI	SI	SI	SI	SI
Mastadenovirus	*Tree shrew adenovirus A*	Tree shrew adenovirus 1	TSA	TSAdV1	51556694	YP_068055.1	*Tupaia belangeri*	NO	NO	NO	SI	NO

A.5. Secuencias de genomas de *Mastadenovirus*

En la Tabla A.5 se listan las secuencias de genomas del género *Mastadenovirus* de la familia *Adenoviridae* recolectadas. En las cuatro primeras columnas se indica la taxonomía correspondiente y la abreviatura utilizada en el alineamiento para su identificación. En la siguiente columna se indica el número de acceso. Aquellos tipos virales que corresponden a una variante viral están identificados con una "v" al lado del serotipo y abreviatura del serotipo. Las secuencias correspondientes a HAdV2 utilizadas como control están señaladas con un asterisco, "*", al lado del serotipo. Por último, se indica si las secuencias como hojas del árbol se incluyeron o no en los análisis filogenéticos.

Tabla A.5: Lista de secuencias de los genomas del género *Mastadenovirus* de la familia *Adenoviridae*.

Especie	Tipo	Abrv. Esp	Abrv. Tipo	Nro de Acceso	Incluida en el análisis
Bat adenovirus A	Bat adenovirus 3	BtA	BtAdV3	GU226970	SI
Bat adenovirus B	Bat adenovirus 2	BtB	BtAdV2	NC_015932	SI
Bovine adenovirus A	Bovine adenovirus 1	BA	BAdV1	AC_000191	SI
Bovine adenovirus B	Bovine adenovirus 3	BB	BAdV3	AC_000002	SI
Canine adenovirus A	Canine adenovirus 1	CA	CAdV1	AC_000003	SI
Canine adenovirus A	Canine adenovirus 2	CA	CAdV2	AC_000020	SI
Equine adenovirus A	Equine adenovirus 1	EA	EAdV1	JN418926	SI
Human adenovirus A	Human adenovirus 12	HA	HAdV12	NC_001460	SI
Human adenovirus A	Human adenovirus 18	HA	HAdV18	GU191019	SI
Human adenovirus A	Human adenovirus 31	HA	HAdV31	AM749299	SI
Human adenovirus A	Human adenovirus 61	HA	HAdV61	JF964962	SI
Human adenovirus B	Human adenovirus 3	HB	HAdV3	DQ086466	SI
Human adenovirus B	Human adenovirus 7	HB	HAdV7	AC_000018	SI
Human adenovirus B	Human adenovirus 11	HB	HAdV11	AY163756	SI
Human adenovirus B	Human adenovirus 14	HB	HAdV14	AY803294	SI
Human adenovirus B	Human adenovirus 16	HB	HAdV16	AY601636	SI
Human adenovirus B	Human adenovirus 21	HB	HAdV21	AY601633	SI
Human adenovirus B	Human adenovirus 34	HB	HAdV34	AY737797	SI
Human adenovirus B	Human adenovirus 35	HB	HAdV35	AC_000019	SI
Human adenovirus B	Human adenovirus 50	HB	HAdV50	AY737798	SI
Human adenovirus B	Human adenovirus 55	HB	HAdV55	FJ643676	SI
Human adenovirus B	Human adenovirus 66	HB	HAdV66	JN860676	NO
Human adenovirus B	Human adenovirus 68	HB	HAdV68	JN860678	NO
Human adenovirus B	Simian adenovirus 21	HB	SAdV21	AC_000010	SI
Human adenovirus B	Simian adenovirus 27	HB	SAdV27	HC084988	SI
Human adenovirus B	Simian adenovirus 28	HB	SAdV28	HC084950	SI
Human adenovirus B	Simian adenovirus 29	HB	SAdV29	HC085020	SI
Human adenovirus B	Simian adenovirus 32	HB	SAdV32	HC085052	SI
Human adenovirus B	Simian adenovirus 33	HB	SAdV33	HC085083	SI
Human adenovirus B	Simian adenovirus 35	HB	SAdV35	HC085115	SI
Human adenovirus B	Simian adenovirus 41	HB	SAdV41	HI964271	SI
Human adenovirus B	Simian adenovirus 46	HB	SAdV46	FJ025930	SI
Human adenovirus B	Simian adenovirus 47	HB	SAdV47	FJ025929	SI
Human adenovirus C	Human adenovirus 1	HC	HAdV1	AF534906	SI
Human adenovirus C	Human adenovirus 2	HC	HAdV2_83	J01917_83	SI
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_84	J01917_84	*
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_85	J01917_85	*
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_86	J01917_86	*
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_87	J01917_87	*
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_88	J01917_88	*
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_89	J01917_89	*
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_90	J01917_90	*
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_91	J01917_91	*
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_92	J01917_92	*
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_93	J01917_93	*
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_94	J01917_94	*
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_95	J01917_95	*
Human adenovirus C	Human adenovirus 2 *	HC	HAdV2_96	J01917_96	*
Human adenovirus C	Human adenovirus 5	HC	HAdV5	AC_000008	SI
Human adenovirus C	Human adenovirus 6	HC	HAdV6	FJ349096	SI
Human adenovirus C	Human adenovirus 57	HC	HAdV57	HQ003817	SI
Human adenovirus C	Simian adenovirus 31	HC	SAdV31	FJ025906	SI
Human adenovirus C	Simian adenovirus 34	HC	SAdV34	HC000847	SI
Human adenovirus C	Simian adenovirus 40	HC	SAdV40	HC000785	SI
Human adenovirus C	Simian adenovirus 42	HC	SAdV42	FJ025903	SI
Human adenovirus C	Simian adenovirus 42 (v)	HC	HAdV42_v	HC191035	NO

continúa en la siguiente hoja

Especie	Tipo	Abrv. Esp	Abrv. Tipo	Nro de Acceso	Incluida en el análisis
Human adenovirus C	Simian adenovirus 43	HC	SAdV43	FJ025900	SI
Human adenovirus C	Simian adenovirus 44	HC	SAdV44	HC191097	SI
Human adenovirus C	Simian adenovirus 45	HC	SAdV45	FJ025901	SI
Human adenovirus D	Human adenovirus 8	HD	HAdV8	AB448767	SI
Human adenovirus D	Human adenovirus 9	HD	HAdV9	AJ854486	SI
Human adenovirus D	Human adenovirus 10	HD	HAdV10	JN226746	SI
Human adenovirus D	Human adenovirus 13	HD	HAdV13	JN226747	SI
Human adenovirus D	Human adenovirus 15	HD	HAdV15	AB562586	SI
Human adenovirus D	Human adenovirus 17	HD	HAdV17	HQ910407	SI
Human adenovirus D	Human adenovirus 17 (v)	HD	HAdV17_v	AC_000006	NO
Human adenovirus D	Human adenovirus 19	HD	HAdV19	AB448774	SI
Human adenovirus D	Human adenovirus 20	HD	HAdV20	JN226749	SI
Human adenovirus D	Human adenovirus 22	HD	HAdV22	FJ619037	SI
Human adenovirus D	Human adenovirus 22 (v)	HD	HAdV22_v	FJ404771	NO
Human adenovirus D	Human adenovirus 23	HD	HAdV23	JN226750	SI
Human adenovirus D	Human adenovirus 24	HD	HAdV24	JN226751	SI
Human adenovirus D	Human adenovirus 25	HD	HAdV25	JN226752	SI
Human adenovirus D	Human adenovirus 26	HD	HAdV26	EF153474	SI
Human adenovirus D	Human adenovirus 27	HD	HAdV27	JN226753	SI
Human adenovirus D	Human adenovirus 28	HD	HAdV28	FJ824826	SI
Human adenovirus D	Human adenovirus 29	HD	HAdV29	AB562587	SI
Human adenovirus D	Human adenovirus 30	HD	HAdV30	JN226755	SI
Human adenovirus D	Human adenovirus 32	HD	HAdV32	JN226756	SI
Human adenovirus D	Human adenovirus 33	HD	HAdV33	JN226758	SI
Human adenovirus D	Human adenovirus 36	HD	HAdV36	GQ384080	SI
Human adenovirus D	Human adenovirus 37	HD	HAdV37	AB448778	SI
Human adenovirus D	Human adenovirus 37 (v)	HD	HAdV37	DQ900900	NO
Human adenovirus D	Human adenovirus 38	HD	HAdV38	JN226759	SI
Human adenovirus D	Human adenovirus 39	HD	HAdV39	JN226760	SI
Human adenovirus D	Human adenovirus 42	HD	HAdV42	JN226761	SI
Human adenovirus D	Human adenovirus 43	HD	HAdV43	JN226762	SI
Human adenovirus D	Human adenovirus 44	HD	HAdV44	JN226763	SI
Human adenovirus D	Human adenovirus 45	HD	HAdV45	JN226764	SI
Human adenovirus D	Human adenovirus 46	HD	HAdV46	AY875648	NO
Human adenovirus D	Human adenovirus 47	HD	HAdV47	JN226757	SI
Human adenovirus D	Human adenovirus 48	HD	HAdV48	EF153473	SI
Human adenovirus D	Human adenovirus 49	HD	HAdV49	DQ393829	SI
Human adenovirus D	Human adenovirus 51	HD	HAdV51	JN226765	SI
Human adenovirus D	Human adenovirus 53	HD	HAdV53	AB605246	SI
Human adenovirus D	Human adenovirus 53 (v)	HD	HAdV53_v	FJ169625	NO
Human adenovirus D	Human adenovirus 54	HD	HAdV54	AB333801	SI
Human adenovirus D	Human adenovirus 56	HD	HAdV56	HM770721	SI
Human adenovirus D	Human adenovirus 58	HD	HAdV58	HQ883276	SI
Human adenovirus D	Human adenovirus 59	HD	HAdV59	JF799911	SI
Human adenovirus D	Human adenovirus 60	HD	HAdV60	HQ007053	SI
Human adenovirus D	Human adenovirus 62	HD	HAdV62	JN162671	SI
Human adenovirus D	Human adenovirus 63	HD	HAdV63	JN935766	SI
Human adenovirus D	Human adenovirus 64	HD	HAdV64	EF121005	SI
Human adenovirus D	Human adenovirus 65	HD	HAdV65	AP012285	SI
Human adenovirus D	Human adenovirus 67	HD	HAdV67	AP012302	SI
Human adenovirus E	Chimpanzee adenovirus Y25	HE	ChAdVY25	NC_017825	SI
Human adenovirus E	Human adenovirus 4	HE	HAdV4	AY487947	SI
Human adenovirus E	Simian adenovirus 22	HE	SAdV22	AY530876	SI
Human adenovirus E	Simian adenovirus 23	HE	SAdV23	AY530877	SI
Human adenovirus E	Simian adenovirus 24	HE	SAdV24	AY530878	SI
Human adenovirus E	Simian adenovirus 25	HE	SAdV25	AC_000011	SI
Human adenovirus E	Simian adenovirus 26	HE	SAdV26	HB426768	SI
Human adenovirus E	Simian adenovirus 30	HE	SAdV30	HB426704	SI
Human adenovirus E	Simian adenovirus 36	HE	SAdV36	FJ025917	SI
Human adenovirus E	Simian adenovirus 36 (v)	HE	SAdV36_v	HC191003	NO
Human adenovirus E	Simian adenovirus 37	HE	SAdV37	FJ025921	SI
Human adenovirus E	Simian adenovirus 37 (v)	HE	SAdV37_v	HB426639	NO
Human adenovirus E	Simian adenovirus 38	HE	SAdV38	HB426671	SI
Human adenovirus E	Simian adenovirus 39	HE	SAdV39	HB426607	SI
Human adenovirus F	Human adenovirus 40	HF	HAdV40	NC_001454	SI
Human adenovirus F	Human adenovirus 41	HF	HAdV41	DQ315364	SI
Human adenovirus G	Human adenovirus 52	HG	HAdV52	DQ923122	SI
Human adenovirus G	Simian adenovirus 1	HG	SAdV1	NC_006879	SI
Human adenovirus G	Simian adenovirus 7	HG	SAdV7	DQ792570	SI
Murine adenovirus A	Murine adenovirus 1	MA	MAdV1	NC_000942	SI
Murine adenovirus B	Murine adenovirus 2	MB	MAdV2	NC_014899	SI
Murine adenovirus C	Murine adenovirus 3	MC	MAdV3	NC_012584	SI
Ovine adenovirus A	Bovine adenovirus 2	OA	BAdV2	AC_000001	SI
Porcine adenovirus A	Porcine adenovirus 3	PA	PAdV3	AJ237815	SI
Porcine adenovirus A	Porcine adenovirus 3 (v)	PA	PAdV3_v	AC_000189	NO
Porcine adenovirus C	Porcine adenovirus 5	PC	PAdV5	AF289262	SI
Simian adenovirus A	Simian adenovirus 3	SA	SAdV3	NC_006144	SI
Simian adenovirus A	Simian adenovirus 6	SA	SAdV6	JQ776547	SI
Simian adenovirus A	Simian adenovirus 20	SA	SAdV20	HQ605912	SI
Simian adenovirus A	Simian adenovirus 48	SA	SAdV48	HQ241818	SI
Simian adenovirus B	Simian adenovirus 49	SB	SAdV49	HQ241819	SI
Simian adenovirus B	Simian adenovirus 50	SB	SAdV50	HQ241820	SI
Simian adenovirus B	Simian adenovirus A1139	SB	A1139_v	JN880448	NO

continúa en la siguiente hoja

Especie	Tipo	Abrv. Esp	Abrv. Tipo	Nro de Acceso	Incluida en el análisis
Simian adenovirus B	Simian adenovirus A1163	SB	A1163_v	JN880449	NO
Simian adenovirus B	Simian adenovirus A1173	SB	A1173_v	JN880450	NO
Simian adenovirus B	Simian adenovirus A1258	SB	A1258_v	JN880451	NO
Simian adenovirus B	Simian adenovirus A1285	SB	A1285_v	JN880452	NO
Simian adenovirus B	Simian adenovirus A1296	SB	A1296_v	JN880453	NO
Simian adenovirus B	Simian adenovirus A1312	SB	A1312_v	JN880454	NO
Simian adenovirus B	Simian adenovirus A1327	SB	A1327_v	JN880455	NO
Simian adenovirus B	Simian adenovirus A1335	SB	A1335_v	JN880456	NO
Simian adenovirus B	Simian adenovirus BaAdV1	SB	SAdV49_v_Ba1	KC693021	NO
Simian adenovirus C	Simian adenovirus BaAdV2	SC	BaAdV2	KC693022	NO
Simian adenovirus C	Simian adenovirus BaAdV3	SC	BaAdV2_v	KC693023	NO
Simian adenovirus F	Simian adenovirus 18	SF	SAdV18	FJ025931	SI
Tree shrew adenovirus A	Tree shrew adenovirus 1	TSA	TSAdV1	AC_000190	SI

A.6. Hospedadores de los serotipos virales de *Mastadenovirus*

En la Tabla A.6 se listan los hospedadores de los serotipos virales para los cuales hay un representante en la base de datos de la proteína E1A de *Mastadenovirus* (Sección A.4) incluyendo el orden al que pertenece, la especie y el nombre común.

Orden	Especie	Nombre común
Artiodactyla	*Bos taurus*	Vaca o Toro
	Ovis aries	Oveja
	Sus scrofa domestica	Chancho
Chiroptera	*Myotis ricketti*	Murciélago ratonero
	Pipistrellus pipistrellus	Murciélago común o enano
Carnívora	*Canis lupus familiaris*	Perro
Perissodactyla	*Equus caballus*	Caballo
Primates	*Homo sapiens*	Humano
	Gorilla gorilla	Gorila
	Pan troglodytes	Chimpancé
	Pan paniscus	Bonobo o chimpanché pigmeo
	Macaca fascicularis	Macaco cangrejero
	Macaca mulatta	Macaco Rhesus
	Chlorocebus pygerythrus	Cercopiteco verde o Vervet
	Chlorocebus aethiops	Cercopiteco verde o tota
Rodentia	*Mus musculus*	Ratón doméstico
	Apodemus agrarius	Ratón listado
Scandentia	*Tupaia belangeri*	Tupaya de Belanger

Tabla A.6: Hospedadores de los serotipos de *Mastadenovirus*. En la primera columna se indica el orden de los hospedadores, en la segunda columna el nombre científico y en la tercera columna el nombre comúnmente utilizado.

Apéndice B

Alineamientos

Representación de secuencias. La estructura primaria de una proteína está determinada por la secuencia de aminoácidos que la conforman. Es decir, una secuencia proteica se caracteriza por número total y el orden de los aminoácidos indicados desde el extremo N-terminal hacia el extremo C-terminal de la proteína. Una manera simple y práctica de representarla es utilizar el código de la Unión Internacional de Química Pura y Aplicada (IUPAC) (en inglés, *International Union of Pure and Applied Chemistry*) de una letra para cada aminoácido indicado en la tercera columna de la Tabla B.1, junto con el nombre del aminoácido correspondiente y el código de tres letras.

Aminoácido	Código de tres letras	Código de una letra
Ácido glutámico	Glu	E
Prolina	Pro	P
Glutamina	Gln	Q
Serina	Ser	S
Arginina	Arg	R
Lisina	Lys	K
Metionina	Met	M
Ácido aspártico	Asp	D
Glicina	Gly	G
Alanina	Ala	A
Asparagina	Asn	N
Treonina	Thr	T
Histidina	His	H
Leucina	Leu	L
Valina	Val	V
Fenilalanina	Phe	F
Isoleucina	Ile	I
Tirosina	Tyr	Y
Triptofano	Trp	W
Cisteína	Cys	C

Tabla B.1: Aminoácidos. Se indican los nombres de los aminoácidos, el código de tres letras y el código IUPAC de una letra, ordenados según presenten mayor a menor tendencia a aparecer en las regiones desordenadas (Brown *et al.*, 2010).

Las secuencias se almacenaron en formato FASTA.

Formato FASTA. Las secuencias son almacenadas individualmente en archivos de texto utilizando el formato FASTA o grupalmente en formato multiFASTA. Como se muestra en la Figura B.1 una secuencia en formato FASTA comienza con una primera línea única, llamada encabezado, seguida por las líneas con los datos de secuencia.

```
>sp|P03129|VE7_HPV16 Protein E7 OS=Human papillomavirus type
16 GN=E7 PE=1 SV=1
MHGDTPTLHEYMLDLQPETTDLYCYEQLNDSSEEEDEIDGPAGQAEPDRAHYNIVTFCCK
CDSTLRLCVQSTHVDIRTLEDLLMGTLGIVCPICSQKP
```

Figura B.1: Formato FASTA. Secuencia de la proteína E7 del serotipo Human papillomavirus 16 (HPV16) de papilomavirus en formato FASTA.

El encabezado identifica a la secuencia y debe estar señalizado por el signo mayor, ">". Todas las líneas siguientes deben tener como máximo 80 caracteres. Si la secuencia es más larga, se continúa en la línea siguiente.

Los alineamientos se encuentran disponibles en el CD que se adjunta y en:
https://bitbucket.org/jglavina/datos_tesis/wiki/Home

B.1. Alineamientos de E7

B.1.1. Alineamientos de E7 construidos a partir de la base de datos 1.

Los archivos correspondientes a los alineamientos del dominio E7N construido a partir de 207 secuencias son:

Alineamiento completo: E7_Dominio_E7N_Completo_207s.fasta
Alineamiento sin sitios vacíos: E7_Dominio_E7N_SinSitiosVacios_207s.fasta

Los archivos correspondientes a los alineamientos del dominio E7C construido a partir de 219 secuencias son:

Alineamiento completo: E7_Dominio_E7C_Completo_219s.fasta
Alineamiento sin sitios vacíos: E7_Dominio_E7C_SinSitiosVacios_219s.fasta

B.1.2. Alineamientos de E7 construidos a partir de la base de datos 2.

Los archivos correspondientes a los alineamientos de E7 construido a partir de 317 secuencias son:

Alineamiento completo: E7_Completo_317s.fasta
Alineamiento sin sitios vacíos: E7_SinSitiosVacios_317s.fasta

B.2. Alineamientos de E1A

Los archivos correspondientes a los alineamientos de E1A de longitud completa son:
Alineamiento completo: `E1A_Completo_116s.fasta`
Alineamiento sin sitios vacíos: `E1A_SinSitiosVacios_116s.fasta`

Los archivos correspondientes a los alineamientos al dominio N-Terminal de E1A son:
Alineamiento completo: `E1A_Dominio_N_Completo_113s.fasta`
Alineamiento sin sitios vacíos: `E1A_Dominio_N_SinSitiosVacios_113s.fasta`

Los archivos correspondientes a los alineamientos al dominio CR1 de E1A son:
Alineamiento completo: `E1A_Dominio_CR1_Completo_112s.fasta`
Alineamiento sin sitios vacíos: `E1A_Dominio_CR1_SinSitiosVacios_112s.fasta`

Los archivos correspondientes a los alineamientos a la región IDR12 de E1A son:
Alineamiento completo: `E1A_IDR12_Completo_12s.fasta`
Alineamiento sin sitios vacíos: `E1A_IDR12_SinSitiosVacios_12s.fasta`

Los archivos correspondientes a los alineamientos al dominio CR2 de E1A son:
Alineamiento completo: `E1A_Dominio_CR2_Completo_114s.fasta`
Alineamiento sin sitios vacíos: `E1A_Dominio_CR2_SinSitiosVacios_114s.fasta`

Los archivos correspondientes a los alineamientos a la región IDR23 de E1A son:
Alineamiento completo: `E1A_IDR23_Completo_89s.fasta`
Alineamiento sin sitios vacíos: `E1A_IDR23_SinSitiosVacios_89s.fasta`

Los archivos correspondientes a los alineamientos al dominio CR3 de E1A son:
Alineamiento completo: `E1A_Dominio_CR3_Completo_116s.fasta`
Alineamiento sin sitios vacíos: `E1A_Dominio_CR3_SinSitiosVacios_116s.fasta`

Los archivos correspondientes a los alineamientos a la región IDR34 de E1A son:
Alineamiento completo: `E1A_IDR34_Completo_12s.fasta`
Alineamiento sin sitios vacíos: `E1A_IDR34_SinSitiosVacios_12s.fasta`

Los archivos correspondientes a los alineamientos al dominio CR4 de E1A son:
Alineamiento completo: `E1A_Dominio_CR4_Completo_107s.fasta`
Alineamiento sin sitios vacíos: `E1A_Dominio_CR4_SinSitiosVacios_107s.fasta`

B.3. Alineamiento de genomas de *Mastadenovirus*

El archivo correspondiente al alineamiento utilizado para la construcción de la filogenia luego de la eliminación de las regiones con alta tendencia a sufrir recombinación homóloga es:
`Mastadenovirus_genomas_152s.fasta`

Apéndice C

Estructuras proteicas

Archivos PDB. Las estructuras tridimensionales de las proteínas están codificadas en archivos *pdb*. Un archivo `pdb` está compuesto por múltiples líneas de registros, cada uno identificado por una etiqueta determinada incluidos dentro de distintas secciones. En la Figura C.1 se muestra un fragmento de la sección de coordenadas que describe la estructura de la proteína dando las coordenadas *x*, *y* y *z* (azul claro) de cada uno de los átomos identificados.

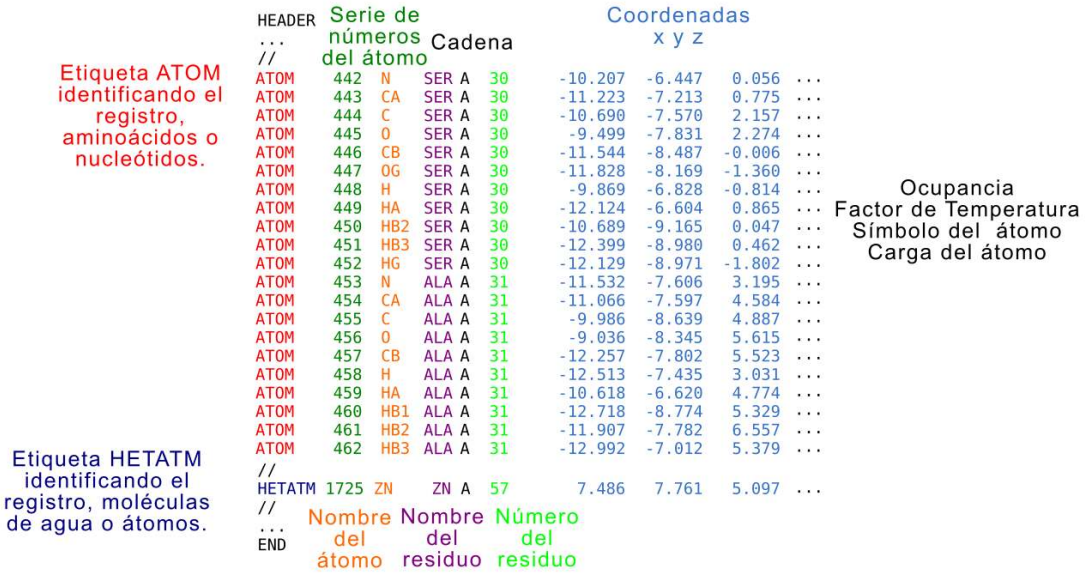


Figura C.1: Fragmento de archivo PDB. Se muestra un fragmento que incluye dos residuos, la serina en la posición 30 y la alanina en la posición 31, de la estructura determinada para el dominio CR3 de la proteína E7 (PDB ID: 2F8B) (Ohlenschläger *et al.*, 2006).

En cada línea de la Figura C.1, además, se identifica si es un átomo (rojo) perteneciente a un aminoácido o nucleótido, o heteroátomo (azul oscuro), la numeración (verde), el nombre del átomo (naranja), el nombre del residuo en el que está incluido el átomo (violeta), la cadena a la que pertenece (negro), el número del residuo al que pertenece (verde). Este archivo puede incluir más columnas para cada átomo con datos relacionados con el espacio que ocupa el átomo, la movilidad del átomo (el factor de temperatura o *B-factor*), el símbolo que representa al átomo y la carga

del mismo (señaladas con "`...`" en la Figura C.1). La descripción del resto del contenido de las secciones del archivo `pdb` puede obtenerse en la sección documentación de `http://www.wwpdb.org/`.

Los complejos estructurales utilizados en esta tesis se indican en la Tabla C.1, junto con la descripción, la técnica utilizada para determinar la estructura y la referencia correspondiente.

PDB ID	**Descripción**	**Técnica**	**Referencia**
2F8B	Dímero dominio globular E7C	RMN	Ohlenschläger *et al.* (2006)
1O9K	Complejo entre pRb y E2F-1	Rayos X	Xiao *et al.* (2003)
1N4M	Complejo entre pRb y E2F-1	Rayos X	Lee *et al.* (2002)
2R7G	Complejo entre pRb y E1A	Rayos X	Liu y Marmorstein (2007)
3N00	Complejo entre Rev-erbα y NCoR	Rayos X	Phelan *et al.* (2010)
2KJE	Complejo entre CBP y E1A	RMN	Ferreon *et al.* (2009)

Tabla C.1: Estructuras PDB utilizadas. Se listan las estructuras proteicas utilizadas en esta tesis.

Apéndice D

Árboles filogenéticos

Almacenamiento de árboles. Los árboles se almacenan en archivos de texto utilizando el formato `NEWICK` o `NEXUS`. En la Figura D.1 se representa el árbol de la Figura 2.13 en formato `NEWICK`.

```
((A:0.95,B:0.23):0.29,((D:0.99,E:0.58):0.27,C:0.89):0.94);
```

Figura D.1: Formato Newick. El árbol filogenético de la Figura 2.13 se muestra en formato `NEWICK` indicando las secuencias actuales con letras y el largo de las ramas correspondientes.

En el formato `NEWICK` los árboles se representan usando paréntesis anidados (parentético) y comas, en una única línea que finaliza con un punto y coma " ; ". Cada par de paréntesis agrupa dos descendientes con un ancestro en común separados por coma. Por ejemplo, observando el árbol de la Figura 2.13 las secuencias D y E descendientes del nodo 9 se representan `(D,E)`. Si agregamos a la secuencia vecina C descendiente del mismo nodo que `(D,E)`, se obtiene `((D,E),C)`. Se puede incluir o no el largo de las ramas a continuación de las hojas o nodos presentes en el extremo de la rama más cercano a las hojas. Por ejemplo, si incluimos en `(D,E)` el largo de las ramas sería `(D:0.99,E:0.58)`, y `((D,E),C)` sería `((D:0.99,E:0.58):0:27,C:0.89)`.

En la Figura D.2 se representa el árbol de la Figura 2.13 en formato `NEXUS`.

El formato `NEXUS` comienza con una línea `#NEXUS` (verde) y está organizada en bloques. Algunos bloques como `TAXA` (rojo) y `TREE` (azul) son reconocidos por todos los programas que utilizan árbol pero otros bloques son exclusivos de determinados programas. El comienzo del bloque se indica con la instrucción `BEGIN BLOQUE`, donde `BLOQUE` corresponde al nombre del mismo y el final del bloque se indica con `END`. Al final de cada instrucción se coloca un punto y coma " ; ". El bloque `TAXA` (rojo) indica el número y nombre de las hojas (o taxones) del árbol. En este caso, la instrucción `dimensions ntax=5` indica que el número de hojas es cinco y `TAXLABELS` el nombre de las hojas: A, B, C, D y E. El bloque `TREE` (azul) contiene el árbol. La instrucción `TRANSLATE` brinda una alternativa abreviada para los nombres de las hojas del árbol y es opcional. La instrucción `TREE` indica el comienzo de la descripción del árbol seguido de un " = " y la topología del árbol en formato `NEWICK`. Al árbol se le puede asignar un nombre

a continuación de la instrucción `TREE`. Por ejemplo, `TREE FICTICIO = ((A,B),C)`. Una ventaja del formato `NEXUS` es que dentro de este bloque se pueden incluir distintos árboles que compartan las mismas secuencias o especies actuales.

```
# NEXUS

BEGIN TAXA;
  DIMENSIONS NTAX = 5;
  TAXLABELS
    A
    B
    D
    E
    C
  ;
END;
BEGIN TREES;
  TRANSLATE
    1 A,
    2 B,
    3 D,
    4 E,
    5 C
  ;
  TREE = ((1:0.95,2:0.23):0.29,((3:0.99,4:0.58):0.27,5:0.89):0.94);
END;
```

Figura D.2: Formato Nexus. El árbol de la Figura 2.13 se muestra en formato `NEXUS`. En verde se indica la primera línea que indica el formato correspondiente. En rojo el bloque que corresponde a los taxones. En azul el bloque que corresponde al árbol.

Los árboles en formato `NEXUS` y `NEWICK` se encuentran disponibles en el CD que se adjunta y en: `https://bitbucket.org/jglavina/datos_tesis/wiki/Home`

D.1. Árbol filogenético de *Mastadenovirus*

Árbol de *Mastadenovirus* en formato `NEXUS`:

`Arbol_Mastadenovirus_152s.nex`

D.2. Árbol filogenético de los hospedadores de los serotipos de *Mastadenovirus*

Árbol de hospedadores de *Mastadenovirus* en formato `NEWICK`:

`Arbol_hospedadores_18s.nwk`

Apéndice E

Proteína E7

E.1. Motivos lineales por secuencia de E7

Tabla E.1: Presencia ausencia de motivos en las secuencias de la proteína E7 de la familia *Papillomaviridae* incluidas en la base de datos 1. Se indica la presencia (1) o ausencia (0) del motivo para cada serotipo.

Especie	Tipo	DYRK1A	pRb_ABGroove	LxCxE	CKII	Región acídica	Pos. Ricas en Cys	NES	PDZ
Alpha1	HPV32	0	0	1	1	1	1	0	1
Alpha1	HPV42	1	0	1	1	1	1	0	1
Alpha2	HPV3	0	0	1	0	1	0	0	1
Alpha2	HPV10	0	0	1	0	1	0	0	1
Alpha2	HPV28	0	0	1	1	1	0	0	1
Alpha2	HPV29	0	0	1	1	0	1	0	1
Alpha2	HPV77	0	0	1	1	0	1	0	1
Alpha2	HPV94	0	0	1	0	1	0	0	1
Alpha2	HPV117	0	0	1	0	1	0	0	1
Alpha2	HPV125	0	0	1	0	1	0	0	1
Alpha3	HPV61	0	1	1	1	1	1	0	1
Alpha3	HPV62	0	1	1	1	1	1	0	1
Alpha3	HPV72	0	1	1	1	1	1	0	1
Alpha3	HPV81	0	1	1	1	1	1	0	1
Alpha3	HPV83	0	1	1	1	1	1	0	1
Alpha3	HPV84	1	1	1	1	1	1	0	1
Alpha3	HPV86	1	0	1	1	1	1	0	1
Alpha3	HPV87	1	1	1	1	1	1	0	1
Alpha3	HPV89	1	1	1	1	0	1	0	1
Alpha3	HPV102	0	1	1	1	1	1	0	1
Alpha3	HPV114	1	1	1	1	1	1	0	1
Alpha4	HPV2	0	0	1	1	1	1	0	1
Alpha4	HPV2a	0	0	1	1	1	1	0	1
Alpha4	HPV27	0	0	1	1	1	1	0	1
Alpha4	HPV27b	0	0	1	1	1	1	0	1
Alpha4	HPV57	0	0	1	1	1	1	0	1
Alpha4	HPV57b	0	0	1	1	1	1	0	1
Alpha4	HPV57c	0	0	1	1	1	1	0	1
Alpha5	HPV26	0	1	1	1	1	1	0	1
Alpha5	HPV51	0	1	1	1	1	1	0	1
Alpha5	HPV69	0	1	1	1	1	1	0	1
Alpha5	HPV82	0	1	1	1	1	1	0	1
Alpha6	HPV30	0	1	1	1	1	1	0	1
Alpha6	HPV53	0	1	1	1	1	1	0	1
Alpha6	HPV56	0	1	1	1	1	1	0	1
Alpha6	HPV66	0	1	1	1	1	1	0	1
Alpha7	HPV18	0	1	1	1	1	1	0	1
Alpha7	HPV39	0	1	1	1	1	1	0	1
Alpha7	HPV45	0	1	1	1	1	1	0	1
Alpha7	HPV59	0	1	1	1	1	1	0	1
Alpha7	HPV68	0	1	1	1	1	1	0	1
Alpha7	HPV68a	0	1	1	1	1	1	0	1
Alpha7	HPV68b	0	1	1	1	1	1	0	1
Alpha7	HPV70	0	1	1	1	1	1	0	1
Alpha7	HPV85	0	1	1	1	1	1	0	1
Alpha7	HPV97	0	1	1	1	1	1	0	1
Alpha7	HPVMe180	0	1	1	1	1	1	0	1

continúa en la siguiente hoja

Especie	Tipo	DYRK1A	pRb_ABGroove	LxCxE	CKII	Región acídica	Pos. Ricas en Cys	NES	PDZ
Alpha8	HPV7	0	1	1	1	1	1	0	1
Alpha8	HPV40	0	1	1	1	1	1	0	1
Alpha8	HPV43	0	1	1	1	1	1	1	1
Alpha8	HPV91	0	1	1	1	1	1	1	1
Alpha9	HPV16	1	1	1	1	1	1	0	1
Alpha9	HPV31	1	1	1	1	1	1	1	1
Alpha9	HPV33	0	1	1	1	1	1	0	1
Alpha9	HPV35	0	1	1	1	1	1	0	1
Alpha9	HPV35H	0	1	1	1	1	1	0	1
Alpha9	HPV52	0	1	1	1	1	1	0	1
Alpha9	HPV58	0	1	1	1	1	1	0	0
Alpha9	HPV67	0	1	1	1	1	1	0	1
Alpha10	HPV6	0	1	1	1	1	1	0	1
Alpha10	HPV6A	0	1	1	1	1	1	0	1
Alpha10	HPV6B	0	1	1	1	1	1	0	1
Alpha10	HPV11	0	1	1	1	1	1	0	1
Alpha10	HPV13	0	1	1	1	1	1	0	1
Alpha10	HPV44	0	1	1	1	1	1	0	1
Alpha10	HPV55	0	1	1	1	1	1	0	1
Alpha10	HPV74	0	1	1	1	1	1	0	1
Alpha10	PpPV1	0	1	1	1	1	1	0	1
Alpha11	HPV34	0	1	1	1	0	1	0	1
Alpha11	HPV73	0	0	1	1	1	1	0	1
Alpha12	MFPV3	0	1	1	1	0	1	0	1
Alpha12	MFPV3b	0	1	1	1	0	1	0	1
Alpha12	MFPV4	0	1	1	1	1	1	0	1
Alpha12	MFPV5	0	1	1	1	1	1	0	1
Alpha12	MFPV6	0	1	1	1	0	1	0	1
Alpha12	MFPV7	0	1	1	1	1	1	0	1
Alpha12	MFPV8	0	1	1	1	1	1	0	1
Alpha12	MFPV9	0	1	1	1	1	1	0	1
Alpha12	MFPV10	0	1	1	1	1	1	0	1
Alpha12	MFPV11	0	1	1	1	1	1	0	1
Alpha12	MmPV1	0	1	1	1	1	1	1	1
Alpha12	RhPV1	0	1	1	1	1	1	0	1
Alpha13	HPV54	0	1	1	1	1	1	1	1
Alpha14	CgPV1	0	1	1	1	1	1	0	1
Alpha14	HPV71	0	1	1	1	0	1	0	1
Alpha14	HPV90	0	1	1	1	1	1	0	1
Alpha14	HPV106	0	1	1	1	1	1	0	1
Beta1	CgPV2	0	1	1	0	1	1	0	1
Beta1	HPV5	0	1	1	1	1	1	0	1
Beta1	HPV5B	0	1	1	1	1	1	0	1
Beta1	HPV8	0	1	1	1	1	1	0	1
Beta1	HPV12	0	0	1	1	1	1	0	1
Beta1	HPV14	0	1	1	1	1	1	0	1
Beta1	HPV19	0	1	1	1	1	1	0	1
Beta1	HPV20	0	1	1	0	1	1	0	1
Beta1	HPV21	0	1	1	1	1	1	0	1
Beta1	HPV24	0	1	0	1	1	1	0	1
Beta1	HPV25	0	0	1	1	1	1	0	1
Beta1	HPV36	0	1	1	1	1	1	0	1
Beta1	HPV47	0	1	1	0	1	1	0	1
Beta1	HPV93	0	0	1	1	0	1	0	1
Beta1	HPV98	0	1	1	1	0	1	0	1
Beta1	HPV99	0	1	1	1	1	1	0	1
Beta1	HPV105	0	1	1	0	1	1	0	1
Beta1	HPV118	0	1	1	1	1	1	0	1
Beta1	HPV124	0	1	1	1	1	1	0	1
Beta1	HPVRTRX7	0	1	1	1	1	1	0	1
Beta1	MFPV1	0	1	1	0	0	1	0	1
Beta2	HPV9	0	1	1	0	0	1	0	0
Beta2	HPV15	0	1	1	1	1	1	0	0
Beta2	HPV17	0	1	1	1	1	1	0	0
Beta2	HPV22	0	1	1	1	0	1	0	1
Beta2	HPV23	0	1	1	0	0	1	0	1
Beta2	HPV37	0	1	1	1	1	1	0	0
Beta2	HPV38	0	1	1	1	0	1	0	1
Beta2	HPV38b	0	1	1	1	0	1	0	1
Beta2	HPV80	0	0	1	1	1	1	0	0
Beta2	HPV100	0	1	1	1	0	1	0	1
Beta2	HPV104	0	1	1	1	0	1	0	0
Beta2	HPV107	0	1	1	1	0	1	0	1
Beta2	HPV110	0	1	1	1	0	1	0	0
Beta2	HPV111	0	1	1	0	0	1	0	0
Beta2	HPV113	0	1	1	0	0	1	0	0
Beta2	HPV120	0	1	1	0	0	1	0	1
Beta2	HPV122	0	1	1	0	0	1	0	0
Beta2	HPV151	0	1	1	0	0	0	0	1
Beta2	HPVFA75/KI88	0	1	1	1	0	1	0	0
Beta2	HPVSIBX-3a	0	1	1	0	0	1	0	1
Beta3	HPV49	0	1	1	0	0	1	0	1
Beta3	HPV75	0	1	1	1	1	1	0	1

continúa en la siguiente hoja

Especie	Tipo	DYRK1A	pRb_ABGroove	LxCxE	CKII	Región acídica	Pos. Ricas en Cys	NES	PDZ
Beta3	HPV76	0	1	1	1	1	1	0	1
Beta3	HPV115	0	1	1	0	0	1	0	1
Beta4	HPV92	0	1	0	0	1	1	0	1
Beta5	HPV96	0	0	1	1	0	1	0	0
Beta5	HPV150	0	0	1	1	0	1	0	1
Beta6	MFPV2	0	1	1	0	0	1	0	1
Chi1	CPV3	0	1	1	1	1	1	0	0
Chi1	CPV5	0	1	1	1	1	1	0	0
Chi2	CPV4	0	1	1	1	1	1	0	0
Delta1	AaPV1	-	-	-	-	1	-	0	1
Delta1	RtPV1	-	-	-	-	1	-	0	1
Delta2	OvPV1	-	-	-	-	0	-	0	1
Delta4	BPV1	-	-	-	-	1	-	0	1
Delta4	BPV2	-	-	-	-	1	-	0	1
Delta5	CcaPV1	-	-	-	-	1	-	0	1
DeltaUNK	BPV	-	-	-	-	1	-	0	1
DeltaUNK	CdPV1	-	-	-	-	1	-	0	1
DeltaUNK	CdPV2	-	-	-	-	0	-	0	1
DyoEta1	EePV1	0	0	1	1	0	0	0	1
DyoIota1	EcPV2	0	1	1	1	0	0	0	1
DyoTheta1	FdPV2	1	0	1	0	0	1	0	1
DyoZeta1	CcPV1	0	1	1	1	1	1	0	1
Epsilon1	BPV5	-	-	-	-	1	-	0	1
Epsilon1	BPV8	-	-	-	-	0	-	0	0
Gamma1	HPV4	0	0	1	1	1	1	0	1
Gamma1	HPV65	0	0	1	1	1	1	0	1
Gamma1	HPV95	0	0	1	1	1	1	0	1
Gamma2	HPV48	0	0	1	1	1	1	0	1
Gamma3	HPV50	0	1	1	0	1	1	0	1
Gamma3	HPV131	0	0	1	1	1	1	0	1
Gamma4	HPV60	0	0	1	0	1	0	0	1
Gamma5	HPV88	0	0	1	1	0	1	0	1
Gamma6	HPV101	0	0	1	1	0	1	0	1
Gamma6	HPV103	0	0	1	0	0	1	0	1
Gamma6	HPV108	0	0	1	0	0	1	0	1
Gamma6	HPV128	0	0	1	1	1	1	0	1
Gamma7	HPV109	0	1	0	0	0	1	0	1
Gamma7	HPV123	0	1	1	0	0	1	0	1
Gamma7	HPV134	0	1	1	0	0	1	0	1
Gamma7	HPV149	0	1	1	0	1	1	0	1
Gamma8	HPV112	0	0	1	0	1	0	0	1
Gamma8	HPV119	0	0	1	0	1	0	0	1
Gamma9	HPV116	0	0	1	1	1	1	0	1
Gamma9	HPV129	0	0	1	1	1	1	0	1
Gamma10	HPV121	0	0	1	1	0	1	0	1
Gamma10	HPV130	0	0	1	1	0	1	0	1
Gamma10	HPV133	0	0	1	1	0	1	0	1
Gamma11	HPV127	0	1	1	1	1	1	0	1
Gamma11	HPV132	0	1	1	0	0	1	0	1
Gamma11	HPV148	0	0	1	0	1	1	0	1
Iota1	MnPV1	0	0	1	0	1	1	0	1
Kappa1	OcPV1	0	1	1	0	1	1	0	1
Kappa2	SfPV1k	0	0	1	1	1	1	0	1
Kappa2	SfPV1w	0	0	1	1	1	1	0	1
Lambda1	FdPV1	0	1	1	0	1	1	0	1
Lambda1	LrPV1	0	1	1	0	1	1	0	1
Lambda1	PcPV1	0	1	1	0	1	1	0	1
Lambda1	PlpPV1	0	1	1	0	1	1	0	1
Lambda1	UuPV1	0	1	1	0	1	1	0	1
Lambda2	CPV1	0	1	1	1	1	1	0	1
Lambda3	CPV6	0	1	1	0	1	1	0	1
Lambda4	PlPV1	0	1	1	1	0	1	0	1
Mu1	HPV1a	0	1	1	0	0	1	0	1
Mu2	HPV63	0	0	1	0	1	1	0	1
Nu1	HPV41	0	0	1	1	1	0	0	1
Phi1	ChPV1	0	0	0	0	0	1	0	1
Pi1	McPV2	0	0	1	1	0	0	0	1
Pi1	MmiPV1	0	1	1	1	0	0	0	0
Pi1	RnPV1	0	0	0	0	0	0	0	0
Psi1	RaPV1	0	0	0	0	1	1	0	1
Rho1	FmPV1	0	1	1	0	0	1	0	1
Rho1	TmPV1	0	1	1	0	0	1	0	1
Sigma1	EdPV1	0	1	1	0	1	0	0	1
Tau1	CPV2	0	0	1	1	0	1	0	1
Tau1	CPV7	0	1	1	1	1	1	0	1
UNK	BPP1	0	1	1	1	1	0	0	0
UNK	BPV7	-	-	-	-	0	-	0	1
UNK	EcPV3	0	0	0	1	0	0	0	1
UNK	MMPV1	0	0	1	1	1	0	0	1
UNK	OPV3	0	1	1	0	0	1	0	0
UNK	ZcPV1	0	1	0	0	1	1	0	1
Xi	BPV3	0	1	1	0	0	1	0	1
Xi	BPV4	0	1	1	0	0	1	0	1

continúa en la siguiente hoja

Especie	Tipo	DYRK1A	pRb_ABGroove	LxCxE	CKII	Región acídica	Pos. Ricas en Cys	NES	PDZ
Xi	BPV6	0	1	1	0	0	1	0	1
Xi	BPV9	0	1	0	0	0	1	0	1
Xi	BPV10	0	0	1	0	0	1	0	1
Xi	BPV11	0	0	1	0	0	1	0	1
Xi	BPVAA5	0	0	1	0	0	1	0	1
Zeta1	EcPV1	1	1	1	1	1	0	0	1

E.2. Conservación de secuencia en la proteína E7

E.2.1. Prueba Shapiro-Wilk. Valores *p*

Dominio	Media	Desvío	Valor *p*
E7N	2.38	0.99	0.11
E7C	2.07	1.15	0.01

Tabla E.2: Valores *p* de la prueba de Shapiro-Wilk para evaluar normalidad en los valores de contenido de información por posición de la proteína E7. No todos los grupos siguen una distribución normal (valor $p < 0.05$).

E.2.2. Prueba de permutación. Valores *p*

Grupo 1	Grupo 2	Valor *p*
E7C	E7N	0.20
E7N	E7C	0.21
E7N	E7	0.39
E7	E7N	0.39
E7	E7C	0.51
E7C	E7	0.52

Tabla E.3: Valores *p* de la prueba de permutación del contenido de información por posición de la proteína E7.

E.3. Predicción de desorden en la proteína E7

E.3.1. Prueba Shapiro-Wilk. Valores *p*

Dominio	Media	Desvío	Valor *p*
E7N	0.38	0.03	0.01
E7C	0.08	0.02	0.02

Tabla E.4: Valores *p* de la prueba de Shapiro-Wilk para evaluar normalidad en los valores promedio de IUPred por posición de la proteína E7. Los grupos no siguen una distribución normal (valor $p < 0.05$).

E.3.2. Método de remuestreo. Intervalos de confianza.

Dominio	Intervalo de confianza	Media estimada	Desvio estimado
E7N	0.33, 0.43	0.38	0.11
E7C	0.07, 0.09	0.08	0.03

Tabla E.5: Intervalos de confianza para el valor promedio de desorden predicho de los dominios de la proteína E7.

E.4. Información directa

HPV16		Alineamiento		
Posición 1	Posición 2	Posición 1	Posición 2	Información directa
2	4	2	4	0.1
2	95	2	82	0.1
3	10	3	10	0.11
3	13	3	13	0.23
7	13	7	13	0.17
7	65	7	52	0.12
7	83	7	70	0.11
7	92	7	79	0.1
8	10	8	10	0.11
8	13	8	13	0.1
10	11	10	11	0.16
10	13	10	13	0.34
10	69	10	56	0.11
11	13	11	13	0.31
12	13	12	13	0.13
15	56	15	43	0.11
16	17	16	17	0.24
21	22	19	20	0.15
21	24	19	22	0.12
21	25	19	23	0.11
21	26	19	24	0.11
21	34	19	31	0.1
22	24	20	22	0.16
22	26	20	24	0.47
24	26	22	24	0.19
27	32	25	29	0.1
27	95	25	82	0.12
32	33	29	30	0.13
37	38	34	35	0.14
37	83	34	70	0.11
39	40	36	37	0.4
51	70	38	57	0.12
54	87	41	74	0.11
54	90	41	77	0.11
64	65	51	52	0.1
66	82	53	69	0.19
67	89	54	76	0.13
67	95	54	82	0.12
69	76	56	63	0.2
72	86	59	73	0.11
75	86	62	73	0.13
77	80	64	67	0.2
89	92	76	79	0.12
97	98	84	85	0.15

Tabla E.6: Valores de información directa para E7 Se muestran los pares de residuos que coevolucionan. Las posiciones correspondientes a la proteína E7 de HPV16 se muestran en las dos primeras columnas. Las posiciones correspondientes al alineamiento sin sitios vacíos de la proteína E7 construido a partir de la base de datos 2 se indican en la tercera y cuarta columna. En la última columna se indican los valores significativos de información directa ($Z \geq 3$ e información directa ≥ 0.1).

Apéndice F

Proteína E1A

F.1. Blancos proteicos de E1A

La lista de blancos proteicos se encuentra disponible en el CD que se adjunta y en:
`https://bitbucket.org/jglavina/datos_tesis/wiki/Home`
Blancos proteicos E1A: `E1A_Blancos_Proteicos.xls`

F.2. Motivos lineales por secuencia de E1A

Tabla F.1: Presencia ausencia de motivos en las secuencias de la proteína E1A del género *Mastadenovirus*. Se indica la presencia (1) o ausencia (0) del motivo para cada serotipo.

Especie	Tipo	IDMBR	CoRNR Box	pRb_ABGroove	TRAM-CBP	MYND	LxCxE	Región Acídica	CKII	Pos. Ricas en Cys	CtBP	NLS
BtA	BtAdV3	0	1	1	0	0	1	1	1	0	0	0
BtB	BtAdV2	0	1	1	0	0	1	1	1	0	1	0
BA	BAdV1	0	0	1	0	0	1	0	1	0	1	0
BB	BAdV3	0	1	0	0	0	0	0	1	1	0	0
CA	CAdV1	0	1	1	0	0	1	1	1	0	1	0
CA	CAdV2	0	1	1	0	0	1	1	1	0	1	0
EA	EAdV1	0	1	1	0	0	1	1	1	0	0	0
HA	HAdV12	0	1	1	1	1	1	1	1	0	1	1
HA	HAdV18	0	1	1	0	1	1	1	1	0	1	1
HA	HAdV31	0	1	1	1	1	1	1	1	0	1	1
HA	HAdV61	0	1	1	1	1	1	1	1	0	1	1
HB	HAdV11	0	1	1	0	0	1	1	1	1	1	1
HB	HAdV14	0	1	1	0	0	1	1	1	1	1	1
HB	HAdV16	0	1	1	0	0	1	1	1	1	1	0
HB	HAdV21	0	1	1	0	0	1	1	1	1	1	0
HB	HAdV3	0	1	1	0	0	1	1	1	1	1	0
HB	HAdV34	0	1	1	0	0	1	1	1	1	1	1
HB	HAdV35	0	1	1	0	0	1	1	1	1	1	1
HB	HAdV50	0	1	1	0	0	1	1	1	1	1	0
HB	HAdV55	0	1	1	0	0	1	1	1	1	1	1
HB	HAdV7	0	1	1	0	0	1	1	1	1	1	0
HB	SAdV21	0	1	1	0	0	1	1	1	1	0	0
HB	SAdV27	0	1	1	0	0	1	1	1	1	1	0
HB	SAdV28	0	1	1	0	0	1	1	1	1	1	0
HB	SAdV29	0	1	1	0	0	1	1	1	1	1	0
HB	SAdV32	0	1	1	0	0	1	1	1	1	1	0
HB	SAdV33	0	1	1	0	0	1	1	1	1	0	0
HB	SAdV35	0	1	1	0	0	1	1	1	1	0	0
HB	SAdV41	0	1	1	0	0	1	1	1	1	1	0
HB	SAdV46	0	1	1	0	0	1	1	1	1	1	0
HB	SAdV47	0	1	1	0	0	1	1	1	1	1	0
HC	HAdV1	1	1	1	1	1	1	1	1	0	1	1
HC	HAdV2	1	1	1	1	1	1	1	1	0	1	1
HC	HAdV5	1	1	1	1	1	1	1	1	0	1	1
HC	HAdV57	1	1	1	1	1	1	1	1	0	1	1

continúa en la siguiente hoja

Especie	Tipo	IDMBR	CoRNR Box	pRb_ABGroove	TRAM-CBP	MYND	LxCxE	Región Acídica	CKII	Pos. Ricas en Cys	CtBP	NLS
HC	HAdV6	1	1	1	1	1	1	1	1	1	1	1
HC	SAdV31	0	1	1	0	1	1	1	1	1	1	1
HC	SAdV34	0	1	1	0	1	1	1	1	1	1	1
HC	SAdV40	0	1	1	0	1	1	1	1	1	1	1
HC	SAdV42	0	1	1	0	1	1	1	1	1	1	1
HC	SAdV43	1	1	1	1	1	1	0	1	0	1	1
HC	SAdV44	0	1	1	0	1	1	1	1	1	1	1
HC	SAdV45	1	1	1	1	1	1	0	1	0	1	1
HD	HAdV10	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV13	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV15	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV17	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV19	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV20	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV22	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV23	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV24	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV25	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV26	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV27	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV28	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV29	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV30	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV32	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV33	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV36	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV37	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV38	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV39	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV42	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV43	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV44	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV45	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV47	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV48	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV49	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV51	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV53	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV54	0	1	1	0	1	1	1	1	0	0	1
HD	HAdV56	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV58	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV59	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV60	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV62	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV63	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV64	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV65	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV67	0	1	1	0	1	1	1	1	0	1	1
HD	HAdV8	0	1	1	0	1	1	1	1	0	0	1
HD	HAdV9	0	1	1	0	1	1	1	1	0	1	1
HE	ChAdVY25	0	1	1	0	0	1	1	1	1	0	1
HE	HAdV4	0	1	1	0	0	1	1	1	1	0	1
HE	SAdV22	0	1	1	0	0	1	1	1	1	0	1
HE	SAdV23	0	1	1	0	0	1	1	1	1	0	1
HE	SAdV24	0	1	1	0	0	1	1	1	1	0	1
HE	SAdV25	0	1	1	0	0	1	1	1	1	0	1
HE	SAdV26	0	1	1	0	0	1	1	1	1	0	1
HE	SAdV30	0	1	1	0	0	1	1	1	1	0	1
HE	SAdV36	0	1	1	0	0	1	1	1	1	0	1
HE	SAdV37	0	1	1	0	0	1	1	1	1	0	1
HE	SAdV38	0	1	1	0	0	1	1	1	1	0	1
HE	SAdV39	0	1	1	0	0	1	1	1	1	0	1
HF	HAdV40	0	1	1	0	0	1	1	1	1	0	1
HF	HAdV41	0	1	1	0	0	1	1	1	1	0	1
HG	HAdV52	0	0	1	0	0	1	1	1	0	1	1
HG	SAdV1	0	0	1	0	0	1	1	1	1	1	1
HG	SAdV7	0	1	1	0	1	1	1	1	1	1	0
MA	MAdV1	0	0	0	0	0	1	0	0	0	0	0
MB	MAdV2	0	0	0	0	0	1	0	0	1	0	0
MC	MAdV3	0	0	0	0	0	1	0	0	0	0	0
OA	BAdV2	0	0	1	0	0	1	1	1	0	1	0
PA	PAdV3	0	0	0	0	0	0	1	1	1	0	0
PC	PAdV5	0	0	1	0	0	1	1	1	0	0	0
SA	SAdV20	0	1	1	0	1	1	1	1	0	1	1
SA	SAdV3	0	1	1	0	0	1	1	1	0	1	1
SA	SAdV48	0	1	1	0	0	1	1	1	0	1	1
SA	SAdV6	0	1	1	0	0	1	1	1	0	1	1
SB	SAdV49	0	0	1	0	0	1	1	1	1	1	1
SB	SAdV50	0	0	1	0	0	1	1	1	1	1	1
SF	SAdV18	0	0	1	0	0	1	1	1	1	0	1
TSA	TSAdV1	0	0	0	0	0	0	1	1	1	0	0

F.3. Conservación de secuencia

F.3.1. A nivel de dominios y regiones

Prueba Shapiro-Wilk. Valores *p*.

Dominio	Media	Desvío	Valor *p*
N-terminal	2.3	0.7	0.02
CR1	2.7	0.8	0.5
CR2	2.9	0.7	0.12
CR3	2.8	0.9	0.005
CR4	2.8	0.8	0.02
IDR12	2.4	0.7	0.01
IDR23	2.2	0.6	0.5
IDR34	2.5	0.7	0.003

Tabla F.2: Valores *p* de la prueba de Shapiro-Wilk para evaluar normalidad en los valores de contenido de información por posición de la proteína E1A. No todos los grupos siguen una distribución normal (valor $p < 0.05$).

Prueba de permutación. Valores *p*.

Grupo 1	Grupo 2	Valor *p*	Valor *p**
N-terminal	CR2	0.0007	0.01
IDR12	CR2	0.001	0.01
IDR12	CR4	0.001	0.01
IDR34	CR2	0.002	0.01
N-terminal	CR4	0.002	0.01
IDR23	CR2	0.003	0.02
IDR34	CR4	0.003	0.02
IDR12	CR1	0.004	0.02
IDR12	TODAS	0.005	0.02
IDR12	CR3	0.007	0.02
IDR34	CR1	0.008	0.02
IDR23	CR4	0.008	0.02
N-terminal	CR1	0.009	0.02
IDR34	TODAS	0.02	0.04
N-terminal	TODAS	0.02	0.04
IDR34	CR3	0.02	0.04

Tabla F.3: Valores *p* de la prueba de permutación del contenido de información. En las dos últimas columnas se indican los valores p de la prueba de permutación sin corregir y corregidos (valor p^*) por Benjamini-Hochberg. Sólo se muestran los valores significativos (valor $p^* < 0.05$). El número total de comparaciones realizadas es 36.

F.3.2. A nivel de posiciones de motivos

Prueba Shapiro-Wilk. Valores *p*.

Posicion	Media	Desvío	Valor *p*
Todas	2.63	0.81	$6.7\ 10^{-6}$
Fijas	3.31	0.75	$5.8\ 10^{-4}$
Comodín	2.35	0.65	0.07
Adyacentes	2.63	0.81	0.3
Otras	2.5	0.65	0.0005

Tabla F.4: Valores *p* de la prueba de Shapiro-Wilk para evaluar normalidad en los valores de contenido de información por posición de motivos de la proteína E1A. No todos los grupos siguen una distribución normal (valor $p < 0.05$).

Prueba de permutación. Valores *p*.

Grupo 1	Grupo 2	Valor *p*	Valor *p**
Fijas	Todas	0	0
Comodín	Fijas	0	0
Otras	Fijas	0	0
Adyacentes	Fijas	0.0005	0.002

Tabla F.5: Valores *p* de la prueba de permutación del contenido de información para las posiciones de motivos. En las dos últimas columnas se indican los valores *p* de la prueba de permutación sin corregir y corregidos (valor *p**) por Benjamini-Hochberg. Sólo se muestran los valores significativos (valor $p^* < 0.05$).

F.4. Predicción de desorden

F.4.1. Prueba Shapiro-Wilk. Valores *p*.

Dominio	Media	Desvío	Valor *p*
N-terminal	0.42	0.13	0.001
CR1	0.56	0.09	0.2
CR2	0.64	0.1	0.1
CR3	0.36	0.16	0.1
CR4	0.67	0.11	0.002

Tabla F.6: Valores *p* de la prueba de Shapiro-Wilk para evaluar normalidad en los valores promedio de IUPred por posición de la proteína E1A. No todos los grupos siguen una distribución normal (valor $p < 0.05$).

F.4.2. Gráfico QQ

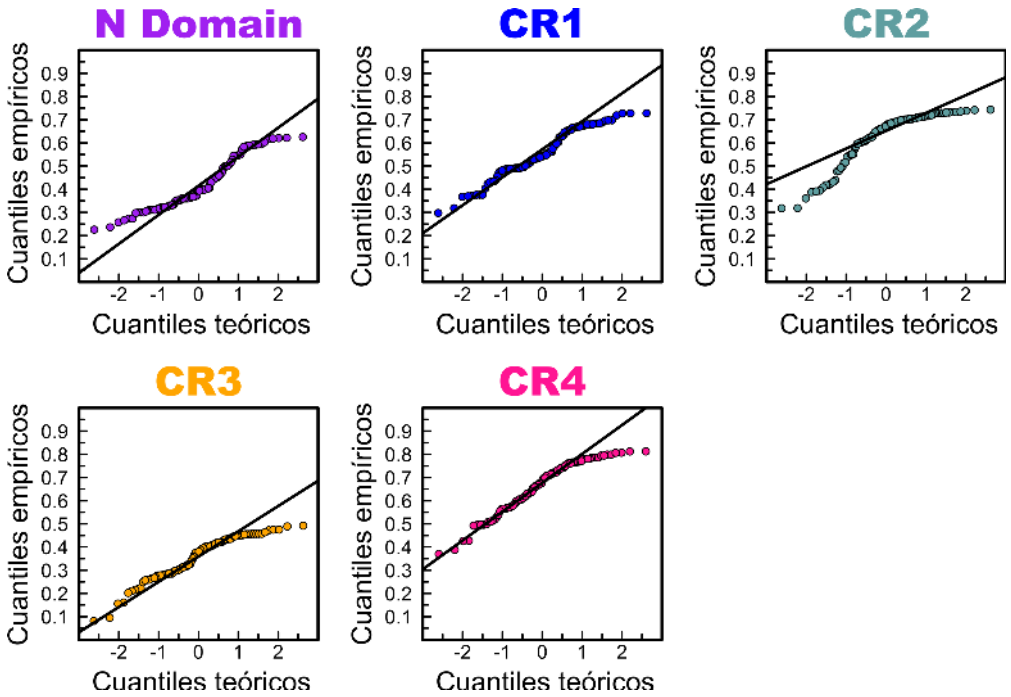


Tabla F.7: Q-Q plot normal para el conjunto de datos de IUPred de la proteína E1A. Se realizó una comparación entre los cuantiles del conjunto de datos en el eje *y* y una población con distribución normal en el eje *x* para cada dominio de E1A. El código de colores es el mismo que se utilizó en la Figura 4.2.

F.4.3. Método de remuestreo. Intervalos de confianza.

Dominio	Intervalo de confianza	Media estimada	Desvio estimado
N-Terminal	0.39, 0.43	0.42	0.1
CR1	0.53, 0.58	0.56	0.1
CR2	0.61, 0.66	0.63	0.1
CR3	0.34, 0.38	0.36	0.09
CR4	0.64, 0.69	0.67	0.1
IDR12	0.73, 0.87	0.79	0.1
IDR23	0.49, 0.56	0.53	0.13
IDR34	0.72, 0.79	0.75	0.05

Tabla F.8: Intervalos de confianza para el el valor promedio de desorden predicho de los dominios y regiones de E1A.

F.5. Coevolucion de secuencia

F.5.1. Información directa

Tabla F.9: Valores de información directa para E1A Se muestran los pares de residuos que coevolucionan. Las posiciones correspondientes a la proteína E1A de HAdV5 se muestran en las dos primeras columnas. Las posiciones correspondientes al alineamiento sin sitios vacíos de la proteína E1A se indican en la tercera y cuarta columna. En la última columna se indican los valores significativos de información directa ($Z \geq 3$ e información directa ≥ 0.04).

HAdV5		Alineamiento		
Posición 1	**Posición 2**	**Posición 1**	**Posición 2**	**Información directa**
3	252	3	208	0.04
8	23	8	23	0.05
9	21	9	21	0.05
11	120	11	99	0.06
14	18	14	18	0.07
14	25	14	25	0.07
14	36	14	35	0.07
14	135	14	114	0.05
18	25	18	25	0.04
18	36	18	35	0.05
25	36	25	35	0.06
25	116	25	90	0.06
25	135	25	114	0.04
30	139h	30	125	0.06
31	117c	31	94	0.07
39	238	38	194	0.06
40	50	39	49	0.04
45	75	44	74	0.07
45	282	44	239	0.06
54	139i	53	126	0.1
54	139j	53	127	0.06
54	139n	53	131	0.05
60	143	59	137	0.04
68	146	67	140	0.06
70	167	69	161	0.07
70	239	69	195	0.04
74	149	73	143	0.04
75	116	74	90	0.04
75	136	74	115	0.07
78	80	77	79	0.14
80	82	79	81	0.05
80	136	79	115	0.08
81	84	80	83	0.04
83	84	82	83	0.12
84	277	83	232	0.05
110	131	84	110	0.04
111	247	85	203	0.05
116	136	90	115	0.04
117a	117b	92	93	0.06
117b	117c	93	94	0.09
117b	117d	93	95	0.1
117b	117f	93	97	0.04
117c	117d	94	95	0.05

continúa en la siguiente hoja

HAdV5		Alineamiento		
Posición 1	**Posición 2**	**Posición 1**	**Posición 2**	**Información directa**
117c	119	94	98	0.05
117d	117f	95	97	0.05
117e	117f	96	97	0.06
139b	139c	119	120	0.06
139b	139f	119	123	0.07
139b	139g	119	124	0.07
139f	139g	123	124	0.05
139h	141	125	135	0.04
139h	190	125	184	0.07
139i	139j	126	127	0.06
139i	139n	126	131	0.06
139n	139o	131	132	0.05
142	144	136	138	0.04
142	145	136	139	0.07
142	147	136	141	0.04
142	156	136	150	0.1
142	164	136	158	0.07
144	145	138	139	0.11
144	147	138	141	0.1
144	148	138	142	0.06
144	151	138	145	0.07
144	164	138	158	0.07
145	147	139	141	0.08
145	148	139	142	0.06
145	151	139	145	0.08
145	156	139	150	0.06
145	164	139	158	0.07
147	151	141	145	0.08
147	156	141	150	0.05
148	164	142	158	0.04
150	167	144	161	0.06
150	245	144	201	0.05
151	164	145	158	0.07
156	164	150	158	0.06
167	178	161	172	0.04
168	179	162	173	0.08
178	179	172	173	0.12
231	232	187	188	0.05
235	264	191	221	0.05
236	253	192	209	0.04
242	244	198	200	0.12
242	245	198	201	0.05
242	246	198	202	0.12
244	245	200	201	0.06
244	246	200	202	0.1
245	246	201	202	0.06
252	267	208	224	0.09
267	274	224	231	0.06
279	280	236	237	0.06
279	282	236	239	0.04
279	283	236	240	0.04
281	282	238	239	0.04

F.6. Teoría de polimeros

F.6.1. Constante intramolecular para distintas ventanas

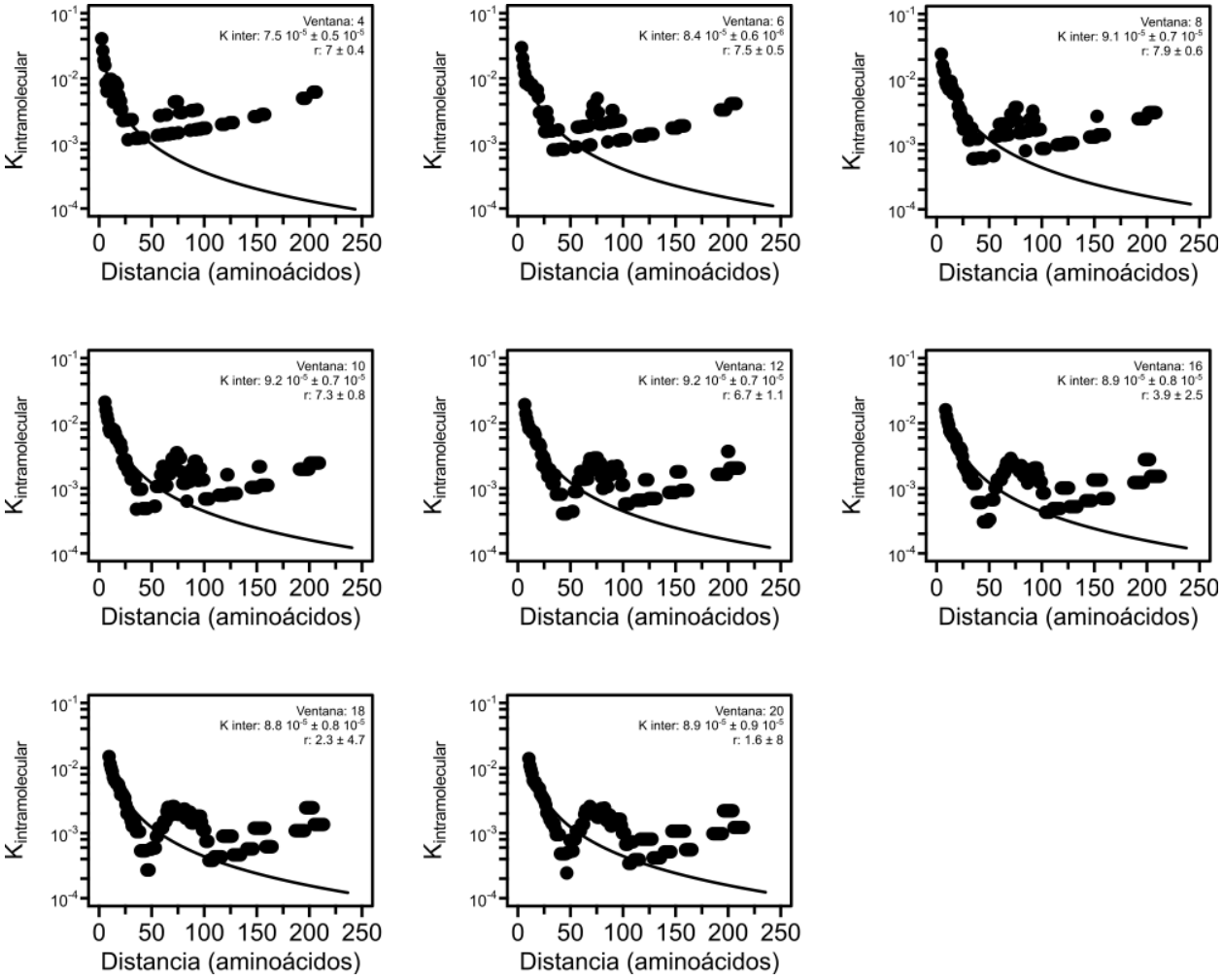


Tabla F.10: Probabilidad de formación de contacto en E1A es función de la separación de secuencia para distintos tamaños de ventana. La constante de equilibrio para la formación de contactos entre aminoácidos $K_{intramolecular}$ fue estimada como función de la separación de secuencia L utilizando el mapa de contactos predicho (puntos) y el modelo de cadena entrópica (Zhou, 2004). El tamaño de ventana y los parámetros que mejor ajustan la ecuación que describe la formación de contactos intramoleculares de una cadena entrópica (línea) están indicados en la parte superior de cada gráfico.

F.7. Modelo estructural del CR3

F.7.1. Modelo alternativo

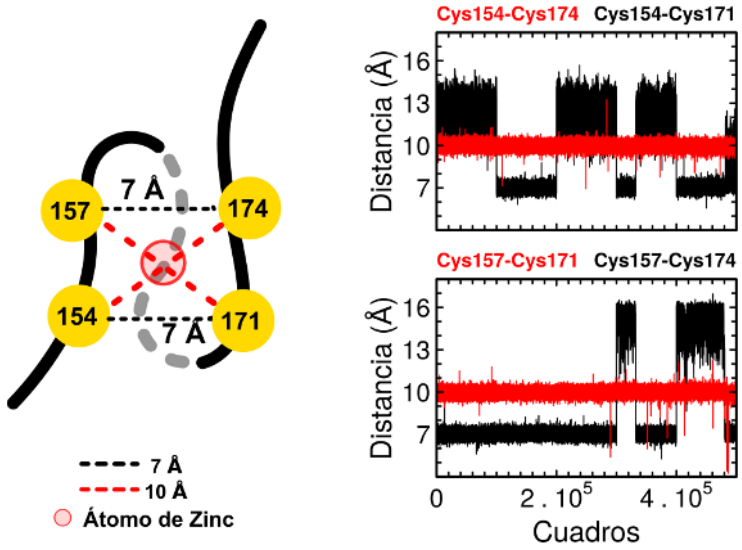


Tabla F.11: Construcción del modelo estructural alternativo del dominio CR3 de E1A. *Derecha.* Configuración del sitio de unión a zinc. Los cuatro C_α de las cisteínas que coordinan el átomo de zinc (círculo rojo claro) están representados como círculos amarillos y se muestran las interacciones utilizadas como restricciones en la simulación (líneas punteadas). *Izquierda.* Distancia entre las cisteínas que coordinan el zinc a lo largo de la trayectoria. Se puede observar que las distancias entre las cisteínas 154-171 y 157-174 no se mantienen dentro del rango del modelo.

F.8. Asociaciones entre motivos y entre motivos y rasgos fenotípicos

F.8.1. Asociación entre motivos

Motivo 1	Motivo 2	valor *p*	valor *p**
MYND	NLS	$9\ 10^{11}$	$5\ 10^{-9}$
TRAM-CBP	IDMBR	$3\ 10^{-9}$	$7\ 10^{-8}$
MYND	CtBP	$7\ 10^{-9}$	$1\ 10^{-7}$
CoRNR Box	MYND	$3\ 10^{-5}$	0.0005
CoRNR Box	pRb_ABGroove	$5\ 10^{-5}$	0.0005
pRb_ABGroove	Región Acídica	$7\ 10^{-5}$	0.0005
pRb_ABGroove	CKII	$8\ 10^{-5}$	0.0005
pRb_ABGroove	LxCxE	$8\ 10^{-5}$	0.0005
Región Acídica	CKII	0.0001	0.0008
pRb_ABGroove	CtBP	0.0002	0.0009
pRb_ABGroove	NLS	0.0002	0.001
TRAM-CBP	MYND	0.003	0.004
CoRNR Box	CKII	0.001	0.005
CoRNR Box	Región Acídica	0.003	0.01
CoRNR Box	NLS	0.004	0.01
MYND	IDMBR	0.008	0.02
pRb_ABGroove	MYND	0.01	0.03
Región Acídica	NLS	0.01	0.03
CKII	CtBP	0.01	0.04
LxCxE	CtBP	0.01	0.04
CKII	NLS	0.02	0.04
LxCxE	NLS	0.02	0.04
CoRNR Box	CtBP	0.02	0.04

Tabla F.12: Co-ocurrencias motivo-motivo. Se muestran los valores *p* de la prueba hipergeométrica y los valores *p* corregidos (valor *p**) por Benjamini-Hochberg. Únicamente se muestran las asociaciones significativas (valor $p^* < 0.05$).

F.8.2. Asociación entre motivo y hospedador

Motivo	Hospedador	valor *p*	valor *p**
NLS	Primates	0	0
pRb_ABGroove	Primates	$1\ 10^{-6}$	$4\ 10^{-5}$
CoRNR Box	Primates	$5\ 10^{-6}$	0.0001
MYND	Primates	$1\ 10^{-5}$	0.0003
Región Acídica	Primates	0.0002	0.003
CtBP	Primates	0.001	0.01
CKII	Primates	0.001	0.01
LxCxE	Primates	0.001	0.01
MYND	Humanos	$3\ 10^{11}$	$3\ 10^{9}$
Pos. Ricas en Cys	P. No Humanos	4.4210^{10}	1.9510^{-8}
NLS	Humanos	$4\ 10^{-7}$	$1\ 10^{-5}$
CtBP	Humanos	$1\ 10^{-6}$	$2\ 10^{-5}$
CoRNR Box	Humanos	0.0001	0.003
Región Acídica	Humanos	0.002	0.04

Tabla F.13: Co-ocurrencias motivo-hospedador. Se muestran los valores *p* de la prueba hipergeométrica y los valores *p* corregidos (valor *p**) por Benjamini-Hochberg. Únicamente se muestran las asociaciones significativas (valor $p^* < 0.05$).

F.8.3. Asociación entre motivo y tropismo

Motivo	Tropismo	valor *p*	valor *p**
TRAM-CBP	SNC	0.0001	0.01

Tabla F.14: Co-ocurrencias motivo-tropismo. Se muestran los valores *p* de la prueba hipergeométrica y los valores *p* corregidos (valor *p**) por Benjamini-Hochberg. Únicamente se muestran las asociaciones significativas (valor $p^* < 0.05$).

F.9. Asociación entre eventos de aparición y desaparición de motivos

F.9.1. Metodo de bayes

Tabla F.15: Co-ocurrencias eventos de aparición, desaparición y cambio de estado entre motivos. Se muestran los valores *p* de la prueba hipergeométrica y los valores *p* corregidos (valor *p**) por Benjamini-Hochberg. Únicamente se muestran las asociaciones significativas (valor $p^* < 0.05$).

Evento Motivo 1	Evento Motivo 2	valor *p*	valor *p**
Aparición CKII	Desaparición pRb_ABGroove	$3.1\ 10^{-7}$	0.0001
Desaparición CtBP	Desaparición NLS	$8.1\ 10^{-7}$	0.0001
Cambio CtBP	Desaparición NLS	$8.1\ 10^{-7}$	0.0001
Cambio CKII	Desaparición pRb_ABGroove	$1.1\ 10^{-6}$	0.0001
Aparición CKII	Cambio pRb_ABGroove	$1.5\ 10^{-6}$	0.0001

continúa en la siguiente hoja

Evento Motivo 1	Evento Motivo 2	valor *p*	valor *p**
Desaparición CtBP	Cambio NLS	$1.6\ 10^{-6}$	0.0001
Cambio CtBP	Cambio NLS	$1.6\ 10^{-6}$	0.0001
Desaparición LxCxE	Desaparición pRb_ABGroove	$2\ 10^{-6}$	0.0001
Cambio LxCxE	Desaparición pRb_ABGroove	$2\ 10^{-6}$	0.0001
Aparición CKII	Desaparición NLS	$5\ 10^{-6}$	0.0002
Desaparición LxCxE	Cambio pRb_ABGroove	$5\ 10^{-6}$	0.0002
Cambio LxCxE	Cambio pRb_ABGroove	$5\ 10^{-6}$	0.0002
Cambio CKII	Cambio pRb_ABGroove	$5.4\ 10^{-6}$	0.0002
Desaparición NLS	Desaparición pRb_ABGroove	$6.3\ 10^{-6}$	0.0002
Aparición CKII	Cambio NLS	$7.7\ 10^{-6}$	0.0002
Desaparición CtBP	Desaparición pRb_ABGroove	$8.8\ 10^{-6}$	0.0002
Cambio CtBP	Desaparición pRb_ABGroove	$8.8\ 10^{-6}$	0.0002
Cambio NLS	Desaparición pRb_ABGroove	$8.8\ 10^{-6}$	0.0002
Cambio CoRNR Box	Cambio pRb_ABGroove	$9\ 10^{-6}$	0.0002
Aparición CKII	Desaparición LxCxE	$1.7\ 10^{-5}$	0.0004
Aparición CKII	Cambio LxCxE	$1.7\ 10^{-5}$	0.0004
Aparición CKII	Desaparición CoRNR Box	$2.1\ 10^{-5}$	0.0005
Cambio MYND	Cambio pRb_ABGroove	$2.3\ 10^{-5}$	0.0005
Cambio CoRNR Box	Cambio MYND	$2.3\ 10^{-5}$	0.0005
Cambio CKII	Desaparición NLS	$2.8\ 10^{-5}$	0.0006
Desaparición NLS	Cambio pRb_ABGroove	$3\ 10^{-5}$	0.0006
Cambio CKII	Desaparición LxCxE	$4.2\ 10^{-5}$	0.0007
Cambio CKII	Cambio LxCxE	$4.2\ 10^{-5}$	0.0007
Desaparición CtBP	Cambio pRb_ABGroove	$4.3\ 10^{-5}$	0.0007
Cambio CtBP	Cambio pRb_ABGroove	$4.3\ 10^{-5}$	0.0007
Cambio NLS	Cambio pRb_ABGroove	$4.3\ 10^{-5}$	0.0007
Cambio CKII	Cambio NLS	$4.3\ 10^{-5}$	0.0007
Desaparición CoRNR Box	Desaparición MYND	$4.6\ 10^{-5}$	0.0007
Cambio CoRNR Box	Desaparición NLS	$5.4\ 10^{-5}$	0.0008
Desaparición CoRNR Box	Cambio MYND	$5.7\ 10^{-5}$	0.0008
Aparición CKII	Cambio CoRNR Box	$6.1\ 10^{-5}$	0.0008
Cambio CKII	Desaparición CoRNR Box	$7.2\ 10^{-5}$	0.0001
Cambio CoRNR Box	Cambio NLS	$8.3\ 10^{-5}$	0.001
Desaparición CoRNR Box	Desaparición pRb_ABGroove	0.0001	0.001
Desaparición LxCxE	Desaparición NLS	0.0001	0.002
Cambio LxCxE	Desaparición NLS	0.0001	0.002
Desaparición MYND	Desaparición pRb_ABGroove	0.0002	0.002
Desaparición CtBP	Desaparición LxCxE	0.0002	0.002
Cambio CtBP	Desaparición LxCxE	0.0002	0.002
Desaparición CtBP	Cambio LxCxE	0.0002	0.002
Cambio CtBP	Cambio LxCxE	0.0002	0.002
Desaparición LxCxE	Cambio NLS	0.0002	0.002
Cambio LxCxE	Cambio NLS	0.0002	0.002
Cambio MYND	Desaparición pRb_ABGroove	0.0002	0.002
Cambio CKII	Cambio CoRNR Box	0.0002	0.002
Cambio IDMBR	Cambio TRAM-CBP	0.0002	0.002
Cambio CoRNR Box	Desaparición pRb_ABGroove	0.0002	0.002
Cambio Pos. Ricas en Cys	Cambio TRAM-CBP	0.0002	0.002
Desaparición CoRNR Box	Cambio pRb_ABGroove	0.0003	0.002
Aparición CKII	Desaparición CtBP	0.0003	0.002
Aparición CKII	Cambio CtBP	0.0003	0.002
Cambio CoRNR Box	Desaparición MYND	0.0003	0.002
Aparición CKII	Desaparición MYND	0.0003	0.003
Desaparición CtBP	Desaparición MYND	0.0004	0.003
Cambio CtBP	Desaparición MYND	0.0004	0.003

continúa en la siguiente hoja

Evento Motivo 1	Evento Motivo 2	valor *p*	valor *p**
Desaparición CoRNR Box	Desaparición NLS	0.0004	0.003
Aparición CKII	Cambio MYND	0.0004	0.003
Desaparición CtBP	Cambio MYND	0.0005	0.004
Cambio CtBP	Cambio MYND	0.0005	0.004
Desaparición CoRNR Box	Desaparición CtBP	0.0005	0.004
Desaparición CoRNR Box	Cambio CtBP	0.0005	0.004
Desaparición CoRNR Box	Cambio NLS	0.0005	0.004
Desaparición MYND	Cambio pRb_ABGroove	0.0008	0.006
Aparición CKII	Desaparición Pos. Ricas en Cys	0.0009	0.006
Cambio CKII	Desaparición CtBP	0.0009	0.006
Cambio CKII	Cambio CtBP	0.0009	0.006
Desaparición Pos. Ricas en Cys	Aparición TRAM-CBP	0.001	0.007
Desaparición Región Acídica	Desaparición CtBP	0.002	0.01
Cambio CoRNR Box	Desaparición CtBP	0.002	0.01
Desaparición Región Acídica	Cambio CtBP	0.002	0.01
Cambio CoRNR Box	Cambio CtBP	0.002	0.01
Desaparición LxCxE	Desaparición MYND	0.002	0.01
Cambio LxCxE	Desaparición MYND	0.002	0.01
Cambio CKII	Desaparición MYND	0.002	0.01
Cambio Pos. Ricas en Cys	Cambio IDMBR	0.002	0.01
Cambio Pos. Ricas en Cys	Aparición TRAM-CBP	0.002	0.01
Desaparición LxCxE	Cambio MYND	0.002	0.01
Cambio LxCxE	Cambio MYND	0.002	0.01
Aparición CKII	Cambio Pos. Ricas en Cys	0.002	0.01
Cambio CKII	Cambio MYND	0.002	0.01
Cambio CKII	Desaparición Pos. Ricas en Cys	0.002	0.01
Desaparición CoRNR Box	Desaparición LxCxE	0.003	0.02
Desaparición CoRNR Box	Cambio LxCxE	0.003	0.02
Aparición MYND	Aparición pRb_ABGroove	0.004	0.02
Cambio CKII	Cambio Pos. Ricas en Cys	0.004	0.02
Cambio CoRNR Box	Desaparición LxCxE	0.005	0.03
Cambio CoRNR Box	Cambio LxCxE	0.005	0.03
Desaparición Pos. Ricas en Cys	Cambio TRAM-CBP	0.006	0.03
Aparición IDMBR	Aparición TRAM-CBP	0.009	0.04
Desaparición IDMBR	Desaparición TRAM-CBP	0.009	0.04
Aparición Pos. Ricas en Cys	Desaparición IDMBR	0.009	0.04
Aparición CoRNR Box	Aparición MYND	0.009	0.04
Aparición CoRNR Box	Aparición pRb_ABGroove	0.009	0.04
Desaparición Región Acídica	Desaparición pRb_ABGroove	0.01	0.04

F.10. Soluciones del análisis de codivergencia de Jane

F.10.1. Solución 3740 de costo total 48

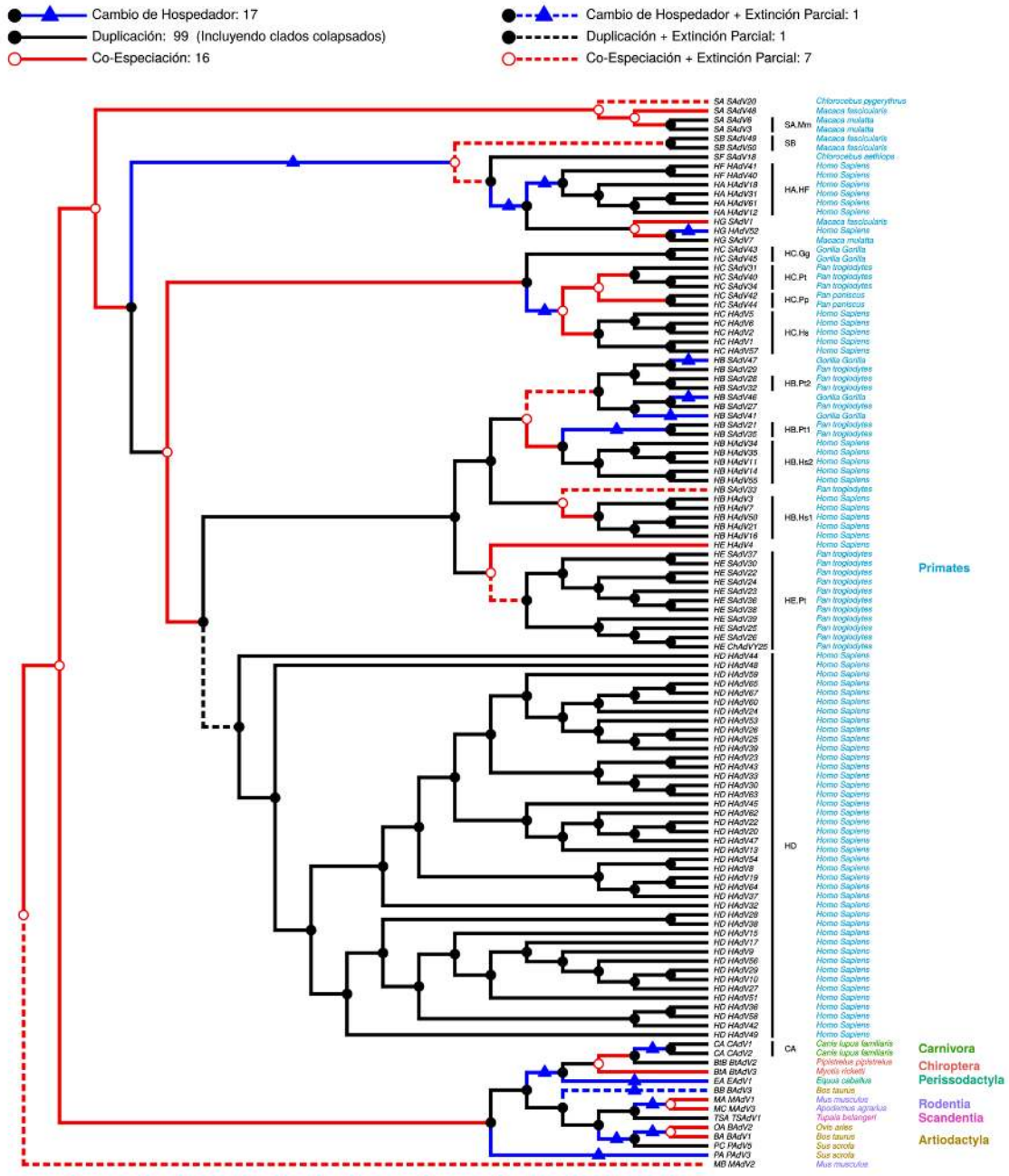


Figura F.1: Mapeo de los procesos evolutivos en la evolución de *Mastadenovirus* sobre su filogenia. Se muestra el mapeo de la reconciliación 3740 creada por Jane (Conow *et al.*, 2010) del árbol de *Mastadenovirus* y el árbol de su hospedador. Los símbolos utilizados para los eventos son: círculo blanco con borde rojo para coespeciación, círculos negros para duplicación, triángulos azules para cambio de hospedador y línea punteada para extinción parcial. En las hojas del árbol se muestran los serotipos con su especie. Se señalan los grupos monofiléticos que compartían hospedador y que fueron colapsados para el análisis. Al lado de los serotipos se indica el hospedador al que infectan. Por último, se indica el orden al que pertenece el hospedador.

F.11. Asociación entre eventos de aparición y desaparición de motivos y eventos evolutivos

F.11.1. Metodo de bayes

Solución 3734

Evento Motivo	Evento Evolutivo	valor *p*	valor *p**
Aparición CKII	Cambio de Hospedador	2.3110^{-5}	0.003
Cambio CKII	Cambio de Hospedador	0.0001	0.008
Desaparición CtBP	Extinción Parcial	0.0009	0.02
Cambio CtBP	Extinción Parcial	0.0009	0.02
Desaparición NLS	Cambio de Hospedador	0.001	0.02
Desaparición Pos. Ricas en Cys	Cambio de Hospedador	0.001	0.02
Desaparición pRb_ABGroove	Cambio de Hospedador	0.001	0.02
Desaparición CtBP	Cambio de Hospedador	0.002	0.02
Cambio CtBP	Cambio de Hospedador	0.002	0.02
Cambio NLS	Cambio de Hospedador	0.002	0.02
Cambio pRb_ABGroove	Cambio de Hospedador	0.003	0.04
Cambio Pos. Ricas en Cys	Cambio de Hospedador	0.003	0.04
Desaparición Región Acídica	Extinción Parcial	0.004	0.04

Tabla F.16: Co-ocurrencias eventos de aparición, desaparición y cambio de estado entre motivos y eventos evolutivos. Se muestran los valores *p* de la prueba hipergeométrica y los valores *p* corregidos (valor *p**) por Benjamini-Hochberg. Únicamente se muestran las asociaciones significativas (valor $p^* < 0.05$).

Solución 3740

Evento Motivo	Evento Evolutivo	valor *p*	valor *p**
Aparición CKII	Cambio de Hospedador	0.00002	0.003
Cambio CKII	Cambio de Hospedador	1.27e-04	0.008
Desaparición NLS	Cambio de Hospedador	0.001	0.03
Desaparición Pos. Ricas en Cys	Cambio de Hospedador	0.001	0.03
Desaparición pRb_ABGroove	Cambio de Hospedador	0.001	0.03
Desaparición CtBP	Cambio de Hospedador	0.002	0.03
Cambio CtBP	Cambio de Hospedador	0.002	0.03
Cambio NLS	Cambio de Hospedador	0.002	0.03
Cambio pRb_ABGroove	Cambio de Hospedador	0.003	0.04
Cambio Pos. Ricas en Cys	Cambio de Hospedador	0.003	0.04

Tabla F.17: Co-ocurrencias eventos de aparición, desaparición y cambio de estado entre motivos y eventos evolutivos. Se muestran los valores *p* de la prueba hipergeométrica y los valores *p* corregidos (valor *p**) por Benjamini-Hochberg. Únicamente se muestran las asociaciones significativas (valor $p^* < 0.05$).

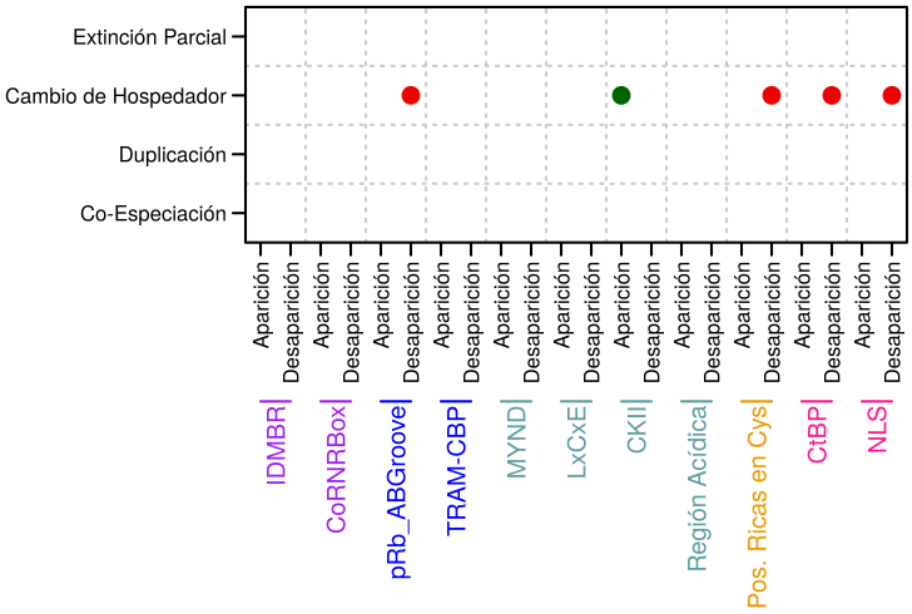


Figura F.2: Asociaciones entre eventos de aparición y desaparición de motivos a lo largo de la filogenia de *Mastadenovirus* y eventos evolutivos. Las asociaciones positivas (valor $p^* < 0.05$) están marcadas con un punto rojo cuando ocurre entre un evento evolutivo y un evento de desaparición. Un punto verde cuando la asociación ocurre entre un evento evolutivo y un evento de aparición. El código de colores para las etiquetas es el utilizado en la Figura 4.2.